1 SPECIFICATIONS OF FROZENLAKE PROBLEM

FrozenLake is a classic benchmark problem for Q-learning, in which an agent controls the movement of a character in an $n \times n$ grid world. Some tiles of the grid are walkable, and others lead to the agent falling into the water. Additionally, the movement direction of the agent is uncertain and only partially depends on the chosen direction. The agent is rewarded for finding a feasible path to a goal tile. As shown in Figure 1 with a Frozenlake-$8 \times 8$ task, “S” is the safe starting point, “F” is the safe frozen surface, “H” stands for the hole that terminates the game, and “G” is the target state that comes with an immediate reward of 1. This forms a problem with the state-space size $n^2$, the action-space size 4 and the reward space $R = \{0, 1\}$. For tabular Q-learning algorithms with finite state-action problems of relatively small dimensions, FrozenLake-$4 \times 4$ and FrozenLake-$8 \times 8$ are two typical benchmark tasks. As the grid world becomes large, e.g., FrozenLake-$128 \times 128$, Q-learning with linear function approximation is then adopted to solve the problem.

![Figure 1: The FrozenLake-$8 \times 8$ task environment.](image)

2 PROOF OF LEMMA 1

We bound the expectation of bias via constructing a new Markov chain and applying some techniques from information theory. Before deriving the bound, we first introduce some technical lemmas.

**Lemma 1.** Suppose Assumptions 1 and 3 hold. Then for $g_k$ defined in (12), we have $\|g_k\|_2 \leq G_{\text{max}}$ for all $k$, where $G_{\text{max}} = 2D_{\text{max}} + R_{\text{max}}$. 

*equal contribution

Supplement for the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021).
where (i) follows from Cauchy-Schwartz inequality and the assumption \( \| \Phi(x, u) \|_2 \leq 1, \| \theta \|_2 \leq D_{\max}, \) and \( \| R(x, u) \|_2 \leq R_{\max}, \) we have

\[
\| g_k \|_2 = \left\| (\Phi(x_k, u_k)^T \theta_k - R(x_k, u_k)) - \gamma \max_{u' \in U(x_{k+1})} \Phi(x_{k+1}, u')^T \theta_k \Phi(x_k, u_k) \right\|_2 \\
\leq \| \Phi(x_k, u_k)^T \theta_k \|_2 + \| R(x_k, u_k) \|_2 + \max_{u' \in U(x_{k+1})} \| \Phi(x_{k+1}, u')^T \theta_k \|_2 \\
\leq 2D_{\max} + R_{\max},
\]

where we use Cauchy-Schwartz inequality and the triangle inequality.

For notational simplicity, throughout this section we use \( O = (x, u, x') \) to denote the sample tuple and \( O_k = (x_k, u_k, x_{k+1}) \) to denote the sample tuple at time \( k \).

**Lemma 2.** Let \( \xi(\theta; O) := (g(\theta; O) - \bar{g}(\theta))^T (\theta - \theta^*) \). Then \( \xi(\theta; O) \) is uniformly bounded by

\[
\| \xi(\theta; O) \|_2 \leq 2D_{\max}G_{\max}, \quad \forall \theta \in B,
\]

and it is Lipschitz continuous with

\[
\| \xi(\theta; O) - \xi(\theta'; O) \|_2 \leq 2((1 + \gamma)D_{\max} + G_{\max}) \| \theta - \theta' \|_2, \quad \forall \theta, \theta' \in B.
\]

**Proof.** The first statement is straightforward based on Assumption 3 and Lemma 1. That is,

\[
\| \xi(\theta; O) \| \leq \| g(\theta; O) - \bar{g}(\theta) \|_2 \| \theta - \theta^* \|_2 \leq 2D_{\max}G_{\max}.
\]

Next to prove the Lipschtz condition, we first prove the Lipschitz condition of \( g(\theta; O_k) \) with respect to \( \theta \).

\[
\| g(\theta; O) - g(\theta'; O) \|_2 \leq \| \Phi(x, u)^T (\theta - \theta') \| + \gamma \max_{u' \in U(x')} \Phi(x', u')^T \theta' - \gamma \max_{u' \in U(x')} \Phi(x', u')^T \theta
\]

\[
\leq \| \Phi(x, u)^T (\theta - \theta') \| + \gamma \max_{u' \in U(x')} \Phi(x', u')^T \theta' - \gamma \max_{u' \in U(x')} \Phi(x', u')^T \theta,
\]

where (i) follows from Cauchy-Schwartz inequality and the assumption \( \| \Phi \|_2 \leq 1 \), and (ii) follows from the triangle inequality.

Now we consider two cases. If the item in the second norm of (ii) is non-negative, we let \( u^* = \arg \max_{u' \in U(x')} \Phi(x', u')^T \theta' \). Then \( \max_{u' \in U(x')} \Phi(x', u')^T \theta \geq \Phi(x', u^*)^T \theta' \). Thus, we continue to bound the above inequality as

\[
\| g(\theta; O) - g(\theta'; O) \|_2 \leq \| \Phi(x, u)^T (\theta - \theta') \| + \gamma \Phi(x', u^*)^T (\theta' - \theta)
\]

\[
| \Phi(x, u)^T (\theta - \theta') | + | \Phi(x', u^*)^T (\theta' - \theta) | \tag{1}
\]

Similarly, if this item is negative, we let \( u^* = \arg \max_{u' \in U(x')} \Phi(x', u')^T \theta' \). Then \( \max_{u' \in U(x')} \Phi(x', u')^T \theta \geq \Phi(x', u^*)^T \theta' \). Thus, we have

\[
\| g(\theta; O) - g(\theta'; O) \|_2 \leq \| \Phi(x, u)^T (\theta - \theta') \| + \gamma \Phi(x', u^*)^T (\theta' - \theta)
\]

\[
| \Phi(x, u)^T (\theta - \theta') | + | \Phi(x', u^*)^T (\theta' - \theta) | \tag{2}
\]

Then it follows from (1) and (2) that

\[
\| g(\theta; O) - g(\theta'; O) \|_2 \leq (1 + \gamma) \| \theta - \theta' \|.
\]

Similarly, we obtain the same result for \( \bar{g}(\theta) \) as follows.

\[
\| \bar{g}(\theta) - \bar{g}(\theta') \|_2 \leq \mathbb{E}_\mu \| g_k(\theta) - g_k(\theta') \|_2 \leq (1 + \gamma) \| \theta - \theta' \|.
\]
Then we focus on obtaining the second statement,

\[
|\xi(\theta; O) - \xi(\theta'; O)|
= |(g(\theta; O) - \tilde{g}(\theta))\xi(\theta - \theta^*) - (g(\theta'; O) - \tilde{g}(\theta'))\xi(\theta' - \theta^*)|
\leq \|g(\theta; O) - \tilde{g}(\theta)\|_2 \|\theta - \theta^*\|_2 + \|\theta' - \theta^*\|_2 \|(g(\theta; O) - \tilde{g}(\theta)) - (g(\theta'; O) - \tilde{g}(\theta'))\|_2
\leq 2G_{\max} \|\theta - \theta^*\|_2 + \max_{i} D_{\max} \|(g(\theta; O) - g(\theta'; O)) - (\tilde{g}(\theta) - \tilde{g}(\theta'))\|_2
\leq 2G_{\max} \|\theta - \theta^*\|_2 + \max_{i} (1 + \gamma) \|\theta - \theta^*\|_2
= 2((1 + \gamma)D_{\max} + G_{\max}) \|\theta - \theta^*\|_2,
\]

where (i) follows from Assumption 3 and Lemma 1, and (ii) follows from triangle inequality and (1).

We use \(X \to Z \to Y\) to indicate that the random variable \(X\) and \(Y\) are independent conditioned on \(Z\).

**Lemma 3.** [Bhandari et al., 2018, Lemma 9] Consider two random variables \(X\) and \(Y\) such that

\[
X \to x_k \to x_{k+r} \to Y,
\]

for fixed \(k\) and \(r > 0\). Suppose Assumption 4 holds. Let \(X', Y'\) are independent copies drawn from the marginal distributions of \(X\) and \(Y\), that is \(P(X' = \cdot, Y' = \cdot) = P(X = \cdot)P(Y = \cdot)\). Then, for any bounded \(v\), we have

\[
|\mathbb{E}[v(X, Y)] - \mathbb{E}[v(X', Y')]| \leq 2 \|v\|_{\infty}(\sigma \rho^r).
\]

We continue the proof of Lemma 3. We first develop the connection between \(\xi(\theta_k; O_k)\) and \(\xi(\theta_{k-r}; O_k)\) via Lemma 2. To do so, we first observe that

\[
\|\theta_{i+1} - \theta_i\|_2 = \|\beta_i(\theta_i - \theta_{i-1}) + a_i(1 + b_i)g_i + a_ib_i g_{i-1}\|_2
\leq \|\beta_i(\theta_i - \theta_{i-1})\|_2 + \|a_i(1 + b_i)g_i\|_2 + \|a_ib_i g_{i-1}\|_2
\leq D_{\max}\beta_i + 3G_{\max} a_i,
\]

where (i) follows from the triangle inequality and (ii) from the Assumptions 5 and 1 and the fact \(b_i < 1\). Then we have

\[
\|\theta_k - \theta_{k-r}\|_2 \leq \sum_{i=k-r}^{k-1} \|\theta_{i+1} - \theta_i\|_2 \leq \max_{i=k-r} \sum_{i=k-r}^{k-1} \beta_i + 3G_{\max} \sum_{i=k-r}^{k-1} a_i.
\]

Thus, we can relate \(\xi(\theta_k; O_k)\) and \(\xi(\theta_{k-r}; O_k)\) by using the Lipschitz property established in Lemma 2 as follows:

\[
\xi(\theta_k; O_k) - \xi(\theta_{k-r}; O_k) \leq \|\theta_k - \theta_{k-r}\|_2 \leq 2((1 + \gamma)D_{\max} + G_{\max}) \|\theta_k - \theta_{k-r}\|_2
\leq 2((1 + \gamma)D_{\max} + G_{\max}) \left(\sum_{i=k-r}^{k-1} \beta_i + 3G_{\max} \sum_{i=k-r}^{k-1} a_i\right). \tag{4}
\]

Next, we bound \(\mathbb{E}[\xi(\theta_{k-r}; O_k)]\) using Lemma 3. Observe that given any deterministic \(\theta \in B\), we have

\[
\mathbb{E}[\xi(\theta; O_k)] = (\mathbb{E}[g(\theta; O_k)] - \tilde{g}(\theta))^T(\theta - \theta^*) = 0.
\]

Since \(\theta_0\) is a fixed constant, we have \(\mathbb{E}[\xi(\theta_0, O_k)] = 0\). Now we are ready to bound \(\mathbb{E}[\xi(\theta_{k-r}, O_k)]\) via Lemma 3 by constructing a random process satisfying (3). To do so, consider random variables \(\theta_{k-r}'\) and \(O_k'\) drawn independently from the marginal distribution of \(\theta_{k-r}\) and \(O_k\), so that \(\mathbb{P}(\theta_{k-r}' = \cdot, O_k' = \cdot) = \mathbb{P}(\theta_{k-r} = \cdot)\mathbb{P}(O_k = \cdot)\). We further obtain \(\mathbb{E}[\xi(\theta_{k-r}', O_k')] = \mathbb{E}[\mathbb{E}[\xi(\theta_{k-r}', O_k')|O_k']] = 0\) since \(\theta_{k-r}'\) and \(O_k'\) are independent. Combining Lemmas 2 and 3, we have

\[
\mathbb{E}[\xi(\theta_{k-r}, O_k)] \leq 2(2D_{\max}G_{\max})(\sigma \rho^r). \tag{5}
\]
Finally, we are ready to bound the bias. We first take expectation for both sides of (4) and obtain

$$E[\xi(\theta_k; O_k)] \leq E[\xi(\theta_{k-\tau}; O_k)] + 2((1 + \gamma) D_{\text{max}} + G_{\text{max}}) \left( D_{\text{max}} \sum_{i=k-\tau}^{k-1} \beta_i + 3G_{\text{max}} \sum_{i=k-\tau}^{k-1} a_i \right).$$

When $k \leq \tau_{\text{mix}}(\kappa)$, we choose $\tau = k$ and have

$$E[\xi(\theta_k; O_k)] \leq E[\xi(\theta_0; O_k)] + 2((1 + \gamma) D_{\text{max}} + G_{\text{max}}) \left( D_{\text{max}} \sum_{i=0}^{k-1} \beta_i + 3G_{\text{max}} \sum_{i=0}^{k-1} a_i \right)$$

$$= 2((1 + \gamma) D_{\text{max}} + G_{\text{max}}) \left( D_{\text{max}} \sum_{i=0}^{k-1} \beta_i + 3G_{\text{max}} \sum_{i=0}^{k-1} a_i \right).$$

When $k > \tau_{\text{mix}}(\kappa)$, we choose $\tau = \tau^* := \tau_{\text{mix}}(\kappa)$ and have

$$E[\xi(\theta_k; O_k)]$$

$$\leq E[\xi(\theta_{k-\tau^*}; O_k)] + 2((1 + \gamma) D_{\text{max}} + G_{\text{max}}) \left( D_{\text{max}} \sum_{i=k-\tau^*}^{k-1} \beta_i + 3G_{\text{max}} \sum_{i=k-\tau^*}^{k-1} a_i \right)$$

$$(i) \leq 4D_{\text{max}} G_{\text{max}} \sigma \rho^{\tau^*} + 2((1 + \gamma) D_{\text{max}} + G_{\text{max}}) \left( D_{\text{max}} \sum_{i=k-\tau^*}^{k-1} \beta_i + 3G_{\text{max}} \sum_{i=k-\tau^*}^{k-1} a_i \right)$$

$$(ii) \leq 4D_{\text{max}} G_{\text{max}} \kappa + 2((1 + \gamma) D_{\text{max}} + G_{\text{max}}) \left( D_{\text{max}} \sum_{i=k-\tau^*}^{k-1} \beta_i + 3G_{\text{max}} \sum_{i=k-\tau^*}^{k-1} a_i \right)$$

$$(iii) \leq 4D_{\text{max}} G_{\text{max}} \kappa + 2((1 + \gamma) D_{\text{max}} + G_{\text{max}}) \left( D_{\text{max}} \sum_{i=k-\tau^*}^{k-1} \beta_i + 3G_{\text{max}} \sum_{i=k-\tau^*}^{k-1} a_i \right) \cdot \kappa,$$

where (i) follows from (5), (ii) follows due to the definition of the mixing time, and (iii) follows because $a_k, \beta_k$ are non-increasing.

### 3 PROOF OF THEOREM 1

Recall that MomentumQ with linear function approximation updates as (12). Given the unique fixed point $\theta^*$ and denoting $b_k + c_k = \beta_k$, we have

$$\|\theta_{k+1} - \theta^*\|^2 = \|\theta_k - \theta^* + \beta_k(\theta_k - \theta_{k-1}) - a_k(1 + b_k)g_k + a_kb_kg_{k-1}\|^2$$

$$= \|\theta_k - \theta^*\|^2 + \|\beta_k(\theta_k - \theta_{k-1}) - a_k(1 + b_k)g_k + a_kb_kg_{k-1}\|^2$$

$$+ 2\langle \theta_k - \theta^*, \beta_k(\theta_k - \theta_{k-1}) - a_k(1 + b_k)g_k + a_kb_kg_{k-1}\rangle$$

$$= \|\theta_k - \theta^*\|^2 + \|\beta_k(\theta_k - \theta_{k-1}) - a_k(1 + b_k)g_k + a_kb_kg_{k-1}\|^2$$

$$+ 2\|\beta_k(\theta_k - \theta_{k-1}) - a_k(1 + b_k)g_k + a_kb_kg_{k-1}\|_{\theta_k - \theta^*, g_k}.$$ 

Next, taking the expectation over all the randomness up to time step $k$ on both sides, we have

$$E \|\theta_k - \theta^*\|^2$$

$$= \|\theta_k - \theta^*\|^2 + E \|\beta_k(\theta_k - \theta_{k-1}) - a_k(1 + b_k)g_k + a_kb_kg_{k-1}\|^2$$

$$+ 2E\langle \theta_k - \theta^*, \beta_k(\theta_k - \theta_{k-1}) - a_k(1 + b_k)g_k + a_kb_kg_{k-1}\rangle$$

$$\leq \|\theta_k - \theta^*\|^2 + 3\|\beta_k(\theta_k - \theta_{k-1}) - a_k(1 + b_k)g_k + a_kb_kg_{k-1}\|^2$$

$$+ 2\|\beta_k(\theta_k - \theta_{k-1}) - a_k(1 + b_k)g_k + a_kb_kg_{k-1}\|_{\theta_k - \theta^*, g_k}.$$
We consider a constant stepsize $\alpha_k = \alpha$. For notational simplicity, we denote $f_k = 5\beta_k D_{\text{max}}^2 + 15\alpha_k^2 G_{\text{max}}^2 + 2a_k b_k D_{\text{max}} G_{\text{max}}$, and $\zeta_k = -2a_k(1 + b_k)E[\xi(\theta_k; O_k)]$. Then for $k > \tau^*$ we have

$$
E[\|\theta_{k+1} - \theta^*\|_2^2] 
\leq (1 - 2\alpha\delta(1 + b_k))E[\|\theta_k - \theta^*\|_2^2] + f_k + \zeta_k 
\leq \ldots
\leq \prod_{i=0}^k (1 - 2\alpha\delta(1 + b_i)) \|\theta_0 - \theta^*\|_2^2 + \sum_{i=0}^k f_i \prod_{j=i+1}^k (1 - 2\alpha\delta(1 + b_j))
\quad + \sum_{i=\tau^*+1}^k \zeta_i \prod_{j=i+1}^\tau (1 - 2\alpha\delta(1 + b_j)) + \tau^* \sum_{i=0}^{\tau^*} \zeta_i \prod_{j=i+1}^\tau (1 - 2\alpha\delta(1 + b_j))
\leq \prod_{i=0}^k (1 - 2\alpha\delta(1 + b_i)) \|\theta_0 - \theta^*\|_2^2 + \sum_{i=0}^k f_i(1 - 2\alpha\delta)^{k-i}
\quad + \sum_{i=\tau^*+1}^k \zeta_i(1 - 2\alpha\delta)^{k-i} + \tau^* \sum_{i=0}^{\tau^*} \zeta_i(1 - 2\alpha\delta)^{k-i},
$$

where the last inequality follows because $b_k > 0$, $\forall k$. Further, we bound the term $\sum_{i=0}^k (1 - 2\alpha\delta)^{k-i} f_i$ as

$$
\sum_{i=0}^k (1 - 2\alpha\delta)^{k-i} f_i = 5D_{\text{max}}^2 \sum_{i=0}^k (1 - 2\alpha\delta)^{k-i} b_i + 15\alpha^2 G_{\text{max}}^2 \sum_{i=0}^k (1 - 2\alpha\delta)^{k-i} + 2a D_{\text{max}} G_{\text{max}} \sum_{i=0}^k (1 - 2\alpha\delta)^{k-i} b_i
$$
\[
\leq 15\alpha^2G_{\max}^2\sum_{i=0}^{k}(1-2\delta\alpha)^{k-i} + (5D_{\max}^2 + 2\alpha D_{\max}G_{\max})\sum_{i=0}^{k}(1-2\delta\alpha)^{k-i}\beta_i \\
\leq \frac{15\alpha^2G_{\max}^2}{2\delta} + (5D_{\max}^2 + 2\alpha D_{\max}G_{\max})\beta(1-2\delta\alpha)^k \sum_{i=0}^{k} \left(\frac{\lambda}{1-2\delta\alpha}\right)^i \\
\leq \frac{15\alpha^2G_{\max}^2}{2\delta} + (5D_{\max}^2 + 2\alpha D_{\max}G_{\max})\beta(1-2\delta\alpha)^k \frac{1}{1-2\delta\alpha - \lambda}, \\
\]

where (i) follows from \(\alpha < \frac{1-\lambda}{2\delta}\). It remains to bound the last two tail terms. From Lemma, we obtain

\[
\zeta_i = \begin{cases} 
2\alpha(1 + b_i) \left( \eta_1 \sum_{i=1}^{k-1} \beta_i + \eta_2 \sum_{i=1}^{k-1} a_i \right) & \leq 4\alpha (\eta_1 \tau^* \beta + \eta_2 \tau^* \alpha), \quad i \leq \tau^*; \\
4\alpha (4D_{\max} G_{\max} \kappa + \eta_1 \tau^* \beta_{i-\tau^*} + \eta_2 \tau^* \alpha), & i > \tau^*,
\end{cases}
\]

where \(\eta_1 = 2D_{\max}(1+\gamma)D_{\max} + G_{\max}\), \(\eta_2 = 6G_{\max}(1+\gamma)D_{\max} + G_{\max}\). Then we obtain

\[
\sum_{i=\tau^*+1}^{k} \zeta_i(1-2\alpha\delta)^{k-i} + \sum_{i=0}^{\tau^*} \zeta_i(1-2\alpha\delta)^{k-i} \\
\leq 4\eta_1 \tau^* \alpha^2 \sum_{i=0}^{k}(1-2\alpha\delta)^{k-i} + 4\alpha \eta_1 \tau^* \beta \sum_{i=0}^{\tau^*}(1-2\alpha\delta)^{k-i} \\
+ 16D_{\max} G_{\max} \kappa \alpha \sum_{i=\tau^*+1}^{k}(1-2\alpha\delta)^{k-i} + 4\alpha \eta_1 \tau^* \sum_{i=\tau^*+1}^{k} \beta_{i-\tau^*}(1-2\alpha\delta)^{k-i} \\
\leq \frac{2\eta_1 \tau^* \alpha}{\delta} + \frac{2\eta_1 \tau^* \beta}{\delta} (1-2\alpha\delta)^{k-\tau^*} + \frac{8D_{\max} G_{\max} \kappa}{\delta} + 4\alpha \beta \eta_1 \tau^* \sum_{i=\tau^*+1}^{k} \lambda^i(1-2\alpha\delta)^{k-i} \\
= \frac{2\eta_1 \tau^* \alpha}{\delta} + \frac{2\eta_1 \tau^* \beta}{\delta} (1-2\alpha\delta)^{k-\tau^*} + \frac{8D_{\max} G_{\max} \kappa}{\delta} + \frac{4\alpha \beta \eta_1 \tau^* \lambda}{1-2\alpha\delta - \lambda} (1-2\alpha\delta)^{k-\tau^*},
\]

where the last inequality follows due to the fact that \(\alpha < \frac{1-\lambda}{2\delta}\). Thus, we can conclude that

\[
E \|\theta_{k+1} - \theta^*\|_2^2 \\
\leq \prod_{i=0}^{k} \left( (1-2\alpha\delta(1+b_i)) \|\theta_0 - \theta^*\|_2^2 + \sum_{i=0}^{k} f_i(1-2\alpha\delta)^{k-i} \right) \\
+ \sum_{i=\tau^*+1}^{k} \zeta_i(1-2\alpha\delta)^{k-i} + \sum_{i=0}^{\tau^*} \zeta_i(1-2\alpha\delta)^{k-i} \\
\leq \prod_{i=0}^{k} \left( (1-2\alpha\delta(1+b_i)) \|\theta_0 - \theta^*\|_2^2 + \frac{15\alpha^2G_{\max}^2}{2\delta} + \beta(5D_{\max}^2 + 2\alpha D_{\max}G_{\max})(1-2\delta\alpha)^k \right) \\
+ \frac{2\eta_1 \tau^* \alpha}{\delta} + \frac{2\eta_1 \tau^* \beta}{\delta} (1-2\alpha\delta)^{k-\tau^*} + \frac{8D_{\max} G_{\max} \kappa}{\delta} + \frac{4\alpha \beta \eta_1 \tau^* \lambda}{1-2\alpha\delta - \lambda} (1-2\alpha\delta)^{k-\tau^*} \\
\leq \prod_{i=0}^{k} \left( (1-2\alpha\delta(1+b_i)) \|\theta_0 - \theta^*\|_2^2 + \frac{15\alpha^2G_{\max}^2}{2\delta} + \frac{2\eta_1 \tau^* \alpha}{\delta} + \frac{8D_{\max} G_{\max} \kappa}{\delta} \right) \\
+ \beta \left( \frac{2\eta_1 \tau^*}{\delta} + \frac{5D_{\max}^2 + 2\alpha D_{\max} G_{\max} + 4\alpha \eta_1 \tau^* \lambda}{1-2\delta\alpha - \lambda} \right) (1-2\delta\alpha)^{k-\tau^*}.
\]
4 PROOF OF THEOREM 2

Before proving this theorem, we introduce two lemmas of series sum that will help to streamline the presentation.

Lemma 4. Let \( a_k = \frac{\alpha}{\sqrt{k}} \) and \( \beta_k = \beta \lambda^k \) with \( \alpha > 0, \beta, \lambda \in (0, 1) \) for \( k = 1, 2, \ldots \). Then

\[
\sum_{k=1}^{T} \frac{\beta_k}{a_k} \leq \frac{\beta}{\alpha (1 - \lambda)^2}.
\]

Proof. The proof is based on taking the standard sum of geometric sequences as follows:

\[
\sum_{k=1}^{T} \frac{\beta_k}{a_k} = \sum_{k=1}^{T} \frac{\beta \lambda^k \sqrt{k}}{\alpha} \leq \sum_{k=1}^{T} \frac{\beta \lambda^k}{\alpha} = \frac{\beta}{\alpha (1 - \lambda)} \left( \sum_{k=1}^{T} \lambda^k - T \lambda^T \right) \leq \frac{\beta}{\alpha (1 - \lambda)^2}.
\]

Lemma 5. Let \( a_k = \frac{\alpha}{\sqrt{k}} \). Then

\[
\sum_{k=1}^{T} a_k \leq 2\alpha \sqrt{T}.
\]

Proof. We use the comparison principle to bound the series sum as follows:

\[
\sum_{k=1}^{T} a_k = \sum_{k=1}^{T} \frac{\alpha}{\sqrt{k}} \leq \int_{1}^{T+1} \frac{\alpha}{\sqrt{t}} \, dt = 2\alpha \sqrt{T - 1} |_{T=1}^{T+1} = 2\alpha \sqrt{T}.
\]

The proof of Theorem 2 is partially similar to that of Theorem 1. The steps are the same until (8), where we have

\[
\mathbb{E} \| \theta_{k+1} - \theta^* \|_2^2 \\
\leq \mathbb{E} \| \theta_k - \theta^* \|_2^2 + 5 \beta_k D_{\text{max}}^2 + 15 a_k^2 G_{\text{max}}^2 + 2 a_k b_k D_{\text{max}} G_{\text{max}} - 2 a_k (1 + b_k) \delta \mathbb{E} \| \theta_k - \theta^* \|_2^2 \\
- 2 a_k (1 + b_k) \mathbb{E} |\xi(\theta_k; O_k)|.
\]

Then we continue the proof with rearranging the previous inequality:

\[
2 \delta \mathbb{E} \| \theta_k - \theta^* \|_2^2 \\
\leq 2 (1 + b_k) \delta \mathbb{E} \| \theta_k - \theta^* \|_2^2 \\
\leq \mathbb{E} \| \theta_k - \theta^* \|_2^2 - \mathbb{E} \| \theta_{k+1} - \theta^* \|_2^2 \\
- \frac{5 \beta_k}{a_k} \mathbb{E} D_{\text{max}}^2 + 15 a_k^2 G_{\text{max}}^2 + 2 b_k D_{\text{max}} G_{\text{max}} + 4 \mathbb{E} |\xi(\theta_k; O_k)|.
\]

Then we sum over time step \( k \) from 1 to \( T(T > \tau^*) \) and obtain

\[
2 \delta \sum_{k=1}^{T} \mathbb{E} \| \theta_k - \theta^* \|_2^2 \\
\leq \sum_{k=1}^{T} \frac{\mathbb{E} \| \theta_k - \theta^* \|_2^2 - \mathbb{E} \| \theta_{k+1} - \theta^* \|_2^2}{a_k} + 4 \sum_{k=1}^{T} \mathbb{E} |\xi(\theta_k; O_k)| + 4 \sum_{k=\tau^*+1}^{T} \mathbb{E} |\xi(\theta_k; O_k)| \\
+ 5 D_{\text{max}}^2 \sum_{k=1}^{T} \frac{\beta_k}{a_k} + 15 G_{\text{max}}^2 \sum_{k=1}^{T} a_k + 2 D_{\text{max}} G_{\text{max}} \sum_{k=1}^{T} b_k \\
= \frac{\| \theta_1 - \theta^* \|_2^2}{a_1} + \sum_{k=2}^{T} \mathbb{E} \| \theta_k - \theta^* \|_2^2 \left( \frac{1}{a_k} - \frac{1}{a_{k-1}} \right) - \frac{\mathbb{E} \| \theta_{T+1} - \theta^* \|_2^2}{a_{T+1}}.
\]
\[
\begin{align*}
&+ 5D_{\max}^2 \sum_{k=1}^{T} \frac{\beta_k}{a_k} + 15G_{\max}^2 \sum_{k=1}^{T} a_k + 2D_{\max}G_{\max} \sum_{k=1}^{T} b_k \\
&+ 4 \sum_{k=1}^{T} |E[\xi(\theta_k; O_k)]| + 4 \sum_{k=T^*+1}^{T} |E[\xi(\theta_k; O_k)]|
\end{align*}
\]

\[
\leq \frac{\|\theta_i - \theta^*\|_2^2}{a_1} + D_{\max}^2 \sum_{k=2}^{T} \left( \frac{1}{a_k} - \frac{1}{a_{k-1}} \right)
\]

\[
+ 5D_{\max}^2 \sum_{k=1}^{T} \frac{\beta_k}{a_k} + 15G_{\max}^2 \sum_{k=1}^{T} a_k + 2D_{\max}G_{\max} \sum_{k=1}^{T} b_k \\
+ 4 \sum_{k=1}^{T} |E[\xi(\theta_k; O_k)]| + 4 \sum_{k=T^*+1}^{T} |E[\xi(\theta_k; O_k)]|
\]

\[
\leq \frac{D_{\max}^2}{\alpha T} + 5D_{\max}^2 \sum_{k=1}^{T} \frac{\beta_k}{a_k} + 15G_{\max}^2 \sum_{k=1}^{T} a_k + 2D_{\max}G_{\max} \sum_{k=1}^{T} \beta_k
\]

\[
+ 4 \sum_{k=1}^{T} |E[\xi(\theta_k; O_k)]| + 4 \sum_{k=T^*+1}^{T} |E[\xi(\theta_k; O_k)]|
\]

\[
\leq D_{\max}^2 \frac{\sqrt{T}}{\alpha} + \frac{5\beta D_{\max}^2}{\alpha(1 - \lambda)^2} + 30\alpha G_{\max}^2 \sqrt{T} + \frac{2D_{\max}G_{\max}\lambda}{1 - \lambda}
\]

\[
+ 4 \sum_{k=1}^{T} |E[\xi(\theta_k; O_k)]| + 4 \sum_{k=T^*+1}^{T} |E[\xi(\theta_k; O_k)]|
\]

where (i) follows from Assumption \[\text{Assumption 3}\] and the fact that \( \alpha_k < \alpha_{k-1} \), and \( E[\|\theta_{T+1} - \theta^*\|_2^2 / a_{T+1}] > 0 \), (ii) holds due to Assumption \[\text{Assumptions 1-4}\] and (iii) follows from Lemmas \[\text{Lemma 1}\] and \[\text{Lemma 5}\].

It remains to bound \( 4 \sum_{k=1}^{T} |E[\xi(\theta_k; O_k)]| + 4 \sum_{k=T^*+1}^{T} |E[\xi(\theta_k; O_k)]| \). We bound the tail term by using Lemma \[\text{Lemma 1}\].

For simplicity, in the following we denote

\[
\eta_1 = 2D_{\max}((1 + \gamma)D_{\max} + G_{\max}), \quad \eta_2 = 6G_{\max}((1 + \gamma)D_{\max} + G_{\max}).
\]

Following from Lemma \[\text{Lemma 4}\] we have

\[
\sum_{k=1}^{T^*} |E[\xi(\theta_k; O_k)]| \leq \sum_{k=1}^{T^*} \eta_1 \sum_{i=1}^{k-1} \beta_i + \sum_{k=1}^{k-1} \eta_2 \sum_{i=1}^{a_i}
\]

\[
\leq T^* \eta_1 \beta + T^* \eta_2 a_k
\]

\[
\leq \frac{T^* \eta_1 \beta \lambda}{1 - \lambda} + 2T^* \eta_2 \alpha \sqrt{T}.
\]

Similarly, we obtain

\[
\sum_{k=T^*+1}^{T} |E[\xi(\theta_k; O_k)]| \leq \sum_{k=T^*+1}^{T} (4D_{\max}G_{\max} \kappa + \eta_1 \beta_{k-T^*} + \eta_2 \alpha_{k-T^*})
\]

\[
\leq 4D_{\max}G_{\max} \kappa T + \eta_1 \beta_{k-T^*} + \eta_2 \alpha_{k-T^*}
\]

\[
\leq 4D_{\max}G_{\max} \kappa T + \frac{T^* \eta_1 \beta \lambda}{1 - \lambda} + 2T^* \eta_2 \alpha \sqrt{T}.
\]
Thus, we have
\[
2\delta \sum_{k=1}^{T} \mathbb{E} \|\theta_k - \theta^*\|_2^2 \\
\leq \frac{D_{\max}^2 \sqrt{T}}{\alpha} + \frac{5\beta D_{\max}^2}{\alpha(1 - \lambda)^2} + 30\alpha G_{\max}^2 \sqrt{T} + \frac{2D_{\max}G_{\max}\beta\lambda}{1 - \lambda} \\
+ 4 \sum_{k=1}^{\tau^*} |\mathbb{E}[\xi(\theta_k; O_k)]| + 4 \sum_{k=\tau^* + 1}^{T} |\mathbb{E}[\xi(\theta_k; O_k)]|
\]
\[
\leq \frac{D_{\max}^2 \sqrt{T}}{\alpha} + \frac{5\beta D_{\max}^2}{\alpha(1 - \lambda)^2} + 30\alpha G_{\max}^2 \sqrt{T} + \frac{2D_{\max}G_{\max}\beta\lambda}{1 - \lambda} \\
+ 16D_{\max}G_{\max}\kappa T + \frac{8\tau^*\eta_1\beta\lambda}{1 - \lambda} + 16\tau^*\eta_2\alpha \sqrt{T}.
\]

Finally, we apply Jensen’s inequality and complete the proof as
\[
\mathbb{E} \|\theta_{\text{out}} - \theta^*\|_2^2 = \mathbb{E} \left\| \frac{1}{T} \sum_{k=1}^{T} \theta_k - \theta^* \right\|_2^2 \\
\leq \frac{1}{T} \sum_{k=1}^{T} \mathbb{E} \|\theta_k - \theta^*\|_2^2 \\
\leq \frac{D_{\max}^2/\alpha + 30\alpha G_{\max}^2 + 16\tau^*\alpha\eta_2}{2\delta\sqrt{T}} + \frac{8D_{\max}G_{\max}\kappa}{\delta(1 - \lambda)^2} \\
+ \frac{5\beta D_{\max}^2}{2\alpha\delta(1 - \lambda)^2} + \frac{D_{\max}G_{\max}\beta\lambda + 4\tau^*\eta_1\beta\lambda}{\delta(1 - \lambda)}.
\]

5 PROOF OF PROPOSITION 1

Proof. For convenience, we denote \(MQ_k(y_k) := \max_{u \in U(y_k)} Q_k(y_k, u)\), then \(\hat{T}_k Q_k = R + MQ_k(y_k)\) and \(\hat{T}_k Q_{k-1} = R + MQ_{k-1}(y_k)\). If \(k = 0\), we have from (20) that
\[
\|D_0 [Q_0, Q_{-1}]\| = \|\hat{T}_0 Q_0\| \leq \|R\| + \gamma \|MQ_0(y_0)\| \\
\leq R_{\max} + \gamma V_{\max}.
\]

Now, considering \(k \geq 1\) we have
\[
\|D_k [Q_k, Q_{k-1}]\| \\
\leq \|R\| + \gamma ||(1 + b_k)MQ_k - b_kMQ_{k-1}|| \\
\leq R_{\max} + \gamma ||(1 + b_k)MQ_{k-1} - \alpha_{k-1}Q_{k-2} + \alpha_{k-1}D_{k-1}[Q_{k-1}, Q_{k-2}] - b_kMQ_{k-1}|| \\
\leq R_{\max} + \gamma ||Q_{k-1}|| + \gamma ||1 + b_k|a_{k-1}||Q_{k-2}|| + \gamma ||1 + b_k|\alpha_{k-1}||D_{k-1}[Q_{k-1}, Q_{k-2}]||. \quad (12)
\]

where (i) follows from the triangle inequality and (ii) follows from the definition of the infinity norm.

To proceed to bound (12), we consider two cases. If \(k < \frac{m}{2}\), there are at most a finite number of \(D_k\)’s, which are obviously bounded. If \(k \geq \frac{m}{2}\), we have \(|1 + b_k|a_{k-1} = \frac{k-m}{k} \leq 1\). It follows from (12) that
\[
\|D_k [Q_k, Q_{k-1}]\| \\
\leq R_{\max} + \gamma ||Q_{k-1}|| + \gamma ||Q_{k-2}|| + \gamma ||D_{k-1}[Q_{k-1}, Q_{k-2}]|| \\
\leq R_{\max} + 2\gamma V_{\max} + \gamma ||D_{k-1}[Q_{k-1}, Q_{k-2}]|| \\
\leq R_{\max} \sum_{i=0}^{k-[m/2]} \gamma^i + 2V_{\max} \sum_{i=1}^{k-[m/2]} \gamma^i + \gamma^{k-[m/2]} ||D_{[m/2]}[Q_{[m/2]}, Q_{[m/2]-1}]|| \quad (13)
\]
where \(|x|\) denotes the largest integer that is no larger than \(x\). Note that (i) follows from the boundedness of \(Q_k\) (Assumption 5), and (ii) follows from applying (i) to \(D_t\) for \(t = k - 1, k - 2, \ldots, \lfloor m/2 \rfloor + 1\) iteratively. Since \(\gamma < 1\), the first two items in (ii) are bounded. Obviously, the third item is also bounded. Therefore, there exists some constant \(\bar{D}\), such that \(\|D_k\| \leq \bar{D}, \forall k \geq 0\).

The bound on \(\epsilon_k\) follows directly from its definition as

\[
\|\epsilon_k\| = \|E \left( D_k \left[ Q_k, Q_{k-1} \right] (x, u) \big| F_{k-1} \right) - D_k \left[ Q_k, Q_{k-1} \right]\| \\
\leq 2 \|D_k \left[ Q_k, Q_{k-1} \right]\| \leq 2\bar{D}.
\]

Thus we conclude our proof.

\[
\square
\]

### 6 PROOF OF THEOREM 3

We first prove two lemmas that will be useful for establishing the main results. The first lemma derives the dynamics of \(Q_k\) in terms of \(E_k\).

**Lemma 6.** Consider MomentumQ as in Algorithm 1. For any \(k \geq 1\), we have

\[
Q_k = \frac{1}{k} (Q_{k-1} - Q_0 + (k - m - 1)TQ_{k-1}) + \frac{1}{k} ((m + 1)TQ_0 - E_{k-1}).
\]

**Proof.** We prove the lemma by substituting the learning rates \(a_k, b_k, c_k\) in Algorithm 1 and using induction. From (19), we see that \(Q_1 = T_1Q_0 = TQ_0 - E_0\). Thus (14) holds when \(k = 1\). Now assume (14) holds for a certain integer \(k > 1\) we prove it also holds for \(k + 1\). To see this, we rewrite (19) as

\[
Q_{k+1} = \frac{1}{k+1} Q_k - \frac{1}{k+1} Q_{k-1} + \frac{k}{k+1} Q_k + \frac{1}{k+1} \left[ (k-m)T_k Q_k - (k-m-1)T_k Q_{k-1} \right] \\
= \frac{1}{k+1} Q_k - \frac{1}{k+1} Q_{k-1} + \frac{1}{k+1} \left[ (k-m)T_k Q_k - (k-m-1)T_k Q_{k-1} \right] \\
+ (m+1)TQ_0 - E_{k-1}) + \frac{1}{k+1} \left[ (k-m)T_k Q_k - (k-m-1)T_k Q_{k-1} \right] \\
= \frac{1}{k+1} Q_k - \frac{1}{k+1} Q_{k-1} + \frac{1}{k+1} \left[ (k-m)T_k Q_k - (k-m-1)T_k Q_{k-1} \right] \\
+ (m+1)TQ_0 - E_{k-1}) + \frac{1}{k+1} \left[ (k-m)T_k Q_k - (k-m-1)T_k Q_{k-1} - \epsilon_k \right] \\
= \frac{1}{k+1} (Q_k - Q_0 + (k-m)TQ_k + (m+1)TQ_0 - E_k),
\]

which shows that (14) holds for \(k + 1\). Therefore, by induction (14) holds for all \(k \geq 1\).

The second lemma derives the propagation of the approximation errors \(\epsilon_k\) in the process of Q-function iteration, which can be proved conveniently using Lemma 6.

**Lemma 7.** Suppose Assumption 5 holds and \(m \geq \frac{1}{\gamma}\) as in Algorithm 1. Then for all \(k \geq m + 1\), we have

\[
\|Q^* - Q_k\| \leq \frac{\text{h}V_{\text{max}}}{k(1 - \gamma)} + \frac{1}{k} \sum_{i=0}^{k - \lfloor m \rfloor - 2} \gamma^i \|E_{k-i}\|,
\]

where \(\text{h} = 2\gamma(m + \lfloor m \rfloor + 2) + 2\).
Proof. For \( k \geq m + 1 \), expand \( Q_k \) using (14) in Lemma 6 iteratively, yielding

\[
\|Q^* - Q_k\| = \frac{1}{k}\|Q_0 - Q_{k-1}\| + (k - m - 1)(TQ^* - TQ_{k-1}) + (m + 1)(TQ^* - TQ_0) + E_k
\]

(i) \( \leq \frac{\gamma(k - m - 1) + 1}{k}\|Q^* - Q_{k-1}\| + \frac{\gamma(m + 1) + 1}{k}\|Q^* - Q_0\| + \frac{\|E_k\|}{k} \)

(ii) \( \leq \frac{\gamma(k-1)}{k}\|Q^* - Q_{k-1}\| + \frac{2h}{k}V_{\max} + \frac{\|E_k\|}{k} \)

(iii) \( \leq \frac{\gamma^k - |m| - 1}{k}(\|m\| + 1)\|Q^* - Q_{|m|+1}\| + \frac{2hV_{\max}}{k}\sum_{i=0}^{k-|m|-2} \gamma^i + \sum_{i=0}^{k-|m|-2} \frac{\gamma^i}{k}\|E_{k-i}\| \)

\( \leq 2\gamma(|m| + 1)h V_{\max} + \frac{1}{k} \sum_{i=0}^{k-|m|-2} \gamma^i\|E_{k-i}\| \)

where (i) follows from the triangle inequality and the contraction property (3), (ii) follows from Assumption 5 and because \( m \geq \frac{1}{\gamma} \), \( h = \gamma(m + 1) + 1 \), and (iii) follows from applying (ii) to \( \|Q^* - Q_t\| \) for \( t = k - 1, k - 2, \ldots, |m| + 2 \) iteratively. Then (15) follows from the definition of \( \hat{h} \).

\[ \square \]

**Lemma 8.** *(Maximal Hoeffding-Azuma Inequality) [Alon and Spencer 2008, Chapter 7]*

Let \( \{M_1, M_2, \ldots, M_T\} \) be a martingale difference sequence with respect to a sequence of random variables \( \{X_1, X_2, \ldots, X_T\} \) (i.e. \( \mathbb{E}[M_{k+1} | X_1, X_2, \ldots, X_k] = 0, \forall 1 \leq k \leq T \)) and uniformly bounded by \( M > 0 \) almost surely. If we define \( S_k = \sum_{i=1}^{k} M_i \), then for any \( \varepsilon > 0 \), we have

\[
\mathbb{P}\left( \max_{1 \leq k \leq T} S_k > \varepsilon \right) \leq \exp\left( -\frac{\varepsilon^2}{2TM^2} \right).
\]

Now we are ready to prove the main results of Theorem 3.

**Proof of Theorem 3.** The proof applies Lemma 7 and the Maximal Hoeffding-Azuma Inequality (Lemma 8).

Applying Lemma 7 with \( k = T \), we obtain

\[
\|Q^* - Q_T\| \leq \frac{\hat{h}V_{\max}}{T(1 - \gamma)} + \frac{1}{T} \sum_{i=0}^{T-|m|-2} \gamma^i\|E_{T-i}\|.
\]

It suffices to bound the second term. For convenience, we denote \( K = T - |m| - 2 \). Observe that

\[
\frac{1}{T} \sum_{i=0}^{K} \gamma^i\|E_{T-i}\| \leq \frac{1}{T} \sum_{i=0}^{K} \gamma^i \max_{0 \leq i \leq K} \|E_{T-i}\| \leq \frac{\max_{0 \leq i \leq K} \|E_{T-i}\|}{(1 - \gamma)T}.
\]

In remains to bound \( \max_{0 \leq i \leq K} \|E_{T-i}\| \). Notice that \( \max_{0 \leq i \leq K} \|E_{T-i}\| = \max_{(x,u)} |E_{T-i}(x,u)| \). For a given \( (x,u) \) and \( \varepsilon > 0 \), we have

\[
\mathbb{P}\left( \max_{0 \leq i \leq K} |E_{T-i}(x,u)| > \varepsilon \right) = \mathbb{P}\left( \left\{ \max_{0 \leq i \leq K} (E_{T-i}(x,u)) > \varepsilon \right\} \cup \left\{ \max_{0 \leq i \leq K} (-E_{T-i}(x,u)) > \varepsilon \right\} \right)
\]

\[
= \mathbb{P}\left( \max_{0 \leq i \leq K} E_{T-i}(x,u) > \varepsilon \right) + \mathbb{P}\left( \max_{0 \leq i \leq K} (-E_{T-i}(x,u)) > \varepsilon \right),
\]

where \( D \) is specified in Proposition 1. Since \( \{\epsilon_k(x,u)\}_{k \geq 0} \) is a martingale difference sequence with respect to the filtration \( \mathcal{F}_k \) as defined previously, we apply the Maximal Hoeffding-Azuma inequality (see Lemma 8) and obtain

\[
\mathbb{P}\left( \max_{0 \leq i \leq K} (E_{T-i}(x,u)) > \varepsilon \right) \leq \exp\left( -\frac{\varepsilon^2}{8(K+1)D^2} \right),
\]

\[
\mathbb{P}\left( \max_{0 \leq i \leq K} (-E_{T-i}(x,u)) > \varepsilon \right) \leq \exp\left( -\frac{\varepsilon^2}{8(K+1)D^2} \right).
\]

Therefore, we have

\[
\mathbb{P}\left( \max_{0 \leq i \leq K} |E_{T-i}(x,u)| > \varepsilon \right) \leq 2 \exp\left( -\frac{\varepsilon^2}{8(K+1)D^2} \right).
\]

The main results of Theorem 3 follow.

\[ \square \]
and
\[ P \left( \max_{0 \leq i \leq K} \left( -E_{T\rightarrow i}(x, u) \right) > \varepsilon \right) \leq \exp \left( \frac{-\varepsilon^2}{8(K + 1)D^2} \right). \]

Then we further bound (17) as
\[ P \left( \max_{0 \leq i \leq K} |E_{T\rightarrow i}(x, u)| > \varepsilon \right) \leq 2 \exp \left( \frac{-\varepsilon^2}{8(K + 1)D^2} \right). \]

Since we consider a finite state-action space where the number of state-action pairs is defined by \( n \), we use the union bound to obtain
\[ P \left( \max_{0 \leq i \leq K} \|E_{T\rightarrow i}\| > \varepsilon \right) \leq 2n \exp \left( \frac{-\varepsilon^2}{8(K + 1)D^2} \right). \]

Letting \( \delta = 2n \exp \left( \frac{-\varepsilon^2}{8(K + 1)D^2} \right) \), and we have
\[ P \left( \max_{0 \leq i \leq K} \|E_{T\rightarrow i}\| \leq \bar{D} \sqrt{8(K + 1) \log \frac{2n}{\delta}} \right) \geq 1 - \delta, \]

where \( K = T - \lfloor m \rfloor - 2 \). By substituting the above bound into (16) yields the desired result.

\section*{7 PROOF OF COROLLARY 1}

In Theorem 3, take \( \delta = \frac{1}{T^2} \), and denote by \( A_T \) the event “inequality (24) holds”. Then the conclusion of Theorem 3 becomes \( P[A_T] \geq 1 - \frac{1}{T^2} \), or equivalently, \( P[A_T^c] \leq \frac{1}{T^2} \), for all \( T > m \), where the superscript \( c \) meaning taking the set complement. It follows that \( \sum_{T=m+1}^{\infty} P[A_T^c] \leq \sum_{T=m+1}^{\infty} \frac{1}{T^2} < \infty \). By the Borel–Cantelli lemma (see, for example, Chapter 2.3, Theorem 2.3.1 of [Durrett, 2019]), this implies \( P[A_T^c \ i.o.] = 0 \), where i.o. stands for infinitely often. This is equivalent to the statement that \( Q_T \) converges to \( Q^* \) almost surely at a rate of at least \( \tilde{O}(\sqrt{(T - \lfloor m \rfloor - 1) \log \frac{nT}{(1-\gamma)^2 T}}) \), where note that in (24) the constant \( \bar{D} \) is proportional to \( \frac{1}{1-\gamma} \). Using the \( \tilde{O} \) notation which ignores the log \( T \) factor, the order of the convergence rate can be written as \( \tilde{O}(\sqrt{(T - \lfloor m \rfloor - 1) \log \frac{n}{(1-\gamma)^2 T}}) \). Thus it completes the proof.

\section*{References}

