Intelligent Trajectory Planning in UAV-mounted Wireless Networks: A Quantum-Inspired Reinforcement Learning Perspective

Yuanjian Li, A. Hamid Aghvami Fellow, IEEE and Daoyi Dong Senior Member, IEEE

Abstract—In this paper, we consider a wireless uplink transmission scenario in which an unmanned aerial vehicle (UAV) serves as an aerial base station collecting data from ground users. To optimise the expected sum uplink transmit rate without any prior knowledge of ground users (e.g., locations, channel state information and transmit power), the trajectory planning problem is optimized via the quantum-inspired reinforcement learning (QiRL) approach. Specifically, the QiRL method adopts novel probabilistic action selection policy and new reinforcement strategy, which are inspired by the collapse phenomenon and amplitude amplification in quantum computation theory. Numerical results demonstrate that the proposed QiRL solution can offer natural balancing between exploration and exploitation via ranking collapse probabilities of possible actions, compared to the popularly-used reinforcement learning approaches which are highly dependent on tuned exploration parameters.

Index Terms—UAV, trajectory planning, Quantum computation, quantum-inspired reinforcement learning (QiRL).

I. INTRODUCTION

Unmanned aerial vehicle (UAV) has been recognised as a promising technique to facilitate wireless communications in recent years, due to its attractive advancements such as flexible mobility, on-demand deployment and cost effectiveness [1]. Compared to terrestrial wireless communication scenarios, the most notable feature of UAV-mounted wireless networks is the controllable adjustments of UAV’s flying trajectory, which can offer favourable wireless channel qualities [2]. This feature encourages the concern of UAV’s trajectory design, which is a key research objective in UAV-aided networks. To solve optimal trajectory planning problem of UAV-based networks, reinforcement learning (RL) has been leveraged, for its ability to learn in a “trial-and-error” manner without explicit knowledge of the environment [3], [4].

Balancing the ratio of exploration and exploitation remains the inherent challenge of RL-based intelligent systems, which poses significant impacts on learning efficiency and quality, e.g., $\epsilon$-greedy and Boltzmann action selection strategies [5], [6]. On one hand, $\epsilon$-greedy method renders that a random action is executed with probability $\epsilon \in [0, 1]$, and the optimal action is selected with probability $(1 - \epsilon)$ according to the developed action selection policy. This method is simple and effective. However, one of its drawbacks is that it selects actions uniformly among all possible actions while exploring, which means that it cannot distinguish the next-to-optimal action from its worse counterparts. On the other hand, Boltzmann (or the Softmax) exploration method introduces a selection probability $\exp(Q(s,a)/\tau)/(\sum_i \exp(Q(s,a^i)/\tau))$ according to the Q function $Q(s,a)$ of state $s$ and action $a$, where the parameter $\tau$ represents the temperature in the Boltzmann distribution. However, finding a good $\tau$ which can properly balance the ratio between exploration and exploitation is difficult. The parameters $\epsilon$ and $\tau$ pose significant impacts on the convergence performance and the quality of learning output, which makes it necessary to develop new action selection strategy for RL.

Recently, with the advancement of quantum computation techniques, introducing quantum mechanism into the field of machine learning is believed to be a promising direction to build advanced machine learning algorithms. Dong et al. [5] proposed the concept of quantum reinforcement learning (QRL), in which QRL was applied to solve the typical gridworld problem. Thereafter, in [7], Dong et al. introduced quantum-inspired reinforcement learning (QiRL) into the field of navigation control of autonomous mobile robots. Fakhari et al. [8] applied QiRL approach into unknown probabilistic environment, in which the robustness of QiRL solution was demonstrated. Li et al. [4] compared QRL with 12 conventional RL (CRL) models in human decision-making scenarios, suggesting that value-based decision-making can be illustrated by QRL at both the behavioural and neural levels. However, QRL is now still in its infancy state, and it is not yet introduced into the field of UAV-aided networks, e.g., solving path planning problems.

In this paper, a novel RL algorithm inspired by quantum mechanism, which is independent on exploration parameters, is applied to tackle the trajectory planning problem in UAV-aided uplink transmission scenario. Specifically, in this proposed QiRL solution, balancing exploration and exploitation is realized in a manner inspired by the collapse phenomenon of quantum superposition and the quantum amplitude amplification. Different from [5] and [7], we extend the quantum explanation of QiRL from the fixed rotation angles to their flexible counterparts, which is an alternative of [6] and [8]. Besides, we also relax the limitation of linear function mapping in [6] and that of empirical rotation angle setting in [8].
II. Problem Statement

This work concentrates on the uplink transmission scenario consisting of a UAV and K ground users (GUs), where the location of each ground user is denoted as \( D_k = (x_k, y_k, 0) \) with \( k \in \{1, 2, \ldots, K\} \). It is assumed that all the GUs are uploading their messages in a frequency division multiplexing manner. Thus, each GU transmits solely on its assigned channel and inner-channel interference can be approximately ignored. Besides, the UAV flies with a constant velocity \( V \) (m/s) with fixed altitude \( H \) (m)\(^7\). A practical network information assumption is applied, in which the UAV cannot obtain any environment knowledges, e.g., transmit power of the GUs, locations of the GUs, and can only observe the received signals from the GUs. The goal of the UAV is to maximize the expected uplink transmit rate (ESUTR) of the GUs via intelligently adjusting its flying trajectory from the start location \( \bar{L}_0 = (x_0, y_0, H) \) to the destination \( \bar{L}_F = (x_F, y_F, H) \), without any environment information available for the UAV. Assume that the feasible region where the UAV can explore is a rectangular area \([x_0, x_F] \times [y_0, y_F]\), which is denoted as \( \Phi \) for clarity. To make the trajectory design tractable, the entire trajectory is discretized into \( F \) equal-spacing steps, via evenly quantifying the time horizon into \( F \) time slots and the length of each time slot is predefined as \( T(s) \) so that the 3-dimensional location at the beginning of each time slot can be given by \( L = \{L_1, L_2, \ldots, L_F\} \). Then, we have \( \bar{L}_0 \leq L_f \leq \bar{L}_F, \forall f \in [0, F] \), where \( \leq \) represents element-wise inequality. Therefore, the problem of ESUTR maximization can be stated as

\[
\begin{align*}
\max_{L \in \mathcal{F}} & \frac{1}{F} \sum_{f=1}^{F} \sum_{k=1}^{K} \omega_k \log \left( 1 + \frac{P_k \delta_0}{\sigma_k^2 ||L_f - D_k||^\eta} \right), \\
\text{s.t.} & \|L_f - L_{f-1}\| = VT, \\
& \bar{L}_0 \leq L_f \leq \bar{L}_F, \\
& FT \leq E, \\
& \sum_{k} \omega_k \leq B,
\end{align*}
\]

(1a)

(1b)

(1c)

(1d)

(1e)

where \( P_k \) represents the uplink transmit power of the GU \( k \), \( \omega_k \) means the bandwidth occupied by the GU \( k \), \( B \) indicates the total bandwidth constraint of the system, \( \sigma_k^2 \) is the power of the Additive White Gaussian Noise (AWGN), \( \delta_0 \) means the reference channel power gain, \( \eta \) is the path loss exponent and \( E \) represents the maximum flight time threshold. Note that the constraint (1b) ensures that the flying distance between arbitrary adjacent time slots is fixed as the UAV’s roaming capacity \( VT \) in each time slot, the constraint (1c) makes sure that the UAV’s trajectory is exclusively within the feasible regime, the constraint (1d) declares that the maximum exploration time \( FT \) is constrained by the on-board power capacity of the UAV and the constraint (1e) limits that the sum of each GU’s occupied bandwidth should lie in the range of available bandwidth resource. The proposed problem (1) cannot be tackled via traditional optimization approaches due to the lack of environment information but can be solved by the conventional model-free RL algorithms in a "trial-and-error" manner, e.g., Q-learning. However, traditional RL suffers from low learning efficiency, difficulty of balancing the ratio between exploration and exploitation, etc. To give a better alternative approach solving problem (1), the novel QiRL technique will be invoked to tackle the proposed optimal trajectory planning problem.

III. QiRL Solution

The above trajectory design problem is a sequential decision making process, which means the UAV’s movement is determined within each time slot merely based on its current situation. Therefore, Markov decision process (MDP) is a suitable candidate for solving the trajectory optimization problem, which can help forge the optimal mapping from the current state to the best action selection.

To formulate the MDP, first we need to clarify the states of the proposed QiRL solution for our considered scenario. The feasible area \( \Phi \) is divided into \( N_1 \times N_2 \) small grids and the side length of each grid equals \( VT \). Besides, we assume that the sum of received signal strength keeps constant within each grid\(^2\). The GUs are located in some of the small squares, which will be specified in the numerical results. According to the discrete tabular form of \( \Phi \), the state set of the UAV can be written as \( S = \{s_1, s_2, \ldots, s_{N_1 N_2}\} \), where \( s_i \in S \) represents a small square in \( \Phi \). Because we focus on the ESUTR maximization problem, it is straightforward to define \( R = \sum_{k=1}^{K} \omega_k \log \left( 1 + \frac{P_k \delta_0}{\sigma_k^2 ||L_{s_i} - D_k||^\eta} \right) \) as the reward function, where \( L_{s_i} \) denotes the location of a possible state \( s_i \). Note that the UAV is only able to observe \( R_{s_i} \), while other network information of \( P_k, \delta_0, \sigma_k^2 \) and \( D_k \) is inaccessible. The UAV aims to find an optimal path in which the ESUTR of the GUs should be the greatest among all possible UAV roaming routes from \( \bar{L}_0 \) to \( \bar{L}_F \). To drive the UAV to the destination \( \bar{L}_F \), the UAV will gain a special reward which is defined as \( R = 10 \times \max_{s_{i,s} \in S} R_{s_i} \), once it reaches \( \bar{L}_F \).

Regarding the UAV’s possible actions, we limit the movement options of the UAV in the following action set \( \mathcal{A} = \{\text{forward, backward, left, right}\} \), which will be denoted as quantum eigenactions in the proposed QiRL solution. The goal of the proposed QiRL algorithm is to learn a mapping from states to actions, i.e., the UAV aims to learn a policy \( \pi : S \to \mathcal{A} \) so that the expected sum of discounted rewards can be maximized. We define the value function of state \( s \) at trial \( t \) as

\[
V_{\pi}(s) = \mathbb{E}_{\pi} \left[ \sum_{t=0}^{F} \gamma^t R(t + l + 1) | S(t) = s \right],
\]

(2)

where \( \gamma \) represents the discount factor. Furthermore, the temporal difference (TD)-based value updating rule of the proposed QiRL can be described as

\[
V(s) \leftarrow V(s) + \alpha [R(s') + \gamma V(s') - V(s)],
\]

(3)

where \( s' \) means the next state after taking an action and \( \alpha \) indicates the learning rate.

1This work focuses on strong LOS path loss channel model and the effects of small scale fading are omitted so that a lower altitude is always preferable to achieve nearer distance between the UAV and the GUs. Hence, the UAV’s altitude \( H \) is assumed as a fixed parameter, which may be the lowest flying height under the regulation of local laws in practice.

2This assumption is reasonable because the acreage of each grid is far less than that of \( \Phi \), in the case of sufficient discretization.
From the above MDP formulation, it is easy to find that the objective function (11) corresponds to the un-discounted expected rewards over one episode, which is a special type of MDP called episodic tasks where a special state named the terminal state separates the agent-environment interactions into episodes. Thus, we set the discounted factor \( \gamma = 1 \) for the considered problem. The terminal state and the start state correspond to \( L_e \) and \( L_0 \), respectively.

According to quantum mechanics [9], a quantum state \( |\psi\rangle \) (Dirac representation) can describe a state of the closed quantum system, which is a unit vector (i.e., \( \langle \Psi | |\Psi\rangle = 1 \)) in Hilbert space. The quantum state \( |\Psi\rangle \) which consists of \( n \) qubits can be expanded as

\[
|\Psi\rangle = \sum_{\sum_{p=00\ldots0}} h_p |p\rangle,
\]

where \( |\psi_i\rangle, i \in [1, n] \) represent the state of qubit \( i \), \( h_p \) are termed as the probability amplitude which are complex coefficients satisfying the normalisation constraint \( \sum_{p=00\ldots0} h_p^2 = 1 \), and \( \otimes \) represents the tensor product. The representation of quantum state \( |\Psi\rangle \) follows the quantum phenomenon called state superposition principle. Note that the superposition \( |\psi_i\rangle \) can be regarded as the superposition of \( 2^n \) eigenstates ranged from \( |00\ldots0\rangle \) to \( |11\ldots1\rangle \).

To represent the four possible actions in QiRL, two qubits are sufficient. Furthermore, eigenactions (i.e., the quantum representation of classical actions) \( |a_1\rangle, |a_2\rangle, |a_3\rangle, |a_4\rangle \) are allocated to denote the actions forward, backward, left and right, respectively. Inspired by the superposition principle of quantum theory, we can integrate all the four eigenactions \( |a_n\rangle \) for each state in quantum superposition form as

\[
|A(l)\rangle = \sum_{n=1}^{4} h_n |a_n\rangle \rightarrow \sum_{a=00}^{11} h_a |a\rangle,
\]

where \( l \) represents a specific trial and \( h_n \) and \( h_a \) are the complex-value probability amplitudes under the normalisation constraints \( \sum_{n=1}^{4} |h_n|^2 = 1 \) and \( \sum_{a=00}^{11} |h_a|^2 = 1 \), respectively. Note that the superposition \( |A(l)\rangle \) is a unit vector in a 4-dimensional Hilbert space (action space) spanned by the four orthogonal bases \( |a_n\rangle, n = 1, \ldots, 4 \). Specifically, the action taken by the UAV before any quantum measurement lies in a superposition stage (4 possibilities in total), which is mapped into the tensor product of two qubits. The quantum representation \( |A(l)\rangle \) establishes a bridge between quantum eigenactions and the physical action set \( A \), which allows us to apply quantum amplitude amplification as a reinforcement strategy.

In quantum theory, when an external agency (e.g., observer, experimenter) measures the quantum state \( |\Psi\rangle = \sum_{n} \theta_n |\psi_n\rangle \) with the eigenbasis \( \{ |\psi_n\rangle \} \), \( |\Psi\rangle \) will collapse from the superposition state to one of its eigenstates \( |\psi_n\rangle \), i.e., \( |\Psi\rangle \rightarrow |\psi_n\rangle \), with probability \( |\langle \psi_n |\Psi\rangle|^2 = |\theta_n|^2 \). Inspired by this quantum collapse phenomenon, the superposition \( |A(l)\rangle \) will collapse onto one of the eigenactions \( |a_n\rangle \) with probability of \( |h_n|^2 \), during the action decision-making period in the proposed QiRL strategy. Moreover, the probability amplitude of each action can be amplified or attenuated according to corresponding reward and value function, via a specific quantum algorithm (e.g., Grover’s iteration method [9]). Therefore, the quantum collapse phenomenon and the probability amplitude technique together can offer a natural action selection option termed as the collapse action selection strategy, which provides us a new way of balancing between exploration and exploitation without modifying exploration parameters [7].

It is clear that the probability amplitude updating lies at the core of the proposed QiRL algorithm, to select a “good” action in the collapse action selection period. To realize this, two unitary operators can be employed for the currently chosen action \( |a_i\rangle \) which is from the \( l \)-th trial \( |A(l)\rangle = \sum_{n=1}^{4} h_n |a_n\rangle = h_i |a_i\rangle + h_{a_i}^+ |a_i^+\rangle \), shown as

\[
U_{|a_i\rangle} = I - (1 - e^{j\phi_1}) |a_i\rangle \langle a_i|,
\]

\[
U_{|A(l)\rangle} = (1 - e^{j\phi_2}) |A(l)\rangle \langle A(l)| - I,
\]

where \( |a_i\rangle \) and \( h_{a_i}^+ \) are Hermitian transposes of \( |a_i\rangle \) and \( |A(l)\rangle \), respectively. Then, the Grover operator is constructed in the form of unitary transformation, given by

\[
G = U_{|A(l)\rangle} U_{|a_i\rangle}.
\]

After \( m \) times of applying \( G \) on \( |A(l)\rangle \), the amplitude vector in the next trial becomes \( |A(l + 1)\rangle = G^m |A(l)\rangle \).

There are mainly two methods to deal with the aforementioned probability amplitude updating task. One is to choose a feasible value of \( m \) with fixed parameters \( \phi_1 \) and \( \phi_2 \) (commonly both of them equal to \( \pi \)); the other is to fix \( m = 1 \) with dynamic parameters \( \phi_1 \) and \( \phi_2 \). Because the former updating approach can only modify the amplitudes in a discrete manner, the latter method is chosen in this work, i.e., Grover iteration with flexible parameters \( \phi_1 \) and \( \phi_2 \). Then, the impacts of \( G \) on the superposition representation \( |A(l)\rangle \) can be given by the following proposition.

**Proposition 1:** The overall effects of \( G \) with free parameters \( \phi_1 \) and \( \phi_2 \) on the superposition representation \( |A(l)\rangle \) in the \( l \)-th trial can be expressed analytically as

\[
G |A(l)\rangle = (Q - e^{j\phi_2}) h_i |a_i\rangle + (Q - 1) h_{a_i}^+ |a_i^+\rangle,
\]

where \( Q = (1 - e^{j\phi_2}) [1 - (1 - e^{j\phi_1}) |h_i|^2] \).

**Proof:** The impacts of \( U_{|a_i\rangle} \) on \( |a_i\rangle \) and \( |a_i^+\rangle \) can be given by

\[
U_{|a_i\rangle} |a_i\rangle = [I - (1 - e^{j\phi_1}) |a_i\rangle \langle a_i|] |a_i\rangle = e^{j\phi_1} |a_i\rangle,
\]

\[
U_{|a_i\rangle} |a_i^+\rangle = [I - (1 - e^{j\phi_1}) |a_i\rangle \langle a_i|] |a_i^+\rangle = |a_i^+\rangle,
\]

respectively. Furthermore, we have

\[
U_{|a_i\rangle} |A(l)\rangle = [I - (1 - e^{j\phi_1}) |a_i\rangle \langle a_i|] |A(l)\rangle = e^{j\phi_1} h_i |a_i\rangle + h_{a_i}^+ |a_i^+\rangle,
\]

in which \( U_{|a_i\rangle} \) plays the role as a conditional phase shift operator in quantum computation. Finally, we obtain

\[
G |A(l)\rangle = U_{|A(l)\rangle} U_{|a_i\rangle} |A(l)\rangle
= (1 - e^{j\phi_2}) [h_i |a_i\rangle + h_{a_i}^+ |a_i^+\rangle] \times
\left[ h_i^+ \langle a_i| + h_{a_i}^+ |a_i^+\rangle \right] U_{|a_i\rangle} |A(l)\rangle - U_{|a_i\rangle} |A(l)\rangle
= (Q - e^{j\phi_1}) h_i |a_i\rangle + (Q - 1) h_{a_i}^+ |a_i^+\rangle,
\]

where \( Q = (1 - e^{j\phi_2}) [1 - (1 - e^{j\phi_1}) |h_i|^2] \).
The impact of changing the basis from the Grover iteration can be given by \(|R|^2|\psi_i|^2\), where \(R\) is the rotation from \(U\) to \(V\). Then, the updated probability of the selected action \(|a_i\rangle\) after the Grover iteration can be given by \(|R'|^2|\psi_i'||^2\).

**Remark 1:** The ratio between the probability amplitudes of \(|a_i\rangle\) after being acted by the Grover operator \(G\) and before that can be expressed as

\[
R = (1 - e^{i\phi_1} - e^{i\phi_2})(1 - e^{i\phi_1})(1 - e^{i\phi_2})|\psi_i|^2.
\]

Then, the updated probability of the selected action \(|a_i\rangle\) after the Grover iteration can be given by \(|R|^2|\psi_i|^2\).

**Remark 2:** For ease of understanding the effect of \(G\), we show its corresponding algebraic visualization. In Fig. 1, the Bloch sphere representation \(|A(l)\rangle\) is reconstructed in the form of polar coordinates, shown as

\[
|A(l)\rangle = e^{i\phi}(\cos \theta/2 |a_i\rangle + e^{i\varphi}\sin \theta/2 |a_i^+\rangle).
\]

where the parameter \(e^{i\phi}\) can be omitted, since a global phase poses no observable effects [3]. Note that the polar angle parameter \(\theta\) and the azimuthal angle variable \(\varphi\) define the unit vector \(|A(l)\rangle\) on the Bloch sphere, as shown in Fig. 1. The impact of \(U_{A(l)}\) can be understood as a clockwise rotation around the \(z\)-axis by \(\phi_1\) (the red circle) on the Bloch sphere, leading to the rotation from \(|A(l)\rangle\) to \(|A(l')\rangle\). Similarly, if we change the basis from \(|A(l')\rangle\) to \(|A(l)\rangle\), \(U_{A(l)}\) makes a clockwise rotation around the new \(z\)-axis \(|A(l')\rangle\) by \(\phi_2\) (the blue circle), which rotates \(|A(l')\rangle\) into \(|A(l)\rangle\). Therefore, the overall effect of \(G\) on \(|A(l)\rangle\) is a two-step rotation which can modify the polar angle \(\theta\) when the basis is locked as \(|\{a_i, a_i^+\}\rangle\). Via controlling parameters \(\phi_1\) and \(\phi_2\), it is possible to realize arbitrary parametric rotation on the Bloch sphere, which acts as the foundation for modifying the probability amplitudes of \(|A(l)\rangle\). The smaller \(\theta\) is, the higher probability \(|A(l)\rangle\) will collapse to \(|a_i\rangle\) when it is measured, which inspires us to apply it as a reinforcement strategy. The core of this reinforcement approach is to achieve a smaller \(\theta\) via manipulating \(\phi_1\) and \(\phi_2\) when \(|a_i\rangle\) is recognized as a "good" action. Otherwise, if \(|a_i\rangle\) is determined as a "bad" action, we should modify \(\phi_1\) and \(\phi_2\) to enlarge \(\theta\).

**Proposition 2:** For arbitrary \(\theta \in [0, \pi]\), we can modify \(\phi_1\) and \(\phi_2\) to realize either reduce or enlarge \(\theta\), resulting in larger or smaller probability of collapsing from the superposition \(|A(l)\rangle\) to the current action representation \(|a_i\rangle\), respectively. After the amplitude amplification, the updated probability of the currently selected action \(|a_i\rangle\) can be alternatively given by \(e^{k[R+V(s')]|\psi_i|^2}\). Note that all the possible probability amplitudes together should be re-normalised after each implementation of amplitude amplification. The proposed QiRL solution is concluded in Algorithm 1 which can be conducted in conventional computers.

**Note:** Remark 1 and Remark 2 give a thorough explanation for amplitude amplification in quantum mechanism. Inspired by this, we can provide the reinforcement strategy for our proposed model. According to Remark 1, it is straightforward to conclude that \(|R|^2\) should be designed to be larger than 1, if the current representation \(|a_i\rangle\) is determined as a "good" action. Otherwise, \(|R|^2\) should be manipulated to be smaller than 1. By selecting feasible \(\phi_1\) and \(\phi_2\), it is possible to manipulate the value of \(|R|^2\) in the manner as mentioned before, which can be interpreted geometrically via Remark 2. Inspired by the quantum amplitude amplification strategy and for the sake to simulate it in the conventional computer, we use \(e^{k[R+V(s')]|\psi_i|^2}\) to represent alternatively the overall effects of \(G\) on probability \(|h_j|^2\), which means the updated probability should be \(e^{k[R+V(s')]|\psi_i'||^2}\). If \(k > 0\), the current action will be rewarded while it will be punished if \(k < 0\). The updating amplification is controlled via \(k [R + V(s')]\).

**Remark 3:** The quantum-inspired reinforcement strategy prioritizes all possible actions in ranked probability sequence which is updated alongside the learning process. Thus, it can naturally balance the exploration and exploitation, in which no tuned exploration parameter is necessary.

**Proposition 3:** The convergence of the proposed QiRL algorithm is guaranteed when the learning rate \(\alpha\) is non-negative and satisfies \(\lim_{T \to \infty} \sum_{k=1}^T \alpha_k = \infty\) and \(\lim_{T \to \infty} \sum_{k=1}^T \alpha_k^2 < \infty\).

**Proof:** The proof is omitted for its simplicity, which is similar to the proof of Proposition 2 in [5].

**Algorithm 1:** The proposed QiRL algorithm

**Input:** Learning parameters: \(\alpha \in [0, 1]\), \(\gamma = 1\); UAV informations: \(\widehat{L}_0\), \(L_F\), \(H\), \(V\), \(T\);

**Output:** The optimal policy \(\pi^*\) = AmpMem;

1 **Initialization:** \(ep = 0\); \(s = \widehat{L}_0\); \(V(s) = 0\), \(\forall s \in S\);

\[
\text{AmpMem} = \text{defaultdict}(\lambda = \{\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\});
\]

2 **while** \(ep \leq \text{NumEp}\) **do**

3 **repeat**

4 Choose a feasible \(a\) for \(s\) via measuring AmpMem[s];

5 Apply \(a\), then observe the reward \(R\) and the next state \(s'\);

6 Update the value function as per

\[
V(s) \leftarrow V(s) + \alpha [R + \gamma V(s') - V(s)];
\]

7 Apply quantum-inspired reinforcement factor \(e^{k[R+V(s')]\alpha}\) on AmpMem[s][a]. When the UAV hits the boundary or \(\Delta V(s) < 0, k < 0\). Otherwise, \(k > 0\);

8 Re-normalise AmpMem[s];

9 \(s \leftarrow s'\);

10 **until** \(ep > E/T\) or \(s' = L_F\);

11 \(ep += 1\);

12 **end**

IV. SIMULATION RESULTS

In this section, experimental results are evaluated for the considered UAV trajectory planning problem via the proposed
QiRL solution. For performance comparison, two CRL methods (i.e., Q-learning with $\epsilon$-greedy and Boltzmann exploration strategies) are also performed. It is assumed that the feasible UAV exploration field $\Phi$ is a square area with side length 200m, where 5 GUs are located on the ground (denoted by the red stars). By default, the length of each time slot is fixed as $T = 2s$ and the constant flying altitude and speed of the UAV are set as $H = 100m$ and $V = 10m/s$, respectively. The area $\Phi$ is divided into 10-by-10 small grids and the side length of each grid equals $VT = 20m$. The start location and the destination are predefined at $\tilde{L}_0 = (10, 190, 100)$ and $\tilde{L}_F = (190, 10, 100)$, respectively. Considering the on-board power capacity of the UAV, the total flying time of the UAV is constrained as $FT \leq 1800s$ so that we set $E = 1800$. Besides, we set $P_k = 1Watt$, $\delta_0/\sigma_k^2 = 1$, $\eta = 2$, $B = 10MHz$ and $\omega_k = 2MHz$, which is in line with [3].

Fig.2 shows the performance comparison of one widely-used CRL approach called Q-learning with two action selection strategies, i.e., $\epsilon$-greedy and Boltzmann, and the proposed QiRL solution. Note that exploration parameters $\epsilon$ and $\tau$ of Q-learning approach keep annealing alongside the learning progress, which controls the ratio of exploration and exploitation and highly affects the overall learning quality and convergence performance of Q-learning approach. In this figure, the learned trajectories of Q-learning and QiRL are also depicted for intuitive comparison. Specifically, subfigure (a) shows the expected reward curves which correspond to subfigure (b). From subfigure (a), it is straightforward to observe that the proposed QiRL solution can converge much faster than Q-learning with $\epsilon$-greedy action selection strategy, while QiRL has faster convergence speed than Q-learning with advanced Boltzmann action selection strategy. Most importantly, the proposed QiRL applies quantum-inspired action selection approach which can offer natural balancing between exploration and exploitation via ranking the collapse probabilities of superposition action to arbitrary possible actions without tuning the exploration parameter like $\epsilon$ or $\tau$ alongside the learning progress, while the other two Q-learning approaches highly depends on the initial exploration parameter setting and their corresponding annealing speeds. The exploration parameters $\epsilon$ and $\tau$ and their corresponding annealing speeds are non-trivial for the overall learning progress, which normally are determined via empirical modification when the learning environment varies. Moreover, from subfigure (b) and (c), we can conclude that all the simulated RL approaches can output proper trajectories in both network environments using three different action selection strategies. Moreover, although Boltzmann strategy can offer faster convergence performance than $\epsilon$-greedy, it leads to sub-optimal trajectory, as shown in subfigure (a) and (b). The proposed QiRL solution can not only enhance convergence performance but also achieve the equivalently optimal trajectory compared with Q-learning with $\epsilon$-greedy action selection strategy, which indicates that the proposed QiRL solution can better deal with the trade-off between convergence speed and learning quality.

V. CONCLUSION

This paper introduced a QiRL solution to tackle the trajectory planning problem which aims to maximize the ESUTR performance for the UAV flying from the start location to the destination. Specifically, the proposed QiRL approach utilizes the novel collapse action selection strategy inspired by quantum mechanism, which can offer a natural way to balance exploration and exploitation via sorting probabilities of action collapse in a ranking sequence. Numerical results compared the convergence performance and the learned trajectories between the proposed QiRL solution and the widely-used Q-learning approach with $\epsilon$-greedy and Boltzmann exploration strategies, validated the effectiveness of the proposed QiRL solution and showed that the QiRL solution can better deal with the trade-off between convergence speed and learning quality than traditional Q-learning approach.

REFERENCES

[1] Y. Zeng, Q. Wu, and R. Zhang, “Accessing from the sky: A tutorial on UAV communications for 5G and beyond,” Proceedings of the IEEE, vol. 107, no. 12, pp. 2327–2375, 2019.
[2] J. Hu, H. Zhang, L. Song, Z. Han, and H. V. Poor, “Reinforcement learning for a cellular internet of UAVs: protocol design, trajectory control, and resource management,” IEEE Wireless Commun., vol. 27, no. 1, pp. 116–123, 2020.
[3] S. Yin, S. Zhao, Y. Zhao, and F. R. Yu, “Intelligent trajectory design in UAV-aided communications with reinforcement learning,” IEEE Trans. Veh. Technol., vol. 68, no. 8, pp. 8227–8231, 2019.
[4] U. Challita, W. Saad, and C. Bettstetter, “Interference management for cellular-connected UAVs: A deep reinforcement learning approach,” IEEE Trans. Wireless Commun., vol. 18, no. 4, pp. 2125–2140, 2019.
[5] D. Dong, C. Chen, H. Li, and T.-J. Tarn, “Quantum reinforcement learning,” IEEE Trans. Syst. Man. Cybern. B, vol. 38, no. 5, pp. 1207–1220, 2008.
[6] J.-A. Li, D. Dong, Z. Wei, Y. Liu, Y. Pan, F. Nori, and X. Zhang, “Quantum reinforcement learning during human decision-making,” Nature Human Behaviour, vol. 4, no. 3, pp. 294–307, 2020.
[7] D. Dong, C. Chen, J. Chu, and T.-J. Tarn, “Robust quantum-inspired reinforcement learning for robot navigation,” IEEE/ASME Trans. Mechatronics, vol. 17, no. 1, pp. 86–97, 2010.
[8] P. Fakharí, K. Rajagopal, S. Balakrishnan, and J. Busemeyer, “Quantum inspired reinforcement learning in changing environment,” New Mathematics and Natural Computation, vol. 9, no. 03, pp. 273–294, 2013.
[9] M. A. Nielsen and I. Chuang, Quantum computation and quantum information. Cambridge University Press, 2010.