Supplementary Information

I. Performance Analysis with Different Parameters

We evaluate the performance with respect to different parameters, i.e., $\alpha, \Delta\varepsilon$ and γ by calculating the cumulative reward per episode and the l_2-norm of the Q-table entries for different combinations of these parameter values. The rewards indicate the convergence process to the final solution. The final converged reward should be a roughly constant value if the algorithm has converged to the final solution. Furthermore, if the final reward value is identical with different parameters, it indicates that the algorithm is successful in avoiding random and locally optimal solutions for a range of parameters. The l_2-norm of the Q-table entries indicates the same behavior as the reward. In addition, it shows the difference in convergence speed with respect to different parameters more clearly than the reward. We fix $\lambda_l = 0$ for these experiments as our main purpose in this sub-section is to study the effect of different parameter on the proposed algorithm's performance.

1) Learning Rate-α: The learning rate determines the extent to which a new information replaces the previous Q-values during the iterative Q-learning process, with a lower α value promoting a higher emphasis on the previous Q-values and vice-versa. As the choice of the α value can influence the learning process, we conduct the experiments with different learning rates of α = 0.05, 0.1, 0.15 and 0.2, keeping $\Delta\varepsilon$ and γ fixed at 0.005 and 0.95, respectively.

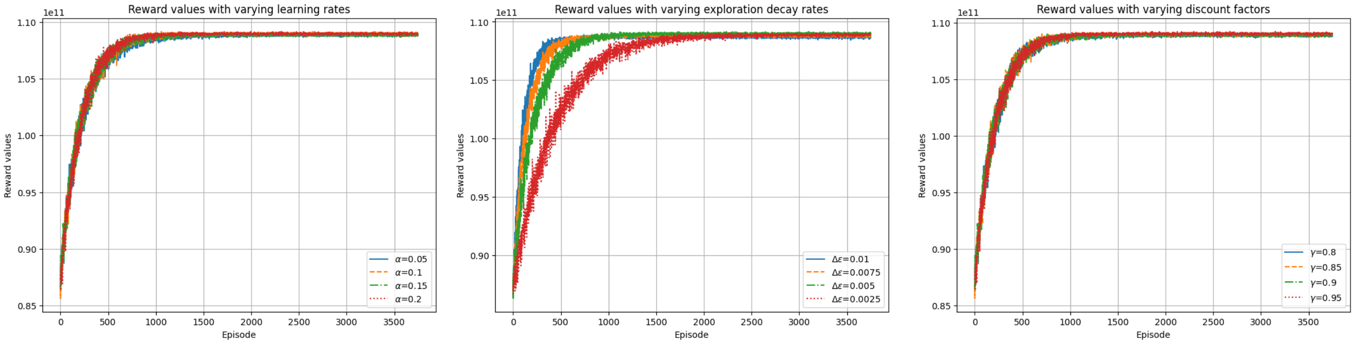
The corresponding results are shown in Figs. 1a and 2a. The former figure shows that the final reward value and the evolution of the reward value versus the episodes are almost similar for all values of α. It can be noted from Fig. 2a that when α changes from 0.05 to 0.1, there is a noticeable change in the convergence speed as well as the maximum l_2-norm value. For α = 0.1, the l_2-norm of the Q-values reach stability with a lower number of episodes compared to that for α = 0.05. For α = 0.15 or 0.2, the convergence speed increases further; however, this increase is less compared to that obtained between the aforementioned α values. The maximum l_2-norm value keeps on increasing with increasing α. As the change in convergence speed is less significant for $\alpha > 0.1$, we choose α = 0.1.

2) Exploration Decay Rate-$\Delta\varepsilon$: Exploration decay rate affects the balance between exploration, i.e., choosing random actions, and exploitation, i.e., choosing actions based on the learned Q-table values. If $\Delta\varepsilon$ value is too high/low, it means that the corresponding ε will decrease quickly or slowly in each episode, impacting the algorithm's ability to explore the environment. Thus, the optimal value of $\Delta\varepsilon$ can facilitate an effective learning while preventing over-exploration by controlling the rate of change of ε in each episode.

We experiment with different values of $\Delta\varepsilon$ = 0.0025, 0.005, 0.0075 and 0.01, keeping α and γ fixed at 0.1 and 0.95, respectively. The corresponding results are shown in Figs. 1b and 2b. Both these figures show that the convergence speed improves as the $\Delta\varepsilon$ value becomes higher, with the most significant improvement exhibited when the $\Delta\varepsilon$ value changes from 0.0025 to 0.005. For further higher values, the improvements in the convergence speed for both the reward values and l_2-norm of Q-table values are less significant, and the final converged reward values are the same for all the aforementioned values of $\Delta\varepsilon$. Therefore, we use $\Delta\varepsilon$ = 0.005 for subsequent simulations.

3) Discount Factor-γ: The discount factor determines the importance of future rewards in the learning process. Varying γ can influence the algorithm's focus on short-term versus long-term gains. We conduct experiments with different γ values of 0.8, 0.85, 0.9 and 0.95, keeping $\Delta\varepsilon$ and α fixed at 0.005 and 0.1, respectively. The results obtained are shown in Figs. 1c and 2c, where the latter figure clearly demonstrates that the convergence speed decreases and the final converged value of the l_2-norm of the Q-table entries increases as the γ values increase. However, it should also be noted that this change in convergence speed is less compared to the changes observed with respect to α and $\Delta\varepsilon$.

A higher value of γ can prioritize long-term rewards, potentially causing the algorithm to overlook the immediate gains, while a lower value can lead to a short-sighted approach, with the algorithm favoring the immediate rewards. For the considered problem and solution, a higher value of γ can ensure that a reward in later episodes can still influence the Q-table value in a particular state, which provides a higher chance to a UD to be associated with the most suitable BS in case this association was not discovered in the earlier episodes. Fig. 2c shows that there is a considerable difference between the final converged value of the curve corresponding to γ = 0.95 compared with those for higher γ values. Furthermore, as Fig. 1c does not depict any major difference in the convergence behavior for the reward, we select γ = 0.95 for subsequent experiments.

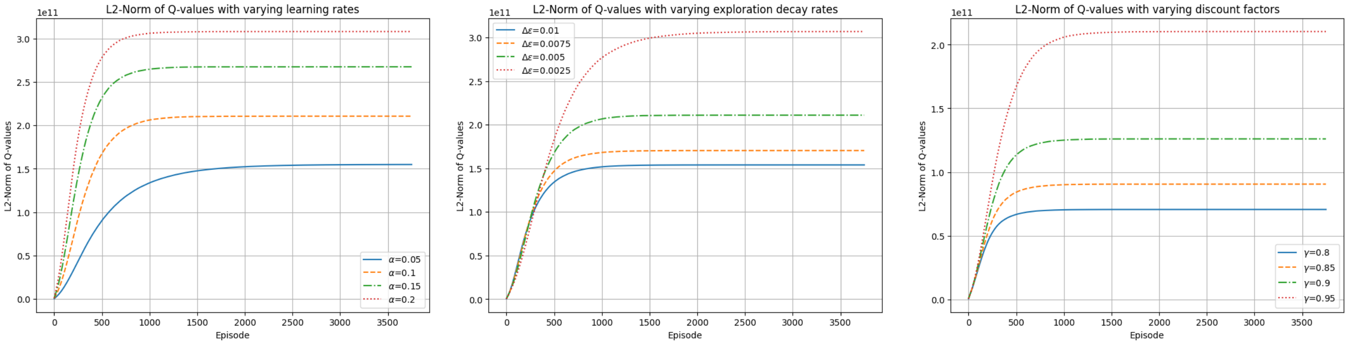


(a) Different α values.

(b) Different $\Delta\epsilon$ values.

(c) Different γ values.

Fig. 1: Reward variation versus episodes for different parameter values.



(a) Different α values.

(b) Different $\Delta\epsilon$ values.

(c) Different γ values.

Fig. 2: L_2-norm variation of the Q-table versus episodes for different parameter values.