Supplementary Appendix

Talha Bozkus and Urbashi Mitra

I. Proof of Proposition 1

Recall the main assumption in the paper:

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

We first prove the expectation.

$$\lim_{t\to\infty} \mathcal{E}_t = \lim_{t\to\infty} \mathbf{Q}_t^{it} - \mathbf{Q}^*. \tag{2}$$

$$= \lim_{t\to\infty} (1-u) \sum_{i=0}^{t} u^{t-i} \sum_{n=1}^{K} \mathbf{w}_i^{(n)} \mathbf{Q}_i^{(n)} - \mathbf{Q}^*. \tag{3}$$

$$= \lim_{t\to\infty} (1-u) \sum_{i=0}^{t} u^{t-i} \sum_{n=1}^{K} \mathbf{w}_i^{(n)} (\mathbf{Q}_i^{(n)} - \mathbf{Q}^*). \tag{4}$$

$$= \lim_{t\to\infty} (1-u) \sum_{i=0}^{t} u^{t-i} \sum_{n=1}^{K} \mathbf{w}_i^{(n)} \mathcal{X}_i^{(n)}, \tag{5}$$

where (3) follows from the explicit expression for $\mathbf{Q}_t^{it}$, which can be obtained by repeatedly plugging the expression of $\mathbf{Q}_{t-1}^{it}$ in $\mathbf{Q}_t^{it}$ in line 9 in Algorithm 1, (4) follows from the fact that $\sum_n w_t^{(n)} = 1$ for all t, and $(1-u)\sum_{i=0}^{t} u^i = 1$ as $t \to \infty$, and (5) follows from (1). If we take the expectation of both sides:

$$\lim_{t\to\infty} \mathbb{E}[\mathcal{E}_t] = \lim_{t\to\infty} (1-u) \sum_{i=0}^{t} u^{t-i} \sum_{n=1}^{K} \mathbf{w}_i^{(n)} \mathbb{E}[\mathcal{X}_i^{(n)}]. \tag{6}$$

$$= 0, \tag{7}$$

where (6) follows from the linearity of expectation, and (7) follows from (1).

We now prove the upper bound on the variance for the modest-independence assumption.

$$\lim_{t\to\infty} \mathbb{V}[\mathcal{E}_t] = \lim_{t\to\infty} \mathbb{V}\Big[(1-u) \sum_{i=0}^{t} u^{t-i} \sum_{n=1}^{K} \mathbf{w}_i^{(n)} \mathcal{X}_i^{(n)}\Big]. \tag{8}$$

$$= \lim_{t\to\infty} (1-u)^2 \Big[\sum_{i=0}^{t} u^{2(t-i)} \big[\sum_{n=1}^{K} (\mathbf{w}_i^{(n)})^2 \mathbb{V}[\mathcal{X}_i^{(n)}] + 2\sum_{n=1}^{K}\sum_{m\neq n}^{K} \mathbf{w}_i^{(n)}\mathbf{w}_i^{(m)} \operatorname{Cov}(\mathcal{X}_i^{(n)}, \mathcal{X}_i^{(m)})\big]\Big]. \tag{9}$$

$$\leq \lim_{t\to\infty} (1-u)^2 \Big[\sum_{i=0}^{t} u^{2(t-i)} \big[\sum_{n=1}^{K} (\mathbf{w}_i^{(n)})^2 \mathbb{V}[\mathcal{X}_i^{(n)}] + 2\sum_{n=1}^{K}\sum_{m\neq n}^{K} \mathbf{w}_i^{(n)}\mathbf{w}_i^{(m)} \sqrt{\mathbb{V}[\mathcal{X}_i^{(n)}]\mathbb{V}[\mathcal{X}_i^{(m)}]}\big]\Big]. \tag{10}$$

$$\leq \lim_{t\to\infty} (1-u)^2 \Big[\sum_{i=0}^{t} u^{2(t-i)} \big[\sum_{n=1}^{K} \mathbf{w}_i^{(n)} \mathbb{V}[\mathcal{X}_i^{(n)}] + 2\sum_{n=1}^{K}\sum_{m=1}^{K} \mathbf{w}_i^{(n)}\mathbf{w}_i^{(m)} \sqrt{\mathbb{V}[\mathcal{X}_i^{(n)}]\mathbb{V}[\mathcal{X}_i^{(m)}]}\big]\Big]. \tag{11}$$

$$\leq \lim_{t\to\infty} (1-u)^2 \Big[\sum_{i=0}^{t} u^{2(t-i)} \big[\sum_{n=1}^{K} \mathbf{w}_i^{(n)} \frac{\lambda^2}{3} + 2\sum_{n=1}^{K}\sum_{m=1}^{K} \mathbf{w}_i^{(n)}\mathbf{w}_i^{(m)} \frac{\lambda^2}{3}\big]. \tag{12}$$

$$\leq \lim_{t\to\infty} (1-u)^2 \Big[\sum_{i=0}^{t} u^{2(t-i)} \lambda^2\Big]. \tag{13}$$

$$\leq \frac{(1-u)}{(1+u)} \lambda^2, \tag{14}$$

where (8) follows from (5), (9) follows from the properties of the variance operator and the modest-independence assumption, (10) follows from the Cauchy-Schwarz inequality for the variance, (11) follows from the fact that $\mathbf{w}_t^{(n)} \leq 1$ and dropping the constraint in the second summation, (12) follows from (1) and $\lambda = \max_n \lambda_n$, (13) follows from the fact that $\sum_{n=1}^K w_t^{(n)} = 1$ for all t, and (14) follows from the infinite geometric sum formula.

When we have the strict-independence assumption, the cross-terms in (12) disappear, and the upper bound is reduced by a factor of 3, which proves the corresponding upper bound.

We now prove the upper bound for the case where no independence is assumed.

$$\lim_{t\to\infty} \mathbb{V}[\mathcal{E}_t] = \lim_{t\to\infty} \mathbb{V}\Big[(1-u)\sum_{i=0}^{t} u^{t-i} \sum_{n=1}^{K} \mathbf{w}_i^{(n)} \mathcal{X}_i^{(n)}\Big]. \tag{15}$$

$$= \lim_{t\to\infty} (1-u)^2 \Big[\sum_{i=0}^{t} u^{2(t-i)} \mathbb{V}\big[\sum_{n=1}^{K} \mathbf{w}_i^{(n)} \mathcal{X}_i^{(n)}\big] + 2\sum_{i=0}^{t}\sum_{j\neq i}^{t} u^{2(t-i)} u^{2(t-j)} \operatorname{Cov}\big[\sum_{n=1}^{K} \mathbf{w}_i^{(n)} \mathcal{X}_i^{(n)}, \sum_{n=1}^{K} \mathbf{w}_j^{(n)} \mathcal{X}_j^{(n)}\big]\Big]. \tag{16}$$

$$\leq \lim_{t\to\infty} (1-u)^2 \Big[\sum_{i=0}^{t} u^{2(t-i)} \mathbb{V}\big[\sum_{n=1}^{K} \mathbf{w}_i^{(n)} \mathcal{X}_i^{(n)}\big] + 2\sum_{i=0}^{t}\sum_{j\neq i}^{t} u^{2(t-i)} u^{2(t-j)} \sqrt{\mathbb{V}\big[\sum_{n=1}^{K} \mathbf{w}_i^{(n)} \mathcal{X}_i^{(n)}\big] \mathbb{V}\big[\sum_{n=1}^{K} \mathbf{w}_j^{(n)} \mathcal{X}_j^{(n)}\big]}\Big]. \tag{17}$$

$$\leq \lim_{t\to\infty} (1-u)^2 \Big[\sum_{i=0}^{t} u^{2(t-i)} + 2\sum_{i=0}^{t}\sum_{j\neq i}^{t} u^{2(t-i)} u^{2(t-j)}\Big]\lambda^2. \tag{18}$$

$$\leq \lim_{t\to\infty} (1-u)^2 \Big[\sum_{i=0}^{t} u^{2(t-i)} + 2\sum_{i=0}^{t}\sum_{j=0}^{t} u^{2(t-i)} u^{2(t-j)}\Big]\lambda^2. \tag{19}$$

$$\leq \frac{2\lambda^2}{(1+u)^2} + \frac{(1-u)}{(1+u)}\lambda^2, \tag{20}$$

where (15) follows from (5), (16) follows from the properties of the variance operator and the no-independence assumption, (17) follows from the Cauchy-Schwarz inequality for the variance operator, (18) follows from the fact that variance is independent of time indices and (13), (19) follows from dropping the constraint in the second summation, and (20) follows from the infinite geometric sum formula.

II. Proof of Corollary 1

Assume the parameter u has the form $u_t = 1 - e^{\frac{-t}{c_4}}$. If we rewrite (8) to (13) using a time-varying update ratio u_t instead of constant u, and remove the limit, we obtain the following:

$$\mathbb{V}[\mathcal{E}_t] \leq (1-u_t)^2 \Big[\sum_{i=0}^{t} u_{t-i}^{2(t-i)} \lambda^2\Big]. \tag{21}$$

Here, we can change the subscript and superscript indices $t-i$ to i and update (21) as follows:

$$\mathbb{V}[\mathcal{E}_t] \leq (1-u_t)^2 \Big[\sum_{i=0}^{t} u_i^{2i} \lambda^2\Big]. \tag{22}$$

Then, we plug the expression for u_t in (22):

$$\mathbb{V}[\mathcal{E}_t] \leq \lambda^2 e^{\frac{-2t}{c_4}} \big[\sum_{i=0}^{t} (1 - e^{\frac{-i}{c_4}})^{2i}\big]. \tag{23}$$

III. Proof of Proposition 2

We firstly bound the weight $\mathbf{w}^{(n)}$ into a more strict interval than [0,1]. We first do the calculations for $n \neq 1$. The maximum value of $\mathbf{w}^{(n)}$ is obtained when $\hat{\mathbf{Q}}^{(1)} = \hat{\mathbf{Q}}^{(n)}$ and $\hat{\mathbf{Q}}^{(i)}$ is maximally different than $\hat{\mathbf{Q}}^{(1)}$ for all $i \neq n$. In this case, $\mathbf{w}^{(n)} = 1$, and $\mathbf{w}^{(i)} = 0$ for all $i \notin \{n, 1\}$. When we apply the softmax operator on the $\mathbf{w}$, we obtain the following:

$$\mathbf{w}^{(n)} = \frac{e}{2e + (K-2)} \leq \frac{e}{K}, \tag{24}$$

which follows as there are $K > 1$ different weights, and $\mathbf{w}^{(1)} = 1$. On the other hand, the maximum value of $\mathbf{w}^{(1)}$ is obtained when $\hat{\mathbf{Q}}^{(i)}$ is maximally different than $\hat{\mathbf{Q}}^{(1)}$ for all $i \neq 1$. In this case, $\mathbf{w}^{(1)} = 1$, and $\mathbf{w}^{(i)} = 0$ for all $i \neq 1$. When we apply the softmax operator on the $\mathbf{w}$, we obtain the following:

$$\mathbf{w}^{(1)} = \frac{e}{e + (K-1)} \leq \frac{e}{K}. \tag{25}$$

Combining (24) and (25), the following holds for all n:

$$\mathbf{w}^{(n)} \leq \frac{e}{K}. \tag{26}$$

We note that this bound is useful for $K > 2$. Then, the term $\sum_{n=1}^{K} \big(\mathbf{w}_i^{(n)}\big)^2$ in (9) can be upper bounded as:

$$\sum_{n=1}^{K} (\mathbf{w}_i^{(n)})^2 \leq \sum_{n=1}^{K} (\frac{e}{K})^2 \leq \frac{e^2}{K}. \tag{27}$$

Using (27), (1), and $\lambda = \max_n \lambda_n$ we can show the following:

$$\mathbb{V}[\sum_{n=1}^{K} \mathbf{w}_i^{(n)} \mathcal{X}_i^{(n)}] \leq \sum_{n=1}^{K} \big(\mathbf{w}_i^{(n)}\big)^2 \mathbb{V}[\mathcal{X}_i^{(n)}] \leq \frac{e^2}{K}\frac{\lambda^2}{3}. \tag{28}$$

Then, we can show the following:

$$\lim_{t\to\infty} \mathbb{V}[\mathcal{E}_t] = \lim_{t\to\infty} \mathbb{V}\Big[(1-u)\sum_{i=0}^{t} u^{t-i} \sum_{n=1}^{K} \mathbf{w}_i^{(n)} \mathcal{X}_i^{(n)}\Big]. \tag{29}$$

$$= \lim_{t\to\infty} (1-u)^2 \Big[\sum_{i=0}^{t} u^{2(t-i)} \mathbb{V}[\sum_{n=1}^{K} \mathbf{w}_i^{(n)} \mathcal{X}_i^{(n)}] + 2\sum_{i=0}^{t}\sum_{j\neq i}^{t} u^{t-i}u^{t-j}\, \mathrm{Cov}(\sum_{n=1}^{K} \mathbf{w}_i^{(n)} \mathcal{X}_i^{(n)}, \sum_{n=1}^{K} \mathbf{w}_j^{(n)} \mathcal{X}_j^{(n)})\Big] \tag{30}$$

$$\leq \lim_{t\to\infty} (1-u)^2 \Big[\sum_{i=0}^{t} u^{2(t-i)} \mathbb{V}[\sum_{n=1}^{K} \mathbf{w}_i^{(n)} \mathcal{X}_i^{(n)}] + 2\sum_{i=0}^{t}\sum_{j\neq i}^{t} u^{t-i}u^{t-j} \sqrt{\mathbb{V}(\sum_{n=1}^{K} \mathbf{w}_i^{(n)} \mathcal{X}_i^{(n)})\mathbb{V}(\sum_{n=1}^{K} \mathbf{w}_j^{(n)} \mathcal{X}_j^{(n)})}\Big] \tag{31}$$

$$\leq \lim_{t\to\infty} (1-u)^2 \frac{e^2}{K}\frac{\lambda^2}{3}\Big[\sum_{i=0}^{t} u^{2(t-i)} + 2\sum_{i=0}^{t}\sum_{j\neq i}^{t} u^{t-i}u^{t-j}\Big] \tag{32}$$

$$\leq \lim_{t\to\infty} (1-u)^2 \frac{e^2}{K}\frac{\lambda^2}{3}\Big[\sum_{i=0}^{t} u^{2(t-i)} + 2\big(\sum_{i=0}^{t} u^{t-i}\big)^2\Big] \tag{33}$$

$$\leq (1-u)^2 \frac{e^2}{K}\frac{\lambda^2}{3}\Big[\frac{1}{1-u^2} + \frac{2}{(1-u)^2}\Big] \tag{34}$$

$$\leq \frac{c}{K} \tag{35}$$

where (29) follows from (8), (30) follows from the properties of the variance operator, (31) follows from the Cauchy-Schwarz inequality for the variance, (32) follows from (28), (33) follows by relaxing the $j \neq i$ condition in the second summation, (34) follows from the sum of infinite geometric series, and (35) follows as c is a constant independent of K.

IV. Proof of Proposition 3

If $D \sim N(0, \frac{\lambda_n^2}{3})$ and the Q-functions of the original system follow a normal distribution as:

$$\mathbf{Q}^{(1)}(s,a) = N(\mu_{sa}, \sigma_{sa}^2), \tag{36}$$

then the distribution of the Q-functions of the n^{th} system can be expressed using (36) and (1) as follows:

$$\mathbf{Q}^{(n)}(s,a) = N(\mu_{sa}, \sigma_{sa}^2 + \frac{\lambda_n^2}{3}), \tag{37}$$

where $\lambda_1 = 0$. We drop the (s,a) notation for simplicity.

We now derive the distribution of $\mathbf{Q}^{it}$ when $\mathbf{Q}^{it}$ converges ($\mathbf{Q}^{it}_{t-1} \to \mathbf{Q}^{it}_{t}$). Recall the update rule of Algorithm 1 from line 9:

$$\mathbf{Q}^{it} = u\mathbf{Q}^{it} + (1-u)\sum_{n=1}^{K} \mathbf{w}^{(n)}\mathbf{Q}^{(n)}. \tag{38}$$

Using the explicit expression of $\mathbf{Q}^{it}$ from (3), (36), (37) and the strict independence assumption, $\mathbf{Q}^{it}$ can be shown to follow a normal distribution. We will now find its mean and variance.

If we take the expectation of both sides in (38), we obtain the following:

$$\begin{aligned}
\mathbb{E}[\mathbf{Q}^{it}] &= \mathbb{E}[u\mathbf{Q}^{it} + (1-u)\sum_{n=1}^{K} \mathbf{w}^{(n)}\mathbf{Q}^{(n)}]. && (39)\\
&= u\mathbb{E}[\mathbf{Q}^{it}] + (1-u)\sum_{n=1}^{K} \mathbf{w}^{(n)}\mathbb{E}[\mathbf{Q}^{(n)}]. && (40)\\
&= u\mathbb{E}[\mathbf{Q}^{it}] + (1-u)\mu_{sa}, && (41)
\end{aligned}$$

where (40) follows from the linearity of expectation, and (41) follows from (37). If we rearrange the terms, we obtain:

$$\mathbb{E}[\mathbf{Q}^{it}] = \mu_{sa}. \tag{42}$$

If we take the variance of both sides in (38), we obtain the following:

$$\begin{aligned}
\mathbb{V}[\mathbf{Q}^{it}] &= \mathbb{V}[u\mathbf{Q}^{it} + (1-u)\sum_{n=1}^{K} \mathbf{w}^{(n)}\mathbf{Q}^{(n)}]. && (43)\\
&= u^2\mathbb{V}[\mathbf{Q}^{it}] + (1-u)^2\mathbb{V}[\sum_{n=1}^{K} \mathbf{w}^{(n)}\mathbf{Q}^{(n)}]. && (44)\\
&= u^2\mathbb{V}[\mathbf{Q}^{it}] + (1-u)^2\sum_{n=1}^{K} \left(\mathbf{w}^{(n)}\right)^2\mathbb{V}[\mathbf{Q}^{(n)}]. && (45)\\
&= u^2\mathbb{V}[\mathbf{Q}^{it}] + (1-u)^2\sum_{n=1}^{K} \left(\mathbf{w}^{(n)}\right)^2[\sigma_{sa}^2 + \frac{\lambda_n^2}{3}], && (46)
\end{aligned}$$

where (44) follows from the properties of variance, and the fact that $\mathbf{Q}^{it}$ is obtained at time $t-1$ and the $\mathbf{Q}^{(n)}$ is obtained at time t; hence, they are independent, (45) follows from the properties of variance and the strict independence assumption, and (46) follows from (37). If we rearrange the terms, we obtain the following:

$$\begin{aligned}
\mathbb{V}[\mathbf{Q}^{it}] &= \frac{1-u}{1+u}\sum_{n=1}^{K}(\mathbf{w}^{(n)})^2[\sigma_{sa}^2 + \frac{\lambda_n^2}{3}]. && (47)\\
&= \sigma_{it}^2, && (48)
\end{aligned}$$

where $\mathbf{w}^{(n)}$ is the final weights of the Algorithm 1.

Using (42) and (48), the distribution of $\mathbf{Q}^{it}$ can be expressed as follows:

$$\mathbf{Q}^{it} \sim N(\mu_{sa}, \sigma_{it}^2). \tag{49}$$

The JSD between $\mathbf{Q}^{(1)}$ and $\mathbf{Q}^{it}$ is defined as follows:

$$\mathrm{JSD}(\mathbf{Q}^{(1)} \| \mathbf{Q}^{it}) = \frac{1}{2}\big(\mathrm{KL}(\mathbf{Q}^{(1)} \| \tilde{\mathbf{Q}}) + \mathrm{KL}(\mathbf{Q}^{it} \| \tilde{\mathbf{Q}})\big), \tag{50}$$

where

$$\tilde{\mathbf{Q}} = \frac{1}{2}(\mathbf{Q}^{(1)} + \mathbf{Q}^{it}) \sim N(\mu_{sa}, \frac{\sigma_{sa}^2 + \sigma_{it}^2}{4}), \tag{51}$$

which follows from the strict independence assumption, and (36) and (37).

The KL divergence between two normal distributions with same means $D_1{\sim}N(\mu,\sigma_1^2)$ and $D_2{\sim}N(\mu,\sigma_2^2)$ can be computed as [1]:

$$\mathrm{KL}(D_1\|D_2)=\frac{1}{2}\Big(\frac{\sigma_1^2}{\sigma_2^2}-1+\ln\big(\frac{\sigma_2^2}{\sigma_1^2}\big)\Big). \tag{52}$$

If we use (36) and (51) with (52), we obtain:

$$\mathrm{KL}(\mathbf{Q}^{(1)}\|\tilde{\mathbf{Q}})=\frac{1}{2}\Big(\frac{4\sigma_{sa}^2}{\sigma_{sa}^2+\sigma_{it}^2}-1+\ln\big(\frac{\sigma_{sa}^2+\sigma_{it}^2}{4\sigma_{sa}^2}\big)\Big). \tag{53}$$

$$\mathrm{KL}(\mathbf{Q}^{(n)}\|\tilde{\mathbf{Q}})=\frac{1}{2}\Big(\frac{4\sigma_{it}^2}{\sigma_{sa}^2+\sigma_{it}^2}-1+\ln\big(\frac{\sigma_{sa}^2+\sigma_{it}^2}{4\sigma_{it}^2}\big)\Big). \tag{54}$$

By combining (53) and (54) using (50), we obtain the following:

$$\mathrm{JSD}(\mathbf{Q}^{(1)}\|\mathbf{Q}^{(n)})=\frac{1}{2}+\frac{1}{2}\ln\left[\frac{\sigma_{sa}^2+\sigma_{it}^2}{16\sigma_{sa}^2\sigma_{it}^2}\right]. \tag{55}$$

V. Proof of the theoretical runtime complexity

We consider the n^{th} chain. The number of different state-action pairs visited in a trajectory of length l can be minimum 1, and let the maximum be l_n, which is a function of l and ϵ. On average, $\frac{(l_n+1)}{2}$ different state-action pairs are visited. The minimum number of iterations for a sufficient exploration of the whole system is $\frac{|\mathcal{S}||\mathcal{A}|v}{\frac{(l_n+1)}{2}}$ as each state-action pair must be visited at least v times. Assuming each visit and corresponding Q-function update takes a unit time, the time for a sufficient exploration can be given by:

$$t_n=\frac{|\mathcal{S}||\mathcal{A}|v}{\frac{(l_n+1)}{2}}l. \tag{56}$$

The minimum number of visit v requirement is significantly alleviated due to the multiple Markov chains running simultaneously. Hence, we can assume the minimum number of visit requirement becomes $v'\approx\frac{v}{K}$. Then, the total time t for sufficient explorations for all n systems is given by:

$$t=\sum_{n=1}^{K}t_n=\sum_{n=1}^{K}\frac{|\mathcal{S}||\mathcal{A}|v'}{\frac{(l_n+1)}{2}}l \tag{57}$$

$$\approx\frac{|\mathcal{S}||\mathcal{A}|v}{K}\sum_{n=1}^{K}\frac{l}{l_n}. \tag{58}$$

Note that the PTTs $\mathbf{P}^n$ converge a fixed rank-1 tensor as $n\to\infty$, and the distribution of the elements in $\mathbf{P}^n$ becomes more uniform (*i.e.* the probabilities between different state-action pairs become closer to each other, and the number of zero probabilities decreases). Hence, it becomes more likely to visit a larger number of different state-action pairs as n increases. Consequently, we can assume that $l_1\ll l_2\ll\ldots\ll l_n$, and (58) can be approximated as:

$$\approx O\left(\frac{|\mathcal{S}||\mathcal{A}|v}{K}\frac{l}{l_1}\right)\approx O\left(\frac{|\mathcal{S}||\mathcal{A}|v}{K}f(l,\epsilon)\right), \tag{59}$$

where f is some function of l, l_1, ϵ.

VI. The practicality and accuracy of the independence assumptions

To assess the practicality of the independence assumptions in Table I , we compute the magnitude of the averaged Pearson's correlation coefficient (APCC) between $\mathcal{X}_{t_1}^{(n_1)}$ and $\mathcal{X}_{t_2}^{(n_2)}$ for (s,a)=(6, 2), with $n_1, n_2\in[2,10]$ and averaging the results over all $t_1\neq t_2\in[0,t]$. The results are shown in Fig.1. For example for $n_1=n_2=2$, the APCC is 0.02, which indicates that the relationship is not statistically significant; hence, Q-function errors of the 2^{nd} chain at different times are independent, and the modest independence assumption holds. However, large n values may lead to weak correlation as can be seen in the right diagonal, suggesting n should not be very large for practical purposes. On the other hand, for $n_1=2$ and $n_2=3$, the ADC equals 0.04, indicating the independence of the Q-function errors of the 2^{nd} and 3^{rd} chains at different times, and the validity of the strict independence assumption. Notably, when n_1 and n_2 are close, the strict independence assumption remains true; otherwise, moderate correlation may exist, as can be observed from the top-left and bottom-right corners. Simulations show that similar results are also valid for different (s,a).

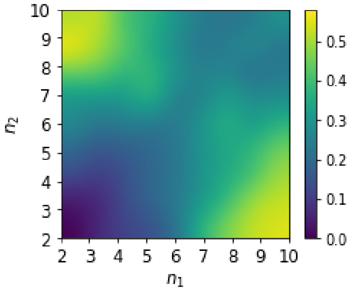

Fig. 1: The magnitude of the averaged Pearson's correlation coefficient (APCC) across different n

VII. The details on the network models and parameter optimizations

We herein explain how to select and optimize the hyper-parameter for different Q-learning algorithms. Since we carry out the simulations across different network sizes, there is no single set of hyper-parameters that fit all simulations. Hence, we give a reasonable range for each parameter across different settings and find the optimal one by trial-error.

- The probability (p) is used in model-2 and model-3 and set to $p = 0.8$, and the discount factor is set to γ = 0.9.
- The trajectory length (l) depends on a variety of parameters, including the underlying structure of the Markov chain. However, the proposed algorithm enables us to have short trajectories while a near-optimal APE performance is achieved, the runtime and number of computations in a single trajectory are minimized, and the necessary information from the neighboring states is sufficiently captured. Hence, it is optimized from the following sets by cross-validation: $l \in [1, 5]$ for small networks, $l \in [5, 10]$ for modest-sized networks, and $l \in [10, 20]$ for large networks.
- The number of visit requirement (v) is independent of the network size as each state-action pair needs to be sufficiently visited regardless of the number of other state-action pairs. The proposed algorithm allows us to choose a small v to achieve a near-optimal performance; in particular, $v \approx 50$ works well.
- The update ratio u is chosen to be a positive constant in Algorithm 1. However, choosing a time-dependent update ratio, u_t is possible and may indeed improve the system's performance and help the algorithm converge. In the beginning, u should be small to exploit multiple VMEs (exploration). Then, with a proper growth rate, u should increase to utilize previously obtained samples (exploitation). We use the following form $u_t = 1 - e^{\frac{-t}{c_4}}$ with the constant c_4 determining the grow rate works well. As the system parameters increase, c_4 should also increase to ensure the adequate exploitation of multiple VMEs. In particular, it is optimized from the following sets by cross-validation: $\{10^2, 10^3\}$ for small networks, $\{10^2, 10^3, 10^4\}$ for modest-sized networks, and $\{10^4, 10^5\}$ for large networks.
- As the number of VMEs (K) increases, the runtime complexity of the algorithm is reduced. However, the high-order node relationships are given more importance, which may result in the loss of the low-order node relationships, the $\mathbf{P}^{(n)}$ matrices converge to a fixed matrix; hence, the samples from the corresponding VMEs bring no new information, and may degrade the overall performance, and the memory needs increase. Thus, K should be small enough to avoid these drawbacks, yet it should also increase with the network size to prevent the increase in other parameters (specifically v, l) and reduce the runtime complexity. To this end, K is optimized from the following sets by cross-validation: $K \in \{2, 3\}$ for small networks, $K \in \{3, 4, 5\}$ for modest-sized networks, and $K \in \{5, 6, 7, 8\}$ for large networks.
- The number of estimators for EGQL and ADQN is selected via cross-validation for different network sizes as follows: for small networks: $\{2, 3, 4\}$, for modest-sized networks: $\{3, 4, 5, 6\}$, and for large networks: $\{5, 6, 7, 8, 9, 10\}$.
- For NQ and ADQN, the following parameters are selected via cross-validation from the following sets: batch size: $\{8, 16, 32, 64, 128\}$, replay buffer memory: $\{5 \cdot 10^3, 10^4, 2 \cdot 10^4\}$. In addition, the following implementations are used: for small networks: 3-layer fully connected NN with 48 nodes in each layer followed by ReLU, for modest-sized networks: 4-layer fully connected NN with 48 nodes in each layer followed by ReLU, for large networks: 4-layer fully connected NN with 96 nodes in each layer followed by ReLU. For model-2 and 3, we use the normalized set of buffer and channel states for different transmitters. The output layer has $|\mathcal{A}|$ nodes, followed by a softmax function, where each node represents the probability of a specific action being the optimal one. We use a dropout with a probability 0.2 after each layer except the final layer. The l_2 regularization is employed with the corresponding weight chosen from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. The Adam optimizer is used to update the weight of neural networks.

References

[1] Thomas M Cover. *Elements of information theory.* John Wiley & Sons, 1999.