Supplementary Appendix

Talha Bozkus and Urbashi Mitra

I. Proof of Proposition 1

The distribution assumption on the Q-function errors between different Markovian environments in [1], [2]:

$$\mathcal{X}_t^{(n)} \stackrel{\text{def}}{=} Q_t^{(n)}(s,a) - Q^*(s,a) \sim D\Big(0, \frac{\lambda_n^2}{3}\Big) \ \forall n, \tag{1}$$

This assumption can be relaxed to a non-zero mean distribution assumption with different conditions. See [1] for details. Since $Q^*(s,a)$ is constant for s, a, we can model the distribution of Q-functions as:

$$Q^{(n)}(s,a) \sim D(Q^*(s,a), \frac{\lambda_n^2}{3}) \ \forall n \tag{2}$$

We assume that the Q-functions of the same state under different actions have the following structure:

$$\frac{1}{\theta} \leq \frac{Q(s,a_{k_1})}{Q(s,a_{k_2})} \leq \theta, \quad \forall k_1 \neq k_2, \theta \geq 1, \tag{3}$$

where we do not assume any greater or smaller relationship between $Q(s,a_{k_1})$ and $Q(s,a_{k_2})$. Note that for $\theta = \infty$, this assumption is trivial. Following the assumption that the policy π always selects the action a, then $d_\pi(s,a) = 1$:

$$\log(c_\pi(s,a)) = \log\Big(\frac{d_\pi(s,a)}{v(s,a)}\Big) = \log(d_\pi(s,a)) - \log(v(s,a)) = -\log(v(s,a)) \tag{4}$$

We set $\pi = \pi^{(n)}$ and $v = v^{(n)}$. Assume there are 2 actions: a_1 and a_2, and we use linear-action selection.

$$v^{(n)}(s,a_k) = \frac{Q^{(n)}(s,a_k)}{Q^{(n)}(s,a_1) + Q^{(n)}(s,a_2)} \tag{5}$$

Let X be a RV with (μ, σ^2) with $\mu > 0$ and $\sigma^2 > 0$. We can approximate the distribution of $\log(X)$ using the second-order Taylor series expansion as:

$$\mathbb{E}(\log(X)) \approx \log(\mu) - \frac{\sigma^2}{2\mu^2}. \tag{6}$$

$$\mathbb{V}(\log(X)) \approx \frac{\sigma^2}{\mu^2}. \tag{7}$$

These approximations are accurate enough when $\frac{\mu}{\sigma} > 2.5$, which almost always holds for Q-learning for the wireless networks considered for numerical simulations. Then we proceed as follows:

$$\mathbb{E}[\log(v^{(n)}(s,a_k))] = \mathbb{E}[\log(Q^{(n)}(s,a_k)) - \log(Q^{(n)}(s,a_1) + Q^{(n)}(s,a_2))] \tag{8}$$

$$= \mathbb{E}[\log(Q^{(n)}(s,a_k))] - \mathbb{E}[\log(Q^{(n)}(s,a_1) + Q^{(n)}(s,a_2))] \tag{9}$$

$$= \left(\log(Q^*(s,a_k)) - \frac{\frac{\lambda_n^2}{3}}{2Q^*(s,a_k)^2}\right) - \left(\log(Q^*(s,a_1) + Q^*(s,a_2)) - \frac{\frac{2\lambda_n^2}{3}}{2(Q^*(s,a_1) + Q^*(s,a_2))^2}\right) \tag{10}$$

$$= \log(\frac{Q^*(s,a_k)}{Q^*(s,a_1) + Q^*(s,a_2)}) + \frac{\lambda_n^2}{3}\Big(\frac{1}{(Q^*(s,a_1) + Q^*(s,a_2))^2} - \frac{1}{2Q^*(s,a_k)^2}\Big) \tag{11}$$

$$= \log(\frac{Q^*(s,a_k)}{Q^*(s,a_1) + Q^*(s,a_2)}) + \frac{\lambda_n^2}{3Q^*(s,a_k)^2}\Big((\frac{Q^*(s,a_k)}{Q^*(s,a_1) + Q^*(s,a_2)})^2 - \frac{1}{2}\Big) \tag{12}$$

$$= \log(\epsilon_k) + \frac{\lambda_n^2}{3Q^*(s,a_k)^2}\Big(\epsilon_k^2 - \frac{1}{2}\Big), \tag{13}$$

where (8) follows from (5), (9) follows from the linearity of expectation, (10) follows from (6) and (7), (11) and (12) follow by algebraic manipulations, (13) follows from the notation that $\frac{Q^*(s,a_k)}{Q^*(s,a_1)+Q^*(s,a_2)} = \epsilon_k$ for the action k.

We take the negative of both sides:

$$-\mathbb{E}[\log(v^{(n)}(s,a_k))] = \log(\frac{1}{\epsilon_k}) + \frac{\lambda_n^2}{3Q^*(s,a_k)^2}\Big(\frac{1}{2} - \epsilon_k^2\Big) \tag{14}$$

We have $\epsilon_k \geq \frac{1}{1+\theta}$, $\frac{1}{\epsilon_k} \leq 1+\theta$, and $\epsilon_k^2 \geq \frac{1}{(1+\theta)^2}$ using (3) and the definition of ϵ_k. If we plug these expressions into (14), we obtain the following bound:

$$\mathbb{E}[\log(c^{(n)}_{\pi^{(n)}}(s,a))] \leq \log(1+\theta) + \frac{\lambda_n^2}{3Q^*(s,a)^2}\Big(\frac{1}{2} + \frac{1}{(1+\theta)^2}\Big), \tag{15}$$

where we simply dropped the subscript k as it is valid for both actions.

We can carry out similar operations for the variance of log-coverage-coefficient:

$$\begin{aligned}
\mathbb{V}[\log(v^{(n)}(s,a_k))] &= \mathbb{V}[\log(Q^{(n)}(s,a_1)) - \log(Q^{(n)}(s,a_1) + Q^{(n)}(s,a_2))] && (16)\\
&\leq \mathbb{V}[\log(Q^{(n)}(s,a_k))] + \mathbb{V}[\log(Q^{(n)}(s,a_1) + Q^{(n)}(s,a_2))] + \\
&\quad 2\sqrt{(\mathbb{V}[\log(Q^{(n)}(s,a_k))]\mathbb{V}[\log(Q^{(n)}(s,a_1) + Q^{(n)}(s,a_2))])} && (17)\\
&= \frac{\frac{\lambda_n^2}{3}}{Q^*(s,a_k)^2} + \frac{\frac{2\lambda_n^2}{3}}{(Q^*(s,a_1)+Q^*(s,a_2))^2} + 2\sqrt{\frac{\frac{2\lambda_n^4}{9}}{Q^*(s,a_k)^2(Q^*(s,a_1)+Q^*(s,a_2))^2}} && (18)\\
&= \frac{\lambda_n^2}{3Q^*(s,a_k)^2}\left(1 + 2\Big(\frac{Q^*(s,a_k)}{Q^*(s,a_1)+Q^*(s,a_2)}\Big)^2 + 2\sqrt{2}\Big(\frac{Q^*(s,a_k)}{Q^*(s,a_1)+Q^*(s,a_2)}\Big)\right) && (19)\\
&= \frac{\lambda_n^2}{3Q^*(s,a_k)}\Big(1 + 2\epsilon_k^2 + 2\sqrt{2}\epsilon_k\Big) && (20)
\end{aligned}$$

Herein, we have $\epsilon_k \leq \frac{\theta}{\theta+1}$. Plugging this into (20) and using $\mathbb{V}[\log(c^{\pi^{(n)}}(s,a))] = \mathbb{V}[\log(v^{(n)}(s,a))]$, and dropping the subscript k, we obtain the following bound:

$$\mathbb{V}[\log(c^{\pi^{(n)}}(s,a))] \leq \frac{\lambda_n^2}{3Q^*(s,a)^2}\Big(1 + \frac{2\theta^2}{(\theta+1)^2} + \frac{2\sqrt{2}\theta}{\theta+1}\Big) \tag{21}$$

II. Proof of Proposition 2

The ensemble Q-function output of nEQL algorithm $Q^{it}(s,a)$ can be modeled as $Q^{it}(s,a) \approx D(Q^*(s,a), \mathbb{V}(Q^{it}(s,a)))$, where $\mathbb{V}(Q^{it}(s,a)) \leq \frac{\lambda^2}{3}\frac{1-u}{1+u}$ from [1]. Note that the bounds (15) and (21) explicitly depends on the variance of $Q^{(n)}(s,a)$, which is $\frac{\lambda_n^2}{3}$. Hence if we replace the $Q^{it}(s,a)$ by $Q^{(n)}(s,a)$, we can replace the variance $\frac{\lambda_n^2}{3}$ by the upper bound on $\mathbb{V}(Q^{it}(s,a)))$, which is $\frac{\lambda^2}{3}\frac{1-u}{1+u}$ as follows:

$$\mathbb{E}[\log(c^{\pi^{it}}(s,a))] \leq \log(1+\theta) + \frac{\lambda^2}{3Q^*(s,a)^2}\Big(\frac{1}{2} - \frac{1}{(1+\theta)^2}\Big)\frac{1-u}{1+u}. \tag{22}$$

$$\mathbb{V}[\log(c^{it}_{\pi^{it}}(s,a))] \leq \frac{\lambda^2}{3Q^*(s,a)^2}\Big(1 + \frac{2\theta^2}{(\theta+1)^2} + \frac{2\sqrt{2}\theta}{\theta+1}\Big)\frac{1-u}{1+u}. \tag{23}$$

III. Proof of Proposition 3

We follow a similar approach to the proof of Proposition 2. The upper bound of the variance $\mathbb{V}(Q^{it}(s,a))$ can also be expressed in terms of K: $\mathbb{V}(Q^{it}(s,a)) \leq \frac{f(\lambda,u)}{K}$ [1]. Hence, we can replace the upper-bound $\frac{\lambda^2}{3}\frac{1-u}{1+u}$ from Proposition 2 by $\frac{f(\lambda,u)}{K}$ and obtain the following bounds:

$$\mathbb{E}[\log(c^{\pi^{it}}(s,a))] \leq \log(1+\theta) + \frac{1}{Q^*(s,a)^2}\Big(\frac{1}{2} - \frac{1}{(1+\theta)^2}\Big)\frac{f(\lambda,u)}{K}. \tag{24}$$

$$\mathbb{V}[\log(c^{\pi^{it}}(s,a))] \leq \frac{1}{Q^*(s,a)^2}\Big(1 + \frac{2\theta^2}{(\theta+1)} + \frac{2\sqrt{2}\theta}{\theta+1}\Big)\frac{f(\lambda,u)}{K} \tag{25}$$

IV. Proof of Proposition 4

Proposition 3 of [1] holds for any policy π, so we use the optimal policy of the original environment $\pi = \pi^*$. We use notation $\mathbf{c}_{\pi^*}^{(n)}$ to denote the cost function of the n^{th} environment under policy π^*. In addition, we express Proposition 3 for a generic state s rather than the whole Q-table:

$$|Q_{\pi^*}^{(1)}(s) - Q_{\pi^*}^{(n)}(s)| < \frac{\gamma}{1-\gamma^n}\frac{1-\gamma^{n-1}}{1-\gamma}|c_{\pi^*}^{(n)}(s)| \tag{26}$$

The Q-functions of the original environment under the optimal policy are the optimal Q-functions Q^*: thus, we can replace $Q_{\pi^*}^{(1)}(s)$ by $Q^*(s)$:

$$|Q^*(s) - Q_{\pi^*}^{(n)}(s)| < \frac{\gamma}{1-\gamma^n}\frac{1-\gamma^{n-1}}{1-\gamma}|c_{\pi^*}^{(n)}(s)| \tag{27}$$

Using the fact that the cost function is always non-negative by definition:

$$-\frac{\gamma}{1-\gamma^n}\frac{1-\gamma^{n-1}}{1-\gamma}c_{\pi^*}^{(n)}(s) < Q^*(s) - Q_{\pi^*}^{(n)}(s) < \frac{\gamma}{1-\gamma^n}\frac{1-\gamma^{n-1}}{1-\gamma}c_{\pi^*}^{(n)}(s) \tag{28}$$

We assume that Q-function errors follow a uniform distribution as:

$$Q^*(s) - Q_{\pi^*}^{(n)}(s) \sim \text{unif}(-\frac{\gamma}{1-\gamma^n}\frac{1-\gamma^{n-1}}{1-\gamma}c_{\pi^*}^{(n)}(s), \frac{\gamma}{1-\gamma^n}\frac{1-\gamma^{n-1}}{1-\gamma}c_{\pi^*}^{(n)}(s)) \tag{29}$$

This is a reasonable assumption because (i) the distribution changes with n and π^*, (ii) the expectation is zero, which is in line with the results in [1], [2], and (iii) uniform Q-function error distribution is commonly employed in RL literature [3], [4], [5]. Using the variance of the uniform distribution:

$$\mathbb{V}[Q^*(s) - Q_{\pi^*}^{(n)}(s)] = \frac{1}{3}\Big[\frac{\gamma}{1-\gamma^n}\frac{1-\gamma^{n-1}}{1-\gamma}\Big]^2\big(c_{\pi^*}^{(n)}(s)\big)^2 \tag{30}$$

$$= \frac{\lambda_n^2}{3}, \tag{31}$$

which follows from the definition of the Q-function error variance in (1), which holds for any policy.

Herein, the term $\left[\frac{\gamma}{1-\gamma^n}\frac{1-\gamma^{n-1}}{1-\gamma}\right]^2$ is monotonically increasing with n and it is 0 when $n = 1$ and (ii) under the policy π^*, $c^{(1)}(s)$ attains its minimum while we do not have information about the cost functions of other environments, i.e. $c_{\pi^*}^{(1)}(s) \leq c_{\pi^*}^{(n)}(s)$ for any n. Hence, the value of λ_n is the minimum for $n = 1$.

On the other hand, as n increases, the term $\left[\frac{\gamma}{1-\gamma^n}\frac{1-\gamma^{n-1}}{1-\gamma}\right]^2$ increases yet the term $c_{\pi^*}^{(n)}(s)$ may increase, decrease or stay the same. Thus, the value of λ_n is non-monotonic across n for $n > 1$.

V. Proof of Proposition 5

Using (30), we have the following:

$$\frac{c_{\pi^*}^{(n)}(s)}{c_{\pi^*}^{(m)}(s)} \underset{m}{\overset{n}{\gtrless}} \frac{\left[\frac{\gamma}{1-\gamma^m}\frac{1-\gamma^{m-1}}{1-\gamma}\right]^2}{\left[\frac{\gamma}{1-\gamma^n}\frac{1-\gamma^{n-1}}{1-\gamma}\right]^2} = \frac{(1-\gamma^n)(1-\gamma^{m-1})}{(1-\gamma^m)(1-\gamma^{n-1})} = f(\gamma, n, m). \tag{32}$$

Assume that the cost function is bounded as: $c_{\pi^*}^{(n)}(s) \in [c_{min}, c_{max}]$ with $c_{min} > 0$ and $c_{max} < \infty$. In this case, the minimum and maximum values of $\frac{c_{\pi^*}^{(n)}(s)}{c_{\pi^*}^{(m)}(s)}$ are $\frac{c_{min}}{c_{max}}$ and $\frac{c_{max}}{c_{min}}$, respectively. Then, we have the following decision rule:

$$\lambda_n < \lambda_m \quad \text{iff} \quad f(\gamma, n, m) > \frac{c_{max}}{c_{min}} \tag{33}$$

$$\lambda_n > \lambda_m \quad \text{iff} \quad f(\gamma, n, m) < \frac{c_{min}}{c_{max}} \tag{34}$$

We note that this rule is inconclusive when $f(\gamma, n, m) < \frac{c_{max}}{c_{min}}$ or $f(\gamma, n, m) > \frac{c_{min}}{c_{max}}$. However, we can get rid of the inconclusive area by a randomized boundary between $\frac{c_{min}}{c_{max}}$ and $\frac{c_{max}}{c_{min}}$ as:

$$\lambda_n < \lambda_m \quad \text{iff} \quad f(\gamma, n, m) > \alpha\frac{c_{max}}{c_{min}} + (1-\alpha)\frac{c_{min}}{c_{max}} \tag{35}$$

$$\lambda_n > \lambda_m \quad \text{iff} \quad f(\gamma, n, m) < \alpha\frac{c_{max}}{c_{min}} + (1-\alpha)\frac{c_{min}}{c_{max}} \tag{36}$$

VI. Generalization to arbitrary $|\mathcal{A}|$

Let's assume $|\mathcal{A}| > 2$. Let $\epsilon_k = \frac{Q^*(s,a_k)}{\sum_{i=1}^{|\mathcal{A}|} Q^*(s,a_i)}$.

$$\mathbb{E}[\log(v^{(n)}(s,a_k))] = \mathbb{E}[\log(Q^{(n)}(s,a_k))] - \mathbb{E}[\log(\sum_{i=1}^{|\mathcal{A}|} Q^{(n)}(s,a_i)] \tag{37}$$

$$= \left(\log(Q^*(s,a_k)) - \frac{\frac{\lambda_n^2}{3}}{2Q^*(s,a_k)^2}\right) - \left(\log\sum_{i=1}^{|\mathcal{A}|} Q^*(s,a_i) - \frac{\frac{|\mathcal{A}|\lambda_n^2}{3}}{2(\sum_{i=1}^{|\mathcal{A}|} Q^*(s,a_i))^2}\right) \tag{38}$$

$$= \log\Big(\frac{Q^*(s,a_k)}{\sum_{i=1}^{|\mathcal{A}|} Q^*(s,a_i)}\Big) + \frac{\lambda_n^2}{6Q^*(s,a_k)^2}\Big(\frac{|\mathcal{A}|Q^*(s,a_k)^2}{(\sum_{i=1}^{|\mathcal{A}|} Q^*(s,a_i))^2} - 1\Big) \tag{39}$$

$$= \log(\epsilon_k) + \frac{\lambda_n^2}{6Q^*(s,a_k)^2}(|\mathcal{A}|\epsilon_k^2 - 1) \tag{40}$$

If we take the negative of both sides:

$$-\mathbb{E}[\log(v^{(n)}(s,a_k))] = \log(\frac{1}{\epsilon_k}) + \frac{\lambda_n^2}{6Q^*(s,a_k)^2}(1 - |\mathcal{A}|\epsilon_k^2) \tag{41}$$

Herein, we have $\frac{1}{\epsilon_k} \leq 1 + (|\mathcal{A}| - 1)\theta$, and $\epsilon_k \geq \frac{1}{1+(|\mathcal{A}|-1)\theta}$. Plugging these into (41) and using $\mathbb{E}[\log(c^{\pi^{(n)}}(s,a))] = -\mathbb{E}[\log(v^{(n)}(s,a_k))]$, we obtain the following bound:

$$\mathbb{E}[\log(c^{\pi^{(n)}}(s,a))] \leq \log(1 + (|\mathcal{A}| - 1)\theta) + \frac{\lambda_n^2}{6Q^*(s,a)^2}(1 + \frac{|\mathcal{A}|}{(1 + (|\mathcal{A}| - 1)\theta)^2}). \tag{42}$$

Note that this bound is the same as (15) for $|\mathcal{A}| = 2$. We can carry out similar operations for the variance:

$$\mathbb{V}[\log(v^{(n)}(s,a_k))] = \mathbb{V}[\log(Q^{(n)}(s,a_1))] - \mathbb{V}[\log(\sum_{i=1}^{|\mathcal{A}|} Q^{(n)}(s,a_i)] \tag{43}$$

$$\leq \mathbb{V}[\log(Q^{(n)}(s,a_k))] + \mathbb{V}[\log(\sum_{i=1}^{|\mathcal{A}|} Q^{(n)}(s,a_i))] + 2\sqrt{(\mathbb{V}[\log(Q^{(n)}(s,a_k))]\mathbb{V}[\log(\sum_{i=1}^{|\mathcal{A}|} Q^{(n)}(s,a_i))])} \tag{44}$$

$$= \frac{\frac{\lambda_n^2}{3}}{Q^*(s,a_k)^2} + \frac{\frac{|\mathcal{A}|\lambda_n^2}{3}}{\sum_{i=1}^{|\mathcal{A}|} Q^*(s,a_i)} + 2\sqrt{\frac{\frac{|\mathcal{A}|\lambda_n^4}{9}}{Q^*(s,a_k)\sum_{i=1}^{|\mathcal{A}|} Q^*(s,a_i)}} \tag{45}$$

$$= \frac{\lambda_n^2}{3Q^*(s,a_k)^2}\Big[1 + |\mathcal{A}|\Big(\frac{Q^*(s,a_k)}{\sum_{i=1}^{|\mathcal{A}|} Q^*(s,a_i)}\Big)^2 + 2\sqrt{|\mathcal{A}|}\Big(\frac{Q^*(s,a_k)}{\sum_{i=1}^{|\mathcal{A}|} Q^*(s,a_i)}\Big)\Big] \tag{46}$$

$$= \frac{\lambda_n^2}{3Q^*(s,a_k)^2}\Big[1 + |\mathcal{A}|\epsilon_k^2 + 2\sqrt{|\mathcal{A}|}\epsilon_k\Big] \tag{47}$$

Herein, we have $\epsilon_k \leq \frac{\theta}{\theta+|\mathcal{A}|-1}$. Plugging this into (47), and using and using $\mathbb{V}[\log(c^{\pi^{(n)}}(s,a))] = \mathbb{V}[\log(v^{(n)}(s,a))]$, we obtain the following bound:

$$\mathbb{V}[\log(c^{\pi^{(n)}}(s,a))] \leq \frac{\lambda_n^2}{3Q^*(s,a)^2}\Big(1 + \frac{|\mathcal{A}|\theta^2}{(\theta + |\mathcal{A}| - 1)^2} + \frac{2\sqrt{|\mathcal{A}|}\theta}{\theta + |\mathcal{A}| - 1}\Big) \tag{48}$$

We can carry out the same operations for the cases in Propositions 2 and 3.

VII. Different assumptions on Q-functions

We employ a softmax-exploration action selection with a parameter τ with 2 actions a_1 and a_2. Let's focus on action a_1:

$$v(s,a_1) = \frac{e^{\frac{Q(s,a_1)}{\tau}}}{e^{\frac{Q(s,a_1)}{\tau}} + e^{\frac{Q(s,a_2)}{\tau}}} \tag{49}$$

We assume $Q(s,a_2) - Q(s,a_1) \geq \theta$ with $\theta > 0$. Herein, we assume that the Q-functions for the action a_2 is strictly larger than that of action a_1. If we derive the expectation bound on $\mathbb{E}[\log(c^{(n)}_{\pi^{(n)}}(s,a))]$, we obtain the same expression as (14). However, the expression for ϵ_1 changes as:

$$\epsilon_1 = \frac{e^{\frac{Q(s,a_1)}{\tau}}}{e^{\frac{Q(s,a_1)}{\tau}} + e^{\frac{Q(s,a_2)}{\tau}}} \tag{50}$$

If we take the inverse of both sides:

$$\frac{1}{\epsilon_1} = 1 + e^{\frac{Q(s,a_1)-Q(s,a_2)}{\tau}} \tag{51}$$

$$\leq 1 + e^{\frac{-\theta}{\tau}} \tag{52}$$

Thus, the upper bound on $\log(\frac{1}{\epsilon_1})$ is $\log(1+e^{\frac{-\theta}{\tau}})$. If $\theta = \tau = 1$, then the upper bound is $\log(1.3678)$, which is much tighter than the previous bound ($\log(1+\theta)$ with $\theta > 1$). Hence, we show that if we have a more strict assumption on Q-functions, we may further tighten the bound on the expectation. Similar operations also apply for the variance bound.

REFERENCES

[1] Talha Bozkus and Urbashi Mitra. Multi-timescale ensemble q-learning for markov decision process policy optimization. *IEEE Transactions on Signal Processing*, 72:1427–1442, 2024.

[2] Talha Bozkus and Urbashi Mitra. Leveraging digital cousins for ensemble q-learning in large-scale wireless networks. *IEEE Transactions on Signal Processing*, 72:1114–1129, 2024.

[3] Sebastian Thrun and Anton Schwartz. Issues in using function approximation for reinforcement learning. In *Proceedings of the 1993 Connectionist Models Summer School Hillsdale, NJ. Lawrence Erlbaum*, volume 6, pages 1–9, 1993.

[4] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[5] Qingfeng Lan, Yangchen Pan, Alona Fyshe, and Martha White. Maxmin Q-learning: Controlling the estimation bias of Q-learning. *CoRR*, abs/2002.06487, 2020.