Supplementary Materials

A More Related Work

In this section, we review a broader set of related work to our study here.

Theory on double Q-learning: Double Q-learning was proposed and proved to converge asymptotically in Hasselt (2010). In Weng et al. (2020c), the authors explored the properties of mean-square errors for double Q-learning both in the tabular case and with linear function approximation, under the assumption that a unique optimal policy exists and the algorithm can converge. The most relevant work to this paper is Xiong et al. (2020), which established the first finite-time convergence rate for tabular double Q-learning with a polynomial learning rate. This paper provides sharper finite-time convergence bounds for double Q-learning, which requires a different analysis approach.

Theory on tabular Q-learning: Proposed in Watkins & Dayan (1992) under finite state-action space, Q-learning has aroused great interest in its theoretical study. Its asymptotic convergence has been established in Tsitsiklis (1994); Jaakkola et al. (1994); Borkar & Meyn (2000); Melo (2001); Lee & He (2019) by requiring the learning rates to satisfy $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$. Another line of research focuses on the finite-time analysis of Q-learning under different choices of the learning rates. Szepesvári (1998) captured the first convergence rate of Q-learning using a linear learning rate (i.e., $\alpha_t = \frac{1}{t}$). Under similar learning rates, Even-Dar & Mansour (2003) provided finite-time results for both synchronous and asynchronous Q-learning with a convergence rate being exponentially slow as a function of $\frac{1}{1-\gamma}$. Another popular choice is the polynomial learning rate which has been studied for synchronous Q-learning in Wainwright (2019b) and for both synchronous/asynchronous Q-learning in Even-Dar & Mansour (2003). With this learning rate, however, the convergence rate still has a gap with the lower bound of $\mathcal{O}(\frac{1}{\sqrt{T}})$ (Azar et al., 2013). To handle this, a more sophisticated rescaled linear learning rate was introduced for synchronous Q-learning (Wainwright, 2019b; Chen et al., 2020) to obtain the state-of-the-art finite bound that scales in $\frac{1}{(1-\gamma)^5\epsilon^2}$. Qu & Wierman (2020) applied a rescaled linear learning rate for asynchronous Q-learning and obtained a complexity that scales in $\frac{t_{\text{mix}}}{\mu_{\min}^2(1-\gamma)^5\epsilon^2}$. Such a bound for asynchronous Q-learning were then improved to $\frac{1}{\mu_{\min}(1-\gamma)^5\epsilon^2} + \frac{t_{\text{mix}}}{\mu_{\min}(1-\gamma)}$ with a constant learning rate in Li et al. (2020).

Compared with the above results for Q-learning, our bounds on double Q-learning match those on Q-learning in terms of the dependence on ϵ for both the synchronous and asynchronous settings, which indicates that double Q-learning can mitigate overestimation without sacrificing the complexity efficiency at the high accuracy regime. Furthermore, our bound on asynchronous double Q-learning also matches that on asynchronous Q-learning in terms of the dependence on $\mu_{\min}$ and L. In terms of the dependence on $1-\gamma$, our bounds on double Q-learning is slightly inferior to vanilla Q-learning. This is not surprising, because double Q-learning by nature takes a more conservative update rule than vanilla Q-learning, which can cause slower convergence in the lower accuracy regime.

In addition to the convergence analysis of vanilla Q-learning reviewed above, another line of theoretical research on Q-learning focuses on the regret bound analysis, e.g., (Jin et al., 2018; Yang et al., 2020). Furthermore, to improve the performance of Q-learning, various variants of Q-learning have been studied, e.g., (Dong et al., 2019; Lee & He, 2020).

In addition to Q-learning, the tabular setting has be analyzed for other RL algorithms, such as primal-dual algorithms (Wang, 2017; Jin & Sidford, 2020), variance reduced value iteration (Sidford et al., 2018b), model-based algorithms (Sidford et al., 2018a; Agarwal et al., 2020; He et al., 2020), to name a few.

Q-learning under large state-action space: When the state-action space is considerably large or even infinite, the Q-function is typically approximated by certain separation schemes of the space (Shah & Xie, 2018; Sinclair et al., 2019), or is approximated by a class of parameterized functions. In the latter case, Q-learning has been shown not to converge in general (Baird, 1995). Strong assumptions are typically needed to establish the convergence of Q-learning with linear function approximation (Bertsekas & Tsitsiklis, 1996; Melo et al., 2008; Zou et al., 2019; Chen et al., 2019; Du et al., 2019; Yang & Wang, 2019; Jia et al., 2019; Weng et al., 2020a;b) or neural network approximation (Cai et al., 2019; Xu & Gu, 2019; Fan et al., 2019). The convergence analysis

of double Q-learning with function approximation raises new technical challenges and can be an interesting topic for future study.

Two-timescale SAs: It is also worth mentioning that the analysis of two time-scale SAs (Xu et al., 2019; Doan, 2019; Dalal et al., 2020; Kaledin et al., 2020) also involves nested SAs. However, the structure is very different from double Q-learning. Specifically, in two time-scale SAs the auxiliary parameter serves as the estimation of the update of the main parameter, and hence its impact on the convergence is captured by its tracking error. In contrast, in double Q-learning, the two parameters (corresponding to two Q-estimators) are symmetric and the update is randomly switched between the two estimators. Handling such switching randomness is one of the key challenges in the convergence analysis of double Q-learning, which does not exist in two time-scale SAs.

B Numerical Experiment

In this section, we compare the performance of double Q-learning with the rescaled linear learning rate and the polynomial rate. We adopt the MDP model employed in Wainwright (2019b). The random reward function is uniformly distributed over $\{R_{sa} - 10, R_{sa} + 10\}$, where $R_{sa} = 1$ is the expected reward. We set the initial conditions as $Q^A = Q^B = -40 \cdot [1]^{|\mathcal{S}|\times|\mathcal{A}|}$, where $[1]^{|\mathcal{S}|\times|\mathcal{A}|}$ denotes the all-one matrix with the dimension of $|\mathcal{S}| \times |\mathcal{A}|$. We run the synchronous double Q-learning algorithm 20 times independently under each learning rate with each execution of the algorithm taking 10^5 iterations. Figure 1 illustrates our experimental result. It can be seen that double Q-learning with a rescaled linear learning rate substantially outperforms that with a polynomial learning rate, which corroborates our theoretical bound on the time complexity.

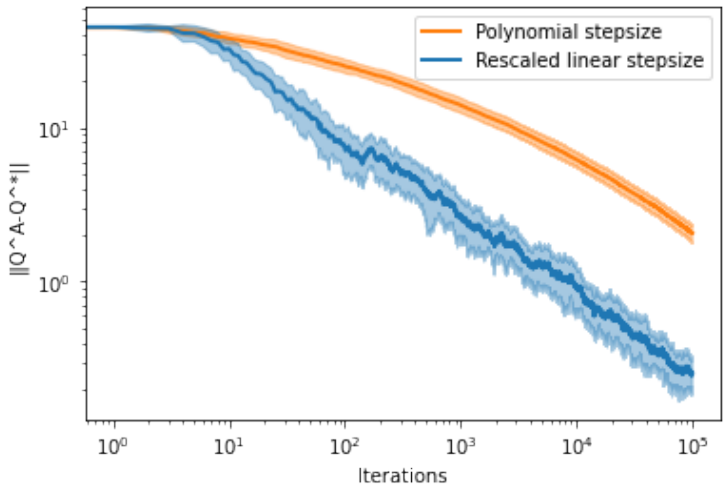

Figure 1: Comparison of the convergence performance of synchronous double Q-learning under rescaled linear learning rate and polynomial learning rate.

C Proof of the Nested SA Formulation and Various Properties

Derivation of the r_t-recursion: For both (3) and (4), we can write the error dynamics as

$$\begin{aligned}
r_{t+1}(s,a) &= (1-\tilde{\alpha}_t(s,a))r_t(s,a) + \tilde{\alpha}_t(s,a)\left(R_t(s,a,s') + \gamma Q_t^B(s',a^*) - Q^*(s,a)\right) \\
&= (1-\tilde{\alpha}_t(s,a))r_t(s,a) + \tilde{\alpha}_t(s,a)\Big(R_t(s,a,s') + \gamma Q_t^A(s',a^*) - \widehat{\mathcal{T}}_t Q^*(s,a) \\
&\quad + \widehat{\mathcal{T}}_t Q^*(s,a) - Q^*(s,a) + \gamma Q_t^B(s',a^*) - \gamma Q_t^A(s',a^*)\Big) \\
&= (1-\tilde{\alpha}_t(s,a))r_t(s,a) + \tilde{\alpha}_t(s,a)\Big(\widehat{\mathcal{T}}_t Q_t^A(s,a) - \widehat{\mathcal{T}}_t Q^*(s,a) \\
&\quad + \widehat{\mathcal{T}}_t Q^*(s,a) - Q^*(s,a) + \gamma Q_t^B(s',a^*) - \gamma Q_t^A(s',a^*)\Big)
\end{aligned}$$

$$\triangleq (1-\tilde{\alpha}_t(s,a))r_t(s,a)+\tilde{\alpha}_t(s,a)\left(\widehat{\mathcal{T}}_t Q_t^A(s,a)-\widehat{\mathcal{T}}_t Q^*(s,a)\right.$$
$$\left.+\varepsilon_t(s,a)+\gamma Q_t^B(s',a^*)-\gamma Q_t^A(s',a^*)\right),$$

where $\tilde{\alpha}_t(s,a)=\begin{cases}\alpha_t\beta_t, & \text{for synchronous version}\\ \alpha_t\beta_t\tau_t(s,a), & \text{for asynchronous version}\end{cases}$, which is exactly (5).

Uniform bound of ε_t: It follows from the definition that

$$\begin{aligned}|\varepsilon_t(s,a)| &= \left|\widehat{\mathcal{T}}_t Q^*(s,a)-\mathcal{T}Q^*(s,a)\right|\\ &= \left|R_t(s,a,s')+\gamma\max_{a'\in\mathcal{A}}Q^*(s',a')-R_{sa}^{s'}-\gamma\mathbb{E}_{s'}\max_{a'\in\mathcal{A}}Q^*(s',a')\right|\\ &\le 2R_{\max}+\gamma\left(\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}Q^*(s,a)-\min_{(s,a)\in\mathcal{S}\times\mathcal{A}}Q^*(s,a)\right)\\ &= 2R_{\max}+\gamma\left\|Q^*\right\|_{\text{span}}.\end{aligned}$$

Therefore, we have $\|\varepsilon_t\|\le 2R_{\max}+\gamma\left\|Q^*\right\|_{\text{span}}:=\kappa$.

Derivation of the ν_t-recursion: Based on (3) or (4), we have $\forall t\ge 1$,

$$\begin{aligned}\nu_{t+1}(s,a) &= Q_{t+1}^B(s,a)-Q_{t+1}^A(s,a)\\ &= (1-\hat{\alpha}_t(s,a)(1-\beta_t))\,Q_t^B(s,a)+\hat{\alpha}_t(s,a)(1-\beta_t)\left(R_t(s,a,s')+\gamma Q_t^A(s',b^*)\right)\\ &\quad-(1-\hat{\alpha}_t(s,a)\beta_t)Q_t^A(s,a)-\hat{\alpha}_t(s,a)\beta_t\left(R_t(s,a,s')+\gamma Q_t^B(s',a^*)\right)\\ &= (1-\hat{\alpha}_t(s,a))\nu_t(s,a)+\hat{\alpha}_t(s,a)\left[(1-\beta_t)\left(R_t(s,a,s')+\gamma Q_t^A(s',b^*)-Q_t^A(s,a)\right)\right.\\ &\quad\left.+\beta_t\left(Q_t^B(s,a)-R_t(s,a,s')-\gamma Q_t^B(s',a^*)\right)\right]\\ &\triangleq (1-\hat{\alpha}_t(s,a))\nu_t(s,a)+\hat{\alpha}_t(s,a)H_t(s,a),\end{aligned}\tag{20}$$

where the definition of H_t is obvious and $\hat{\alpha}_t(s,a)=\begin{cases}\alpha_t, & \text{for synchronous version}\\ \alpha_t\tau_t(s,a), & \text{for asynchronous version}\end{cases}$. Further define $\mathcal{H}_t=\mathbb{E}\left(H_t|\mathcal{F}_t\right)$ and $\mu_t=H_t-\mathcal{H}_t$. Then we immediately see that (7) follows from (20).

Quasi-contractive Property of $\mathcal{H}_t(\nu_t)$: By direct calculation using the definition of H_t and $\mathcal{F}_t$, we have

$$\mathbb{E}\left(H_t(s,a)|\mathcal{F}_t\right)=\frac{1}{2}\nu_t(s,a)+\frac{\gamma}{2}\mathbb{E}_{s'}\left(Q_t^B(s',a^*)-Q_t^A(s',b^*)\right),\tag{21}$$

where we used the fact that $\mathbb{E}(\beta_t)=0.5$. It follows from (21) that

$$\begin{aligned}\left|\mathbb{E}\left(H_t(s,a)|\mathcal{F}_t\right)\right| &\le \frac{1}{2}\left|\nu_t(s,a)\right|+\frac{\gamma}{2}\mathbb{E}_{s'}\left|Q_t^B(s',a^*)-Q_t^A(s',b^*)\right|\\ &\le \frac{1}{2}\left\|\nu_t\right\|+\frac{\gamma}{2}\mathbb{E}_{s'}\begin{cases}Q_t^B(s',a^*)-Q_t^A(s',b^*) & \text{if } Q_t^B(s',a^*)\ge Q_t^A(s',b^*)\\ Q_t^A(s',b^*)-Q_t^B(s',a^*) & \text{if } Q_t^B(s',a^*)<Q_t^A(s',b^*)\end{cases}\\ &\le \frac{1}{2}\left\|\nu_t\right\|+\frac{\gamma}{2}\mathbb{E}_{s'}\begin{cases}Q_t^B(s',b^*)-Q_t^A(s',b^*) & \text{if } Q_t^B(s',a^*)\ge Q_t^A(s',b^*)\\ Q_t^A(s',a^*)-Q_t^B(s',a^*) & \text{if } Q_t^B(s',a^*)<Q_t^A(s',b^*)\end{cases}\\ &\le \frac{1+\gamma}{2}\left\|\nu_t\right\|,\end{aligned}$$

which implies that

$$\left\|\mathbb{E}\left(H_t|\mathcal{F}_t\right)\right\|\le\frac{1+\gamma}{2}\left\|\nu_t\right\|.$$

Boundedness of μ_t: By the definition of H_t in (20) and $\mathcal{H}_t$ in (21), we have if $\beta_t = 0$,

$$
\begin{aligned}
&|\mu_t(s,a)| \\
&= \left| R_t(s,a,s') + \gamma Q_t^A(s', b^*) - Q_t^A(s,a) - \frac{1}{2}\nu_t(s,a) + \frac{\gamma}{2}\mathbb{E}_{s'}\left(Q_t^B(s', a^*) - Q_t^A(s', b^*)\right) \right| \\
&\le |R_t(s,a,s')| + \gamma \left|Q_t^A(s', b^*)\right| + \frac{1}{2}\left(\left|Q_t^A(s,a)\right| + \left|Q_t^B(s,a)\right|\right) + \frac{\gamma}{2}\left(\left|Q_t^B(s', a^*)\right| + \left|Q_t^A(s', b^*)\right|\right) \\
&\le R_{\max} + \frac{\gamma R_{\max}}{1-\gamma} + \frac{R_{\max}}{1-\gamma} + \frac{\gamma R_{\max}}{1-\gamma} \\
&= (1 + \frac{\gamma}{2}) V_{\max}.
\end{aligned}
$$

The case of $\beta_t = 1$ follows similarly and we omit the detailed proof here. Therefore, we conclude that $|\mu_t(s,a)| \le (1 + \frac{\gamma}{2}) V_{\max}$.

D Proof of Proposition 1

To proceed the proof, we first construct some useful recursions:

$$
\begin{aligned}
W_{t+1} &= (1-\alpha_t) W_t + \alpha_t \varepsilon_t, \quad \text{with initialization } W_1 = \mathbf{0}, \\
b_{t+1} &= (1 - (1-\gamma)\alpha_t)\, b_t, \quad \text{with initialization } b_1 = \|\theta_1\|\, I, \\
g_{t+1} &= (1 - (1-\gamma)\alpha_t)\, g_t + \gamma\alpha_t \left(\|W_t\| + \|\nu_t\|\right) I, \quad \text{with initialization } g_1 = \mathbf{0},
\end{aligned}
$$

where I denotes all-ones vector and $\mathbf{0}$ denotes all-zeros vector with appropriate dimensions. Note that $\{b_t\}_{t\ge 1}$ and $\{g_t\}_{t\ge 1}$ are both non-negative sequences satisfying $b_t = \|b_t\|\, I$ and $g_t = \|g_t\|\, I$ for all $t \ge 1$.

Then we have the following claim which gives a sandwich bound of θ_t which is given by

$$
-b_t - g_t + W_t \preceq \theta_t \preceq b_t + g_t + W_t, \tag{22}
$$

where $\preceq$ denotes the elementwise $\le$ relation.

We will prove (22) by induction. For $t = 1$, we have $-b_1 \preceq \theta_1 \preceq b_1$, which holds easily since $b_1 = \|\theta_1\|\, I$. Now suppose (22) holds for some $t \ge 1$, and we prove it holds for $t+1$.

We first note that

$$
\begin{aligned}
\|\theta_t\|\, I &\preceq \max\left\{\|b_t + g_t + W_t\|\, I, \|-b_t - g_t + W_t\|\, I\right\} \\
&\preceq b_t + g_t + \|W_t\|\, I,
\end{aligned} \tag{23}
$$

since $x_t = \|x_t\|\, I$ for $x \in \{b, g\}$.

For the upper bound, we have

$$
\begin{aligned}
\theta_{t+1} &= (1-\alpha_t)\theta_t + \alpha_t\left(\mathcal{G}_t(\theta_t) + \varepsilon_t + \gamma\nu_t\right) \\
&\overset{\text{(i)}}{\preceq} (1-\alpha_t)(b_t + g_t + W_t) + \alpha_t\left(\gamma\|\theta_t\|\, I + \varepsilon_t + \gamma\|\nu_t\|\, I\right) \\
&\overset{\text{(ii)}}{\preceq} (1-\alpha_t)(b_t + g_t + W_t) + \alpha_t\left[\gamma\left(b_t + g_t + \|W_t\|\, I\right) + \varepsilon_t + \gamma\|\nu_t\|\, I\right] \\
&= \underbrace{(1 - (1-\gamma)\alpha_t) b_t}_{b_{t+1}} + \underbrace{(1 - (1-\gamma)\alpha_t) g_t + \gamma\alpha_t\left(\|W_t\| + \|\nu_t\|\right) I}_{g_{t+1}} \\
&\quad + \underbrace{(1-\alpha_t) W_t + \alpha_t \varepsilon_t}_{W_{t+1}},
\end{aligned}
$$

where (i) follows from the induction assumption and the quasi-contractive property of $\mathcal{G}_t$, and (ii) follows from (23).

For the lower bound, we have

$$\begin{aligned}
\theta_{t+1} &= (1-\alpha_t)\theta_t + \alpha_t\left(\mathcal{G}_t(\theta_t) + \varepsilon_t + \gamma\nu_t\right) \\
&\overset{\text{(i)}}{\succeq} (1-\alpha_t)(-b_t - g_t + W_t) + \alpha_t\left(-\gamma\left\|\theta_t\right\| I + \varepsilon_t - \gamma\left\|\nu_t\right\| I\right) \\
&\overset{\text{(ii)}}{\succeq} (1-\alpha_t)(-b_t - g_t + W_t) + \alpha_t\left[-\gamma\left(b_t + g_t + \left\|W_t\right\| I\right) + \varepsilon_t - \gamma\left\|\nu_t\right\| I\right] \\
&= \underbrace{-(1-(1-\gamma)\alpha_t)b_t}_{-b_{t+1}} + \underbrace{-\left[(1-(1-\gamma)\alpha_t)g_t + \gamma\alpha_t\left(\left\|W_t\right\| + \left\|\nu_t\right\|\right) I\right]}_{-g_{t+1}} \\
&\quad + \underbrace{(1-\alpha_t)W_t + \alpha_t\varepsilon_t}_{W_{t+1}},
\end{aligned}$$

where (i) follows from the induction assumption and the quasi-contractive property of $\mathcal{G}_t$, and (ii) follows from (23).

Thus we proved (22) holds for $t+1$. By induction, it holds for all $t \geq 1$. Finally, we immediately have

$$\begin{aligned}
\left\|\theta_t\right\| \leq \prod_{k=1}^{t-1}(1-(1-\gamma)\alpha_k)\left\|\theta_1\right\| + \gamma\alpha_{t-1}\left(\left\|W_{t-1}\right\| + \left\|\nu_{t-1}\right\|\right) \\
+ \gamma\sum_{k=1}^{t-2}\left\{\prod_{l=k+1}^{t-1}(1-(1-\gamma)\alpha_l)\right\}\alpha_k\left(\left\|W_k\right\| + \left\|\nu_k\right\|\right) + \left\|W_t\right\|,
\end{aligned}$$

where the first term on the right hand side is $\|b_t\|$ and the sum of the next two terms correspond to $\|g_t\|$. □

E Proof of Theorem 1

We start the proof with some important properties of the rescaled learning rate $\alpha_t = \frac{3}{3+(1-\gamma)t}$. To be specific, we can show that the following inequalities hold $\forall k \geq 1$:

$$1 - \frac{\alpha_k}{2} \leq 1 - \frac{1-\gamma}{2}\alpha_k \leq \frac{\alpha_k}{\alpha_{k-1}}, \tag{24}$$

$$(1-\alpha_{k+1})^2 \leq \frac{\alpha_{k+1}^N}{\alpha_k^N}, \quad \text{where } N = \frac{18}{(1-\gamma)(4-\gamma)}. \tag{25}$$

$$\left(1 - \frac{\alpha_{k+1}}{2}\right)^2 \leq \frac{\alpha_{k+1}^N}{\alpha_k^N}, \quad \text{where } N = \frac{9}{(1-\gamma)(4-\gamma)}. \tag{26}$$

It is straightforward to verify (24). The proofs of (25) and (26) need significantly more efforts which we will show in Lemma 1 by taking $b=3, c=1$ and $b=3, c=2$ respectively there. It is very important to have such a tight characterization as it leads to a better $1/(1-\gamma)$ dependency in the finite-time error bound (see Proposition 3 and the proof therein).

Step I: Bounding outer SA dynamics $\mathbb{E}\left\|r_t\right\|$ by inner SA dynamics $\mathbb{E}\left\|\nu_t\right\|$

We shall apply Proposition 1 to the error dynamics (5) of r_t. Recall that $\mathcal{G}_t$ in (5) is quasi-contractive, which satisfies $\left\|\mathcal{G}_t(r_t)\right\| \leq \gamma\left\|r_t\right\|$. Now construct the following recursion:

$$W_{t+1} = (1-\tilde{\alpha}_t)W_t + \tilde{\alpha}_t\varepsilon_t, \quad \text{with initialization } W_1 = \mathbf{0}. \tag{27}$$

Further define $\tilde{\nu}_t(s,a) = \nu_t(s', a^*)$ and notice that $\|\tilde{\nu}_t\| \leq \|\nu_t\|$ since all the elements of $\tilde{\nu}_t$ come from ν_t. Then it follows from applying Proposition 1 to the SA (5) that

$$\begin{aligned}
\left\|r_t\right\| \leq \prod_{k=1}^{t-1}(1-(1-\gamma)\tilde{\alpha}_k)\left\|r_1\right\| + \gamma\tilde{\alpha}_{t-1}\left(\left\|W_{t-1}\right\| + \left\|\tilde{\nu}_{t-1}\right\|\right) \\
+ \gamma\sum_{k=1}^{t-2}\left\{\prod_{l=k+1}^{t-1}(1-(1-\gamma)\tilde{\alpha}_l)\right\}\tilde{\alpha}_k\left(\left\|W_k\right\| + \left\|\tilde{\nu}_k\right\|\right) + \left\|W_t\right\|.
\end{aligned}$$

Further taking the expectation for both sides yields

$$\begin{aligned}\mathbb{E}\left\|r_t\right\| &\overset{\text{(i)}}{\leq} \prod_{k=1}^{t-1}\left(1-\frac{1-\gamma}{2}\alpha_k\right)\left\|r_1\right\| + \frac{\gamma}{2}\alpha_{t-1}\left(\mathbb{E}\left\|W_{t-1}\right\| + \mathbb{E}\left\|\tilde{\nu}_{t-1}\right\|\right) \\ &\quad + \frac{\gamma}{2}\sum_{k=1}^{t-2}\left\{\prod_{l=k+1}^{t-1}\left(1-\frac{1-\gamma}{2}\alpha_l\right)\right\}\alpha_k\left(\mathbb{E}\left\|W_k\right\| + \mathbb{E}\left\|\tilde{\nu}_k\right\|\right) + \mathbb{E}\left\|W_t\right\| \\ &\overset{\text{(ii)}}{\leq} \alpha_{t-1}\left\|r_1\right\| + \frac{\gamma}{2}\alpha_{t-1}\sum_{k=1}^{t-1}\left(\mathbb{E}\left\|W_k\right\| + \mathbb{E}\left\|\nu_k\right\|\right) + \mathbb{E}\left\|W_t\right\|, \end{aligned} \tag{28}$$

where (i) follows because $\{\tilde{\alpha}_t\}_{t\geq 1}$ is a sequence of independent random variables, $\{\tilde{\alpha}_s\}_{s\geq t}$ is independent of W_t and $\tilde{\nu}_t$, and $\mathbb{E}\tilde{\alpha}_t = \frac{\alpha_t}{2}$, and (ii) follows by applying (24) repeatedly and noticing that $\|\tilde{\nu}_t\| \leq \|\nu_t\|$.

We bound $\mathbb{E}\|W_k\|$ and $\mathbb{E}\|\nu_k\|$ separately in the next two Steps.

Step II: Bounding $\mathbb{E}\|W_t\|$

We provide the bound on the expectation of the sup norm of W_{t+1} in the following Proposition. Recall $D = |\mathcal{S}|\,|\mathcal{A}|$ is the dimension of the state-action space.

Proposition 2. *Consider the sequence $\{W_{t+1}\}_{t\geq 1}$ generated by the recursion (27). We have*

$$\mathbb{E}\left\|W_{t+1}\right\| \leq \kappa\tilde{C}\sqrt{\alpha_t}, \tag{29}$$

where $\tilde{C} = 6\sqrt{\ln 2D} + 3\sqrt{\pi}$ and κ is defined in (6).

Proof. The key to this proof is to constructing a $\mathcal{F}_t$-martingale sequence $\{\tilde{W}_i\}_{1\leq i\leq t+1}$ with $\tilde{W}_{t+1} = W_{t+1}$ and $\tilde{W}_1 = \mathbf{0}$. We next use Lemma 1 to bound the squared difference sequence $\left(\tilde{W}_{i+1} - \tilde{W}_i\right)^2$ by $4V_{\max}^2\alpha_t^N/\alpha_i^{N-2}$, for $1 \leq i \leq t$, which is important for a better order dependence of $1-\gamma$. Then we apply the Azuma-Hoeffding inequality (see Lemma 5) to $\{\tilde{W}_i\}_{1\leq i\leq t+1}$ and further use Lemma 6 to obtain the bound. The details of this proof can be found in Appendix F. □

Step III: Bounding inner SA dynamics $\mathbb{E}\|\nu_t\|$

Now our goal is to bound $\mathbb{E}\|\nu_t\|$. Recall that in the ν_t-recursion (7), the operator $\mathcal{H}_t$ is quasi-contractive, which satisfies $\|\mathcal{H}_t(\nu_t)\| \leq \frac{1+\gamma}{2}\|\nu_t\|$. Then by constructing the following recursion:

$$M_{t+1} = (1-\alpha_t)M_t + \alpha_t\mu_t, \quad \text{with initialization } M_1 = \mathbf{0}, \tag{30}$$

we have from Proposition 1 that

$$\|\nu_t\| \leq \prod_{k=1}^{t-1}(1-(1-\tilde{\gamma})\alpha_k)\|\nu_1\| + \tilde{\gamma}\alpha_{t-1}\|M_{t-1}\| + \tilde{\gamma}\sum_{k=1}^{t-2}\left\{\prod_{l=k+1}^{t-1}(1-(1-\tilde{\gamma})\alpha_l)\right\}\alpha_k\|M_k\| + \|M_t\|,$$

where $\tilde{\gamma} := \frac{1+\gamma}{2}$ is the quasi-contractive coefficient of $\mathcal{H}_t$.

Applying inequality (24) and noticing that $M_1 = \mathbf{0}$, we further have $\forall t \geq 2$,

$$\|\nu_t\| \leq \alpha_{t-1}\|\nu_1\| + \tilde{\gamma}\alpha_{t-1}\sum_{k=2}^{t-1}\|M_k\| + \|M_t\|. \tag{31}$$

We provide the bound on $\mathbb{E}\|M_t\|$ in the following proposition, which is further proved in Appendix G.

Proposition 3. *Consider the sequence $\{M_{t+1}\}_{t\geq 1}$ generated by the recursion (30). We have*

$$\mathbb{E}\left\|M_{t+1}\right\| \leq \tilde{V}_{\max}\tilde{D}\sqrt{\alpha_t},$$

where $\tilde{V}_{\max} = (1+\frac{\gamma}{2})V_{\max}$ and $\tilde{D}$:=$2\sqrt{\ln 2D}$+$\sqrt{\pi}$.

Taking expectation for both sides of (31) and using Proposition 3 yields

$$\begin{aligned}\mathbb{E}\left\|\nu_t\right\| &\leq \alpha_{t-1}\left\|\nu_1\right\| + \tilde{\gamma}\alpha_{t-1}\sum_{k=2}^{t-1}\mathbb{E}\left\|M_k\right\| + \mathbb{E}\left\|M_t\right\| \\ &\leq \alpha_{t-1}\left\|\nu_1\right\| + \tilde{\gamma}\tilde{V}_{\max}\tilde{D}\alpha_{t-1}\sum_{k=2}^{t-1}\sqrt{\alpha_{k-1}} + \tilde{V}_{\max}\tilde{D}\sqrt{\alpha_{t-1}}.\end{aligned}$$

Since $\left\{\sqrt{\alpha_k}\right\}_{k\geq 1}$ is a decreasing sequence, we have

$$\sum_{k=2}^{t-1}\sqrt{\alpha_{k-1}} \leq \int_1^{t-1}\frac{1}{\sqrt{1+\frac{1-\gamma}{3}(s-1)}}ds < \frac{6}{1-\gamma}\sqrt{1+\frac{1-\gamma}{3}(t-1)} = \frac{6}{(1-\gamma)\sqrt{\alpha_{t-1}}}. \quad (32)$$

It follows that $\forall t \geq 2$,

$$\mathbb{E}\left\|\nu_t\right\| \leq \alpha_{t-1}\left\|\nu_1\right\| + \frac{(4+2\gamma)\tilde{V}_{\max}\tilde{D}}{1-\gamma}\sqrt{\alpha_{t-1}}.$$

Since $\nu_1 = \mathbf{0}$, we simplify the bound and obtain

$$\mathbb{E}\left\|\nu_t\right\| \leq \frac{(4+2\gamma)\tilde{V}_{\max}\tilde{D}}{1-\gamma}\sqrt{\alpha_{t-1}} < \frac{6\tilde{V}_{\max}\tilde{D}}{1-\gamma}\sqrt{\alpha_{t-1}}. \quad (33)$$

The above bound suggests that $\mathbb{E}\left\|\nu_t\right\|$ converges to 0 with a rate of order $\mathcal{O}(\frac{1}{\sqrt{t}})$.

Step IV: Deriving the finite-time bound

We have from (29) that $\mathbb{E}\left\|W_{t+1}\right\| \leq \kappa\tilde{C}\sqrt{\alpha_t}$. Then we have

$$\begin{aligned}\frac{\gamma}{2}\alpha_{t-1}\sum_{k=1}^{t-1}\mathbb{E}\left\|W_k\right\| + \mathbb{E}\left\|W_t\right\| &\leq \kappa\tilde{C}\left(\frac{\gamma}{2}\alpha_{t-1}\sum_{k=2}^{t-1}\sqrt{\alpha_{k-1}} + \sqrt{\alpha_{t-1}}\right) \\ &\leq \frac{\kappa\tilde{C}(2+\gamma)}{1-\gamma}\sqrt{\alpha_{t-1}} < \frac{3\kappa\tilde{C}}{1-\gamma}\sqrt{\alpha_{t-1}}, \quad (34)\end{aligned}$$

where we used the fact that $W_1 = \mathbf{0}$ and the inequality (32).

In addition, using (33) we have

$$\frac{\gamma}{2}\alpha_{t-1}\sum_{k=1}^{t-1}\mathbb{E}\left\|\nu_k\right\| \leq \frac{6\tilde{V}_{\max}\tilde{D}}{1-\gamma}\frac{\gamma}{2}\alpha_{t-1}\sum_{k=2}^{t-1}\sqrt{\alpha_{k-1}} \leq \frac{36V_{\max}\tilde{D}}{(1-\gamma)^2}\sqrt{\alpha_{t-1}}, \quad (35)$$

where we used the fact that $\nu_1 = \mathbf{0}$.

Using (34)-(35) in (28) we have $\forall t \geq 3$,

$$\begin{aligned}\mathbb{E}\left\|r_{t+1}\right\| &\leq \left\|r_1\right\|\alpha_t + \frac{\gamma}{2}\alpha_t\sum_{k=2}^{t}\left(\mathbb{E}\left\|W_k\right\| + \mathbb{E}\left\|\nu_k\right\|\right) + \mathbb{E}\left\|W_t\right\| \\ &\leq \left\|r_1\right\|\alpha_t + \frac{3\kappa\tilde{C}}{1-\gamma}\sqrt{\alpha_t} + \frac{36V_{\max}\tilde{D}}{(1-\gamma)^2}\sqrt{\alpha_t} \\ &< \frac{3\left\|r_1\right\|}{(1-\gamma)t} + \frac{3\sqrt{3}\kappa\tilde{C}}{(1-\gamma)^{3/2}}\frac{1}{\sqrt{t}} + \frac{36\sqrt{3}V_{\max}\tilde{D}}{(1-\gamma)^{5/2}}\frac{1}{\sqrt{t}},\end{aligned}$$

where $\tilde{C} = 6\sqrt{\ln 2D} + 3\sqrt{\pi}$, $\tilde{D} = 2\sqrt{\ln 2D} + \sqrt{\pi}$. Thus it completes the proof. □

E.1 Supporting Lemmas

The following lemma provides a sharp bound on $(1-\frac{1}{c}\alpha_k)^2$.

Lemma 1. *Consider the rescaled linear step sizes $\alpha_k = \frac{b}{b+(1-\gamma)k}$. Fix a positive constant c satisfying $1 \leq c \leq b$. We have $\forall k \geq 0$,*

$$\left(1 - \frac{1}{c}\alpha_{k+1}\right)^2 \leq \frac{\alpha_{k+1}^n}{\alpha_k^n}, \quad \textit{where } n := \frac{2b^2}{c(1-\gamma)(b+1-\gamma)}. \tag{36}$$

Proof. For convenience, we denote $a = (1-\gamma)(k+1)$ and $d = 1 - \frac{1}{c}$, and write

$$\left(1 - \frac{1}{c}\alpha_{k+1}\right)^2 = \left(\frac{(1-\frac{1}{c})b + (1-\gamma)(k+1)}{b + (1-\gamma)(k+1)}\right)^2 = \left(\frac{db+a}{b+a}\right)^2,$$
$$\frac{\alpha_{k+1}^n}{\alpha_k^n} = \left(\frac{b + (1-\gamma)k}{b + (1-\gamma)(k+1)}\right)^n = \left(\frac{b-1+\gamma+a}{b+a}\right)^n.$$

Then (36) becomes $\left(\frac{b-1+\gamma+a}{b+a}\right)^n \left(\frac{db+a}{b+a}\right)^{-2} \geq 1$. Further take the natural logarithm of both sides, yielding

$$n \ln\left(\frac{b-1+\gamma+a}{b+a}\right) - 2\ln\left(\frac{db+a}{b+a}\right) \geq 0. \tag{37}$$

Or equivalently,

$$n \leq \frac{2\ln\left(\frac{b+a}{db+a}\right)}{\ln\left(\frac{b+a}{b-1+\gamma+a}\right)} := f(a). \tag{38}$$

Therefore, it is sufficient to show that

$$\min_{a \in [1-\gamma, \infty)} f(a) \geq \frac{2(1-d)b^2}{(1-\gamma)(b+1-\gamma)}.$$

To this end, we will lower bound $f(a)$ for 3 cases: 1) $a = 1-\gamma$, 2) $a \to \infty$, and 3) at the extrema a^* which satisfy $f'(a^*) = 0$.

For case 1) where $a = 1-\gamma$,we have

$$f(a) = \frac{2\ln\left(1 + \frac{(1-d)b)}{db+1-\gamma}\right)}{\ln\left(1 + \frac{1-\gamma}{b}\right)} \overset{\text{(i)}}{>} \frac{2(1-d)b}{b+1-\gamma}\frac{b}{1-\gamma} = \frac{2(1-d)b^2}{(1-\gamma)(b+1-\gamma)},$$

where (i) follows from Lemma 2.

For case 2) $a \to \infty$, we have

$$\begin{aligned}
\lim_{a\to\infty} \frac{2\ln\left(\frac{b+a}{db+a}\right)}{\ln\left(\frac{b+a}{b-1+\gamma+a}\right)} &\overset{\text{(i)}}{=} 2\lim_{a\to\infty} \frac{\frac{db+a}{b+a}\frac{(d-1)b}{(db+a)^2}}{\frac{b-1+\gamma+a}{b+a}\frac{\gamma-1}{(b-1+\gamma+a)^2}} \\
&= 2\lim_{a\to\infty} \frac{b(1-d)(b-1+\gamma+a)}{(db+a)(1-\gamma)} \\
&= \frac{2b(1-d)}{1-\gamma},
\end{aligned}$$

where (i) follows from the L'Hôpital's Rule.

For case 3), denote $f(a) = 2\frac{g(a)}{h(a)}$, and directly calculating $f'(a^*) = 0$ yields

$$\frac{g(a^*)}{h(a^*)} = \frac{g'(a^*)}{h'(a^*)} = \frac{b(1-d)(b-1+\gamma+a^*)}{(1-\gamma)(db+a^*)}.$$

Using the above equation in $f(a^*)$, we have

$$f(a^*) = \frac{2b(1-d)(b-1+\gamma+a^*)}{(1-\gamma)(db+a^*)} = \frac{2b(1-d)}{1-\gamma}\left(1 + \frac{(1-d)b-1+\gamma}{db+a^*}\right).$$

Since $a^* \geq 1-\gamma$ and $(1-d)b = \frac{b}{c} \geq 1 > 1-\gamma$, we have

$$f(a^*) \geq \frac{2b(1-d)}{1-\gamma}.$$

Now combining the 3 cases, we have

$$\begin{aligned}
\min_{a\in[1-\gamma,\infty)} f(a) &= \min\left\{f(\frac{1-\gamma}{2}), \lim_{a\to\infty} f(a), f(a^*)\right\} \\
&\geq \min\left\{\frac{2(1-d)b^2}{(1-\gamma)(b+1-\gamma)}, \frac{2b(1-d)}{1-\gamma}\right\} \\
&= \frac{2(1-d)b^2}{(1-\gamma)(b+1-\gamma)} \\
&= \frac{2b^2}{c(1-\gamma)(b+1-\gamma)},
\end{aligned}$$

which completes the proof. □

The inequalities in the next lemma are useful for bounding logarithm functions. It can be easily proved, for example, using the properties of exponential functions. Therefore we omit the proof here.

Lemma 2. $\frac{x}{1+x} < \ln(1+x) < x$ *for* $-1 < x \neq 0$.

F Proof of Proposition 2

Recall the definition of $\mathcal{F}_t$ in (8), and we have

$$\begin{aligned}
\mathbb{E}\left(W_{t+1}\,|\mathcal{F}_t\right) &= \mathbb{E}\left((1-\tilde{\alpha}_t)W_t + \tilde{\alpha}_t\varepsilon_t\,|\mathcal{F}_t\right) \\
&\overset{\text{(i)}}{=} (1-\frac{\alpha_t}{2})W_t + \frac{\alpha_t}{2}\mathbb{E}\left(\varepsilon_t\,|\mathcal{F}_t\right) \\
&\overset{\text{(ii)}}{=} (1-\frac{\alpha_t}{2})W_t + \underbrace{\frac{\alpha_t}{2}\mathbb{E}\left(\varepsilon_t\right)}_{=0} \\
&= (1-\frac{\alpha_t}{2})W_t,
\end{aligned}$$

where (i) follows since $\beta_t, W_t, \varepsilon_t$ are independent, $\sigma(W_t) \subset \mathcal{F}_t$ (because W_t is a measurable function of $\{\beta_{k-1}, s_k\}_{2\leq k\leq t}$), and $\mathbb{E}(\tilde{\alpha}_t) = \alpha_t\mathbb{E}(\beta_t) = \frac{\alpha_t}{2}$, (ii) follows because $\sigma(\varepsilon_t) = \sigma(s_{t+1})$ which is independent of $\mathcal{F}_t$ (as a result of the i.i.d. sampling). We immediately have $\mathbb{E}\left(W_{t+1}\right) = \mathbb{E}\left(W_{t+1}\,|\mathcal{F}_1\right) = \mathbf{0}$.

Therefore, if we define

$$\tilde{W}_i := \begin{cases} W_{t+1}, & i = t+1, \\ \prod_{k=i}^{t}(1-\frac{\alpha_k}{2})W_i, & 1 \leq i \leq t, \end{cases} \tag{39}$$

then $\{\tilde{W}_i\}_{1\leq i\leq t+1}$ is a martingale sequence with $\tilde{W}_1 = \mathbf{0}$, for any $t \geq 1$.

Now from Lemma 4, we have

$$d_i := \left|\tilde{W}_{i+1} - \tilde{W}_i\right| \leq \begin{cases} 2\prod_{k=i+1}^{t}\left(1-\frac{\alpha_k}{2}\right)\alpha_i\kappa, & 1 \leq i < t, \\ 2\alpha_t\kappa, & i = t. \end{cases}$$

Then we have

$$d_i^2 \leq \begin{cases} 4\prod_{k=i+1}^{t}\left(1-\frac{\alpha_k}{2}\right)^2\alpha_i^2\kappa^2 \overset{\text{(i)}}{\leq} 4\frac{\alpha_t^N}{\alpha_i^{N-2}}\kappa^2, & 1 \leq i < t, \\ 4\alpha_t^2\kappa^2, & i = t, \end{cases}$$

where $N := \frac{9}{(1-\gamma)(4-\gamma)}$ and (i) follows from (26).

Then using the Azuma-Hoeffding Inequality (see Lemma 5), we have for $t \geq 1$,

$$\mathbb{P}\left(|W_{t+1}| \geq \rho\right) \leq 2\exp\left(-\frac{\rho^2}{2\sum_{i=1}^{t} d_i^2}\right) \leq 2\exp\left(-\frac{\rho^2}{8\kappa^2\alpha_t^N \sum_{i=1}^{t} \alpha_i^{-(N-2)}}\right). \tag{40}$$

Since $N > 2$, $\{\alpha_i^{-(N-2)}\}_{1\leq i\leq t}$ is a monotonically increasing sequence. We have

$$\begin{aligned}
\alpha_t^N \sum_{i=1}^{t} \alpha_i^{-(N-2)} &= \alpha_t^N \sum_{i=1}^{t-1} \alpha_i^{-(N-2)} + \alpha_t^2 \\
&\leq \alpha_t^N \int_1^t \left(1 + \frac{(1-\gamma)}{3} s\right)^{N-2} ds + \alpha_t^2 \\
&\leq \frac{3}{1-\gamma} \frac{\alpha_t}{N-1} + \alpha_t^2 \\
&= \left(\frac{3(4-\gamma)}{9-(1-\gamma)(4-\gamma)} + \alpha_t\right) \alpha_t \\
&< (3+1)\,\alpha_t \\
&= 4\alpha_t.
\end{aligned}$$

Using the above bound in (40) yields

$$\mathbb{P}\left(|W_{t+1}| \geq \rho\right) \leq 2\exp\left(-\frac{\rho^2}{32\kappa^2\alpha_t}\right).$$

By the union bound for the max operator, we have

$$\mathbb{P}\left(\|W_{t+1}\| \geq \rho\right) = \mathbb{P}\left(\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} |W_{t+1}| \geq \rho\right) \leq D\mathbb{P}\left(|W_{t+1}| \geq \rho\right).$$

Then it follows that

$$\begin{aligned}
\mathbb{E}\,\|W_{t+1}\| &= \int_0^\infty \mathbb{P}(\|W_{t+1}\| \geq \rho) d\rho, \\
&\leq 2D \int_0^\infty \exp\left(-\frac{\rho^2}{32\kappa^2\alpha_t}\right) d\rho \\
&\overset{\text{(i)}}{\leq} 6\kappa\sqrt{\alpha_t}\left(\sqrt{\ln 2D} + \frac{\sqrt{\pi}}{2}\right) \\
&:= \kappa\tilde{C}\sqrt{\alpha_t},
\end{aligned}$$

where (i) follows from Lemma 6 and $\tilde{C} := 6\sqrt{\ln 2D} + 3\sqrt{\pi}$.

□

F.1 Supporting Lemmas

Lemma 3. *Consider the sequence $\{W_{t+1}\}_{t\geq 1}$ generated by the recursion (27). We have $|W_{t+1}| \leq \kappa$ where κ is the uniform bound of the i.i.d. noise sequence $\{\varepsilon_t\}_{t\geq 1}$ defined in (6).*

Proof. We prove it by induction. For $t = 1$, we have

$$|W_2| = |\tilde{\alpha}_1\varepsilon_1| \leq \alpha_1\kappa < \kappa.$$

Suppose $|W_t| \leq \kappa$ for some $t \geq 2$. Then it follows that,

$$|W_{t+1}| \leq (1-\tilde{\alpha}_t)\,|W_t| + |\tilde{\alpha}_t\varepsilon_t| \leq (1-\tilde{\alpha}_t)\kappa + \tilde{\alpha}_t\kappa = \kappa.$$

Thus by induction $|W_{t+1}| \leq \kappa$ for all $t \geq 1$. □

Lemma 4. *Consider the martingale sequence $\{\tilde{W}_i\}_{1\le i\le T+1}$, $T\ge 1$ defined in (39). We have the corresponding difference sequence bounded by*

$$\left|\tilde{W}_{i+1}-\tilde{W}_i\right| \le \begin{cases} 2\alpha_T\kappa, & i=T, \\ 2\prod_{k=i+1}^{T}(1-\frac{\alpha_k}{2})\alpha_i\kappa, & 1\le i<T, \end{cases} \tag{41}$$

where κ is the uniform bound of the i.i.d. noise sequence $\{\varepsilon_t\}_{t\ge 1}$ defined in (6).

Proof. By the definition of $\{\tilde{W}_i\}_{1\le i\le T+1}$, we have,

$$\tilde{W}_{i+1}-\tilde{W}_i = \begin{cases} \alpha_T\Gamma_T, & i=T, \\ \prod_{k=i+1}^{T}(1-\frac{\alpha_k}{2})\alpha_i\Gamma_i, & 1\le i<T, \end{cases} \tag{42}$$

where $\Gamma_i := (\frac{1}{2}-\beta_i)W_i+\beta_i\varepsilon_i$, for all $1\le i\le T$. Since $W_1=\mathbf{0}$,we easily have $|\Gamma_1|\le\kappa$. For $i\ge 2$, we have

$$|\Gamma_i| = \begin{cases} \left|-\frac{1}{2}W_i+\varepsilon_i\right| & \text{if } \beta_t=1 \\ \frac{1}{2}|W_i| & \text{if } \beta_t=0 \end{cases} \le \frac{1}{2}|W_i|+|\varepsilon_i| \overset{\text{(i)}}{\le} \frac{3}{2}\kappa < 2\kappa,$$

where (i) follows from Lemma 3. Plugging the above bound in (42) completes the proof. □

Lemma 5. *(Azuma-Hoeffding Inequality) Suppose $\{S_n\}_{n\ge 1}$ is a martingale such that $S_0=0$ and $|S_i-S_{i-1}|\le d_i$ almost surely for some constants d_i, $1\le i\le n$. Then, for all $t\ge 0$,*

$$\mathbb{P}\left(|S_n|\ge\rho\right) \le 2\exp\left(-\frac{\rho^2}{2\sum_{i=1}^{n}d_i^2}\right).$$

The following lemma slightly extends Wainwright (2019a, Exercise 2.8 (a)) to handle the case where $b=0$. The proof is similar and we include it here for completeness.

Lemma 6. *Suppose Z is a non-negative random variable satisfying the concentration inequality $\mathbb{P}\left(Z\ge\rho\right)\le C\exp\left(-\rho^2/\sigma^2\right)$, $\forall\rho>0$, for some $C>1,\sigma>0$. Then we have $\mathbb{E}\left(Z\right)\le\sigma\left(\sqrt{\ln C}+\frac{\sqrt{\pi}}{2}\right)$.*

Proof. By the expectation formula of non-negative random variables, we have

$$\begin{aligned}
\mathbb{E}\left(Z\right) = \int_0^\infty \mathbb{P}(Z\ge\rho)d\rho &\le \int_0^\infty 1\wedge C\exp\left(-\frac{\rho^2}{\sigma^2}\right)d\rho \\
&= \int_0^{\sigma\sqrt{\ln C}} 1d\rho + \int_{\sigma\sqrt{\ln C}}^\infty C\exp\left(-\frac{\rho^2}{\sigma^2}\right)d\rho \\
&= \sigma\sqrt{\ln C} + \int_{\sigma\sqrt{\ln C}}^\infty \exp\left(-\frac{\rho^2-\sigma^2\ln C}{\sigma^2}\right)d\rho \\
&\overset{\text{(i)}}{\le} \sigma\sqrt{\ln C} + \int_{\sigma\sqrt{\ln C}}^\infty \exp\left(-\frac{\left(\rho-\sigma\sqrt{\ln C}\right)^2}{\sigma^2}\right)d\rho \\
&= \sigma\sqrt{\ln C} + \int_0^\infty \exp\left(-\frac{z^2}{\sigma^2}\right)dz \\
&= \sigma\left(\sqrt{\ln C}+\frac{\sqrt{\pi}}{2}\right),
\end{aligned}$$

where (i) follows because $\rho^2-a^2\ge(\rho-a)^2$ for $\rho\ge a\ge 0$. □

G PROOF OF PROPOSITION 3

Since $\{\mu_t\}_{t\geq 1}$is a martingale difference sequence, we have

$$\begin{aligned}\mathbb{E}\left(M_{t+1}\,|\mathcal{F}_t\right) &= \mathbb{E}\left((1-\alpha_t)M_t + \alpha_t\mu_t\,|\mathcal{F}_t\right)\\ &= (1-\alpha_t)M_t + \mathbb{E}\left(\alpha_t\mu_t\,|\mathcal{F}_t\right)\\ &= (1-\alpha_t)M_t.\end{aligned}$$

Therefore, if we define

$$\tilde{M}_i := \begin{cases} M_{t+1}, & i = t+1,\\ \prod_{k=i}^{t}(1-\alpha_k)M_i, & 1\leq i\leq t,\end{cases}$$

then $\{\tilde{M}_i\}_{1\leq i\leq t+1}$ is a martingale sequence with $\mathbb{E}\left(\tilde{M}_1\right) = \mathbf{0}$, for any $t\geq 1$.

To utilize the Azuma-Hoeffding inequality to bound $\mathbb{P}\left(\left|\tilde{M}_{t+1}\right|\geq\rho\right)$, $\forall\rho\geq 0$, we first need to bound $\left|\tilde{M}_{i+1}-\tilde{M}_i\right|$, $1\leq i\leq t$. By the definition of $\{\tilde{M}_i\}_{1\leq i\leq t+1}$, we have

$$\tilde{M}_{i+1}-\tilde{M}_i = \begin{cases}\prod_{k=i+1}^{t}(1-\alpha_k)\alpha_i\mu_i, & 1\leq i<t,\\ \alpha_t\mu_t, & i=t.\end{cases}$$

Since $|\mu_t|\leq(1+\frac{\gamma}{2})V_{\max} := \tilde{V}_{\max}$, we further have

$$d_i := \left|\tilde{M}_{i+1}-\tilde{M}_i\right| \leq \begin{cases}\prod_{k=i+1}^{t}(1-\alpha_k)\alpha_i\tilde{V}_{\max}, & 1\leq i<t,\\ \alpha_t\tilde{V}_{\max}, & i=t,\end{cases}$$

Then we have

$$d_i^2 = \begin{cases}\prod_{k=i+1}^{t}(1-\alpha_k)^2\alpha_i^2\tilde{V}_{\max}^2 \overset{\text{(i)}}{\leq} \frac{\alpha_t^N}{\alpha_i^{N-2}}\tilde{V}_{\max}^2, & 1\leq i<t,\\ \alpha_t^2\tilde{V}_{\max}^2, & i=t,\end{cases}$$

where $N := \frac{18}{(1-\gamma)(4-\gamma)}$ and (i) follows from (25).

Then using the Azuma-Hoeffding Inequality (see Lemma 5), we have for $t\geq 1$,

$$\mathbb{P}\left(|M_{t+1}|\geq\rho\right) \leq 2\exp\left(-\frac{\rho^2}{2\sum_{i=1}^{t}d_i^2}\right) \leq 2\exp\left(-\frac{\rho^2}{2\tilde{V}_{\max}^2\alpha_t^N\sum_{i=1}^{t}\alpha_i^{-(N-2)}}\right). \tag{43}$$

Since $N>2$, $\{\alpha_i^{-(N-2)}\}_{1\leq i\leq t}$ is a monotonically increasing sequence. We have

$$\begin{aligned}\alpha_t^N\sum_{i=1}^{t}\alpha_i^{-(N-2)} &= \alpha_t^N\sum_{i=1}^{t-1}\alpha_i^{-(N-2)} + \alpha_t^2\\ &\leq \alpha_t^N\int_1^t\left(1+\frac{(1-\gamma)}{3}s\right)^{N-2}ds + \alpha_t^2\\ &\leq \frac{3}{1-\gamma}\frac{\alpha_t}{N-1} + \alpha_t^2\\ &= \left(\frac{3(4-\gamma)}{18-(1-\gamma)(4-\gamma)} + \alpha_t\right)\alpha_t\\ &< (1+1)\,\alpha_t\\ &= 2\alpha_t.\end{aligned}$$

Using the above bound in (43) yields

$$\mathbb{P}\left(|M_{t+1}|\geq\rho\right) \leq 2\exp\left(-\frac{\rho^2}{4\tilde{V}_{\max}^2\alpha_t}\right).$$

By the union bound for the max operator, we have

$$\mathbb{P}\left(\|M_{t+1}\| \geq \rho\right) = \mathbb{P}\left(\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} |M_{t+1}| \geq \rho\right) \leq D\mathbb{P}\left(|M_{t+1}| \geq \rho\right).$$

Then it follows that

$$\begin{aligned}
\mathbb{E}\,\|M_{t+1}\| &= \int_0^\infty \mathbb{P}(\|M_{t+1}\| \geq \rho) d\rho, \\
&\leq 2D \int_0^\infty \exp\left(-\frac{\rho^2}{4\tilde{V}_{\max}^2 \alpha_t}\right) d\rho \\
&\overset{\text{(i)}}{\leq} 2\tilde{V}_{\max}\sqrt{\alpha_t}\left(\sqrt{\ln 2D} + \frac{\sqrt{\pi}}{2}\right) \\
&:= \tilde{V}_{\max}\tilde{D}\sqrt{\alpha_t},
\end{aligned}$$

where (i) follows from Lemma 6 and $\tilde{D} := 2\sqrt{\ln 2D} + \sqrt{\pi}$. □

H Proof of Theorem 2

In the following, we consider a constant learning rate, i.e., $\alpha_t = \alpha$. We keep the notations $r_t = Q_t^A - Q^*, \nu_t = Q_t^B - Q_t^A$.

To proceed the proof, we first introduce the following definition of valid iterations when using a fixed state-action pair (s, a) to update.

Definition 1. *We denote by $T(s,a)$ the set of all the iteration indices at which the state-action pair (s,a) is updated for either Q-estimator Q^A or Q^B. In addition, we denote by $T_{t_1}^{t_2}(s,a) \subseteq T(s,a)$ the set of indices that are between time t_1 and t_2, that is,*

$$T_{t_1}^{t_2}(s,a) = \left\{t : t \in [t_1, t_2] \text{ and } t \in T(s,a)\right\}.$$

The number of iterations updating (s,a) between time t_1 and t_2 is thus given by $|T_{t_1}^{t_2}(s,a)|$, i.e., the cardinally of $T_{t_1}^{t_2}(s,a)$.

Based on Definition 1, it is easy to observe that τ_t defined in (4) can be rewritten as

$$\tau_t(s,a) = \mathbb{1}_{t\in T(s,a)}.$$

In the remaining, we show how to use the constant learning rate and the above definition to derive the convergence result.

We first continue with the dynamics of $r_t(s,a)$ derived in Appendix C and obtain

$$\begin{aligned}
r_{t+1}(s,a) &= (1-\tilde{\alpha}_t(s,a))r_t(s,a) + \tilde{\alpha}_t(s,a)\left(\widehat{\mathcal{T}}_t Q_t^A(s,a) - \widehat{\mathcal{T}}_t Q^*(s,a)\right) \\
&\quad + \tilde{\alpha}_t(s,a)\varepsilon_t(s,a) + \tilde{\alpha}_t(s,a)\gamma\nu_t(s',a^*) \\
&= \prod_{i=1}^t (1-\tilde{\alpha}_i(s,a))r_1(s,a) + \sum_{i=1}^t \prod_{j=i+1}^t (1-\tilde{\alpha}_j(s,a))\tilde{\alpha}_i(s,a)\varepsilon_i(s,a) \\
&\quad + \sum_{i=1}^t \prod_{j=i+1}^t (1-\tilde{\alpha}_j(s,a))\tilde{\alpha}_i(s,a)\left(\widehat{\mathcal{T}}_i Q_t^A(s,a) - \widehat{\mathcal{T}}_i Q^*(s,a)\right) \\
&\quad + \sum_{i=1}^t \prod_{j=i+1}^t (1-\tilde{\alpha}_j(s,a))\tilde{\alpha}_i(s,a)\gamma\nu_i(s',a^*).
\end{aligned}$$

Then we have

$$
\begin{aligned}
|r_{t+1}(s,a)| = \Bigg| & \prod_{i=1}^{t}(1-\tilde{\alpha}_i(s,a))r_1(s,a) + \sum_{i=1}^{t}\prod_{j=i+1}^{t}(1-\tilde{\alpha}_j(s,a))\tilde{\alpha}_i(s,a)\varepsilon_i(s,a) \\
& + \sum_{i=1}^{t}\prod_{j=i+1}^{t}(1-\tilde{\alpha}_j(s,a))\tilde{\alpha}_i(s,a)\left(\widehat{\mathcal{T}}_i Q_t^A(s,a) - \widehat{\mathcal{T}}_i Q^*(s,a)\right) \\
& + \sum_{i=1}^{t}\prod_{j=i+1}^{t}(1-\tilde{\alpha}_j(s,a))\tilde{\alpha}_i(s,a)\gamma\nu_i(s',a^*)\Bigg| \\
\leq & \underbrace{\prod_{i=1}^{t}(1-\tilde{\alpha}_i(s,a))\left\|r_1\right\|}_{P_{1,t}(s,a)} + \underbrace{\left|\sum_{i=1}^{t}\prod_{j=i+1}^{t}(1-\tilde{\alpha}_j(s,a))\tilde{\alpha}_i(s,a)\varepsilon_i(s,a)\right|}_{P_{2,t}(s,a)} \\
& + \sum_{i=1}^{t}\prod_{j=i+1}^{t}(1-\tilde{\alpha}_j(s,a))\tilde{\alpha}_i(s,a)\gamma\left\|r_i\right\| \\
& + \sum_{i=1}^{t}\prod_{j=i+1}^{t}(1-\tilde{\alpha}_j(s,a))\tilde{\alpha}_i(s,a)\gamma\left\|\nu_i\right\|. \qquad (44)
\end{aligned}
$$

Next, we will bound the first two terms and also $\|\nu_i\|$, respectively. First, we give a high probability bound for term $P_{1,t}(s,a)$.

Proposition 4. *Fix any $\delta \in (0,1)$. Suppose $T > \frac{886 t_{\text{mix}}}{\mu_{\min}} \ln\left(\frac{4DT}{\delta}\right) := t_{\text{frame}}$ and Assumption 1 holds. Then for any t satisfying $t_{\text{frame}} \leq t \leq T$ and for any (s,a), we have*

$$
P_{1,t}(s,a) \leq (1-\alpha)^{\frac{1}{2}t\mu_{\min}}\left\|r_1\right\|, \qquad (45)
$$

with probability at least $1-\delta$, where $D = |\mathcal{S}||\mathcal{A}|$.

Proposition 4 immediately follows from Lemma 7 whose proof is different from Q-learning since we need to additionally handle the switching parameter β_t.

The next proposition is the bound for term $P_{2,t}(s,a)$, whose proof can be found in Appendix I.

Proposition 5. *Fix any $\delta \in (0,1)$. Then for any $t \in [1,T]$ and for any (s,a), we have*

$$
P_{2,t}(s,a) \leq \sqrt{2\alpha \ln\left(\frac{2DT}{\delta}\right)}\kappa, \qquad (46)
$$

with probability at least $1-\delta$, where $D = |\mathcal{S}||\mathcal{A}|$ and κ is defined in (6).

We next bound $\|\nu_t\|$, whose proof can be found in Appendix J.

Proposition 6. *Fix any $\delta \in (0,1)$. Then for any $t \in [1,T]$, we have*

$$
\left\|\nu_t\right\| \leq \sqrt{8\alpha \ln\left(\frac{2DT}{\delta}\right)}\frac{V_{\max}}{1-\gamma}, \qquad (47)
$$

with probability at least $1-\delta$, where $D = |\mathcal{S}||\mathcal{A}|$.

Then, we apply the above propositions to (44) and obtain that $\forall t \in [1,T], \forall (s,a)$, with probability at least $1-3\delta$ we have

$$
\begin{aligned}
|r_{t+1}(s,a)| \leq & \prod_{i=1}^{t}(1-\tilde{\alpha}_i(s,a))\left\|r_1\right\| + \left|\sum_{i=1}^{t}\prod_{j=i+1}^{t}(1-\tilde{\alpha}_j(s,a))\tilde{\alpha}_i(s,a)\varepsilon_i(s,a)\right| \\
& + \sum_{i=1}^{t}\prod_{j=i+1}^{t}(1-\tilde{\alpha}_j(s,a))\tilde{\alpha}_i(s,a)\gamma\left\|r_i\right\|
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{i=1}^{t}\prod_{j=i+1}^{t}(1-\tilde{\alpha}_j(s,a))\tilde{\alpha}_i(s,a)\gamma\left\|\nu_i\right\| \\
&\leq(1-\alpha)^{\frac{1}{2}t\mu_{\min}}\left\|r_1\right\|+\sum_{i=1}^{t}\prod_{j=i+1}^{t}(1-\tilde{\alpha}_j(s,a))\tilde{\alpha}_i(s,a)\gamma\left\|r_i\right\| \\
&\quad+\sqrt{2\alpha\ln\left(\frac{2DT}{\delta}\right)}\kappa+\sqrt{8\alpha\ln\left(\frac{2DT}{\delta}\right)}\frac{\gamma V_{\max}}{1-\gamma}\sum_{i=1}^{t}\prod_{j=i+1}^{t}(1-\tilde{\alpha}_j(s,a))\tilde{\alpha}_i(s,a) \\
&\overset{\text{(i)}}{\leq}(1-\alpha)^{\frac{1}{2}t\mu_{\min}}\left\|r_1\right\|+\sum_{i=1}^{t}\prod_{j=i+1}^{t}(1-\tilde{\alpha}_j(s,a))\tilde{\alpha}_i(s,a)\gamma\left\|r_i\right\| \\
&\quad+\sqrt{2\alpha\ln\left(\frac{2DT}{\delta}\right)}\kappa+\sqrt{8\alpha\ln\left(\frac{2DT}{\delta}\right)}\frac{\gamma V_{\max}}{1-\gamma} \\
&\triangleq(1-\alpha)^{\frac{1}{2}t\mu_{\min}}\left\|r_1\right\|+\sum_{i=1}^{t}\prod_{j=i+1}^{t}(1-\tilde{\alpha}_j(s,a))\tilde{\alpha}_i(s,a)\gamma\left\|r_i\right\|+C, \qquad (48)
\end{aligned}
$$

where (i) follows from Lemma 8 and

$$
C=\sqrt{2\alpha\ln\left(\frac{2DT}{\delta}\right)}\kappa+\sqrt{8\alpha\ln\left(\frac{2DT}{\delta}\right)}\frac{\gamma V_{\max}}{1-\gamma}. \qquad (49)
$$

We further define more quantities for the ease of presentation:

$$
\mu_{\text{frame}}:=\frac{1}{2}\mu_{\min}t_{\text{frame}}, \qquad (50)
$$

$$
t_{\text{th}}:=\max\left\{\frac{2\ln\frac{1}{(1-\gamma)\epsilon}}{\alpha\mu_{\min}},t_{\text{frame}}\right\}, \qquad (51)
$$

$$
\rho:=(1-\gamma)\left(1-(1-\alpha)^{\mu_{\text{frame}}}\right), \qquad (52)
$$

where t_{frame} is defined in Proposition 4. Then we have the following upper bound.

Proposition 7. *(Li et al., 2020, Lemma 3,4) Fix any $\delta\in(0,1/2),\epsilon\in(0,1/(1-\gamma))$. Given the dynamics of $r_t(s,a)$ in (48), then with probability at least $1-\delta$, we have*

$$
\left\|r_t\right\|\leq\frac{C}{1-\gamma}+(1-\rho)^k\frac{\left\|r_1\right\|}{1-\gamma}+\epsilon, \qquad (53)
$$

where $k=\max\left\{0,\lfloor\frac{t-t_{\text{th}}}{t_{\text{frame}}}\rfloor\right\}$.

The proof is the same as the combining proofs of Lemma 3-4 in Li et al. (2020) except different constants C,t_{frame}, and is thus omitted here.

Combining Propositions 4-7, we have for any $\delta\in(0,1/6)$ and with probability at least $1-6\delta$,

$$
\left\|r_t\right\|\leq\frac{C}{1-\gamma}+(1-\rho)^k\frac{\left\|r_1\right\|}{1-\gamma}+\epsilon. \qquad (54)
$$

First, we have $\frac{C}{1-\gamma}\leq\epsilon$ when letting

$$
\alpha=\frac{(1-\gamma)^6\epsilon^2}{c'\ln\frac{2DT}{\delta}}, \qquad (55)
$$

where c' is some positive constant and is derived by observing $\kappa\leq\frac{c_1}{1-\gamma},V_{\max}\leq\frac{c_2}{1-\gamma}$ for some positive constant c_1,c_2.

Next, we have $(1-\rho)^k \frac{\|r_1\|}{1-\gamma} \leq \exp(-\rho k)\frac{\|r_1\|}{1-\gamma} \leq \epsilon$ if $k \geq \ln \frac{\|r_1\|}{(1-\gamma)\epsilon}/\rho$. By the definition of k in Proposition 7, we have

$$t \geq t_{\text{th}} + t_{\text{frame}} + \frac{t_{\text{frame}}}{\rho} \ln \frac{\|r_1\|}{(1-\gamma)\epsilon}. \tag{56}$$

Further, since $(1-\alpha)^{\mu_{\text{frame}}} \leq 1 - \frac{\alpha\mu_{\text{frame}}}{2}$ when $\alpha < \frac{1}{\mu_{\text{frame}}}$, we have

$$\rho = (1-\gamma)\left(1-(1-\alpha)^{\mu_{\text{frame}}}\right) \geq \frac{\alpha\mu_{\text{frame}}(1-\gamma)}{2}. \tag{57}$$

Last, we know (56) holds as long as

$$\begin{aligned}
t &\geq t_{\text{th}} + t_{\text{frame}} + \frac{2t_{\text{frame}}}{\alpha\mu_{\text{frame}}(1-\gamma)} \ln \frac{\|r_1\|}{v} \\
&= t_{\text{th}} + t_{\text{frame}} + \frac{4}{\alpha\mu_{\min}(1-\gamma)} \ln \frac{\|r_1\|}{(1-\gamma)\epsilon} \\
&\geq t_{\text{th}} + t_{\text{frame}} + \frac{4}{\mu_{\min}(1-\gamma)} \ln \frac{\|r_1\|}{(1-\gamma)\epsilon} \cdot \max\left\{\frac{c' \ln \frac{2DT}{\delta}}{(1-\gamma)^6\epsilon^2}, \mu_{\text{frame}}\right\} \\
&= t_{\text{th}} + t_{\text{frame}} + \frac{4}{\mu_{\min}(1-\gamma)} \ln \frac{\|r_1\|}{(1-\gamma)\epsilon} \cdot \max\left\{\frac{c' \ln \frac{2DT}{\delta}}{(1-\gamma)^6\epsilon^2}, c'' t_{\text{mix}} \ln \frac{DT}{\delta}\right\}.
\end{aligned}$$

Thus, we can continue with (54) and conclude that for any $\delta \in (0, 1/6)$ and with probability at least $1-6\delta$, we have $\|r_t\| \leq 3\epsilon$ as long as

$$T = \tilde{\Omega}\left(\frac{1}{\mu_{\min}\epsilon^2(1-\gamma)^7} \ln \frac{1}{\epsilon(1-\gamma)^2} + \frac{t_{\text{mix}}}{\mu_{\min}(1-\gamma)} \ln \frac{1}{\epsilon(1-\gamma)^2}\right).$$

H.1 Supporting Lemmas

The following lemma characterizes the probability of the number of non-zero $\tilde{\alpha}_i$'s after a sufficient number of iterations.

Lemma 7. *Let $\beta_t, \tau_t(s,a)$ be as defined in (5). Suppose Assumption 1 holds. Fix any $\delta \in (0,1)$ and $T \geq t > \frac{886 t_{\text{mix}}}{\mu_{\min}} \ln\left(\frac{4DT}{\delta}\right) := t_{\text{frame}}$. Then*

$$\forall (s_1, a_1), \qquad \mathbb{P}_{(s_1,a_1)}\left(\sum_{i=1}^{t} \beta_t \tau_t(s,a) \leq \frac{1}{2} t \mu_\pi(s,a)\right) \leq \delta. \tag{58}$$

Proof. The proof is an application of Lemma 5 of Li et al. (2020), where $\mu_{\min}$ is taken to be half of it. The idea is to construct an auxiliary Markov chain which has the same mixing time as the original MDP under the behavioral policy but only has half of its $\mu_{\min}$.

The construction is inspired by the following intuition. Since $\{\beta_i\}$ is a Bernoulli random variable with expectation $\frac{1}{2}$, intuitively, double-Q learning should take two times of the iterations needed by vanilla Q-learning to visit all the states of Q^A the same amount of times (with the same high probability).To show this formally, we construct an auxiliary Markov chain by augmenting the states with β_t, namely, $\bar{M} := \{\bar{X}_t\}_{t\geq 1} = \{s_t, a_t, \beta_t\}_{\forall t \geq 1}$ with state space $\bar{\mathcal{X}} := \mathcal{S} \times \mathcal{A} \times \mathcal{B}$, where $\mathcal{B} = \{0,1\}$. It is easy to see that this auxiliary Markov chain is aperiodic and irreducible (and thus uniformly ergodic) given that the original Markov chain $M_o := \{X_t\}_{t\geq 1} = \{s_t, a_t\}_{t\geq 1}$ is aperiodic and irreducible. The transition probability can be calculated by

$$\mathbb{P}\left(\bar{X}_{t+1} | \bar{X}_t\right) \overset{\text{(i)}}{=} \mathbb{P}\left(\beta_t\right) \mathbb{P}\left(s_{t+1}, a_{t+1} | s_t, a_t\right) \overset{\text{(ii)}}{=} \frac{1}{2}\pi(a_{t+1}|s_{t+1}) \mathbb{P}\left(s_{t+1}|s_t, a_t\right), \tag{59}$$

where (i) follows from the fact that $\{\beta_t\}_{t\geq 1}$ are i.i.d Bernoulli random variables which are also independent of $\{s_t, a_t\}_{t\geq 1}$, and in (ii) we denote by π the underlying behavior policy of the Markov chain we sampled from. Let $\bar{P}$ denote the transition probability matrix of $\bar{M}$ where the

$((s,a,\beta),(s',a',\beta'))$th entry of $\bar{P}$ is $\frac{1}{2}\pi(a'|s')\mathbb{P}(s'|s,a)$. For the ease of discussion, assume that the top left $|\mathcal{S}||A| \times |\mathcal{S}||A|$ submatrix of $\bar{P}$ corresponds to the transitions between $(s,a,1)$'s. Furthermore, let $\bar{\mu} \in \Delta(\mathcal{S} \times \mathcal{A} \times \mathcal{B})$ denote the stationary distribution of $\bar{M}$.

Let P_o denote the transition probability matrix of M_o where the $((s,a),(s',a'))$th entry of P_o is $\pi(a'|s')\mathbb{P}(s'|s,a)$. Let $\mu \in \Delta(\mathcal{S} \times \mathcal{A})$ be the stationary distribution of M_o, and thus we have $\mu P_o = \mu$, assuming that μ is a row vector. Let $P^t(\cdot|x)$ denote the distribution of X_t (assuming a row vector), conditioned on $X_1 = x \in \mathcal{X}$, and we have $P^t(\cdot|x)P_o = P^{t+1}(\cdot|x)$. By (59), we have for $\bar{M}$ that

$$\bar{P} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \otimes \frac{1}{2} P_o, \tag{60}$$

where $\otimes$ denotes the Kronecker product. Similarly, we call $\bar{P}^t(\cdot|\bar{x})$ the distribution of $\bar{X}_t$, conditioned on $\bar{X}_1 = \bar{x}$. It is easy to verify (using (60)) that $\bar{P}^t(\cdot|\bar{x}) = [\frac{1}{2}P^t(\cdot|x), \frac{1}{2}P^t(\cdot|x)]$ with either $\bar{x} = (x,1)$ or $\bar{x} = (x,1)$. Let $t \to \infty$, and we have the stationary distribution of $\bar{M}$as $\bar{\mu} = [\frac{1}{2}\mu, \frac{1}{2}\mu]$. It follows that $\bar{\mu}_{\min} = \frac{1}{2}\mu_{\min}$.

We claim that the mixing times for $\bar{M}$ and M_o are the same. To see this, we calculate the variation distances $\forall x \in \mathcal{X}$,

$$d_{\text{TV}}(P^t(\cdot|x), \mu) = \frac{1}{2}\sum_{y \in \mathcal{X}} \left|P^t(y|x) - \mu(y)\right|,$$

$$d_{\text{TV}}(\bar{P}^t(\cdot|\bar{x}), \bar{\mu}) = d_{\text{TV}}([\frac{1}{2}P^t(\cdot|x), \frac{1}{2}P^t(\cdot|x)], [\frac{1}{2}\mu, \frac{1}{2}\mu]) = \frac{1}{2}\sum_{y \in \mathcal{X}} \left|P^t(y|x) - \mu(y)\right|,$$

which are the same. Therefore we conclude the claim by the definition of the mixing time.

Finally, applying Lemma 5 of Li et al. (2020) to the auxiliary Markov chain, which has $\bar{\mu}_{\min} = \frac{1}{2}\mu_{\min}$ and the same t_{mix} as the original sampled Markov process, we immediately have (58). □

The next lemma is used to deal with the term about learning rates and corresponding randomness.

Lemma 8. *Let $\tilde{\alpha}_t(s,a) = \alpha\beta_t\tau_t(s,a)$ as defined in (5). Then,*

$$\sum_{i=1}^{t}\prod_{j=i+1}^{t}(1-\tilde{\alpha}_j(s,a))\tilde{\alpha}_i(s,a) \leq 1.$$

Proof. Based on the definition of $\tau_i(s,a)$, we have

$$\begin{aligned}\sum_{i=1}^{t}\prod_{j=i+1}^{t}(1-\tilde{\alpha}_j(s,a))\tilde{\alpha}_i(s,a) &= \sum_{i \in T_1^t(s,a)}\prod_{j \in T_{i+1}^t(s,a)}(1-\alpha\beta_j)\alpha\beta_i \\ &\triangleq \sum_{i=1}^{|T_1^t(s,a)|}\prod_{j \in T_{t_i+1}^t(s,a)}(1-\alpha\beta_j)\alpha\beta_{t_i},\end{aligned}$$

where t_i denotes the time stamp when (s,a) is visited for the i^{th} time in the window $[1,t]$.

Fix any $m \in [0, |T_1^t(s,a)|]$. If there are m non-zero β_i's in the set $\left\{\beta_{t_1}, \ldots, \beta_{t_{|T_1^t(s,a)|}}\right\}$, i.e., $\sum_{i=1}^{|T_1^t(s,a)|}\beta_{t_i} = m$, then we have

$$\begin{aligned}\sum_{i=1}^{t}\prod_{j=i+1}^{t}(1-\tilde{\alpha}_j(s,a))\tilde{\alpha}_i(s,a) &= \sum_{i=1}^{|T_1^t(s,a)|}\prod_{j \in T_{t_i+1}^t(s,a)}(1-\alpha\beta_j)\alpha\beta_{t_i} \\ &= \sum_{i=1}^{m}(1-\alpha)^{m-i}\alpha \\ &\leq 1.\end{aligned}$$

Since the above bound holds for any $|T_1^t(s,a)|$ and any $m \in [0, |T_1^t(s,a)|]$, we can conclude the lemma.

□

I PROOF OF PROPOSITION 5

Recall that

$$\begin{aligned} P_{2,t}(s,a) &= \left| \sum_{i=1}^{t} \prod_{j=i+1}^{t} (1-\tilde{\alpha}_j(s,a))\tilde{\alpha}_i(s,a)\varepsilon_i(s,a) \right| \\ &= \left| \sum_{i \in T_1^t(s,a)} \prod_{j \in T_{i+1}^j(s,a)} (1-\alpha\beta_j)\alpha\beta_i\varepsilon_i(s,a) \right| \\ &= \left| \sum_{i=1}^{|T_1^t(s,a)|} \prod_{j \in T_{t_i+1}^j(s,a)} (1-\alpha\beta_j)\alpha\beta_{t_i}\varepsilon_{t_i}(s,a) \right|, \end{aligned}$$

where t_i denotes the time stamp when (s,a) is sampled for the i^{th} time in the window $[1,t]$.

It suffices to show that for any fixed $m = |T_1^t(s,a)| \in [0,t]$, we have

$$\begin{aligned} \mathbb{P}\left(P_{2,t}(s,a) \geq \rho\right) &= \mathbb{P}\left(\left| \sum_{i=1}^{m} \prod_{j \in T_{t_i+1}^j(s,a)} (1-\alpha\beta_j)\alpha\beta_{t_i}\varepsilon_{t_i}(s,a) \right| \geq \rho \right) \\ &\leq 2\exp\left(-\frac{\rho^2}{2\alpha\kappa^2}\right) := \frac{\delta}{DT}. \end{aligned}$$

To this end, we observe that

$$\begin{aligned} \mathbb{P}\left(P_{2,t}(s,a) \geq \rho\right) &= \mathbb{P}\left(P_{2,t}(s,a) \geq \rho \middle| \sum_{i=1}^{m} \beta_{t_i} = 0\right) \mathbb{P}\left(\sum_{i=1}^{m} \beta_{t_i} = 0\right) \\ &\quad + \mathbb{P}\left(P_{2,t}(s,a) \geq \rho \middle| \sum_{i=1}^{m} \beta_{t_i} = 1\right) \mathbb{P}\left(\sum_{i=1}^{m} \beta_{t_i} = 1\right) \\ &\quad + \cdots \\ &\quad + \mathbb{P}\left(P_{2,t}(s,a) \geq \rho \middle| \sum_{i=1}^{m} \beta_{t_i} = m\right) \mathbb{P}\left(\sum_{i=1}^{m} \beta_{t_i} = m\right). \end{aligned} \tag{61}$$

Next, we state the following claim for further proof.

Claim: For any $k \in [0,m]$, we have

$$\mathbb{P}\left(P_{2,t}(s,a) \geq \rho \middle| \sum_{i=1}^{m} \beta_{t_i} = k\right) \leq 2\exp\left(-\frac{\rho^2}{2\alpha\kappa^2}\right).$$

Thus, we obtain

$$\mathbb{P}\left(P_{2,t}(s,a) \geq \rho\right) \leq 2\exp\left(-\frac{\rho^2}{2\alpha\kappa^2}\right) \sum_{k=0}^{m} \mathbb{P}\left(\sum_{i=1}^{m} \beta_{t_i} = k\right) = 2\exp\left(-\frac{\rho^2}{2\alpha\kappa^2}\right). \tag{62}$$

It remains to prove the claim, which can be done by observing

$$\begin{aligned}
&\mathbb{P}\left(P_{2,t}(s,a) \geq \rho \middle| \sum_{i=1}^{m} \beta_{t_i} = k\right) \\
&= \mathbb{P}\left(\left|\sum_{i=1}^{m} \prod_{j \in T^j_{t_{i+1}}(s,a)} (1-\alpha\beta_j)\alpha\beta_{t_i}\varepsilon_{t_i}(s,a)\right| \geq \rho \middle| \sum_{i=1}^{m} \beta_{t_i} = k\right) \\
&= \mathbb{P}\left(\left|\sum_{i=1}^{k} (1-\alpha)^{k-i}\alpha\varepsilon_{t'_i}(s,a)\right| \geq \rho\right) \\
&\overset{\text{(i)}}{\leq} 2\exp\left(-\frac{\rho^2}{2\alpha\kappa^2}\right),
\end{aligned} \tag{63}$$

where t'_i denotes the time stamp of the i^{th} non-zero β_{t_i} is the sequential array $(\beta_{t_1}, \beta_{t_2}, \ldots, \beta_{t_m})$, and (i) follows from (9) with the fact that $\mathbb{E}(\varepsilon_i(s,a)) = 0$ and $|\varepsilon_i(s,a)| \leq \kappa$.

I.1 Supporting Lemmas

The following lemma is useful to bound the sum of a sequence of discounted random variables (not necessarily independent).

Lemma 9. *Fix $k > 0$ and $\alpha \in (0,1)$. Given a sequence of random variables $\{X_i\}$ and a filtration $\{\mathcal{F}_i\}$ satisfying $\mathbb{E}(X_i|\mathcal{F}_i) = 0$ and $|X_i| \leq \bar{c}$, then*

$$\mathbb{P}\left(\left|\sum_{i=1}^{k} (1-\alpha)^{k-i}\alpha X_i\right| \geq \rho\right) \leq 2\exp\left(-\frac{\rho^2}{2\alpha\bar{c}^2}\right).$$

Proof. Define $\{M_i\}_{1\leq i\leq k}$ as

$$M_{i+1} = (1-\alpha)M_i + \alpha X_i, \qquad \text{with } M_1 = 0.$$

Clearly we have $M_{k+1} = \sum_{i=1}^{k}(1-\alpha)^{k-i}\alpha X_i$, and

$$\begin{aligned}
\mathbb{E}(M_{i+1}|\mathcal{F}_i) &= \mathbb{E}((1-\alpha)M_i + \alpha X_i|\mathcal{F}_i) \\
&= (1-\alpha)M_i + \mathbb{E}(\alpha X_i|\mathcal{F}_i) \\
&= (1-\alpha)M_i.
\end{aligned}$$

Next, we construct $\{\tilde{M}_i\}$ as

$$\tilde{M}_i := \begin{cases} M_{k+1}, & i = k+1, \\ (1-\alpha)^{k-i+1}M_i, & 1 \leq i \leq k. \end{cases}$$

Then $\{\tilde{M}_i\}_{1\leq i\leq k+1}$ is a martingale sequence with $\mathbb{E}\left(\tilde{M}_1\right) = 0$.

Observe that

$$d_i := \left|\tilde{M}_{i+1} - \tilde{M}_i\right| \leq \begin{cases} M_{k+1} - (1-\alpha)M_k = \alpha X_k, & i = k, \\ (1-\alpha)^{k-i}M_{i+1} - (1-\alpha)^{k-i+1}M_i = (1-\alpha)^{k-i}\alpha X_i, & 1 \leq i < k. \end{cases}$$

Then we have

$$d_i^2 \leq (1-\alpha)^{2(k-i)}\alpha^2|X_i|^2 \leq (1-\alpha)^{k-i}\alpha^2\bar{c}^2,$$

where the last inequality follows because $(1-\alpha)^2 < 1-\alpha$ and $|X_i| \leq \bar{c}$.

Applying the Azuma-Hoeffding Inequality (see Lemma 5) yields

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{i=1}^{k}(1-\alpha)^{k-i}\alpha X_i\right|\geq\rho\right)&=\mathbb{P}\left(M_{k+1}\geq\rho\right)\\&\leq 2\exp\left(-\frac{\rho^2}{2\sum_{i=1}^{k}d_i^2}\right)\\&\leq 2\exp\left(-\frac{\rho^2}{2\sum_{i=1}^{k}(1-\alpha)^{k-i}\alpha^2\bar{c}^2}\right)\\&\leq 2\exp\left(-\frac{\rho^2}{2\alpha\bar{c}^2}\right).\end{aligned}$$

□

J Proof of Proposition 6

We start with the dynamics of ν_t derived in Appendix C and have

$$\begin{aligned}\nu_{t+1}(s,a)&=Q_{t+1}^{B}(s,a)-Q_{t+1}^{A}(s,a)\\&=(1-\hat{\alpha}_t(s,a))\nu_t(s,a)+\hat{\alpha}_t(s,a)\mathcal{H}_t(s,a)+\hat{\alpha}_t(s,a)\mu_t(s,a)\\&=(1-\hat{\alpha}_t(s,a))\nu_t(s,a)+\hat{\alpha}_t(s,a)\left(\frac{1}{2}\nu_t(s,a)+\frac{\gamma}{2}\underbrace{\mathbb{E}_{s'}\left(Q_t^B(s',a^*)-Q_t^A(s',b^*)\right)}_{J_t(s,a)}\right)\\&\quad+\hat{\alpha}_t(s,a)\mu_t(s,a)\\&=\left(1-\frac{\hat{\alpha}_t(s,a)}{2}\right)\nu_t(s,a)+\frac{\gamma\hat{\alpha}_t(s,a)}{2}J_t(s,a)+\hat{\alpha}_t(s,a)\mu_t(s,a)\\&=\prod_{i=1}^{t}\left(1-\frac{\hat{\alpha}_i(s,a)}{2}\right)\nu_1(s,a)+\sum_{i=1}^{t}\prod_{j=i+1}^{t}\left(1-\frac{\hat{\alpha}_j(s,a)}{2}\right)\frac{\gamma\hat{\alpha}_i(s,a)}{2}J_i(s,a)\\&\quad+\sum_{i=1}^{t}\prod_{j=i+1}^{t}\left(1-\frac{\hat{\alpha}_j(s,a)}{2}\right)\hat{\alpha}_i(s,a)\mu_i(s,a)\\&\overset{\text{(i)}}{=}\sum_{i=1}^{t}\prod_{j=i+1}^{t}\left(1-\frac{\hat{\alpha}_j(s,a)}{2}\right)\frac{\gamma\hat{\alpha}_i(s,a)}{2}J_i(s,a)\\&\quad+\sum_{i=1}^{t}\prod_{j=i+1}^{t}\left(1-\frac{\hat{\alpha}_j(s,a)}{2}\right)\hat{\alpha}_i(s,a)\mu_i(s,a),\end{aligned}$$

where (i) follows because $\|\nu_1\|=0$.

Next, we have

$$\begin{aligned}|\nu_{t+1}(s,a)|&\leq\left|\sum_{i=1}^{t}\prod_{j=i+1}^{t}\left(1-\frac{\hat{\alpha}_j(s,a)}{2}\right)\frac{\hat{\alpha}_i(s,a)}{2}\gamma J_i(s,a)\right|\\&\quad+\left|\sum_{i=1}^{t}\prod_{j=i+1}^{t}\left(1-\frac{\hat{\alpha}_j(s,a)}{2}\right)\hat{\alpha}_i(s,a)\mu_i(s,a)\right|\\&\overset{\text{(i)}}{\leq}\sum_{i=1}^{t}\prod_{j=i+1}^{t}\left(1-\frac{\hat{\alpha}_j(s,a)}{2}\right)\frac{\hat{\alpha}_i(s,a)}{2}\gamma\left\|\nu_i\right\|\end{aligned}$$

$$+\left|\sum_{i=1}^{t}\prod_{j=i+1}^{t}\left(1-\frac{\hat{\alpha}_j(s,a)}{2}\right)\hat{\alpha}_i(s,a)\mu_i(s,a)\right|, \tag{64}$$

where (i) follows from the property $|J_i(s,a)| \le \|\nu_i\|$ derived in Appendix C.

Next, from Lemma 10 we have $\forall t \in [1,T], \forall (s,a)$, with probability at least $1-\delta$,

$$|\nu_{t+1}(s,a)| \le \sum_{i=1}^{t}\prod_{j=i+1}^{t}\left(1-\frac{\hat{\alpha}_j(s,a)}{2}\right)\frac{\hat{\alpha}_i(s,a)}{2}\gamma\|\nu_i\| + \sqrt{8\alpha\ln\left(\frac{2DT}{\delta}\right)}V_{\max}.$$

In the remaining, we close this proof by induction. The base case holds trivially since $\|\nu_1\| = 0$. Suppose that $\|\nu_t\| \le \sqrt{8\alpha\ln\left(\frac{2DT}{\delta}\right)}\frac{V_{\max}}{1-\gamma}$. Then we have

$$\begin{aligned}
\|\nu_{t+1}\| &\le \sum_{i=1}^{t}\prod_{j=i+1}^{t}\left(1-\frac{\hat{\alpha}_j(s,a)}{2}\right)\frac{\hat{\alpha}_i(s,a)}{2}\gamma\|\nu_i\| + \sqrt{8\alpha\ln\left(\frac{2DT}{\delta}\right)}V_{\max} \\
&\le \sum_{i=1}^{t}\prod_{j=i+1}^{t}\left(1-\frac{\hat{\alpha}_j(s,a)}{2}\right)\frac{\hat{\alpha}_i(s,a)}{2}\sqrt{8\alpha\ln\left(\frac{2DT}{\delta}\right)}\frac{\gamma V_{\max}}{1-\gamma} + \sqrt{8\alpha\ln\left(\frac{2DT}{\delta}\right)}V_{\max} \\
&\overset{\text{(i)}}{\le} \sqrt{8\alpha\ln\left(\frac{2DT}{\delta}\right)}\frac{\gamma V_{\max}}{1-\gamma} + \sqrt{8\alpha\ln\left(\frac{2DT}{\delta}\right)}V_{\max} \\
&= \sqrt{8\alpha\ln\left(\frac{2DT}{\delta}\right)}\frac{V_{\max}}{1-\gamma},
\end{aligned}$$

where (i) follows from Lemma 8 by replacing $\beta_i = \frac{1}{2}$ without affecting the upper bound.

J.1 Supporting Lemmas

Lemma 10. *Fix $\delta \in (0,1)$. Then for any $t \in [1,T]$ and for any (s,a), we have*

$$\left|\sum_{i=1}^{t}\prod_{j=i+1}^{t}\left(1-\frac{\hat{\alpha}_j(s,a)}{2}\right)\hat{\alpha}_i(s,a)\mu_i(s,a)\right| \le \sqrt{8\alpha\ln\left(\frac{2DT}{\delta}\right)}V_{\max}, \tag{65}$$

with probability at least $1-\delta$, where $D = |\mathcal{S}||\mathcal{A}|$.

Proof. Observe that

$$\begin{aligned}
&\sum_{i=1}^{t}\prod_{j=i+1}^{t}\left(1-\frac{\hat{\alpha}_j(s,a)}{2}\right)\hat{\alpha}_i(s,a)\mu_i(s,a) \\
&\quad= \sum_{i\in T_1^t(s,a)}\left(1-\frac{\alpha}{2}\right)^{|T_1^t(s,a)|-i}\alpha\mu_i(s,a) \\
&\quad= \sum_{i=1}^{|T_1^t(s,a)|}\left(1-\frac{\alpha}{2}\right)^{|T_1^t(s,a)|-i}\alpha\mu_{t_i}(s,a),
\end{aligned}$$

where t_i denotes the time stamp when (s,a) is sampled for the i^{th} time in the window $[1,t]$.

It suffices to show that for any $m = |T_1^t(s,a)| \in [0,t]$, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{i=1}^{t}\prod_{j=i+1}^{t}\left(1-\frac{\hat{\alpha}_j(s,a)}{2}\right)\hat{\alpha}_i(s,a)\mu_i(s,a)\right| \ge \rho\right) \\
&= \mathbb{P}\left(\left|\sum_{i=1}^{m}\left(1-\frac{\alpha}{2}\right)^{m-i}\alpha\mu_{t_i}(s,a)\right| \ge \rho\right) \\
&\le 2\exp\left(-\frac{\rho^2}{8\alpha V_{\max}^2}\right) := \frac{\delta}{DT}.
\end{aligned}$$

The last inequality can be shown based on Lemma 9 by observing that $|\mu_i(s,a)| \leq 2V_{\max}$ and $\mathbb{E}(\mu_{t_i}(s,a)|\mathcal{F}'_i) = \mathbb{E}(\mu_{t_i}(s,a)|\mathcal{F}_{t_i}) = 0$ as derived in Appendix C, which completes the proof.

□

K PROOF OF THEOREM 3

The adaption from Theorem 2 to Theorem 3 is the same as the proof of Theorem 2 in Li et al. (2020), and is thus omitted here.