# Supplementary Appendix

Talha Bozkus and Urbashi Mitra

## I. Proof of Corollary 1

The following expressions are valid for all $(s,a)$ pairs; hence, we drop the $(s,a)$ notation for simplicity.

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

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

$$\leq \lim_{t\to\infty}(1-u)^2\Big[\sum_{i=0}^{t-1}u^{2(t-i-1)}\mathbb{V}\big[\sum_{n=1}^{K}\mathbf{w}_i^{(n)}\mathcal{X}_i^{(n)}\big]+ 2\sum_{i=0}^{t-1}\sum_{j=i+1}^{t-1}u^{2(t-i-1)}u^{2(t-j-1)}\sqrt{\mathbb{V}\big[\sum_{n=1}^{K}\mathbf{w}_i^{(n)}\mathcal{X}_i^{(n)}\big]\mathbb{V}\big[\sum_{n=1}^{K}\mathbf{w}_j^{(n)}\mathcal{X}_j^{(n)}\big]}\Big]. \tag{3}$$

$$\leq \lim_{t\to\infty}(1-u)^2\Big[\sum_{i=0}^{t-1}u^{2(t-i-1)}+2\sum_{i=0}^{t-1}\sum_{j=i+1}^{t-1}u^{2(t-i-1)}u^{2(t-j-1)}\Big]\mathbb{V}\big[\sum_{n=1}^{K}\mathbf{w}_i^{(n)}\mathcal{X}_i^{(n)}\big]. \tag{4}$$

$$\leq \lim_{t\to\infty}(1-u)^2\Big[\sum_{i=0}^{t-1}u^{2(t-i-1)}+2\sum_{i=0}^{t-1}\sum_{j=0}^{t-1}u^{2(t-i-1)}u^{2(t-j-1)}\Big]\lambda^2. \tag{5}$$

$$\leq \frac{(1-u)}{(1+u)}\lambda^2+\frac{2\lambda^2}{(1+u)^2}, \tag{6}$$

where (1) follows from (19), (2) follows from the properties of the variance operator, (3) follows from the Cauchy-Schwarz inequality for the variance operator, (4) follows from the fact that variance is independent of time indices from (7), (5) follows from (26), and (6) follows from the infinite geometric sum formula.

## II. The details on the network models

### A. Model 1: SISO wireless network model with independent channels

We use the network model in [1]. The parameters $C_0$, $\eta$, $l$, $P_{rcv}$, $p$ are chosen the same as in [1]. The buffer size $Q$ and the number of channel partitions $H$ are chosen close to each other. The network size is chosen up to 20000 and formed by increasing $Q$ and $H$ comparably.

### B. Model 2: MISO energy harvesting wireless network with Gaussian interference channels and multiple relays

We consider the second wireless network model in [2]. The number of transmitters is chosen from $\{2,3,4,5,6,7,8\}$, and the number of relays is chosen from $\{1,2,3,4\}$. The input parameters $x_i = 1$ for $i \in \{2,3,4,5\}$. The probabilities $p_i = 0.2$ and $q_i = 0.8$ for $i \in \{2,3,4,5\}$. The weight parameters in the cost function $\alpha_1 = \frac{1}{9}$, $\alpha_2 = \frac{1}{6}$ and $\alpha_3 = \frac{1}{3}$, and these are chosen such that different cost components have comparable contributions on the overall cost function. These parameters are also kept constant for different network sizes. The network size of this network is formed as follows:

- Interval [0, 2000]: formed by increasing $N$ with 2 transmitters and 1 relay
- Interval [2000, 4000]: formed by increasing $N$ with 3 transmitters and 1 relay
- Interval [4000, 6000]: formed by increasing $N$ with 3 transmitters and 2 relay
- Interval [6000, 8000]: formed by increasing $N$ with 4 transmitters and 2 relay
- Interval [8000, 10000]: formed by increasing $N$ with 5 transmitters and 2 relay

- Interval [10000, 12000]: formed by increasing $N$ with 5 transmitters and 3 relay
- Interval [12000, 14000]: formed by increasing $N$ with 6 transmitters and 3 relay
- Interval [14000, 16000]: formed by increasing $N$ with 7 transmitters and 3 relay
- Interval [16000, 18000]: formed by increasing $N$ with 8 transmitters and 3 relay
- Interval [18000, 20000]: formed by increasing $N$ with 8 transmitters and 4 relay
- For all algorithms, the same structure for the learning rate ($\alpha_t$) and epsilon-probability ($\epsilon_t$) are used, and the parameters $c_1, c_2, c_3$ are optimized.
- For MMQ, EBQ, ADQN, EGQL, and TRPO, the number of estimators/models is selected via cross-validation for different network sizes: for small networks ($|\mathcal{S}| < 10^3$): $\{2, 3, 4\}$, for modest-sized networks ($10^3 < |\mathcal{S}| < 10^4$): $\{3, 4, 5\}$, and for large networks ($|\mathcal{S}| > 10^4$): $\{4, 5, 6, 7, 8\}$.
- For ADQN and TRPO, the neural network implementations given in the papers are used, and the parameters (batch size, replay memory, dropout, learning rate, regularization rate, etc) are chosen by cross-validation through a grid search. For model 3, the pair of the normalized buffer and channel states is used as input to neural networks (*i.e.* the input size is 2). For model 4, the pair of the normalized battery and channel states for different transmitters are concatenated and used as input (*i.e.* the input size is two times the number of 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.

## III. The performance of different distance metrics in Algorithm 1

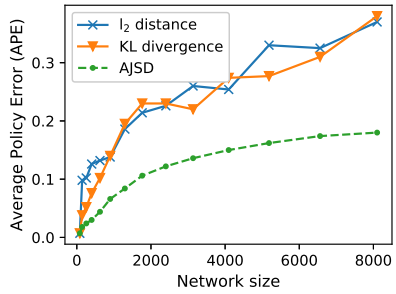


Fig. 1: APE vs network size across different similarity metrics

We firstly use the $l_2$ distance between the $Q$-functions as a similarity measure, and update the line-8 of Algorithm as follows:

$$\mathbf{w}_t^{(n)} \leftarrow 1 - \frac{1}{|\mathcal{S}|} \sum_{s=1}^{\mathcal{S}} \|\mathbf{Q}_t^{(1)}(s,:) - \mathbf{Q}_t^{(n)}(s,:)\|^2. \tag{7}$$

Herein, a larger distance leads to a smaller weight.

We also use the KL divergence between the $Q$-functions as a similarity measure, and update the line 8 of Algorithm as follows:

$$\mathbf{w}_t^{(n)} \leftarrow 1 - \frac{1}{|\mathcal{S}|} \sum_{s=1}^{\mathcal{S}} \mathrm{KL}\left(\hat{\mathbf{Q}}_t^{(1)}(s,:) \| \hat{\mathbf{Q}}_t^{(n)}(s,:)\right). \tag{8}$$

Herein, a larger KL divergence leads to a smaller weight.

We carry out the simulations with the same settings in Section IV-B. The APE of the proposed algorithm across network size is shown in Fig.1. Clearly, using AJSD as a distance metric to compare the $Q$-functions in Algorithm 1 results in %50 less APE. Moreover, slight changes in the network size may cause significant variations in APE when $l_2$ distance or KL divergence is used, whereas AJSD provides a more robust distance measure.

## References

[1] Libin Liu, Arpan Chattopadhyay, and Urbashi Mitra. On solving mdps with large state space: Exploitation of policy structures and spectral properties. *IEEE Transactions on Communications*, 67(6):4151–4165, 2019.

[2] Talha Bozkus and Urbashi Mitra. Link analysis for solving multiple-access mdps with large state spaces. *IEEE Transactions on Signal Processing*, 71:947–962, 2023.