Intelligent UAV Navigation: A DRL-QiER Solution

Yuanjian Li, *Student Member, IEEE* and A. Hamid Aghvami, *Fellow, IEEE*
Centre for Telecommunications Research (CTR), King's College London, London WC2R 2LS, U.K.
Email: {yuanjian.li, hamid.aghvami}@kcl.ac.uk

ICC 2022 - IEEE International Conference on Communications | 978-1-5386-8347-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICC45855.2022.9838566

***Abstract*—In cellular-connected unmanned aerial vehicle (UAV) network, a minimization problem on the weighted sum of time cost and expected outage duration is considered. Taking advantage of UAV's adjustable mobility, an intelligent UAV navigation approach is formulated to achieve the aforementioned optimization goal. Specifically, after mapping the navigation task into a Markov decision process (MDP), a deep reinforcement learning (DRL) solution with novel quantum-inspired experience replay (QiER) framework is proposed to help the UAV find the optimal flying direction within each time slot. Via relating experienced transition's importance to its associated quantum bit (qubit) and applying Grover-iteration-based amplitude amplification technique, the proposed DRL-QiER solution commits a better trade-off between sampling priority and diversity. Compared to several representative baselines, the effectiveness and supremacy of the proposed DRL-QiER solution are demonstrated and validated in numerical results.**

***Index Terms*—Drone, trajectory design, deep reinforcement learning, quantum-inspired experience replay.**

I. Introduction

With flexible mobility, low cost and on-demand deployment, unmanned aerial vehicles (UAVs) have been widely used in civilian scenarios, e.g., building safety inspections, disaster management and material transport. In practice, point-to-point (P2P) wireless links over unlicensed spectrum are commonly utilized to support the communications between UAVs and ground nodes, leading to limited communication quality [1]. To further enhance wireless transmission between UAVs and ground transceivers, cellular-connected UAV technique is deemed as a promising solution, via adopting widely-existing terrestrial base stations (BSs) to help establish high-quality ground-to-air (G2A) transmission links [2].

Current cellular networks are genuinely established for serving user equipments on the ground, via downtilting the main lobe of BS's antenna towards the earth [2]. More severe inter-cell interferences (ICIs) introduced by line-of-sight (LoS)-dominated G2A links can further deteriorate the aerial coverage issue, compared to terrestrial communication scenario where non line-of-sight (NLoS) channels are most likely experienced. The controllable mobility feature of UAV makes it possible to tackle the aforementioned aerial coverage obstacles via UAV trajectory planning, either by on-board algorithms or remote pilots. However, standard off-line optimization approaches solving trajectory design problem suffer from inefficiency due to non-convex nature of the formulated optimization objective and the corresponding constraints, even under impractical assumptions where perfect knowledge of wireless environment is available. Fortunately, reinforcement learning (RL) serves as a good complement to traditional off-line optimization solutions, which is famous for learning unknown environment. Cui *et al.* [3] investigated a real-time design on resource allocation for multiple-UAV network, in which multi-agent reinforcement learning (MARL) framework was proposed to realize optimal user selection, power allocation and sub-channel association. Zeng *et al.* [2] solved an optimal UAV trajectory planning problem on minimizing the weighted sum of mission completion time and expected transmission outage duration, via deep reinforcement learning (DRL)-aided approaches.

Meanwhile, quantum theory has been proven to pose a positive impact on improving learning efficiency for artificial intelligence algorithms in general, and RL-related approaches in particular. Dong *et al.* [4] proposed quantum-inspired reinforcement learning (QiRL) to solve intelligent navigation problem for autonomous mobile robots, where probabilistic action selection method and novel reinforcement approach inspired by quantum phenomenon were integrated into standard RL frameworks. Paparo *et al.* [5] showed that quadratic speed-up is achievable for intelligent agents, with the help of quantum mechanics. In the field of wireless communications, Li *et al.* [6] investigated an optimal path planning problem for UAV-mounted networks, in which QiRL solution was demonstrated to offer better learning performance than conventional RL methods with ϵ-greedy or Boltzmann action selection policy.[1]

In this paper, we integrate several ideas in quantum mechanics and DRL techniques to solve intelligent trajectory planning problem for cellular-connected UAV networks. The main contributions of this paper are summarized as follows.

- A cellular-connected UAV trajectory planning problem is formulated to minimize the weighted sum of flight time cost and the corresponding expected outage duration. Without knowledge of wireless environment, the path planning problem is challenging to be tackled via conventional optimization techniques. Alternatively, it is solved by the proposed DRL solution with quantum-inspired experience replay (QiER).
- A novel QiER framework is coined to help the learning agent achieve better training performance, via a three-phase quantum-inspired process.
- Compared to DRL approach with standard experience relay (DRL-ER) or prioritized ER (DRL-PER), deep curriculum reinforcement learning (DCRL) method and

[1]Compared to our prior work [6], we extend the quantum aid from enhancing action selection quality for RL framework to improving experience replay performance for DRL counterpart, breaking the curse of dimensionality and enabling the agent to practically solve problems with continuous state space.

simultaneous navigation and radio mapping (SNARM) strategy, simulation results demonstrate that the proposed DRL-QiER solution can achieve more efficient and steady learning performance.

II. System Model

A downlink transmission scenario inside cellular-connected UAV network is considered, where a set $\mathcal{U}$ of U UAVs is served by a set $\mathcal{B}$ of B BSs within cellular coverage. These UAVs are supposed to reach a common destination from their respective initial locations, for accomplishing their own missions.[2] Without loss of generality, an arbitrary UAV (denoted as u hereafter) out of these U drones are concentrated for investigating the navigation task.[3] For clarity, the UAV's exploration environment is defined as a cubic subregion $\mathbb{A} : [x_{\text{lo}}, x_{\text{up}}] \times [y_{\text{lo}}, y_{\text{up}}] \times [z_{\text{lo}}, z_{\text{up}}]$, where the subscripts "lo" and "up" represent the lower and upper boundaries of this 3D airspace, respectively. Furthermore, the coordinate of the focused UAV at time t should locate in the range of $\vec{q}_{\text{lo}} \preceq \vec{q}_u(t) \preceq \vec{q}_{\text{up}}$, where $\vec{q}_{\text{lo}} = (x_{\text{lo}}, y_{\text{lo}}, z_{\text{lo}})$, $\vec{q}_{\text{up}} = (x_{\text{up}}, y_{\text{up}}, z_{\text{up}})$ and $\preceq$ denotes the element-wise inequality. The initial location and the destination are given by $\vec{q}_u(I) \in \mathbb{R}^{1*3}$ and $\vec{q}_u(D) \in \mathbb{R}^{1*3}$, respectively. Then, the overall trajectory of this UAV's flight can be fully traced by $\vec{q}_u(t) = (x_u(t), y_u(t), z_u(t))$, starting from $\vec{q}_u(I)$ and ending at $\vec{q}_u(D)$.

Following standard sectorization, each BS is portioned to cover three sectors. Therefore, there are $3B$ sectors in total within the interested airspace $\mathbb{A}$. Denote $i \in \{1, \ldots, 3B\}$ as the label of sectors and assume that the UAV is associated with sector $\hat{i}$ at time t, the signal-to-interference-plus-noise ratio (SINR) at the UAV can be derived as

$$\Gamma_u(t) = \frac{P_{\hat{i}} 10^{\frac{G^{\hat{i}}[\vec{q}_u(t)] - \text{PL}^{\hat{i}}[\vec{q}_u(t)]}{10}} |h_{\hat{i}u}|^2}{I_u(t) + \sigma^2}, \tag{1}$$

where $I_u(t) = \sum_{i \neq \hat{i}} P_i 10^{\frac{G^i[\vec{q}_u(t)] - \text{PL}^i[\vec{q}_u(t)]}{10}} |h_{iu}|^2$ means the ICIs from un-associated sectors, $G^i[\vec{q}_u(t)]$ indicates antenna gain, $\text{PL}^i[\vec{q}_u(t)]$ means G2A pathloss, P_i is the average transmit power of sector i, h_{iu} represents the corresponding small-scale fading channel and σ^2 denotes the variance of additive complex Gaussian noise (AWGN). The received SINR (1) is a random variable because of the randomness introduced by small-scale fadings, with given UAV coordinate $\vec{q}_u(t)$ and cell association $\hat{i}(t)$. Therefore, the corresponding transmission outage probability (TOP) can be formulated as a function of $\vec{q}_u(t)$ and $\hat{i}(t)$, i.e., $TOP_u\{\vec{q}_u(t), \hat{i}(t)\} = \Pr[\Gamma_u(t) < \Gamma_{th}]$, where $\Pr$ outputs the probability calculated with respect to (w.r.t.) the aforementioned small-scale fadings. Then, the ergodic outage duration (EOD) of the UAV u travelling with trajectory $\vec{q}_u(t), \forall t \in [0, T_u]$ from $\vec{q}_u(I)$ to $\vec{q}_u(D)$ can be expressed as

$$EOD_u\{\vec{q}_u(t), \hat{i}(t)\} = \int_0^{T_u} TOP_u\{\vec{q}_u(t), \hat{i}(t)\} dt. \tag{2}$$

In this paper, we focus on minimizing the weighted sum of T_u and $EOD_u\{\vec{q}_u(t), \hat{i}(t)\}$ via designing $\vec{q}_u(t)$ and $\hat{i}(t)$. Unfortunately, continuous time t implies infinite amount of velocity constraints and location possibilities, leading the UAV path planning task too sophisticated to be handled. Alternatively, the flight period T_u is uniformly divided into N time slots. The duration of each time slot $\Delta_t = T_u/N$ is controlled to be sufficiently small so that the distance, pathloss and antenna gain from each sector towards the UAV can be considered as approximately static within arbitrary time slot. Besides, sector assignment is commonly dependent on pathloss to avoid non-stop handover in practice, and thus the associated sector within each time slot is assumed unchanged. Therefore, (2) can be approximated as $EOD_u\{\vec{q}_u(t), \hat{i}(t)\} \approx \sum_{n=1}^{N} \Delta_t TOP_u\{\vec{q}_u(n), \hat{i}(n)\}$. With given $\vec{q}_u(n)$ and $\hat{i}(n)$ for each time slot, $TOP_u\{\vec{q}_u(n), \hat{i}(n)\}$ can be obtained via numerical signal measurement at the UAV. Then, we have

$$TOP_u\{\vec{q}_u(n), \hat{i}(n)\} \simeq \frac{1}{L} \sum_{\iota=1}^{L} ITOP\{\vec{q}_u(n), \hat{i}(n) | h(\iota)\}, \tag{3}$$

where $h(\iota)$ indicates one realization of the involved small-scale fading components, L represents the amount of signal measurements, the TOP indicator $ITOP\{\vec{q}_u(n), \hat{i}(n) | h(\iota)\} = 1$ if $\Gamma_u\{\vec{q}_u(n), \hat{i}(n) | h(\iota)\} < \Gamma_{th}$ and $ITOP\{\vec{q}_u(n), \hat{i}(n) | h(\iota)\} = 0$ otherwise. Note that $L \gg 1$ stands in practice, which means that the approximation (3) is feasible to be treated as an equation. Then, the corresponding optimization problem can be stated as

$$\text{(P1)}: \min_{\vec{v}_u(n)} \frac{\tau \Delta_t}{L} \sum_{n=1}^{N} \sum_{\iota=1}^{L} ITOP\{\vec{q}_u(n), \hat{i}(n) | h(\iota)\} + N, \tag{4a}$$

$$\text{s.t.} \hat{i}(n) = \underset{i \in \{1,2,\cdots,3B\}}{\arg\min} PL^i[\vec{q}_u(n)], \tag{4b}$$

$$\vec{q}(n+1) = \vec{q}(n) + V_u \Delta_t \vec{v}_u(n), \|\vec{v}_u(n)\| = 1, \tag{4c}$$

$$\vec{q}_{\text{lo}} \preceq \vec{q}_u(n) \preceq \vec{q}_{\text{up}}, \vec{q}_u(0) = \vec{q}_u(I), \vec{q}_u(N) = \vec{q}_u(D), \tag{4d}$$

where τ is the weight balancing the minimization objective, V_u represents the UAV's flying velocity and $\vec{v}_u(n)$ specifies the mobility direction. The constraint (4b) holds because the sector association strategy is dependent sorely on pathlosses from all the sectors within each time slot and it is clear that the UAV should always pair with the sector which can offer the least degree of pathloss.

It is straightforward to conclude that antenna gain and LoS/NLoS condition from each sector to the UAV are dependent on the UAV's location with given building and BS distribution, which further impacts the corresponding pathloss and type of small-scale fading. This makes it extremely sophisticated to solve problem (P1) via standard optimization methods, if not impossible. To provide a better alternative solving the proposed optimization problem (P1), a DRL-aided solution with a novel QiER framework will be proposed.

[2]For example, one typical UAV application case is parcel collection. Various UAVs are launched from different costumers' properties carrying parcels to the local distribution centre of delivery firm.

[3]These UAVs share the same airspace and common location-dependent database, which means that the trained DRL model can be downloaded by the remaining UAVs, helping them accomplish their navigation tasks.

III. DRL-QiER Algorithm

In this section, a DRL-QiER solution[4] is developed to solve optimization problem (P1).

A. The MDP Formulation

To solve the optimal trajectory planning problem (P1) via DRL-aided technique, the first step is to map it into a Markov decision process (MDP), which can be described as follows.

- $\mathcal{S}$: The state space consists of possible UAV locations $\vec{q}_u$ under constraint $\vec{q}_{\text{lo}} \preceq \vec{q}_u \preceq \vec{q}_{\text{up}}$, which means that the state space is continuous.
- $\mathcal{A}$: The continuous action space involves all the feasible flying directions $\vec{v}_u$ under constraint $\|\vec{v}_u\| = 1$. To break the curse of dimensionality caused by continuous state and action spaces, the action space is discretized as $\mathcal{A} = \{[1,0,0],[0,1,0],[-1,0,0],[0,-1,0],[-\sqrt{2}/2,\sqrt{2}/2,0],[\sqrt{2}/2,\sqrt{2}/2,0],[\sqrt{2}/2,-\sqrt{2}/2,0],[-\sqrt{2}/2,-\sqrt{2}/2,0]\}$, corresponding to flying directions right, forward, left, backward, left-forward, right-forward, right-backward and left-backward, respectively.
- $\mathcal{T}$: The state transition is deterministic and controlled by the mobility constraint (4c).
- r: Our goal is to minimize the weighted sum of time cost and EOD. Thus, we may design the reward function as $r(\vec{q}_u) = -1 - \frac{\tau\Delta_t}{L}\sum_{\iota=1}^{L} ITOP\{\vec{q}_u|h(\iota)\}$. The formulation of $r(\vec{q}_u)$ can be interpreted as follows: 1) for each time of state transition, the agent will receive a movement penality 1, encouraging the UAV to use less steps to generate the trajectory; and 2) on top of the movement penality, the UAV will get a weighted outage duration penality $\frac{\tau\Delta_t}{L}\sum_{\iota=1}^{L} ITOP\{\vec{q}_u|h(\iota)\}$ as well, pushing the UAV to visit locations with stronger wireless coverage quality. Besides, two special cases are considered as follows: 1) once the UAV reaches the predefined destination $\vec{q}_u(D)$, the training episode terminates and a positive value r_D will replace the reward function; and 2) once the UAV crashes onto the boundary of the considered airspace, the training episode terminates and a negative value r_{ob} will replace the reward function instead. In summary, the aforementioned design of reward function aims to encourage the UAV to reach $\vec{q}_u(D)$ with as fewer steps as possible, while avoiding hitting the boundary and visiting areas with weak wireless coverage.
- γ: To connect the objective function of (P1) and the discounted accumulated-rewards over each learning episode, the discount factor is chosen as $\gamma = 1$.

B. Quantum-Inspired Representation of Experience's Priority

In the proposed DRL-QiER solution, the priority of experienced transition is represented by the k-th qubit, where the scalar index k indicates this transition's location index in the QiER buffer. Specifically, the quantum representation of stored transition's priority can be given by

$$|\Psi_k\rangle = \alpha_k |0\rangle + \beta_k |1\rangle , \tag{5}$$

where the complex-valued probability amplitudes α_k and β_k follow the normalization constraint $|\alpha_k|^2+|\beta_k|^2 = 1$. It is worth noting that the eigenstates $|0\rangle$ and $|1\rangle$ in (5) mean accepting and denying this transition, respectively. After quantum measurement, the superposition $|\Psi_k\rangle$ will collapse onto eigenstate $|0\rangle$ with probability $|\langle 0|\Psi_k\rangle|^2 = |\alpha_k|^2$ or eigenstate $|1\rangle$ with probability $|\langle 1|\Psi_k\rangle|^2 = |\beta_k|^2$. The complex coefficients α_k and β_k are of importance in the QiER system, influencing the occurrence probability of accepting or denying the corresponding transition when $|\Psi_k\rangle$ is observed.

C. QiER Framework

The proposed QiER framework consists of the following three phases.

1) Quantum Initialization Phase: When one transition is stored into the QiER buffer with finite capacity C, a label $k \in \{1,\ldots,C\}$ will be assigned to it, which specifies the location of this transition being recorded within the QiER buffer. When a new transition is recorded into the QiER buffer and before being sampled out to feed the training agent, its associated qubit $|\Psi_k\rangle$ should be initialized as eigenstate $|0\rangle$, i.e., $|\Psi_k\rangle \leftarrow |0\rangle$. The reason is that the agent has never been trained with these un-sampled transitions that may have unimaginable potentials to help the agent learn the characteristics of environment with which the agent is interacting. Thus, we assign these newly-recorded transitions with the highest priority, encouraging the agent to more likely learn from them.

2) Quantum Preparation Phase: After an experience is sampled from the QiER buffer to train the agent, the quantum preparation phase should be performed on its associated qubit, updating the corresponding priority. This is due to two reasons: 1) the temporal difference (TD) error of this transition is updated; and 2) the experience becomes older for the agent.

The uniform quantum state is defined as

$$|+\rangle = \frac{\sqrt{2}}{2}\left(|0\rangle + |1\rangle\right). \tag{6}$$

The absolute value of TD error $|\delta_t|$ is chosen to reflect priority of the corresponding transition. Once a recorded transition is sampled, its associated qubit $|\Psi_k\rangle$ should first be reset to the uniform quantum state, i.e., $|\Psi_k\rangle \leftarrow |+\rangle$. Then, to map the updated priority into $|\Psi_k\rangle$, one time of Grover iteration with flexible parameters ϕ_1 and ϕ_2 will be applied on the uniform quantum state, shown as

$$|\Psi_k\rangle = \boldsymbol{U}_{|+\rangle}\boldsymbol{U}_{|0\rangle}|+\rangle \overset{(a)}{=} (\mathcal{P} - e^{j\phi_1})\frac{\sqrt{2}}{2}|0\rangle + (\mathcal{P}-1)\frac{\sqrt{2}}{2}|1\rangle , \tag{7}$$

where $\mathcal{P} = (1-e^{j\phi_2})\left[1-0.5(1-e^{j\phi_1})\right]$ and the derivation (a) is based on ***Proposition 1*** of [7].

In practical applications, some experiences may be sampled for training with undesired high frequency, leading to over-training issue. Besides, the finite size of QiER buffer could further deteriorate this disservice [8], which will cause unfair and biased sampling performance. To circumvent this issue, the replay time of each stored transition should be taken into consideration for the quantum preparation phase, which enables it

[4]For detailed information regarding the design of DRL-QiER approach, please refer to [7] which is the full version of this conference paper.

to enrich sample diversity to improve the learning performance. In the early stage of training the agent, the importance of each experience is ambiguous. However, alongside the learning process, the absolute TD errors of some transitions remain relatively large, despite many times they have been sampled for training. Hence, it is necessary to relate training episode to the quantum preparation phase.

The quantum preparation phase aims to modify the collapse probability onto eigenstate $|0\rangle$, via one time of Grover iteration with free parameters ϕ_1 and ϕ_2. To quantify the amplification step of quantum preparation phase, we let

$$\phi_1 = \frac{e^{\frac{|\delta_t|\pi}{\delta_{\max}}} - e^{-\frac{|\delta_t|\pi}{\delta_{\max}}}}{e^{\frac{|\delta_t|\pi}{\delta_{\max}}} + e^{-\frac{|\delta_t|\pi}{\delta_{\max}}}} \frac{\pi}{2} = \frac{\pi}{2} tanh\left(\frac{|\delta_t|\pi}{\delta_{\max}}\right) \in \left[0, \frac{\pi}{2}\right), \quad (8)$$

$$\phi_2 = \frac{rt_k}{rt_{\max}} \frac{te}{te_{\max}} \pi + \frac{\pi}{2} \in \left(\frac{\pi}{2}, \frac{3\pi}{2}\right]. \quad (9)$$

With (8) and (9), the quantum amplitude amplification is related with the corresponding absolute TD error $|\delta_t|$, maximum TD error $\delta_{\max}$, replay times rt_k, maximum replay time $rt_{\max}$, current training episode te and the total training episode $te_{\max}$, which means that the quantum preparation phase updates the priority of experience into its associated k-th qubit $|\Psi_k\rangle$.

3) Quantum Measurement Phase: After the QiER buffer is fully occupied by recorded transitions, a mini-batch of experiences will be sampled to perform network training for the agent, via standard gradient descent method. To prepare the mini-batch sampling procedure under constraint of priorities, quantum measurement on the associated qubits should be accomplished first. Specifically, the probability of the k-th qubit collapsing onto eigenstate $|0\rangle$ can be calculated as $|\langle 0|\Psi_k\rangle|^2$. Then, the probability of the corresponding experience being picked up during the mini-batch sampling process can be defined as $bp_k = |\langle 0|\Psi_k\rangle|^2 / \sum_{e=1}^{C} |\langle 0|\Psi_e\rangle|^2$.

D. The Proposed DRL-QiER Solution

The proposed DRL-QiER algorithm is summarized in **Algorithm 1**, of which the flow chart is illustrated in Fig. 1. To solve the formulated MDP in Section III-A, double deep Q network (DQN) with duelling architecture (D3QN) is adopted to approximate the Q function $Q(\vec{q}_u, \vec{v}_u)$. To further speed up and stabilize the learning process, N_{ms}-step learning and target network techniques are adopted for updating parameters of the online D3QN. Specifically, the online D3QN aims to minimize the following loss function

$$\mathcal{L}(\boldsymbol{\theta}_{D3}) = \Big[r_{t:t+N_{ms}} + \gamma^{N_{ms}} Q(\vec{q}_{u(t+N_{ms})}, \vec{v}_u^*|\boldsymbol{\theta}_{D3}^-) - Q(\vec{q}_{u(t)}, \vec{v}_{u(t)}|\boldsymbol{\theta}_{D3})\Big]^2, \quad (10)$$

where $r_{t:t+N_{ms}} = \sum_{n_{ms}=0}^{N_{ms}-1} \gamma^{n_{ms}} r_{t+n_{ms}+1}$ indicates the N_{ms}-step discounted accumulated-reward, $\boldsymbol{\theta}_{D3}$ is the parameter vector of the online D3QN and $\boldsymbol{\theta}_{D3}^-$ means the parameter vector of the target D3QN. The selected action $\vec{v}_u^*$ in (10) is chosen from the online D3QN rather than the target D3QN, i.e., $\vec{v}_u^* = \arg\max_{\vec{v}_u \in \mathcal{A}} Q(\vec{q}_{u(t+N_{ms})}, \vec{v}_u|\boldsymbol{\theta}_{D3})$, which completes the double DQN procedure.

Algorithm 1: The Proposed DRL-QiER Solution

Initialization: Initialize the online D3QN network $Q_{D3}(s, a|\boldsymbol{\theta}_{D3})$ and its target network $Q_{D3}(s, a|\boldsymbol{\theta}_{D3}^-)$, with $\boldsymbol{\theta}_{D3}^- \leftarrow \boldsymbol{\theta}_{D3}$. Initialize the QiER buffer R with capacity C. Initialize the vector of replay time as $\vec{rt} = [rt_1, rt_2, \ldots, rt_C] = \vec{0}$. Set the size of mini-batch as N_{mb}. Set the order index of R as $k = 1$. Set the flag indicating whether the QiER buffer is fully occupied or not as $LF = False$. Set the maximum TD error as $\delta_{\max} = 1$;
for $te = [1, te_{\max}]$ **do**
  Set time step $n = 0$. Randomly set the the UAV's initial location as $\vec{q}_u(n) \in \mathcal{S}$. Initialize a sliding buffer $\hat{\text{R}}$ with capacity N_{ms};
  repeat
    Select and execute action a_n, then observe the next state $\vec{q}_u(n+1)$ and the immediate reward $r_n = r_n[\vec{q}_u(n+1)]$;
    if $LF == True$ **then**
      Perform quantum measurement on all stored experiences' qubits and get the vector of their replaying probabilities $[bp_1, bp_2, \ldots, bp_C]$;
      for $n_{mb} = [1, N_{mb}]$ **do**
        Sample a transition according to $[bp_1, bp_2, \ldots, bp_C]$ and get its location index $d \in \{1, 2, \ldots, C\}$;
        Reset the d-th qubit back to uniform quantum state $|\Psi_d\rangle = |+\rangle$;
        Update the corresponding replay time $rt_d += 1$ and $rt_{\max} = \max(\vec{rt})$;
        Calculate the sampled transition's absolute N_{ms}-step TD error $|\delta_{N_{ms}}|$ and update the maximum TD error $\delta_{\max} = \max(\delta_{\max}, |\delta_{N_{ms}}|)$;
        Perform quantum preparation phase on the d-th qubit;
      end
      Update the online D3QN network $Q_{D3}(s, a|\boldsymbol{\theta}_{D3})$ via gradient descent method using the mini-batch of sampled N_{mb} transitions from R;
    end
    Get and record transition $exp_n = \{\vec{q}_u(n), a_n, r_n, \vec{q}_u(n+1)\}$ into $\hat{\text{R}}$;
    if $n \geq N_{ms}$ **then**
      Generate the N_{ms}-step reward $r_{n-N_{ms}:n}$ from $\hat{\text{R}}$ and record N_{ms}-step experience $exp_{n-N_{ms}:n} = \{\vec{q}_u(n - N_{ms}), a_{n-N_{ms}}, r_{n-N_{ms}:n}, \vec{q}_u(n)\}$ into R with order index k;
      Perform quantum initialization phase on the k-th qubit as $|\Psi_k\rangle = |0\rangle$. Reset $rt_k = 0$ and let $k+ = 1$;
      if $k > C$ **then**
        Set $LF = True$ and reset $k = 1$;
      end
    end
    Let $n+ = 1$;
  until $\vec{q}_u(n) = \vec{q}_u(D)$ || $\vec{q}_u(n) \notin \mathcal{S}$ || $n = N_{\max}$;
  Update $\epsilon \leftarrow \epsilon \times dec_\epsilon$. Update the target D3QN $Q_{D3}(s, a|\boldsymbol{\theta}_{D3}^-)$ every Υ_{D3} episodes, i.e., $\boldsymbol{\theta}_{D3}^- \leftarrow \boldsymbol{\theta}_{D3}$;
end

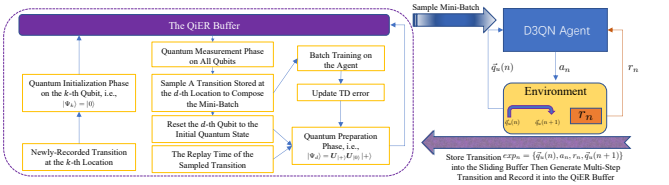


Figure 1: Flow chart of the proposed DRL-QiER algorithm

IV. Numerical Results

In this section, simulation results for the proposed DRL-QiER solution and the corresponding performance comparison against several baselines are performed. For conducting the simulation, the UAV's exploration airspace is set as $\mathbb{A} : [0, 1] \times [0, 1] \times [0, 0.1]$ km. Fig. 2 delivers the top view of $\mathbb{A}$, in which the locations of involved BSs and the direction of each ULA's boresight are specified. To generate building distribution within $\mathbb{A}$, one realization of statistical model suggested by the International Telecommunication Union (ITU) [9] is invoked, which is subject to the following three parameters: 1) $\hat{\alpha}$ indicates the ratio of region covered by buildings to the whole land; 2) $\hat{\beta}$ represents average amount of buildings; and 3) $\hat{\gamma}$ determines building heights' distribution (say, Rayleigh distribution with mean $\hat{\gamma} > 0$). Besides, the pathloss and small-scale fading components of G2A link are assumed to follow 3GPP urban Macro (UMa) [10] and block Nakagami-m channel models, respectively. For BSs' antenna model, vertically-placed uniform linear array (ULA) suggested by 3GPP [11] is applied to serve each sector, with fixed 3-dimensional (3D) radiation pattern. The common destination's location is fixed at $\vec{q}_u(D) = (0.8, 0.8, 0.1)$ km. Unless otherwise mentioned, the parameter setups regarding simulation environment are in line with Table I.

Four DRL-aided baselines are considered for performance

Table I: Parameter Settings for Simulation

Parameters	Values	Parameters	Values	Parameters	Values
Amount of BSs B	4	Amount of sectors $3B$	12	Capacity of QiER buffer C	20000
Horizontal side-length of $\mathbb{A}$ D	1 km	Amount of each ULA's array elements M	8	Size of mini-batch N_{mb}	128
Half-power beamwidth Θ_{3dB}/Φ_{3dB}	$65^{\circ}/65^{\circ}$	Speed of light c	3×10^8 m/s	Initial ϵ-greedy factor ϵ	0.5
Carrier frequency f_c	2 GHz	Wave length λ	15 cm	Annealing speed dec_ϵ	0.994/episode
ULA's element spacing distance d_v	7.5 cm	ULA's electrically titled angle θ_{etilt}	100°	Target D3QN update frequency Υ_{D3}	5
Antenna height of BS	25 m	Flying altitude of UAV	100 m	Length of sliding buffer N_{mls}	30
ITU building distribution parameter $\hat{\alpha}$	0.3	ITU building distribution parameter $\hat{\beta}$	118	Positive special reward r_D	400
ITU building distribution parameter $\hat{\gamma}$	25	Amount of buildings $\hat{\beta}D^2$	118	Negative special reward r_{ob}	-10000
Expected size of each building $\hat{\alpha}/\hat{\beta}$	0.0025 km^2	Maximum height of buildings	70 m	Learning rate α_{lr}	Adam's default
Transmit power of each sector P_i	20 dBm	Nakagami shape factor m for LoS/NLoS	3/1	Discount factor γ	1
Transmission outage threshold Γ_{th}	0 dB	Average power of AWGN σ^2	-90 dBm	Maximum training episodes $te_{\max}$	2000
Duration of time slot Δ_t	0.5 s	Velocity of the UAV V_u	30 m/s	Step threshold $N_{\max}$	400
Amount of signal Measurements L	1000	Weight balancing the minimization τ	50		

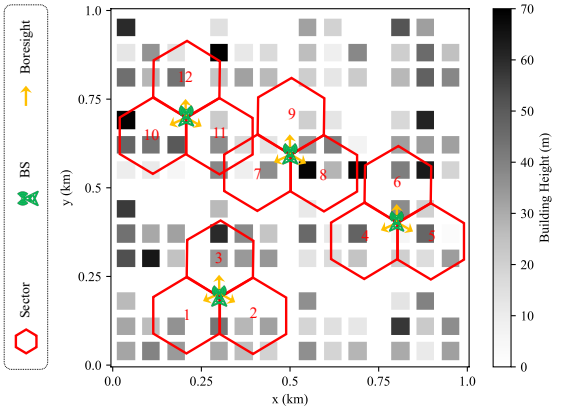


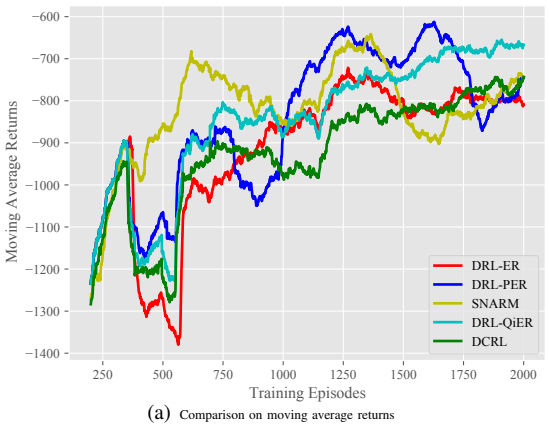
Figure 2: The simulation environment

comparison, i.e., DRL-ER [12], DRL-PER [13], DCRL [14] and SNARM [2].[5] For fair comparison, the structures of online and target D3QNs for all baselines are the same as those of the proposed DRL-QiER solution, while the hyper-parameter settings of DRL-QiER algorithm and these baselines are in line with Table I. Besides, the construction of radio map's DNN and the corresponding hyper-parameter settings of baseline SNARM are in accordance to [2], while the complexity index function, the curriculum evaluation function, the self-paced prioritized function, the coverage penalty function and the corresponding DCRL hyper-parameter settings are in line with [14]. Furthermore, the additional hyper-parameters regarding PER in DRL-PER baseline are set as $\alpha_{\text{PER}} = 1$, $\xi = 0.01$ and $\beta_{\text{PER}} = 0.4$. All the baselines are altered to involve multi-step learning and start training after their replay buffers are fully exploited. Moreover, all the baselines share the same randomly-generated initial UAV locations with the proposed DRL-QiER solution, for each training episode.

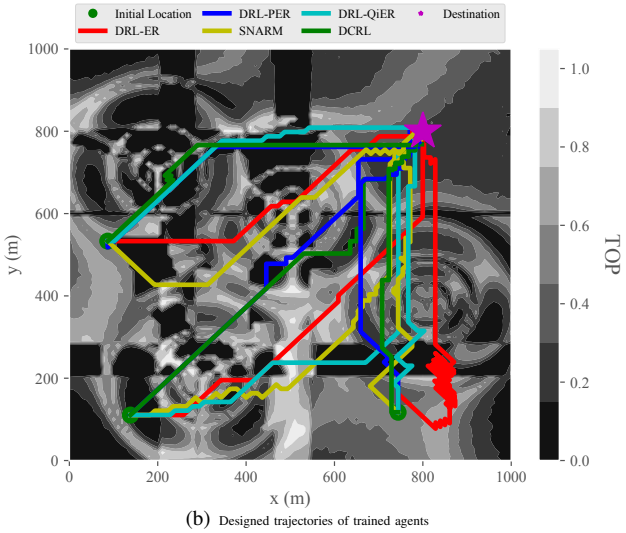
Fig. 3(a) delivers the performance comparison on moving average returns of the proposed DRL-QiER solution and considered baselines, versus training episodes. From this subfigure, it is easy to find that SNARM approach can offer satisfactory learning performance, thanks to the simulated trajectories enabled by the extra DNN (i.e., the radio map), especially in the range of training episode from 400 to 1000, despite that the radio map is getting well trained as the training process going. Besides, DRL-PER, DRL-QiER and DCRL approaches can achieve better moving average returns than DRL-ER method, in the early training stage (e.g., episodes 500-750). The reason is that DRL-ER solution samples transitions uniformly without considering their priorities, which leads transitions with higher importance to have less opportunities for training the online D3QN. However, DRL-PER method experiences server fluctuations than DRL-QiER and DCRL (e.g., episodes 1250-2000), which is because DRL-PER does not take transitions' replay time into account and thus some transitions are sampled with undesired high frequency while their absolute TD errors remain relatively large. The proposed DRL-QiER solution showcases more steady learning ability, with less amplification of fluctuation and overall raising trend, thanks to the QiER technique which balances sampling priority and diversity in a better manner. Although SNARM and DCRL approaches can offer satisfactory learning performances, their respective shortcomings are: 1) SNARM framework needs to train an extra DNN, which thus introduces heavy training burden, and 2) it is difficult to set up feasible complexity index function, curriculum evaluation function, self-paced prioritized function, coverage penalty function and the corresponding DCRL hyper-parameters, which limits the robustness of DCRL solution. The proposed DRL-QiER method requires less hyper-parameters tuning and contains no extra DNN, and therefore is easier and more robust for implementation. To deliver more insights, Fig. 3(b) depicts the comparison on designed trajectories of the implemented algorithms, over three representative starting locations chosen from episodes 1910-2000. It is straightforward to observe that the proposed DRL-QiER and the considered baselines direct the UAV to hit the common destination with different trajectories.

Fig. 4(a) demonstrates comparison on average time cost of designed trajectories and the corresponding EOD for the considered algorithms, over four episode slots 1-1400, 1401-1600, 1601-1800 and 1801-2000. From this figure, one can find that the proposed DRL-QiER solution can help achieve both lower average EOD and average time cost, within each episode slot. Especially, in the late training state (e.g., episode slot 1800-2000), the proposed DRL-QiER method outperforms other baselines, in terms of both average EOD and average time cost. Furthermore, Fig. 4(b) illustrates comparison on average duration and average weighted sum of EOD and time cost over the last 200 training episodes, for all the DRL-aided approaches and non-learning-based strategy termed as straight line. From this figure, it is easy to find that while the straight line solution offers the cheapest average time cost, it leads the UAV to suffer from the highest average EOD, which is extremely non-preferable and thus unveils the benefits provided by DRL-aided

[5]Due to page-length regulation, the detailed explanations of these baselines are omitted, which can be found in [7].



(a) Comparison on moving average returns



(b) Designed trajectories of trained agents

Figure 3: Performance comparison on moving average returns and designed trajectories

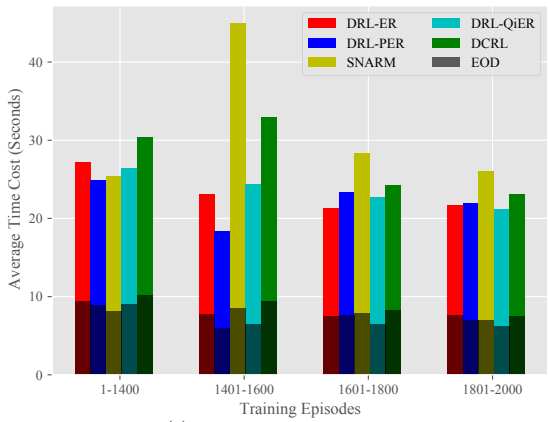
approaches. On the contrary, the proposed DRL-QiER solution can not only help the UAV experience the lowest average EOD, compared to both other DRL-aided approaches and the straight line strategy, but also direct the UAV to reach the common destination with the cheapest average time cost, against other DRL-aided solutions.

V. Conclusion

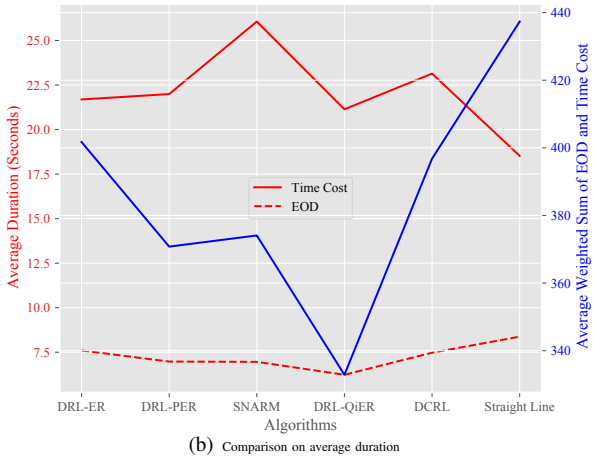
In this work, an intelligent navigation task for cellular-connected UAV networks was investigated, aiming at minimizing the weighted sum of time cost and expected outage duration alongside UAVs' flying trajectories towards the common destination with randomly-generated initial UAV locations. To navigate the UAV, a DRL-QiER solution was proposed, in which the innovative QiER technique can help the DRL agent hit a better learning efficiency. Simulation results validated the effectiveness of the proposed DRL-QiER solution, while performance comparison against both several DRL-aided baselines and straight line strategy showcased DRL-QiER method's superiority.

References

[1] L. Liu, S. Zhang, and R. Zhang, "Multi-beam UAV communication in cellular uplink: Cooperative interference cancellation and sum-rate maximization," *IEEE Trans. Wireless Commun.*, vol. 18, no. 10, pp. 4679–4691, 2019.



(a) Comparison on average time cost



(b) Comparison on average duration

Figure 4: Performance comparison on average time costs and EODs

[2] Y. Zeng, X. Xu, S. Jin, and R. Zhang, "Simultaneous navigation and radio mapping for cellular-connected UAV with deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4205–4220, 2021.

[3] J. Cui, Y. Liu, and A. Nallanathan, "Multi-agent reinforcement learning-based resource allocation for UAV networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 729–743, 2019.

[4] 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.

[5] G. D. Paparo, V. Dunjko, A. Makmal, M. A. Martin-Delgado, and H. J. Briegel, "Quantum speedup for active learning agents," *Phys. Rev. X*, vol. 4, no. 3, p. 031002, 2014.

[6] Y. Li, A. H. Aghvami, and D. Dong, "Intelligent trajectory planning in UAV-mounted wireless networks: A quantum-inspired reinforcement learning perspective," *IEEE Wireless Commun. Lett.*, 2021.

[7] ——, "Path planning for cellular-connected UAV: A DRL solution with quantum-inspired experience replay," *preprint arXiv:2108.13184*, 2021.

[8] T. De Bruin, J. Kober, K. Tuyls, and R. Babuška, "The importance of experience replay database composition in deep reinforcement learning," in *Proc. Deep reinforcement learning workshop, NIPS*, 2015.

[9] P. Series, "Propagation data and prediction methods required for the design of terrestrial broadband radio access systems operating in a frequency range from 3 to 60 GHz," *Recommendation ITU-R*, pp. 1410–1415, 2013.

[10] 3GPP TR 36.777, "Enhanced LTE support for aerial vehicles," Dec. 2017.

[11] 3GPP TR 36.873, "Study on 3D channel model for LTE," Dec. 2017.

[12] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[13] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," in *Proc. IEEE Int. Conf. Learn. Represent.*, 2016.

[14] Z. Ren, D. Dong, H. Li, and C. Chen, "Self-paced prioritized curriculum learning with coverage penalty in deep reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2216–2226, 2018.