A. Analysis: finite-horizon MDPs

In this section, we present the proof of Theorem 3 — a more general version of Theorem 2 — which accounts for the full ε-range.

Theorem 3. *Consider any* $\varepsilon \in \big(0, \frac{1}{1-\gamma}\big]$. *Theorem 2 continues to hold if*

$$T \geq \frac{c_3 H^4 \big(\log^3 T\big)\big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{\min\{\varepsilon^2, \varepsilon\}} \tag{27}$$

for some sufficiently large universal constant $c_3 > 0$.

A.1. Preliminaries

Let us first introduce several vector and matrix notations that are adopted for the finite-horizon case.

Vector and matrix notation. We use vector $\boldsymbol{r}_h \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ to represent the reward function r_h at step h. The vectors $\boldsymbol{V}_h^\pi \in \mathbb{R}^{|\mathcal{S}|}$, $\boldsymbol{V}_h^\star \in \mathbb{R}^{|\mathcal{S}|}$, $\boldsymbol{Q}_h^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and $\boldsymbol{Q}_h^\star \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ are defined in an analogous manner. Let $\boldsymbol{Q}_{t,h} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ (resp. $\boldsymbol{V}_{t,h} \in \mathbb{R}^{|\mathcal{S}|}$) be the estimate $Q_{t,h}$ (resp. $V_{t,h}$) in the t-th iteration at step h, namely

$$\boldsymbol{Q}_{t,h} = (1-\eta_t)\boldsymbol{Q}_{t-1,h} + \eta_t(\boldsymbol{r}_h + \boldsymbol{P}_{t,h}\boldsymbol{V}_{t,h+1}), \qquad \forall 1 \leq h \leq H, \tag{28a}$$

$$\boldsymbol{V}_{t,h} = \max_a \boldsymbol{Q}_{t,h}, \qquad \forall 1 \leq h \leq H. \tag{28b}$$

Here, the maximum in (28b) is taken in an entry-wise manner (cf. (8)). We also use matrix $\boldsymbol{P}_h \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$ to represent the probability transition kernel P_h at step h. Moreover, let the matrix $\boldsymbol{P}_{t,h} \in \{0,1\}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$ be

$$\boldsymbol{P}_{t,h}\big((s,a),s'\big) := \begin{cases} 1, & \text{if } s' = s_{t,h}(s,a), \\ 0, & \text{otherwise.} \end{cases} \tag{29}$$

Similar to the infinite-horizon case, let us first collect a couple of basic facts that will be useful in the proof.

Ranges of $\boldsymbol{Q}_{t,h}$ and $\boldsymbol{V}_{t,h}$. We shall start with some simple bounds for $\boldsymbol{Q}_{t,h}$ and $\boldsymbol{V}_{t,h}$. Lemma 3 (below) demonstrates that the estimates for the Q-function and the value function are bounded as long as they are properly initialized.

Lemma 3. *Suppose that* $0 \leq \eta_t \leq 1$ *for all* $t \geq 0$. *Assume that* $\boldsymbol{Q}_{H+1}^0 = \boldsymbol{V}_{H+1}^0 = \boldsymbol{0}$. *Then for all* $t \geq 0$ *and* $1 \leq h \leq H+1$, *one has*

$$\boldsymbol{0} \leq \boldsymbol{Q}_{t,h} \leq (H+1-h)\boldsymbol{1} \qquad \textit{and} \qquad \boldsymbol{0} \leq \boldsymbol{V}_{t,h} \leq (H+1-h)\boldsymbol{1}. \tag{30}$$

Proof. We can use the induction argument to prove this. First, our initialization obeys (30) for $t = 0$ and $h = H+1$. Next, suppose that (30) is true for $t-1$ and $h+1$. By the update rule (19), it is straightforward to compute

$$\boldsymbol{Q}_{t,h} = (1-\eta_t)\boldsymbol{Q}_{t-1,h} + \eta_t(\boldsymbol{r}_h + \boldsymbol{P}_{t,h}\boldsymbol{V}_{t,h+1}) \geq \boldsymbol{0},$$

and

$$\begin{aligned} \boldsymbol{Q}_{t,h} &= (1-\eta_t)\boldsymbol{Q}_{t-1,h} + \eta_t(\boldsymbol{r}_h + \boldsymbol{P}_{t,h}\boldsymbol{V}_{t,h+1}) \\ &\leq (1-\eta_t)\|\boldsymbol{Q}_{t-1,h}\|_\infty \boldsymbol{1} + \eta_t(\|\boldsymbol{r}_h\|_\infty + \|\boldsymbol{P}_{t,h}\|_1\|\boldsymbol{V}_{t,h+1}\|_\infty)\boldsymbol{1} \\ &\leq (1-\eta_t)(H+1-h)\boldsymbol{1} + \eta_t\big(1+(H-h)\big)\boldsymbol{1} = (H+1-h)\boldsymbol{1}. \end{aligned}$$

where we use the facts $\boldsymbol{r}_h \leq \boldsymbol{1}$ and $\boldsymbol{P}_h^t$ is a probability transition kernel. In addition, since $V_h^t(s) := \max_a Q_h^t(s,a)$ for all $t \geq 0$, $1 \leq h \leq H+1$ and $s \in \mathcal{S}$, it is easy to see that $\boldsymbol{0} \leq \boldsymbol{V}_{t,h} \leq (H+1-h)\boldsymbol{1}$. This completes the proof for (30). □

It immediately follows from Lemma 3 that for all $t \geq 0$ and $1 \leq h \leq H+1$,

$$-H\boldsymbol{1} \leq -\boldsymbol{Q}_h^\star \leq \boldsymbol{Q}_{t,h} - \boldsymbol{Q}_h^\star \leq \boldsymbol{Q}_{t,h} \leq H\boldsymbol{1}.$$

Combined with the fact $\|\boldsymbol{Q}^\star\|_\infty \le H$, one further has

$$\max_{t\ge 0,\, 1\le h\le H+1} \|\boldsymbol{Q}_{t,h} - \boldsymbol{Q}^\star\|_\infty \le H.$$

This suggests we can focus on the case where $\varepsilon \le H$ and the claimed iteration number in (27) satisfies

$$T = \frac{c_3 H^4 \log^3 T \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{\min\{\varepsilon^2, \varepsilon\}} \ge \frac{c_3 H^4 \log^3 T \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{\varepsilon} \ge c_3 H^3 \log^3 T \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}. \tag{31}$$

Several facts regarding the learning rates. Next, we present a few useful bounds regarding the learning rates $\eta_i^{(t)}$ defined in the same way as (25). From the assumption (11a) and the bound (31), it is easily seen that the step size obeys

$$\frac{H\log^2 T}{2c_1 T} \le \frac{1}{1+\frac{c_1 T}{H\log^2 T}} \le \eta_t \le \frac{1}{1+\frac{c_2 t}{H\log^2 T}} \le \frac{H\log^2 T}{c_2 t}. \tag{32}$$

Let us set

$$\beta := \frac{c_4}{H} \tag{33}$$

for some sufficiently small constant $c_4 > 0$. In what follows, we present two upper bounds of $\eta_i^{(t)}$ for any $t \ge \frac{T}{c_2 \log H}$.

- For any $0 \le i \le (1-\beta)t$, one know from (32) and $T \ge H^2$ (cf. (11b)) that

$$\eta_i^{(t)} \le \left(1 - \frac{H\log^2 T}{2c_1 T}\right)^{\beta t} \le \left(1 - \frac{H\log^2 T}{2c_1 T}\right)^{\frac{c_4 T}{c_2 H(\log H)}} < \frac{1}{2T^2} \tag{34}$$

  where the last step holds provided $c_1 c_2 \le c_4/4$.

- Turning to $i > (1-\beta)t \ge t/2$, we can use the condition $t \ge \frac{T}{c_2 \log H}$ to bound

$$\eta_i^{(t)} \le \eta_i \le \frac{H\log^2 T}{c_2 i} < \frac{2H\log^2 T}{c_2 t} \le \frac{2H\log^2 T}{T/\log H} \le \frac{2H\log^3 T}{T}. \tag{35}$$

Moreover, the sum of $\eta_i^{(t)}$ continues to satisfy

$$\sum_{i=0}^{t} \eta_i^{(t)} = \prod_{j=1}^{t}(1-\eta_j) + \eta_1 \prod_{j=2}^{t}(1-\eta_j) + \eta_2 \prod_{j=3}^{t}(1-\eta_j) + \cdots + \eta_{t-1}(1-\eta_t) + \eta_t = 1. \tag{36}$$

A.2. Proof of Theorem 3

We are now in the position to prove Theorem 3. For convenience of notation, we shall define

$$\boldsymbol{\Delta}_{t,h} := \boldsymbol{Q}_{t,h} - \boldsymbol{Q}_h^\star.$$

In addition, let π_t denote the policy such that for any state-action-horizon pair (s,a,h),[1]

$$\pi_t(a\,|\,s,h) := \begin{cases} 1, & \text{if } a = \min\left\{a' \,|\, Q_{t,h}(s,a') = \max_{a''} Q_{t,h}(s,a'')\right\}, \\ 0, & \text{else.} \end{cases} \tag{37}$$

Namely, for any $s \in \mathcal{S}$ and $1 \le h \le H+1$, the policy π_t chooses the smallest indexed action that achieves the largest Q-value in the estimate $Q_{t,h}(s,\cdot)$. It immediately follows that

$$Q_{t,h}\big(s, \pi_t(s,h)\big) = V_{t,h}(s) \qquad \text{and} \qquad \boldsymbol{P}_h \boldsymbol{V}_{t,h+1} = \boldsymbol{P}_h^{\pi_t} \boldsymbol{Q}_{t,h+1} \ge \boldsymbol{P}_h^{\pi} \boldsymbol{Q}_{t,h+1} \quad \text{for any } \pi, \tag{38}$$

where $\boldsymbol{P}^\pi$ is defined in (14).

[1] If there is only a single action that satisfies $Q_{t,h}(s,a') = \max_{a''} Q_{t,h}(s,a'')$, then $\pi_t(a\,|\,s,h) = 1$ if and only if $a = \arg\min_{a'} Q_{t,h}(s,a')$ and 0 otherwise.

A.2.1. KEY DECOMPOSITION

We first make the following elementary decomposition:

$$\begin{aligned}\boldsymbol{\Delta}_{t,h} &= \boldsymbol{Q}_{t,h} - \boldsymbol{Q}^\star_h = (1-\eta_t)\boldsymbol{Q}_{t-1,h} + \eta_t(\boldsymbol{r}_h + \boldsymbol{P}_{t,h}\boldsymbol{V}_{t,h+1}) - \boldsymbol{Q}^\star_h \\ &= (1-\eta_t)(\boldsymbol{Q}_{t-1,h} - \boldsymbol{Q}^\star_h) + \eta_t(\boldsymbol{r}_h + \boldsymbol{P}_{t,h}\boldsymbol{V}_{t,h+1} - \boldsymbol{Q}^\star_h) \\ &= (1-\eta_t)\boldsymbol{\Delta}_{t-1,h} + \eta_t(\boldsymbol{P}_{t,h}\boldsymbol{V}_{t,h+1} - \boldsymbol{P}_h\boldsymbol{V}^\star_{h+1}) \\ &= (1-\eta_t)\boldsymbol{\Delta}_{t-1,h} + \eta_t\big\{\boldsymbol{P}_h(\boldsymbol{V}_{t,h+1} - \boldsymbol{V}^\star_{h+1}) + (\boldsymbol{P}_{t,h} - \boldsymbol{P}_h)\boldsymbol{V}_{t,h+1}\big\}. \end{aligned} \tag{39}$$

Similar to (22), one can use (38) to control the quantity $\boldsymbol{P}_h(\boldsymbol{V}_{t,h+1} - \boldsymbol{V}^\star_{h+1})$ by

$$\boldsymbol{P}_h(\boldsymbol{V}_{t,h+1} - \boldsymbol{V}^\star_{h+1}) = \boldsymbol{P}^{\pi_t}_h\boldsymbol{Q}_{t,h+1} - \boldsymbol{P}^{\pi^\star}_h\boldsymbol{Q}^\star_{h+1} \le \boldsymbol{P}^{\pi_t}_h\boldsymbol{Q}_{t,h+1} - \boldsymbol{P}^{\pi_t}_h\boldsymbol{Q}^\star_{h+1} = \boldsymbol{P}^{\pi_t}_h\boldsymbol{\Delta}_{t,h+1}, \tag{40a}$$

$$\boldsymbol{P}_h(\boldsymbol{V}_{t,h+1} - \boldsymbol{V}^\star_{h+1}) = \boldsymbol{P}^{\pi_t}_h\boldsymbol{Q}_{t,h+1} - \boldsymbol{P}^{\pi^\star}_h\boldsymbol{Q}^\star_{h+1} \ge \boldsymbol{P}^{\pi^\star}_h\boldsymbol{Q}_{t,h+1} - \boldsymbol{P}^{\pi^\star}_h\boldsymbol{Q}^\star_{h+1} = \boldsymbol{P}^{\pi^\star}_h\boldsymbol{\Delta}_{t,h+1}, \tag{40b}$$

Combining (40) with (39) yields

$$\begin{aligned}\boldsymbol{\Delta}_{t,h} &\le (1-\eta_t)\boldsymbol{\Delta}_{t-1,h} + \eta_t\big\{\boldsymbol{P}^{\pi_t}_h\boldsymbol{\Delta}_{t,h+1} + (\boldsymbol{P}_{t,h} - \boldsymbol{P}_h)\boldsymbol{V}_{t,h+1}\big\}; \\ \boldsymbol{\Delta}_{t,h} &\ge (1-\eta_t)\boldsymbol{\Delta}_{t-1,h} + \eta_t\big\{\boldsymbol{P}^{\pi^\star}_h\boldsymbol{\Delta}_{t,h+1} + (\boldsymbol{P}_{t,h} - \boldsymbol{P}_h)\boldsymbol{V}_{t,h+1}\big\}.\end{aligned}$$

We can then apply this relation recursively to reach

$$\boldsymbol{\Delta}_{t,h} \le \eta^{(t)}_0\boldsymbol{\Delta}_{0,h} + \sum_{i=1}^{t}\eta^{(t)}_i(\boldsymbol{P}^{\pi_i}_h\boldsymbol{\Delta}_{i,h+1} + (\boldsymbol{P}_{i,h} - \boldsymbol{P}_h)\boldsymbol{V}_{i,h+1}), \tag{41a}$$

$$\boldsymbol{\Delta}_{t,h} \ge \eta^{(t)}_0\boldsymbol{\Delta}_{0,h} + \sum_{i=1}^{t}\eta^{(t)}_i(\boldsymbol{P}^{\pi^\star}_h\boldsymbol{\Delta}_{i,h+1} + (\boldsymbol{P}_{i,h} - \boldsymbol{P}_h)\boldsymbol{V}_{i,h+1}). \tag{41b}$$

In the following, we shall use (41) to upper and lower bound $\boldsymbol{\Delta}_{t,h}$ individually.

A.2.2. UPPER BOUNDING $\boldsymbol{\Delta}_{t,h}$

Let us first upper bound $\boldsymbol{\Delta}_{t,h}$ for $t \ge \frac{T}{c_2\log H}$. In view of (41a), we further decompose its right-hand side as

$$\begin{aligned}\boldsymbol{\Delta}_{t,h} \le\ & \underbrace{\eta^{(t)}_0\boldsymbol{\Delta}_{0,h} + \sum_{i=1}^{(1-\beta)t}\eta^{(t)}_i\big(\boldsymbol{P}^{\pi_i}_h\boldsymbol{\Delta}_{i,h+1} + (\boldsymbol{P}_{i,h} - \boldsymbol{P}_h)\boldsymbol{V}_{i,h+1}\big)}_{=:\,\boldsymbol{\zeta}_{t,h}} \\ &+ \underbrace{\sum_{i=(1-\beta)t+1}^{t}\eta^{(t)}_i(\boldsymbol{P}_{i,h} - \boldsymbol{P}_h)\boldsymbol{V}_{i,h+1}}_{=:\,\boldsymbol{\xi}_{t,h}} + \sum_{i=(1-\beta)t+1}^{t}\eta^{(t)}_i\boldsymbol{P}^{\pi_i}_h\boldsymbol{\Delta}_{i,h+1}\end{aligned} \tag{42}$$

where we recall that $\beta := \frac{c_4}{H}$ defined in (33).

Step 1: bounding $\boldsymbol{\zeta}_{t,h}$. From the upper bounds (34) for $\eta^{(t)}_i$, it is straightforward to control $\boldsymbol{\zeta}_{t,h}$ as follows:

$$\begin{aligned}\|\boldsymbol{\zeta}_{t,h}\|_\infty &\le \eta^{(t)}_0\|\boldsymbol{\Delta}_{0,h}\|_\infty + t\max_{i\le(1-\beta)t}\eta^{(t)}_i\max_{1\le i\le(1-\beta)t}\big(\|\boldsymbol{P}^{\pi_i}_h\boldsymbol{\Delta}_{i,h+1}\|_\infty + \|\boldsymbol{P}_{i,h}\boldsymbol{V}_{i,h+1}\|_\infty + \|\boldsymbol{P}_h\boldsymbol{V}_{i,h+1}\|_\infty\big) \\ &\le \eta^{(t)}_0\|\boldsymbol{\Delta}_{0,h}\|_\infty + t\max_{i\le(1-\beta)t}\eta^{(t)}_i\max_{1\le i\le(1-\beta)t}\Big\{\|\boldsymbol{P}^{\pi_i}_h\|_1\|\boldsymbol{\Delta}_{i,h+1}\|_\infty + \big(\|\boldsymbol{P}_{i,h}\|_1 + \|\boldsymbol{P}_h\|_1\big)\|\boldsymbol{V}_{i,h+1}\|_\infty\Big\} \\ &\overset{\text{(i)}}{=} \eta^{(t)}_0\|\boldsymbol{\Delta}_{0,h}\|_\infty + t\max_{i\le(1-\beta)t}\eta^{(t)}_i\max_{1\le i\le(1-\beta)t}\big(\|\boldsymbol{\Delta}_{i,h+1}\|_\infty + 2\,\|\boldsymbol{V}_{i,h+1}\|_\infty\big) \\ &\overset{\text{(ii)}}{\le} \frac{1}{2T^2}\cdot H + \frac{1}{2T^2}\cdot t\cdot 3H \\ &\le \frac{2H}{T}.\end{aligned}$$

Here, (i) relies on $\|\boldsymbol{P}^{\pi_i}_h\|_1 = \|\boldsymbol{P}_{i,h}\|_1 = \|\boldsymbol{P}_h\|_1 = 1$ since they are all probability transition matrices; (ii) holds due to (34).

Step 2: bounding $\boldsymbol{\xi}_{t,h}$. Observe that $\boldsymbol{\xi}_{t,h}$ is a sum of martingale differences, namely

$$\boldsymbol{\xi}_{t,h} = \sum_{i=(1-\beta)t+1}^{t} \boldsymbol{z}_{i,h} \qquad \text{with } \boldsymbol{z}_{i,h} \coloneqq \eta_i^{(t)}(\boldsymbol{P}_{i,h} - \boldsymbol{P}_h)\boldsymbol{V}_{i,h+1},$$

where the $\boldsymbol{z}_{i,h}$'s satisfy

$$\mathbb{E}\left[\boldsymbol{z}_{i,h} \,|\, \boldsymbol{V}_{i,h+1}, \cdots, \boldsymbol{V}_{0,h+1}\right] = \boldsymbol{0}.$$

This suggests we can invoke Freedman's inequality (see Lemma 4) to control $\boldsymbol{\xi}_{t,h}$ for any t such that $\frac{T}{c_2 \log H} \le t \le T$.

- First, it is straightforward to bound

$$\begin{aligned} B &\coloneqq \max_{(1-\beta)t<i\le t} \|\boldsymbol{z}_{i,h}\|_\infty \le \max_{(1-\beta)t<i\le t} \|\eta_i^{(t)}(\boldsymbol{P}_{i,h} - \boldsymbol{P}_h)\boldsymbol{V}_{i,h+1}\|_\infty \\ &\le \max_{(1-\beta)t<i\le t} \eta_i^{(t)}\big(\|\boldsymbol{P}_{i,h}\|_1 + \|\boldsymbol{P}_h\|_1\big)\|\boldsymbol{V}_{i,h+1}\|_\infty \le \frac{4H^2\log^3 T}{T}. \end{aligned} \tag{43}$$

  where the last step arises from (35), Lemma 3, and the fact $\|\boldsymbol{P}_{i,h}\|_1 = \|\boldsymbol{P}_h\|_1 = 1$.

- Next, recall the notation $\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{z})$ in (16). One can compute

$$\begin{aligned} \boldsymbol{W}_t &\coloneqq \sum_{i=(1-\beta)t+1}^{t} \mathsf{Var}\big(\boldsymbol{z}_{i,h} \,|\, \boldsymbol{V}_{i,h+1}, \cdots, \boldsymbol{V}_{0,h+1}\big) = \sum_{i=(1-\beta)t+1}^{t} \big(\eta_i^{(t)}\big)^2 \mathsf{Var}_{\boldsymbol{P}_h}\big(\boldsymbol{V}_{i,h+1}\big) \\ &\le \Big(\max_{(1-\beta)t\le i\le t} \eta_i^{(t)}\Big)\Big(\sum_{i=(1-\beta)t+1}^{t} \eta_i^{(t)}\Big) \max_{(1-\beta)t\le i<t} \mathsf{Var}_{\boldsymbol{P}_h}\big(\boldsymbol{V}_{i,h+1}\big) \\ &\le \frac{2H\log^3 T}{T} \max_{(1-\beta)t\le i<t} \mathsf{Var}_{\boldsymbol{P}_h}\big(\boldsymbol{V}_{i,h+1}\big), \end{aligned} \tag{44}$$

  where the last inequality relies on (35) and (36).

- Additionally, we can use Lemma 3 to further bound $\boldsymbol{W}_t$

$$\big|\boldsymbol{W}_t\big| \le \frac{2H\log^3 T}{T} \cdot H^2\boldsymbol{1} = \frac{2H^3\log^3 T}{T}\boldsymbol{1} =: \sigma^2\boldsymbol{1}.$$

  In particular, we know that

$$\frac{\sigma^2}{2^K} \le \frac{2H\log^3 T}{T} \tag{45}$$

  where $K \coloneqq \lceil 2\log H \rceil$.

With the above bounds in place, we apply the Freedman inequality in Lemma 4 and the union bound to find: with probability at least $1 - \frac{\delta}{TH}$,

$$\begin{aligned} |\boldsymbol{\xi}_{t,h}| &\le \sqrt{8\Big(\boldsymbol{W}_t + \frac{\sigma^2}{2^K}\boldsymbol{1}\Big)\log\frac{|\mathcal{S}||\mathcal{A}|THK}{\delta}} + \Big(\frac{4}{3}B\log\frac{|\mathcal{S}||\mathcal{A}|THK}{\delta}\Big)\cdot\boldsymbol{1} \\ &\overset{\text{(i)}}{\le} \sqrt{16\Big(\boldsymbol{W}_t + \frac{2H\log^3 T}{T}\boldsymbol{1}\Big)\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}} + \Big(3B\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\Big)\cdot\boldsymbol{1} \\ &\overset{\text{(ii)}}{\le} \sqrt{\frac{32H\big(\log^3 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{T}\Big(\max_{(1-\beta)t\le i<t} \mathsf{Var}_{\boldsymbol{P}_h}\big(\boldsymbol{V}_{i,h+1}\big) + \boldsymbol{1}\Big)} + \frac{12H^2\big(\log^3 T\big)\big(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{T}\boldsymbol{1} \end{aligned}$$

Here, (i) arises from (45) and $\log\frac{|\mathcal{S}||\mathcal{A}|THK}{\delta} \le 2\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}$ (which holds due to (31)); (ii) makes use of the relation (43) and (44).

Step 3: using the bounds on $\boldsymbol{\zeta}_{t,h}$ and $\boldsymbol{\xi}_{t,h}$ to control $\boldsymbol{\Delta}_{t,h}$. Let us define

$$\boldsymbol{\varphi}_{t,h} \coloneqq c_5 \frac{H \log^3 T \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T} \left(\max_{\frac{t}{2} \le i < t} \mathsf{Var}_{\boldsymbol{P}_h}(\boldsymbol{V}_{i,h+1}) + \mathbf{1} \right) \tag{46}$$

for some sufficiently large constant $c_5 > 0$. In view of bounds for $\boldsymbol{\zeta}_{t,h}$ and $\boldsymbol{\xi}_{t,h}$, the following holds with probability exceeding $1-\delta$: for all $\frac{2t}{3} \le k \le t$ and $1 \le h \le H$,

$$|\boldsymbol{\zeta}_{k,h}| + |\boldsymbol{\xi}_{k,h}| \le \sqrt{\boldsymbol{\varphi}_{t,h}}. \tag{47}$$

Inserting (47) into (42) reveals

$$\boldsymbol{\Delta}_{t,h} \le \sqrt{\boldsymbol{\varphi}_{t,h}} + \sum_{i=(1-\beta)t+1}^{t} \eta_i^{(t)} \boldsymbol{P}_h^{\pi_i} \boldsymbol{\Delta}_{i,h+1}. \tag{48}$$

We now define a sequence $\left\{\alpha_i^{(t)}\right\}_i$ as follows

$$\alpha_i^{(t)} := \frac{\eta_i^{(t)}}{\sum_{j=(1-\beta)t+1}^{t} \eta_j^{(t)}}, \qquad 0 \le i \le t.$$

It is easy to check that for any t, the sequence $\left\{\alpha_i^{(t)}\right\}_i$ satisfies

$$\alpha_i^{(t)} \ge \eta_i^{(t)} \qquad \text{and} \qquad \sum_{i=(1-\beta)t+1}^{t} \alpha_i^{(t)} = 1 \tag{49}$$

where the first inequality results from (36). This enables us to rewrite (48) as

$$\boldsymbol{\Delta}_{k,h} \le \sqrt{\boldsymbol{\varphi}_{t,h}} + \sum_{i_h=(1-\beta)k+1}^{k} \eta_{i_1}^{(k)} \boldsymbol{P}_h^{\pi_{i_1}} \boldsymbol{\Delta}_{i_1,h+1} = \sum_{i_h=(1-\beta)k+1}^{k} \left(\alpha_{i_h}^{(k)} \sqrt{\boldsymbol{\varphi}_{t,h}} + \eta_{i_h}^{(k)} \boldsymbol{P}_h^{\pi_{i_h}} \boldsymbol{\Delta}_{i_h,h+1} \right). \tag{50}$$

for all $2t/3 \le k \le t$. By the definition of β (cf. (33)), one has $(1-\beta)t \ge 2t/3$. We can then exploit this relation recursively to obtain

$$\begin{aligned}
\boldsymbol{\Delta}_{t,h} &\le \sum_{i_h=(1-\beta)t+1}^{t} \left(\alpha_{i_h}^{(t)} \sqrt{\boldsymbol{\varphi}_{t,h}} + \eta_{i_h}^{(t)} \boldsymbol{P}_h^{\pi_{i_h}} \boldsymbol{\Delta}_{i_h,h+1} \right) \\
&\le \sum_{i_h=(1-\beta)t+1}^{t} \left\{ \alpha_{i_h}^{(t)} \sqrt{\boldsymbol{\varphi}_{t,h}} + \eta_{i_h}^{(t)} \boldsymbol{P}_h^{\pi_{i_h}} \sum_{i_{h+1}=(1-\beta)i_h+1}^{i_h} \left(\alpha_{i_{h+1}}^{(i_h)} \sqrt{\boldsymbol{\varphi}_{t,h+1}} + \eta_{i_{h+1}}^{(i_h)} \boldsymbol{P}_{h+1}^{\pi_{i_{h+1}}} \boldsymbol{\Delta}_{i_{h+1},h+2} \right) \right\} \\
&\overset{\text{(i)}}{\le} \sum_{i_h=(1-\beta)t+1}^{t} \alpha_{i_h}^{(t)} \sqrt{\boldsymbol{\varphi}_{t,h}} + \sum_{i_h=(1-\beta)t+1}^{t} \sum_{i_{h+1}=(1-\beta)i_h+1}^{i_h} \alpha_{i_h}^{(t)} \alpha_{i_{h+1}}^{(i_h)} \boldsymbol{P}_h^{\pi_{i_h}} \sqrt{\boldsymbol{\varphi}_{t,h+1}} \\
&\qquad + \sum_{i_h=(1-\beta)t+1}^{t} \sum_{i_{h+1}=(1-\beta)i_h+1}^{i_h} \eta_{i_h}^{(t)} \eta_{i_{h+1}}^{(i_h)} \prod_{k=h}^{h+1} \boldsymbol{P}_k^{\pi_{i_k}} \boldsymbol{\Delta}_{i_{h+1},h+2} \\
&\overset{\text{(ii)}}{=} \sum_{i_h=(1-\beta)t+1}^{t} \sum_{i_{h+1}=(1-\beta)i_h+1}^{i_h} \alpha_{i_h}^{(t)} \alpha_{i_{h+1}}^{(i_h)} \left\{ \sqrt{\boldsymbol{\varphi}_{t,h}} + \boldsymbol{P}_h^{\pi_{i_h}} \sqrt{\boldsymbol{\varphi}_{t,h+1}} \right\} \\
&\qquad + \sum_{i_h=(1-\beta)t+1}^{t} \sum_{i_{h+1}=(1-\beta)i_h+1}^{i_h} \eta_{i_h}^{(t)} \eta_{i_{h+1}}^{(i_h)} \prod_{k=h}^{h+1} \boldsymbol{P}_k^{\pi_{i_k}} \boldsymbol{\Delta}_{i_{h+1},h+2},
\end{aligned} \tag{51}$$

where (i) relies on $\eta_{i_h}^{(t)} \le \alpha_{i_h}^{(t)}$ in (49), and (ii) holds since $\sum_{i_{h+1}=(1-\beta)i_h+1}^{i_h} \alpha_{i_{h+1}}^{(i_h)} = 1$ by (49).

Our proof strategy is applying (50) recursively to control $\boldsymbol{\Delta}_{t,h}$ for all $1 \le h \le H$. Towards this, we need some preparation beforehand. First, let us define

$$\alpha_{\{i_k\}_{k=h}^H} := \alpha_{i_h}^{(t)} \alpha_{i_{h+1}}^{(i_h)} \cdots \alpha_{i_H}^{(i_{H-1})} \ge 0, \qquad 1 \le h \le H \tag{52}$$

for any $t \ge i_h \ge i_{h+1} \ge \cdots \ge i_H$. By (49), one has

$$\alpha_{\{i_k\}_{k=h}^H} \ge \eta_{i_h}^{(t)} \eta_{i_{h+1}}^{(i_h)} \cdots \eta_{i_H}^{(i_{H-1})}. \tag{53}$$

Next, let us define the index set

$$\mathcal{I}_{t,h} := \Big\{ (i_h, \cdots, i_H) \,|\, (1-\beta)t \le i_h \le t-1,\ (1-\beta) i_j \le i_{j+1} \le i_j - 1,\ \forall h \le j < H \Big\}, \tag{54}$$

which satisfies

$$\sum_{(i_h,\cdots,i_H)\in\mathcal{I}_{t,h}} \alpha_{\{i_k\}_{k=h}^H} = 1. \tag{55}$$

In addition, as $\beta := c_4/H$ for some sufficiently small constant $c_4 > 0$, we know that

$$(1-\beta)^H = \Big(1 - \frac{c_4}{H}\Big)^H \ge \frac{2}{3},$$

and consequently

$$i_h \ge i_{h+1} \ge \cdots \ge i_H \ge (1-\beta)^H t \ge 2t/3 \qquad \text{for all } (i_h, \cdots, i_H) \in \mathcal{I}_{t,h}.$$

With these in place, we now invoke the relation (50) in a recursive manner to obtain

$$\begin{aligned} \boldsymbol{\Delta}_{t,h} &\le \sum_{(i_h,\cdots,i_H)\in\mathcal{I}_{t,h}} \alpha_{\{i_k\}_{k=h}^H} \Big\{ \sqrt{\boldsymbol{\varphi}_{t,h}} + \sum_{j=h+1}^{H} \prod_{k=h}^{j-1} \boldsymbol{P}_k^{\pi_{i_k}} \sqrt{\boldsymbol{\varphi}_{t,j}} \Big\} \\ &\le \max_{(i_h,\cdots,i_H)\in\mathcal{I}_{t,h}} \Big\{ \sum_{j=h}^{H} \prod_{k=h}^{j-1} \boldsymbol{P}_k^{\pi_{i_k}} \sqrt{\boldsymbol{\varphi}_{t,j}} \Big\} \end{aligned} \tag{56}$$

for all $t \ge \frac{T}{c_2 \log \frac{1}{1-\gamma}}$. Here, the last step arises from the fact that $\sum_{(i_h,\cdots,i_H)\in\mathcal{I}_{t,h}} \alpha_{\{i_k\}_{k=h}^H} = 1$ (cf. (55)). One can upper bound the entrywise square of the quantity in curly braces as follows

$$\begin{aligned} \Big| \sum_{j=h}^{H} \prod_{k=h}^{j-1} \boldsymbol{P}_k^{\pi_{i_k}} \sqrt{\boldsymbol{\varphi}_{t,j}} \Big|^2 &\overset{\text{(i)}}{\le} \Big| \sum_{j=h}^{H} \sqrt{\prod_{k=h}^{j-1} \boldsymbol{P}_k^{\pi_{i_k}} \boldsymbol{\varphi}_{t,j}} \Big|^2 \overset{\text{(ii)}}{\le} H \sum_{j=h}^{H} \prod_{k=h}^{j-1} \boldsymbol{P}_k^{\pi_{i_k}} \boldsymbol{\varphi}_{t,j} \\ &\overset{\text{(iii)}}{\lesssim} H \sum_{j=h}^{H} \prod_{k=h}^{j-1} \boldsymbol{P}_k^{\pi_{i_k}} \frac{H \log^3 T \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{T} \Big(\max_{\frac{t}{2} \le i < t} \mathsf{Var}_{\boldsymbol{P}_j}(\boldsymbol{V}_{i,j+1}) + 1 \Big) \\ &\overset{\text{(iv)}}{\le} \frac{H^2 \big(\log^3 T\big) \big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{T} \sum_{j=h}^{H} \prod_{k=h}^{j-1} \boldsymbol{P}_k^{\pi_{i_k}} \max_{\frac{t}{2} \le i < t} \mathsf{Var}_{\boldsymbol{P}_j}(\boldsymbol{V}_{i,j+1}) + \frac{H^3 \big(\log^3 T\big) \big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{T} \mathbf{1} \\ &\overset{\text{(v)}}{\lesssim} \frac{H^4 \big(\log^3 T\big) \big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{T} \Big(1 + \max_{\frac{t}{2} \le i \le t, k > h} \|\boldsymbol{\Delta}_{i,k}\|_\infty \Big) \mathbf{1}. \end{aligned} \tag{57}$$

where (i) arises from the Jensen inequality and the fact $\prod_{k=h}^{j-1} \boldsymbol{P}_k^{\pi_{i_k}}$ is a probability transition matrix; (ii) relies on the Cauchy-Schwarz inequality; (iii) is due to the definition of $\boldsymbol{\varphi}_{t,h}$; (iv) holds since $\prod_{k=h}^{j-1} \boldsymbol{P}_k^{\pi_{i_k}} \mathbf{1} = \mathbf{1}$; (v) is valid as long as Lemma 4 holds.

Lemma 4. *One has*

$$\sum_{j=h}^{H} \prod_{k=h}^{j-1} \boldsymbol{P}_k^{\pi_{i_k}} \max_{\frac{t}{2} \le i < t} \mathsf{Var}_{\boldsymbol{P}_j}(\boldsymbol{V}_{i,j+1}) \lesssim H^2 \Big(1 + \max_{\frac{t}{2} \le i \le t,\, k > h} \|\boldsymbol{\Delta}_{i,k}\|_\infty \Big) \mathbf{1}. \tag{58}$$

Proof. See Section A.2.5. □

Plugging (57) back into (56) reveals that the following holds simultaneously for all $t \geq \frac{T}{c_2 \log H}$ with probability at least $1-\delta$:

$$\boldsymbol{\Delta}_{t,h} \lesssim \sqrt{\frac{H^4\big(\log^3 T\big)\big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{T}\Big(1 + \max_{\frac{t}{2} \leq i \leq t, k>h} \|\boldsymbol{\Delta}_{i,k}\|_\infty\Big)}\, \mathbf{1}. \tag{59}$$

A.2.3. Lower bounding $\boldsymbol{\Delta}_{t,h}$

In this section, we proceed to lower bound $\boldsymbol{\Delta}_{t,h}$. Invoking a similar argument for (56) and replacing π_i with $\pi^\star$, we can derive

$$\boldsymbol{\Delta}_{t,h} \geq -\max_{(i_h,\cdots,i_H)\in\mathcal{I}_{t,h}} \Big\{\sum_{j=h}^{H}\prod_{k=h}^{j-1} \boldsymbol{P}_k^{\pi^\star}\sqrt{\boldsymbol{\varphi}_{t,j}}\Big\}.$$

One can further apply an analogous argument for (57) to bound the right-hand side as follows

$$\Big|\sum_{j=h}^{H}\prod_{k=h}^{j-1} \boldsymbol{P}_k^{\pi^\star}\sqrt{\boldsymbol{\varphi}_{t,j}}\Big|^2 \lesssim \frac{H^4\big(\log^3 T\big)\big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{T}\Big(1 + \max_{\frac{t}{2} \leq i \leq t, k>h} \|\boldsymbol{\Delta}_{i,k}\|_\infty\Big)\mathbf{1}.$$

Consequently, we find that with probability at least $1-\delta$,

$$\boldsymbol{\Delta}_{t,h} \gtrsim -\sqrt{\frac{H^4\big(\log^3 T\big)\big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{T}\Big(1 + \max_{\frac{t}{2} \leq i \leq t, k>h} \|\boldsymbol{\Delta}_{i,k}\|_\infty\Big)}\, \mathbf{1} \tag{60}$$

holds simultaneously for all $t \geq \frac{T}{c_2 \log H}$.

A.2.4. Combining our upper and lower bounds on $\boldsymbol{\Delta}_{t,h}$

Taking (59) and (60) together, we know that with probability greater than $1-2\delta$,

$$\|\boldsymbol{\Delta}_{t,h}\|_\infty \lesssim \sqrt{\frac{H^4\big(\log^3 T\big)\big(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\big)}{T}\Big(1 + \max_{\frac{t}{2} \leq i \leq t, k>h} \|\boldsymbol{\Delta}_{i,k}\|_\infty\Big)}, \tag{61}$$

holds simultaneously for all $t \geq \frac{T}{c_2 \log H}$. As a result, the claim in Theorem 2 immediately follows from the same argument for the infinite-horizon case in Appendix 4.2.3, which we omit for the sake of conciseness.

A.2.5. Proof of Lemma 4

We shall invoke a similar argument for Li et al. (2021a, Lemma 5) to establish Lemma 4. For the sake of conciseness, we omit some details of proof.

To begin with, we can argue analogously as for Li et al. (2021a, (64)) to show that for any $1 \leq j \leq H$,

$$\max_{\frac{t}{2} \leq i < t} \mathsf{Var}_{\boldsymbol{P}_j}(\boldsymbol{V}_{i,j+1}) - \mathsf{Var}_{\boldsymbol{P}_j}(\boldsymbol{V}^\star_{j+1}) \leq 4H \max_{\frac{t}{2} \leq i < t} \|\boldsymbol{\Delta}_{i,j+1}\|_\infty \mathbf{1}. \tag{62}$$

As a consequence, one can bound the left-hand side of (58) by

$$\sum_{j=h}^{H}\prod_{k=h}^{j-1} \boldsymbol{P}_k^{\pi_{i_k}} \max_{\frac{t}{2} \leq i < t} \mathsf{Var}_{\boldsymbol{P}_j}(\boldsymbol{V}_{i,j+1}) \leq \sum_{j=h}^{H}\prod_{k=h}^{j-1} \boldsymbol{P}_k^{\pi_{i_k}} \mathsf{Var}_{\boldsymbol{P}_j}(\boldsymbol{V}^\star_{j+1}) + 4H^2 \max_{\frac{t}{2} \leq i \leq t, k>h} \|\boldsymbol{\Delta}_{i,k}\|_\infty \mathbf{1}. \tag{63}$$

Now it remains to control the first term on the right-hand side of (63). Towards this, we can first bound similarly as in Li et al. (2021a, (67)) to obtain

$$\begin{aligned}\mathsf{Var}_{\boldsymbol{P}_j}(\boldsymbol{V}_{j+1}^\star) &= \boldsymbol{P}_j(\boldsymbol{V}_{j+1}^\star \circ \boldsymbol{V}_{j+1}^\star) - (\boldsymbol{P}_j \boldsymbol{V}_{j+1}^\star)\circ(\boldsymbol{P}_j \boldsymbol{V}_{j+1}^\star) \\ &\le \big(\boldsymbol{P}_j^{\pi_{i_j}}(\boldsymbol{Q}_{j+1}^\star \circ \boldsymbol{Q}_{j+1}^\star) - \boldsymbol{Q}_j^\star \circ \boldsymbol{Q}_j^\star\big) + 2\boldsymbol{Q}_j^\star \circ \boldsymbol{r}_j + 4H \max_{\frac{t}{2}\le i<t} \|\boldsymbol{\Delta}_{i,j+1}\|_\infty \mathbf{1}. \end{aligned} \tag{64}$$

With the estimate for $\mathsf{Var}_{\boldsymbol{P}_j}(\boldsymbol{V}_{j+1}^\star)$ in place, one can invoke the same argument for Li et al. (2021a, (68)) to reach

$$\begin{aligned}\sum_{j=h}^{H}\prod_{k=h}^{j-1} \boldsymbol{P}_k^{\pi_{i_k}} \mathsf{Var}_{\boldsymbol{P}_j}(\boldsymbol{V}_{j+1}^\star) &\le \sum_{j=h}^{H}\prod_{k=h}^{j-1} \boldsymbol{P}_k^{\pi_{i_k}} \big(\boldsymbol{P}_j^{\pi_{i_j}}(\boldsymbol{Q}_{j+1}^\star \circ \boldsymbol{Q}_{j+1}^\star) - \boldsymbol{Q}_j^\star \circ \boldsymbol{Q}_j^\star\big) \\ &\quad + \sum_{j=h}^{H}\prod_{k=h}^{j-1} \boldsymbol{P}_k^{\pi_{i_k}} \big(2\boldsymbol{Q}_j^\star \circ \boldsymbol{r}_j + 4H \max_{\frac{t}{2}\le i<t} \|\boldsymbol{\Delta}_{i,j+1}\|_\infty \mathbf{1}\big) \\ &\le 4H^2\big(1 + 4 \max_{\frac{t}{2}\le i\le t, k>h} \|\boldsymbol{\Delta}_{i,k}\|_\infty\big)\mathbf{1}. \end{aligned}$$

Plugging the above bounds back into (64) immediately establishes the claimed bound (58).

B. Freedman's inequality

The analysis of this work relies heavily on Freedman's inequality (Freedman, 1975), which is an extension of the Bernstein inequality and allows one to establish concentration results for martingales. For ease of presentation, we include a user-friendly version of Freedman's inequality as follows.

Theorem 4. *Suppose that $Y_n = \sum_{k=1}^n X_k \in \mathbb{R}$, where $\{X_k\}$ is a real-valued scalar sequence obeying*

$$|X_k| \le R \qquad \text{and} \qquad \mathbb{E}\Big[X_k \mid \{X_j\}_{j:j<k}\Big] = 0 \qquad \text{for all } k \ge 1.$$

Define

$$W_n \coloneqq \sum_{k=1}^{n} \mathbb{E}_{k-1}\left[X_k^2\right],$$

where we write $\mathbb{E}_{k-1}$ for the expectation conditional on $\{X_j\}_{j:j<k}$. Then for any given $\sigma^2 \ge 0$, one has

$$\mathbb{P}\left\{|Y_n| \ge \tau \text{ and } W_n \le \sigma^2\right\} \le 2\exp\left(-\frac{\tau^2/2}{\sigma^2 + R\tau/3}\right). \tag{65}$$

In addition, suppose that $W_n \le \sigma^2$ holds deterministically. For any positive integer $K \ge 1$, with probability at least $1-\delta$ one has

$$|Y_n| \le \sqrt{8\max\left\{W_n, \frac{\sigma^2}{2^K}\right\}\log\frac{2K}{\delta}} + \frac{4}{3}R\log\frac{2K}{\delta}. \tag{66}$$

Proof. See (Freedman, 1975; Tropp, 2011) for the proof of (65). As an immediate consequence of (65), one has

$$\mathbb{P}\left\{|Y_n| \ge \sqrt{4\sigma^2\log\frac{2}{\delta}} + \frac{4}{3}R\log\frac{2}{\delta} \text{ and } W_n \le \sigma^2\right\} \le \delta. \tag{67}$$

Next, we turn attention to (66). Consider any positive integer K. As can be easily seen, the event

$$\mathcal{H}_K \coloneqq \left\{|Y_n| \ge \sqrt{8\max\left\{W_n, \frac{\sigma^2}{2^K}\right\}\log\frac{2K}{\delta}} + \frac{4}{3}R\log\frac{2K}{\delta}\right\}$$

is contained within the union of the following K events

$$\mathcal{H}_K \subseteq \bigcup_{0 \le k < K} \mathcal{B}_k,$$

where we define

$$\mathcal{B}_k \coloneqq \left\{ |Y_n| \ge \sqrt{\frac{4\sigma^2}{2^{k-1}} \log \frac{2K}{\delta}} + \frac{4}{3} R \log \frac{2K}{\delta} \text{ and } \frac{\sigma^2}{2^k} \le W_n \le \frac{\sigma^2}{2^{k-1}} \right\}, \qquad 1 \le k \le K-1,$$

$$\mathcal{B}_0 \coloneqq \left\{ |Y_n| \ge \sqrt{\frac{4\sigma^2}{2^{K-1}} \log \frac{2K}{\delta}} + \frac{4}{3} R \log \frac{2K}{\delta} \text{ and } W_n \le \frac{\sigma^2}{2^{K-1}} \right\}.$$

Invoking inequality (67) with σ^2 set to be $\frac{\sigma^2}{2^{k-1}}$ and δ set to be $\frac{\delta}{K}$, we arrive at $\mathbb{P}\{\mathcal{B}_k\} \le \delta/K$. Taken this fact together with the union bound gives

$$\mathbb{P}\{\mathcal{H}_K\} \le \sum_{k=0}^{K-1} \mathbb{P}\{\mathcal{B}_k\} \le \delta.$$

This concludes the proof. □