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

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

***Abstract*—In this letter, 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 optimize 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, respectively. 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 traditional reinforcement learning approaches that 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 delightful advancements such as flexible mobility, on-demand deployment and cost effectiveness [1], [2]. Compared to terrestrial wireless communication scenarios, one of the most notable features of UAV-mounted wireless networks is the controllable adjustments of UAV's flying trajectory, which can offer favourable wireless channel quality [3]. 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 [4], [5].

Balancing exploration and exploitation remains the inherent challenge of RL-based intelligent systems, which poses significant impacts on learning efficiency and quality, e.g., ϵ-greedy and Boltzmann action selection strategies [6]–[8]. On one hand, ϵ-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 uniformly among all possible actions while exploring, which means that it cannot distinguish the next-to-optimal action from other possible counterparts. On the other hand, Boltzmann (or the Softmax) exploration method introduces an action selection probability $\exp(Q(s,a)/\tau)/(\sum_i \exp(Q(s,a^i)/\tau))$ based on the Q function $Q(s, a)$ of state s and action a, where the parameter τ represents the *temperature* in the Boltzmann distribution. However, finding a good τ which can properly balance exploration and exploitation is difficult. The parameters ϵ and τ 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, it is believed to be a promising direction to adopt quantum mechanism into the field of machine learning [9]. Dong *et al.* [6] proposed the concept of quantum reinforcement learning (QRL), in which QRL was applied to solve the typical grid-world problem. Thereafter, in [10], Dong *et al.* introduced quantum-inspired reinforcement learning (QiRL) into the field of navigation control of autonomous mobile robots. Fakhari *et al.* [11] applied QiRL approach into unknown probabilistic environment, in which the robustness of QiRL solution was demonstrated. Li *et al.* [8] compared QRL with several conventional RL (CRL) models[1] in human decision-making scenarios, suggesting that value-based decision-making can be illustrated by QRL at both the behavioral and neural levels. However, QiRL is now still in its infancy, and has not been yet introduced into the field of UAV-aided networks.

In this letter, 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.[2] Different from [6] and [10], we extend the quantum explanation of QiRL from fixed rotation angles to their flexible counterparts, which is an alternative of [8] and [11]. Besides, we also relax the limitation of linear function mapping in [8] and that of empirical rotation angle setting in [11]. We aim at providing the first exploration of emerging QiRL for UAV-aided wireless networks.

II. System Model

This work concentrates on the uplink transmission scenario consisting of one UAV[3] and K ground users (GUs),

Manuscript received May 12, 2021; accepted June 8, 2021. Date of publication June 16, 2021; date of current version September 9, 2021. This work was supported by the K-CSC scholarship funded jointly by China Scholarship Council and King's College London, under Grant CSC201908350102. The associate editor coordinating the review of this article and approving it for publication was Z. Chang. *(Corresponding author: Yuanjian Li.)*

Yuanjian Li and A. Hamid Aghvami are with the Centre for Telecommunications Research, King's College London, London WC2R 2LS, U.K. (e-mail: yuanjian.li@kcl.ac.uk; hamid.aghvami@kcl.ac.uk).

Daoyi Dong is with the School of Engineering and Information Technology, University of New South Wales, Canberra, ACT 2600, Australia (e-mail: d.dong@adfa.edu.au).

Digital Object Identifier 10.1109/LWC.2021.3089876

[1] The abbreviation "CRL" denotes the RL methods without involving neural networks, distinguishing itself from deep reinforcement learning (DRL).

[2] In QRL, it is expected to implement real quantum computation on practical quantum computers, while QiRL algorithm invokes several ideas from quantum theory and is still in the frame of CRL which can be directly conducted on traditional computers.

[3] Without loss of generality, we focus on the system model with one single UAV, while the proposed QiRL algorithm can be similarly applied to other UAVs. The multi-UAV scenario is of importance to be evaluated, which is out of the scope of this letter and left as one of future research directions.

in which the location of each ground user is denoted as $\vec{D}_k = (x_k, y_k, 0)$ where $k \in \{1, 2, \ldots, K\}$. It is assumed that all the GUs are uploading their messages in a frequency division multiplexing manner. Thus, each GU transmits sorely on its assigned channel and inner-channel interference can be approximately ignored. Besides, the UAV is assumed to fly with constant velocity V (m/s) and fixed altitude H (m).[4] A practical assumption on the availability of network information is applied, in which the UAV cannot obtain any environment knowledge, 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 sum uplink transmit rate (ESUTR) of the GUs via intelligently adjusting its flying trajectory from the start location $\vec{L}_0 = (x_0, y_0, H)$ to the destination $\vec{L}_F = (x_F, y_F, H)$. Assume that the feasible region where the UAV can explore is a rectangular area $[x_0, x_F] \times [y_0, y_F]$, denoted as Φ 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, where the length of each time slot is predefined as T (s). Furthermore, the 3-dimensional Cartesian coordinate at the beginning of each time slot can be given by $\mathcal{L} = \{\vec{L}_0, \vec{L}_1, \ldots, \vec{L}_F\}$, in which $\vec{L}_0 \preceq \vec{L}_f \preceq \vec{L}_F, \forall f \in [0, F]$, where $\preceq$ represents element-wise inequality.

The large-scale path loss model on the sub-6 GHz band is considered to characterize the channel gains for wireless links between the UAV and all GUs, which can be given by $PL_{fk}(\text{dB}) = 20\lg(d_{fk}) + 20\lg(\varpi) - 147.55$, where $d_{fk} = \|\vec{L}_f - \vec{D}_k\|$ denotes the Euclidean distance between the UAV at sampled location $\vec{L}_f$ and the GU k, and ϖ represents the carrier frequency. Note that we herein take line-of-sight (LoS)-dominated channel gain as an example to evaluate the proposed system model, which is suitable for suburban or rural scenario, i.e., the channel gain between the drone and GUs can be characterized by the distance-based fading channel model.[5]

The received signal-to-noise ratio (SNR) at the UAV from the GU k can be derived as $\Gamma_{fk} = P_k/(\sigma_k^2 10^{PL_{fk}/10})$, where P_k represents the uplink transmit power of the GU k and σ_k^2 denotes the power of additive white Gaussian noise.

III. Problem Formulation

In this letter, we focus on maximizing the ESUTR for the UAV travelling from the predefined start location to the destination, via finding its optimal trajectory. It is straightforward to conclude that, at each sampled UAV coordinate $\vec{L}_f$, the sum uplink transmission rate can be characterized by $\sum_{k=1}^{K} \omega_k \log(1 + \Gamma_{fk})$ where ω_k means the bandwidth occupied by the GU k. Furthermore, the problem of ESUTR maximization can be stated as

$$\max_{\mathcal{L}} \ \frac{1}{F} \sum_{f=1}^{F} \sum_{k=1}^{K} \omega_k \log(1 + \Gamma_{fk}), \tag{1a}$$

$$\text{s.t.} \ \|\vec{L}_f - \vec{L}_{f-1}\| = VT, \tag{1b}$$

$$\vec{L}_0 \preceq \vec{L}_f \preceq \vec{L}_F, \tag{1c}$$

$$FT \leq E, \tag{1d}$$

$$\sum_k \omega_k \leq B, \tag{1e}$$

where B indicates bandwidth capacity of the system 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*, 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 model-free RL algorithms in a "trial-and-error" manner, e.g., Q-learning. However, CRL with tuned exploration parameters (e.g., hyperparameters ϵ and τ) may suffer from difficulty of balancing exploration and exploitation, which can further affect its learning quality and convergence performance. To give a better alternative for solving problem (1), the QiRL technique will be invoked to tackle the proposed optimal trajectory planning problem.

IV. QiRL Solution

The above trajectory design problem can be interpreted as a sequential decision-making process following Markov property, which means that the UAV's movement decision for the current time slot can be sorely determined according to the information of the previous time slot, regardless those of time slots before the previous time slot. Therefore, Markov decision process (MDP) is a suitable candidate for solving the trajectory optimization problem, forging the optimal mapping (i.e., the optimal action selection policy) from the state space to the corresponding action selections.

A. The MDP Formulation

To formulate the MDP, we need to clarify the *states* of the proposed QiRL solution for the considered scenario. The feasible area Φ is divided into N_1 by 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.[6] 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 Φ, the state set of the UAV can be written as $\mathcal{S} = \{s_1, s_2, \ldots, s_{N_1 N_2}\}$, where $s_i \in \mathcal{S}$ represents a small square in Φ. Because we focus on the ESUTR maximization problem, it is straightforward to define $R(s_i) = \sum_{k=1}^{K} \omega_k \log(1 + \Gamma_{s_i k})$ as the *reward* function for state s_i (also denoting $R(s_i)$ as R for simplicity), where L_{s_i} in $\Gamma_{s_i k}$ denotes the location of s_i. In the case of reaching the boundary of Ψ, the UAV will be rebounded back and the reward for this trial is set to zero.[7] Note that the UAV is only

[4] The UAV's altitude H is assumed as a fixed parameter, which may correspond to the lowest altitude required for terrain or building avoidance, under the regulation of local laws in practice.

[5] This work focuses on strong LoS path loss channel model and the effects of small-scale fading (e.g., Rician fading or Nakagami-m fading) is omitted. Besides, non-line-of-sight (NLoS) channel gain can also be easily integrated into the proposed model via involving extra NLoS fading component, which means the proposed algorithm is still applicable for NLoS case and this case is omitted for conciseness.

[6] This assumption is reasonable because the acreage of each grid is far less than that of Φ, in the case of sufficient discretization.

[7] Hereby, we take zero reward for crashing into the boundary as an example. Of course, one can let this kind of scenario be punished by minus reward.

able to observe *R* while other network information is inaccessible, i.e., P_k, ω_k, σ_k^2 and $\vec{D}_k$. 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 $\vec{L}_0$ to $\vec{L}_F$. To drive the UAV to the destination $\vec{L}_F$, the UAV will gain a special reward which is defined as $\hat{R} = 10 \times \max_{s_i \in \mathcal{S}} R(s_i)$, once it reaches $\vec{L}_F$. 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 $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ so that the expected sum of discounted rewards for each episode can be maximized. We define the value function of state *s* at trial *t* as $V_\pi(s) = \mathbb{E}_\pi[\sum_{l=0}^{F} \gamma^l R(t + l + 1)|\mathcal{S}(t) = s]$, where γ represents the discount factor. Furthermore, the temporal difference (TD)-based value updating rule [10] of the proposed QiRL can be described as $V(s) \leftarrow V(s) + \alpha[R(s') + \gamma V(s') - V(s)]$, where s' means the next state after taking an action and α indicates the learning rate.

B. Collapsing Action Selection

According to quantum mechanics [12], a quantum state $|\Psi\rangle$ (Dirac representation) can describe the state of a closed quantum system, which is a unit vector (i.e., $\langle\Psi|\Psi\rangle = 1$) in Hilbert space. The quantum state $|\Psi\rangle$ consisting of *n* quantum bits (qubits) can be expanded as $|\Psi\rangle = |\psi_1\rangle \otimes |\psi_2\rangle \otimes \cdots \otimes |\psi_n\rangle = \sum_{p=00\cdots0}^{11\cdots1} h_p|p\rangle$, where $|\psi_i\rangle, i \in [1, n]$ represents the i-th qubit which is a two-state quantum system and the basic unit of quantum information, the complex coefficient h_p (subject to $\sum_{p=00\cdots0}^{11\cdots1} |h_p|^2 = 1$) denotes the probability amplitude for eigenstate $|p\rangle$ of $|\Psi\rangle$ and $\otimes$ represents the tensor product. The representation of *n*-qubit quantum state $|\Psi\rangle$ follows the quantum phenomenon called *state superposition principle*. Note that h_p can take 2^n complex values so that the *n*-qubit quantum state $|\Psi\rangle$ can be regarded as the superposition of 2^n eigenstates, in the range from $|00\cdots0\rangle$ to $|11\cdots1\rangle$.

To represent the four possible actions in QiRL, two qubits are sufficient. Furthermore, eigenactions (i.e., the quantum representation of physical 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 represent the four egienactions in their quantum superposition form, given by $|A(l)\rangle = |\psi_1\rangle \otimes |\psi_2\rangle = \sum_{a=00}^{11} h_a|a\rangle \to \sum_{n=1}^{4} h_n|a_n\rangle$, where l represents a specific trial and the complex coefficients h_n and h_a are the 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 two-qubit superposition $|A(l)\rangle$ is a unit vector in a 4-dimensional Hilbert space spanned by the four orthogonal bases $|00\rangle$, $|01\rangle$, $|10\rangle$ and $|11\rangle$. Specifically, the action taken by the UAV before any quantum measurement lies in a superposition state (four options in total, i.e., $|a_1\rangle$, $|a_2\rangle$, $|a_3\rangle$ and $|a_4\rangle$), which is mapped into the tensor product of two qubits.

In quantum theory, when an external agency (e.g., experimenter) measures the quantum state $|\Psi\rangle = \sum_n \varrho_n|\psi_n\rangle$ with the eigenbasis $\{\psi_n\}$, $|\Psi\rangle$ will collapse from the superposition state to one of its eigenstates $|\psi_n\rangle$, i.e., $|\Psi\rangle \to |\psi_n\rangle$, with probability $|\langle\psi_n||\Psi\rangle|^2 = |\varrho_n|^2$. Inspired by this *quantum collapse phenomenon*, the superposition $|A(l)\rangle$ is supposed to collapse onto one of its eigenactions $|a_n\rangle$ with probability of $|h_n|^2$, during action picking in the proposed QiRL algorithm.

C. Grover Iteration

The quantum representation $|A(l)\rangle$ establishes a bridge between quantum eigenactions and the physical action set $\mathcal{A}$, which allows us to apply quantum amplitude amplification as a reinforcement strategy. The probability amplitude of each eigenaction can be amplified or attenuated via specific quantum algorithm (e.g., Grover's iteration [12]), gradually modifying the probability distribution of collapsing. To realize this, two unitary operators can be employed for the currently chosen action $|a_i\rangle$ which is from the *l*-th trial $|A(l)\rangle = \sum_{n=1}^{4} h_n|a_n\rangle = h_i|a_i\rangle + h_{a_i^\perp}|a_i^\perp\rangle$, shown as $\boldsymbol{U}_{|a_i\rangle} = \boldsymbol{I} - (1 - e^{j\phi_1})|a_i\rangle\langle a_i|$ and $\boldsymbol{U}_{|A(l)\rangle} = (1 - e^{j\phi_2})|A(l)\rangle\langle A(l)| - \boldsymbol{I}$, where $|a_i^\perp\rangle = \sum_{n \neq i} \frac{h_n}{h_{a_i^\perp}}|a_n\rangle$ means the vector orthogonal to $|a_i\rangle$, $h_{a_i^\perp} = \sqrt{\sum_{n \neq i} |h_n|^2} = \sqrt{1 - |h_i|^2}$, $\boldsymbol{I}$ represents the identity matrix, and $\langle a_n|$ and $\langle A(l)|$ are Hermitian transposes of $|a_n\rangle$ and $|A(l)\rangle$, respectively. Then, the Grover operator can be constructed as unitary transformation $\boldsymbol{G} = \boldsymbol{U}_{|A(l)\rangle}\boldsymbol{U}_{|a_i\rangle}$. After *m* times of applying $\boldsymbol{G}$ on $|A(l)\rangle$, the amplitude vector in the next trial becomes $|A(l + 1)\rangle = \boldsymbol{G}^m|A(l)\rangle$.

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 ϕ_1 and ϕ_2 (commonly both of them equal to π); the other is to fix $m = 1$ with dynamic parameters ϕ_1 and ϕ_2. Because the former updating approach can only modify the amplitudes in a discrete manner, the later method is chosen in this work, i.e., Grover iteration with flexible parameters ϕ_1 and ϕ_2. Then, the impacts of $\boldsymbol{G}$ on the superposition representation $|A(l)\rangle$ can be given by the following proposition.

Proposition 1: The overall effects of $\boldsymbol{G}$ with free parameters ϕ_1 and ϕ_2 on the superposition representation $|A(l)\rangle$ at the *l*-th trial can be expressed analytically as $\boldsymbol{G}|A(l)\rangle = (\mathcal{Q} - e^{j\phi_1})h_i|a_i\rangle + (\mathcal{Q} - 1)h_{a_i^\perp}|a_i^\perp\rangle$, where $\mathcal{Q} = (1 - e^{j\phi_2})[1 - (1 - e^{j\phi_1})|h_i|^2]$.

Proof: The impacts of $\boldsymbol{U}_{|a_i\rangle}$ on $|a_i\rangle$ and $|a_i^\perp\rangle$ can be given by $\boldsymbol{U}_{|a_i\rangle}|a_i\rangle = [\boldsymbol{I} - (1 - e^{j\phi_1})|a_i\rangle\langle a_i|]|a_i\rangle = e^{j\phi_1}|a_i\rangle$ and $\boldsymbol{U}_{|a_i\rangle}|a_i^\perp\rangle = [\boldsymbol{I} - (1 - e^{j\phi_1})|a_i\rangle\langle a_i|]|a_i^\perp\rangle = |a_i^\perp\rangle$, respectively. Furthermore, we have $\boldsymbol{U}_{|a_i\rangle}|A(l)\rangle = [\boldsymbol{I} - (1 - e^{j\phi_1})|a_i\rangle\langle a_i|]|A(l)\rangle = e^{j\phi_1}h_i|a_i\rangle + h_{a_i^\perp}|a_i^\perp\rangle$, in which $\boldsymbol{U}_{|a_i\rangle}$ plays the role as a conditional phase shift operator in quantum computation. At the end, we can obtain $\boldsymbol{G}|A(l)\rangle = \boldsymbol{U}_{|A(l)\rangle}\boldsymbol{U}_{|a_i\rangle}|A(l)\rangle = (1 - e^{j\phi_2})[h_i|a_i\rangle + h_{a_i^\perp}|a_i^\perp\rangle][h_i^\dagger\langle a_i| + h_{a_i^\perp}^\dagger\langle a_i^\perp|]\boldsymbol{U}_{|a_i\rangle}|A(l)\rangle - \boldsymbol{U}_{|a_i\rangle}|A(l)\rangle = (\mathcal{Q} - e^{j\phi_1})h_i|a_i\rangle + (\mathcal{Q} - 1)h_{a_i^\perp}|a_i^\perp\rangle$, where $\mathcal{Q} = (1 - e^{j\phi_2})[1 - (1 - e^{j\phi_1})|h_i|^2]$. ■

Remark 1: The ratio between the probability amplitudes of $|a_i\rangle$ after being acted by the Grover operator $\boldsymbol{G}$ and before that can be expressed as $\Lambda = (1 - e^{j\phi_1} - e^{j\phi_2}) - (1 - e^{j\phi_1})(1 - e^{j\phi_2})|h_i|^2$. Then, the updated occurrence probability of the selected action $|a_i\rangle$ can be given by $|\Lambda|^2|h_i|^2$.

Remark 2: For ease of understanding the effect of $\boldsymbol{G}$, we depict its algebraic visualization. In Fig. 1, $|A(l)\rangle$ is reconstructed via polar coordinates on the Bloch sphere, shown as $|A(l)\rangle = e^{j\zeta}(\cos\frac{\theta}{2}|a_i\rangle + e^{j\varphi}\sin\frac{\theta}{2}|a_i^\perp\rangle) \simeq \cos\frac{\theta}{2}|a_i\rangle +$

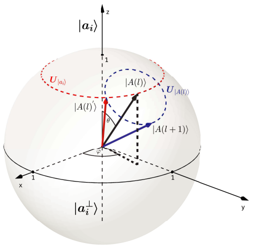


Fig. 1. Geometric explanation of the Grover rotation.

$e^{j\varphi}\sin\frac{\theta}{2}|a_i^{\perp}\rangle$, where $e^{j\zeta}$ can be omitted because a global phase poses no observable effects [8]. Note that the polar angle parameter θ and the azimuthal angle variable φ define the unit vector $|A(l)\rangle$ on the Bloch sphere, as shown in Fig. 1. The impact of $\boldsymbol{U}_{|a_i\rangle}$ can be understood as a clockwise rotation around the z-axis by ϕ_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 $\{|a_i\rangle, |a_i^{\perp}\rangle\}$ to $\{|A(l)\rangle, |A(l)^{\perp}\rangle\}$, $\boldsymbol{U}_{|A(l)\rangle}$ makes a clockwise rotation around the new z-axis $|A(l)\rangle$ by ϕ_2 (the blue circle), which rotates $|A(l)'\rangle$ to $|A(l+1)\rangle$. Therefore, the overall effect of $\boldsymbol{G}$ on $|A(l)\rangle$ is a two-step rotation which can modify the polar angle θ, when the basis is locked as $\{|a_i\rangle, |a_i^{\perp}\rangle\}$. Via controlling parameters ϕ_1 and ϕ_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 θ 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 θ via manipulating ϕ_1 and ϕ_2 when $|a_i\rangle$ is recognized as a "good" action. Otherwise, if $|a_i\rangle$ is determined as a "bad" action, ϕ_1 and ϕ_2 should be modified to enlarge θ.

D. The Proposed QiRL Algorithm

Remark 1 and Remark 2 give an explanation for amplitude amplification in quantum mechanism, which can be applied as the quantum-inspired reinforcement strategy for our proposed QiRL approach. According to Remark 1, it is straightforward to conclude that $|\Lambda|^2$ should be designed to be larger than 1, if the current representation $|a_i\rangle$ is determined as a "good" action. Otherwise, $|\Lambda|^2$ should be manipulated to be smaller than 1. By selecting feasible ϕ_1 and ϕ_2, it is possible to manipulate the value of $|\Lambda|^2$ in the manner as mentioned before, which can be interpreted geometrically via Remark 2. For the sake to simulate it on a conventional computer, we use $e^{k*[R+V(s')]}$ to alternatively represent the overall effects of $\boldsymbol{G}$ on probability $|h_i|^2$, which means the updated occurrence probability of the selected action $|a_i\rangle$ should be $e^{k*[R+V(s')]}|h_i|^2$. If $k > 0$, the current action will be rewarded while it will be punished if $k < 0$. The updating amplification is controlled via $k * [R + V(s')]$.[8]

Note that all the possible probability amplitudes together should be re-normalized after each implementation of amplitude amplification, which is subject to the normalization constraint of quantum superposition. The proposed QiRL solution is concluded in Algorithm 1, which can be conducted in conventional computers.

Algorithm 1: The Proposed QiRL Algorithm

```
Input: Learning parameters: $\alpha \in [0,1]$, $\gamma = 1$; UAV
       informations: $\vec{L}_0$, $\vec{L}_F$, $H$, $V$, $T$, $E$;
Output: The optimal policy $\pi^*$=AmpMem;
Initialization: $ep = 0$; s = $\vec{L}_0$; $V(s) = 0, \forall s \in \mathcal{S}$;
  AmpMem = defaultdict($lambda$: [$\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{4}$]);
while $ep \leq NumEp$ do
    repeat
        Pick $a$ for $s$ via measuring AmpMem[s];
        Apply $a$ and observe reward $R$ and next state $s'$;
        Update the value function as per
        $V(s) \leftarrow V(s) + \alpha\big[R + \gamma V\big(s'\big) - V(s)\big]$;
        Apply quantum-inspired reinforcement factor
         $e^{k*[R+V(s')]}$ on AmpMem[s][a]. When the
         UAV hits the boundary or value difference
         $\Delta V(s) < 0$, $k < 0$. Otherwise, $k > 0$;
        Re-normalize AmpMem[s] and set $s \leftarrow s'$;
    until $F > E/T$ or $s' == \vec{L}_F$;
    $ep$ += 1;
end
```

Remark 3: The quantum-inspired reinforcement strategy prioritizes all possible actions in ranked probability sequence which is gradually updated alongside the learning process. Thus, it can naturally balance the exploration and exploitation, in which no tuned exploration parameter is necessary. This enhancement has the potential to help realize faster convergence and satisfactory learning quality, which will be later illustrated in the simulation results.

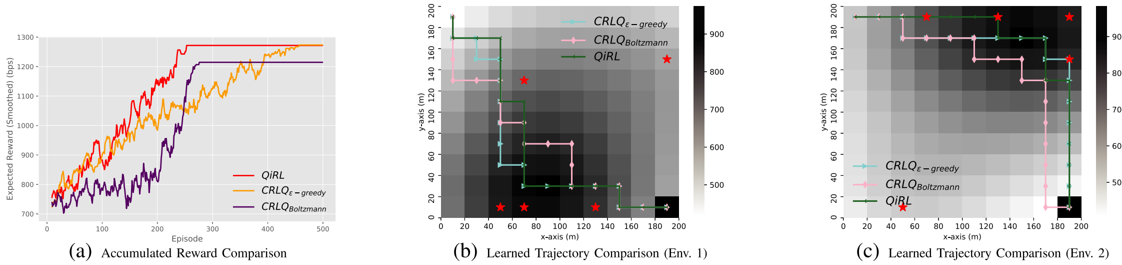
Proposition 2: The convergence of the proposed QiRL algorithm is guaranteed when the learning rate α is non-negative and satisfies $\lim_{T\to\infty}\sum_{k=1}^{T}\alpha_k = \infty$ and $\lim_{T\to\infty}\sum_{k=1}^{T}\alpha_k^2 < \infty$.

Proof: The proof is omitted for its simplicity, which is similar to the proof of Proposition 2 in [6]. ■

V. Simulation Results

In this section, experimental results are evaluated for the considered UAV trajectory planning problem via the proposed QiRL solution. For comparison, two CRL methods (i.e., Q-learning with ϵ-greedy and Boltzmann exploration strategies) are performed as benchmarks. It is assumed that the feasible UAV exploration field Φ is a square area with side length 200 m, where 5 GUs are located on the ground (denoted by the red stars). By default, the length of each time slot is fixed as $T = 2$ s and the constant flying altitude and speed of the UAV are set as $H = 100$ m and $V = 10$ m/s, respectively. The area Φ is divided into 10-by-10 small grids and the side length of each grid equals $VT = 20$ m. The start location and the destination are predefined at $\vec{L}_0 = (10, 190, 100)$ and $\vec{L}_F = (190, 10, 100)$, respectively. Considering the on-board power capacity of the UAV, the total flying time of the UAV is constrained as $FT \leq 1800$ s so that we set $E = 1800$. Besides, we set $P_k = 1$ Watt, $\sigma_k^2 = 1$, $\varpi = 2$ GHz, $B = 10$ MHz and $\omega_k = 2$ MHz, which is in line with [4].

[8] The absolute value of constant hyper-parameter k should be chosen as per the environment, to avoid over-updating issue on occurrence probability of the selected action. Then, the updating amplification is dynamically steered by $R + V(s')$ with constant k, because the state values are being modified alongside the learning process.



(a) Accumulated Reward Comparison (b) Learned Trajectory Comparison (Env. 1) (c) Learned Trajectory Comparison (Env. 2)

Fig. 2. Performance Comparison of Two Q-Learning Approaches and the Proposed QiRL Solution.

Fig. 2 shows the performance comparison of one widely-used CRL approach called Q-learning with two action selection strategies, i.e., ϵ-greedy and Boltzmann, and the proposed QiRL solution. Note that exploration parameters ϵ and τ of Q-learning approach keep annealing alongside the learning progress, which controls the ratio of exploration and exploitation and significantly affects the overall learning quality and convergence performance. 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 corresponds to subfigure (b).

From subfigure (a), it is straightforward to observe that the proposed QiRL solution can converge much faster than Q-learning with ϵ-greedy action selection strategy, while it has relatively faster convergence speed than Q-learning with advanced Boltzmann action selection strategy, which illustrates that the proposed QiRL algorithm can offer better convergence performance. Moreover, from subfigures (b) and (c), we can observe that all the simulated RL approaches can output proper trajectories in these two different network environments. However, while Boltzmann strategy can offer faster convergence performance than ϵ-greedy, it leads to sub-optimal trajectory, as shown in subfigures (a) and (b). According to Fig. 2, the proposed QiRL solution can not only enhance convergence performance but also achieve the equivalently optimal trajectory compared to Q-learning with ϵ-greedy action selection strategy. Note that the balancing between exploration and exploitation in ϵ-greedy or Boltzmann aided Q-learning approach is controlled by the pickings of initial exploration parameter (i.e., ϵ or τ, respectively) and their corresponding annealing speeds, which directly and inherently influences convergence performance and learning quality. Generally speaking, the initial exploration parameters and their corresponding annealing speeds are modified via empirical knowledge when the learning environment varies. However, simply decaying exploration parameter (linearly or non-linearly) alongside the learning progress could easily lead to insufficient learning or low speed of convergence. To deal with this unsatisfactoriness, the proposed QiRL algorithm applies quantum-inspired action selection approach, offering natural balancing between exploration and exploitation alongside the learning progress and thus can better deal with the trade-off between convergence speed and learning quality.

VI. Conclusion

This letter introduced a QiRL solution to tackle the trajectory planning problem which aims to optimize 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 ϵ-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 approaches.

References

[1] Y. Zeng, Q. Wu, and R. Zhang, "Accessing from the sky: A tutorial on UAV communications for 5G and beyond," *Proc. IEEE*, vol. 107, no. 12, pp. 2327–2375, Dec. 2019.

[2] J. Wang, C. Jiang, Z. Han, Y. Ren, R. G. Maunder, and L. Hanzo, "Taking drones to the next level: Cooperative distributed unmanned-aerial-vehicular networks for small and mini drones," *IEEE Veh. Technol. Mag.*, vol. 12, no. 3, pp. 73–82, Sep. 2017.

[3] 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, Feb. 2020.

[4] 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, Aug. 2019.

[5] Y. Zeng and X. Xu, "Path design for cellular-connected UAV with reinforcement learning," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2019, pp. 1–6.

[6] D. Dong, C. Chen, H. Li, and T.-J. Tarn, "Quantum reinforcement learning," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 5, pp. 1207–1220, Oct. 2008.

[7] J. Wang, C. Jiang, H. Zhang, Y. Ren, K.-C. Chen, and L. Hanzo, "Thirty years of machine learning: The road to pareto-optimal wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1472–1514, 3rd Quart., 2020.

[8] J.-A. Li, D. Dong, Z. Wei, Y. Liu, Y. Pan, F. Nori, and X. Zhang, "Quantum reinforcement learning during human decision-making," *Nat. Human Behav.*, vol. 4, no. 3, pp. 294–307, 2020.

[9] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, "Quantum machine learning," *Nature*, vol. 549, no. 7671, pp. 195–202, 2017.

[10] 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, Feb. 2012.

[11] P. Fakhari, K. Rajagopal, S. Balakrishnan, and J. Busemeyer, "Quantum inspired reinforcement learning in changing environment," *New Math. Nat. Comput.*, vol. 9, no. 3, pp. 273–294, 2013.

[12] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*. Cambridge, U.K.: Cambridge Univ. Press, 2010.