Intelligent UAV Navigation: A DRL-QiER Solution

Authors: Yuanjian Li & Prof. Hamid Aghvami *Fellow, IEEE*

Department of Enginnering
Faculty of Natural, Mathematical & Engineering Sciences
King's College London

20th March 2022

Outline

Brief Overview

System Model

- Antenna Model
- Pathloss Model
- Building Distribution
- The Considered Optimization Problem

Quantum Basics

The Proposed DRL-QiER Algorithm

- Flow chart of the proposed DRL-QiER algorithm

Simulation Results

Brief Overview

In this presentation, quantum mechanics and deep reinforcement learning (DRL) techniques are invoked to help solve intelligent trajectory optimization problem for cellular-connected UAV networks. The main contributions are summarized as follows.

- Different from the vast majority of existing literature, more practical G2A pathloss model based on one realization of local building distribution and directional antenna with fixed 3-dimensional (3D) radiation pattern are considered in this paper. Then, 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 prior knowledge of the wireless environment, the focused path planning problem is challenging to be tackled via conventional optimization techniques. Alternatively, the proposed optimization problem is mapped into Markov decision process (MDP) and solved by the proposed DRL solution with novel quantum-inspired experience replay (QiER).

Brief Overview

- A novel QiER framework is coined to help the learning agent achieve better training performance, via a three-phase quantum-inspired process. Specifically, the quantum initialization phase allocates initial priority for the newly-recorded experiences, the quantum preparation phase generates the updated priority for the sampled transitions with the help of Grover iteration, and the quantum measurement phase outputs distribution of sampling probabilities to help accomplish the mini-batch training procedure.

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.[1] Intuitively, each UAV should be navigated with a feasible trajectory, alongside which the corresponding time consumption should be the shortest and wireless transmission quality provided by the cellular network should be maintained satisfactorily.[2]

[1] 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. Besides, collision avoidance during UAVs' flights needs to be guaranteed, via separating UAV's operation spaces and keeping their flying altitudes higher than the tallest building.

[2] This paper concentrates on UAV navigation task within coverage of cellular networks, while global positioning system (GPS)-supported UAV navigation is beyond the scope of this paper and left as one of future research directions.

System Model

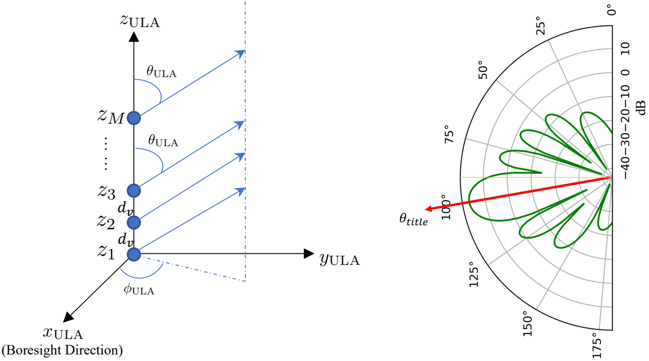
Without loss of generality, an arbitrary UAV (denoted as u hereafter) out of these U drones are concentrated for investigating the navigation task.[3] For clarity, the UAV's exploration environment is defined as a cubic sub-region $\mathbb{A} : [x_{\text{lo}}, x_{\text{up}}] \times [y_{\text{lo}}, y_{\text{up}}] \times [z_{\text{lo}}, z_{\text{up}}]$, 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. Therefore, 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 $\vec{q}_u(D)$. Besides, the location of arbitrary BS $b \in \mathcal{B}$ is indicated as $\vec{q}_b = (x_b, y_b, z_b)$, where $\vec{q}_{\text{lo}} \preceq \vec{q}_b \preceq \vec{q}_{\text{up}}$.

[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.

Antenna Model

In compliance with BS's antenna modelling of current cellular networks, directional antenna with fixed 3D radiation pattern is assumed to be equipped at each BS. Following standard sectorization, each BS is portioned to cover three sectors. Therefore, there are $3B$ sectors in total within the interested airspace $\mathbb{A}$. Specifically, it is assumed that three vertically-placed M-element uniform linear arrays (ULAs) are equipped by each BS with boresights directed to their corresponding sectors covered by this BS, subject to the 3GPP specification on cellular BS's antenna model.

Antenna Model



(a) Coordinate system of ULA

(b) Vertical pattern at boresight

Figure 1: Demonstration of ULA's coordinate system and vertical radiation pattern

Pathloss Model

Different from terrestrial transmissions, G2A links are more likely to experience LoS pathloss. In this subsection, the adopted G2A channel model will be interpreted.

According to 3GPP urban-macro (UMa) pathloss model, the G2A pathloss in dB from sector i to the UAV at time t is given by

$$\mathrm{PL}^{i}\left[\vec{q}_u(t)\right] = \begin{cases} 28.0 + 22\log_{10}\left(d_{iu}\right) + 20\log_{10}\left(f_c\right), & \text{if LoS} \\ -17.5 + \left[46 - 7\log_{10}\left(z_u(t)\right)\right]\log_{10}\left(d_{iu}\right) + 20\log_{10}\left(\frac{40\pi f_c}{3}\right), & \text{if NLoS} \end{cases}, \quad (1)$$

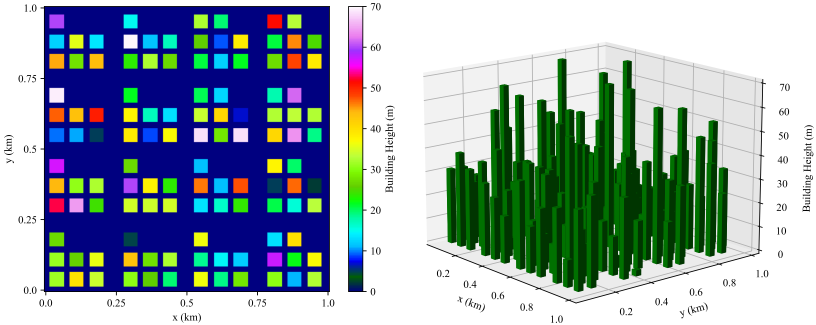
where $d_{iu} = ||\vec{q}_u(t) - \vec{q}_i||_2$ outputs the Euclidean distance between the UAV and the location of ULA for sector i.

Pathloss Model

To practically trace the type of G2A pathlosses, building distribution in the interested airspace $\mathbb{A}$ should be taken into consideration. Fig. 2 illustrates an example of local building distribution, including their horizontal locations on the ground and heights (Fig. 2a), as well as the corresponding 3D view (Fig. 2b). With given building distribution, the type of large-scale pathloss of G2A channels for UAV at arbitrary location $\vec{q}_u(t)$, i.e., LoS or NLoS in (1), can be accurately determined via checking the potential blockages between the UAV and sectors.[4]

[4] Note that our method generating G2A pathloss is more practical than the widely-used probabilistic G2A channel model in current literature because the later can only characterize the average G2A pathloss rather than its real counterpart.

Building Distribution



(a) Local building distribution

(b) 3D view of local building distribution

Figure 2: The building distribution under consideration

The Considered Optimization Problem

In short, the corresponding optimization problem is stated as

$$\text{(P1)}: \min_{\vec{v}_u(n)} \frac{\tau\Delta_t}{L}\sum_{n=1}^{N}\sum_{\iota=1}^{L} ITOP\{\vec{q}_u(n), \hat{i}(n)|h(\iota)\} + N, \tag{2a}$$

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

$$\vec{q}(n+1) = \vec{q}(n) + V_u\Delta_t\vec{v}_u(n), \|\vec{v}_u(n)\| = 1, \tag{2c}$$

$$\vec{q}_{\text{lo}} \preceq \vec{q}_u(n) \preceq \vec{q}_{\text{up}}, \vec{q}_u(0) = \vec{q}_u(I), \vec{q}_u(N) = \vec{q}_u(D), \tag{2d}$$

where τ is the weight balancing the minimization dilemma, V_u represents the UAV's flying velocity and $\vec{v}_u(n)$ specifies the mobility direction. The constraint (2b) 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.

Quantum State

In quantum mechanics, a quantum state of a closed quantum system can be described by a unit vector in Hilbert space. Specifically, a quantum state $|\Psi_c\rangle$ (Dirac notation) comprised of $\hat{n}$ quantum bits (qubits[5]) can be expressed as

$$|\Psi_c\rangle = |\Psi_1\rangle \otimes |\Psi_2\rangle \otimes \cdots \otimes |\Psi_{\hat{n}}\rangle = \sum_{p=00...0}^{\overbrace{11...1}^{\hat{n}}} h_p \, |p\rangle \,, \tag{3}$$

where $|\Psi_e\rangle \,, e \in [1, \hat{n}]$ represents the e-th qubit, h_p means the complex coefficient (i.e., probability amplitude) of eigenstate $|p\rangle$ subject to $\sum_{p=00...0}^{11...1} |h_p|^2 = 1$ and $\otimes$ denotes the tensor product. The representation of $\hat{n}$-qubit quantum state $|\Psi_c\rangle$ follows the quantum phenomenon known as *state superposition principle*. That is, the $|\Psi_c\rangle$ can be regarded as the superposition of $2^{\hat{n}}$ eigenstates, ranged from $|00...0\rangle$ to $|11...1\rangle$.

Quantum State

As a special case, a two-eigenstate quantum system (say, a single qubit) can be described as an arbitrary superposition state of eigenstates $|0\rangle$ and $|1\rangle$, given by

$$|\Psi\rangle = \alpha\,|0\rangle + \beta\,|1\rangle\,, \tag{4}$$

where the complex coefficients $\alpha = \langle 0|\Psi\rangle$ and $\beta = \langle 1|\Psi\rangle$ denote the probability amplitudes for eigenstates $|0\rangle$ and $|1\rangle$, respectively. Note that the single-qubit superposition $|\Psi\rangle$ is a unit vector (i.e., $\langle\Psi|\Psi\rangle = 1$) in Hilbert space spanned by orthogonal bases $|0\rangle$ and $|1\rangle$, subject to $|\alpha|^2 + |\beta|^2 = 1$. According to *quantum collapse phenomenon*, after measurement or observation of an external experimenter, $|\Psi\rangle$ will collapse from its superposition state onto one of its eigenstates $|0\rangle$ and $|1\rangle$ with probabilities $|\alpha|^2$ and $|\beta|^2$, respectively.

Quantum Amplitude Amplification

For a two-eigenstate qubit $|\Psi\rangle$, the probability amplitudes of each eigenstate can be changed via a quantum operation (e.g., Grover iteration), gradually modifying the collapse probability distribution. Two unitary reflections are applied to achieve Grover iteration, given by

$$\boldsymbol{U}_{|0\rangle} = \boldsymbol{I} - (1 - e^{j\phi_1}) |0\rangle \langle 0| , \quad (5)$$

$$\boldsymbol{U}_{|\Psi\rangle} = (1 - e^{j\phi_2}) |\Psi\rangle \langle \Psi| - \boldsymbol{I}, \quad (6)$$

where $\{\phi_1, \phi_2\} \in [0, 2\pi]$, $\boldsymbol{I}$ indicates identity matrix, and $\langle 0|$ and $\langle \Psi|$ are Hermitian transposes of $|0\rangle$ and $|\Psi\rangle$, respectively. Then, the Grover iterator can be formulated as $\boldsymbol{G} = \boldsymbol{U}_{|\Psi\rangle}\boldsymbol{U}_{|0\rangle}$, which remains unitary. After m times of acting $\boldsymbol{G}$ on $|\Psi\rangle$, the two-eigenstate qubit with updated probability amplitudes can be given by $|\Psi\rangle \leftarrow \boldsymbol{G}^m |\Psi\rangle$.

Quantum Amplitude Amplification

Two updating approaches can be used to accomplish quantum amplitude amplification task: 1) $m = 1$ with dynamic parameters ϕ_1 and ϕ_2; and 2) dynamic m with fixed parameters ϕ_1 and ϕ_2 (e.g., π). The latter updating method can only change the probability amplitudes in a discrete manner, and thus the former solution is chosen in this demonstration.

Proposition

For Grover iteration with flexible parameters, the overall effects of $\boldsymbol{G}$ *on the superposition* $|\Psi\rangle$ *can be derived analytically as* $\boldsymbol{G}|\Psi\rangle = (\mathcal{Q} - e^{j\phi_1})\alpha|0\rangle + (\mathcal{Q}-1)\beta|1\rangle$, *where* $\mathcal{Q} = (1 - e^{j\phi_2})\left[1 - (1 - e^{j\phi_1})|\alpha|^2\right]$ *and* $|(\mathcal{Q} - e^{j\phi_1})|^2|\alpha|^2 + |(\mathcal{Q}-1)|^2|\beta|^2 = 1$.

Proof.

The effects of $\boldsymbol{U}_{|0\rangle}$ on $|0\rangle$ and $|1\rangle$ are expressed as

$$\boldsymbol{U}_{|0\rangle}\,|0\rangle = \Big[\boldsymbol{I} - (1 - e^{j\phi_1})\,|0\rangle\,\langle 0|\Big]\,|0\rangle = e^{j\phi_1}\,|0\rangle\,, \tag{7}$$

$$\boldsymbol{U}_{|0\rangle}\,|1\rangle = \Big[\boldsymbol{I} - (1 - e^{j\phi_1})\,|0\rangle\,\langle 0|\Big]\,|1\rangle = |1\rangle\,, \tag{8}$$

respectively. Then, we obtain

$$\boldsymbol{U}_{|0\rangle}\,|\Psi\rangle = \Big[\boldsymbol{I} - (1 - e^{j\phi_1})\,|0\rangle\,\langle 0|\Big]\,|\Psi\rangle = e^{j\phi_1}\alpha\,|0\rangle + \beta\,|1\rangle\,, \tag{9}$$

where $\boldsymbol{U}_{|0\rangle}$ plays the role as a *conditional phase shift operator*.
Furthermore, we get

$$\begin{aligned}\boldsymbol{G}\,|\Psi\rangle = \boldsymbol{U}_{|\Psi\rangle}\boldsymbol{U}_{|0\rangle}\,|\Psi\rangle &= (1 - e^{j\phi_2})\,[\alpha\,|0\rangle + \beta\,|1\rangle]\Big[\alpha^\dagger\,\langle 0| + \beta^\dagger\,\langle 1|\Big]\,\boldsymbol{U}_{|0\rangle}\,|\Psi\rangle - \boldsymbol{U}_{|0\rangle}\,|\Psi\rangle \\ &= (\mathcal{Q} - e^{j\phi_1})\alpha\,|0\rangle + (\mathcal{Q} - 1)\beta\,|1\rangle\,, \end{aligned} \tag{10}$$

where $\mathcal{Q} = (1 - e^{j\phi_2})(e^{j\phi_1}|\alpha|^2 + |\beta|^2) = (1 - e^{j\phi_2})\Big[1 - (1 - e^{j\phi_1})|\alpha|^2\Big]$.
Because Grover operator $\boldsymbol{G}$ is unitary, the updated superposition $|\Psi\rangle \leftarrow \boldsymbol{G}\,|\Psi\rangle$ still follows the normalization rule of probability amplitudes, i.e., $|(\mathcal{Q} - e^{j\phi_1})|^2|\alpha|^2 + |(\mathcal{Q} - 1)|^2|\beta|^2 = 1$. □

Corollary

The ratio between collapse probabilities of $|\Psi\rangle \rightarrow |0\rangle$ before and after being impacted by G can be given by $|\mathcal{R}|^2 = |(1-e^{j\phi_1}-e^{j\phi_2})-(1-e^{j\phi_1})(1-e^{j\phi_2})|\alpha|^2|^2$, which is symmetric w.r.t. $\phi_1 = \phi_2$ and $\phi_1 = 2\pi - \phi_2$. Then, the updated collapse probabilities onto eigenstates $|0\rangle$ and $|1\rangle$ can be given by $|\mathcal{R}|^2|\alpha|^2$ and $1-|\mathcal{R}|^2|\alpha|^2$, respectively.

Proof.

Based on (4) and (10), the ratio between the probability amplitudes of $|0\rangle$ after being acted by $\boldsymbol{G}$ and before that can be derived as $\mathcal{R} = (1-e^{j\phi_1}-e^{j\phi_2})-(1-e^{j\phi_1})(1-e^{j\phi_2})|\alpha|^2$, which completes the proof. □

The proposed QiER framework consists of the following three phases.

- *Quantum Initialization Phase*: When transition exp_t is stored into the QiER buffer with finite capacity C, a label $k \in \{1, \ldots, C\}$ will be assigned to exp_t, which specifies the location of exp_t being recorded within the QiER buffer.[6] Then, experience exp_t and the k-th qubit $|\Psi_k\rangle$ together will be stored into the QiER buffer, which can be regarded as a collection of $(exp_t, |\Psi_k\rangle)$. 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 set these newly-recorded transitions with the highest priority, encouraging the agent to more likely learn from them.

- *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 TD error of this transition is updated; and 2) the experience becomes older for the agent. The uniform quantum state is defined as

$$|+\rangle = \frac{\sqrt{2}}{2}\left(|0\rangle + |1\rangle\right). \tag{11}$$

The absolute value of TD error $|\delta_t|$ is chosen to reflect priority of the corresponding transition exp_t. 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 of exp_t into $|\Psi_k\rangle$, one time of Grover iteration with flexible parameters will be applied on the uniform quantum state, shown as

$$|\Psi_k\rangle = \boldsymbol{U}_{|+\rangle}\boldsymbol{U}_{|0\rangle}\,|+\rangle \overset{(a)}{=} (\mathcal{P} - e^{j\phi_1})\frac{\sqrt{2}}{2}\,|0\rangle + (\mathcal{P} - 1)\frac{\sqrt{2}}{2}\,|1\rangle\,, \quad (12)$$

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*.

- *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 $bp_k = |\langle 0|\Psi_k\rangle|^2 / \sum_{e=1}^{C} |\langle 0|\Psi_e\rangle|^2$, in which the denominator means the sum of collapse probabilities onto eigenstate $|0\rangle$ of qubits that are associated with all stored experiences.

  During the mini-batch sampling period, several times of picking recorded experiences from the QiER buffer will be executed, following the generated picking probability vector $\vec{bp} = [bp_1, bp_2, \ldots, bp_C]$ after quantum measurement phase. Note that the total sampling time is equal to the size of mini-batch, which will be specified in the numerical result section later.

Remark

Although the QiER buffer involves quantum representations and operations, the corresponding processes, i.e., the quantum initialization phase, the quantum preparation phase and the quantum measurement phase, can be imitated on conventional computing devices without implementing real quantum computations on practical quantum computers.

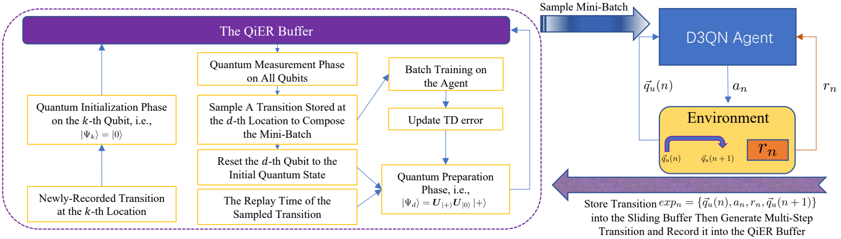
Remark

The associated qubit of sampled experience should be reset to the uniform quantum state, which means that the corresponding quantum preparation phase starts from the uniform quantum state rather than the previous counterpart. This is in line with the quantum phenomenon where a quantum system will collapse onto one of its eigenstates after an observation. Note that the sampled transitions are still remained in the QiER buffer until they are discarded.

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 R̂ 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 R̂;
        if $n \geq N_{ms}$ then
            Generate the $N_{ms}$-step reward $r_{n-N_{ms}:n}$ from R̂ 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 $\Upsilon_{D3}$ episodes, i.e., $\boldsymbol{\theta}^{-}_{D3} \leftarrow \boldsymbol{\theta}_{D3}$;
end
```

Flow chart of the proposed DRL-QiER algorithm



The QiER Buffer

Quantum Measurement Phase on All Qubits

Sample A Transition Stored at the d-th Location to Compose the Mini-Batch

Reset the d-th Qubit to the Initial Quantum State

The Replay Time of the Sampled Transition

Batch Training on the Agent

Update TD error

Quantum Preparation Phase, i.e., $|\Psi_d\rangle = \boldsymbol{U}_{|+\rangle}\boldsymbol{U}_{|0\rangle}\,|+\rangle$

Quantum Initialization Phase on the k-th Qubit, i.e., $|\Psi_k\rangle = |0\rangle$

Newly-Recorded Transition at the k-th Location

Sample Mini-Batch

D3QN Agent

$\vec{q}_u(n)$

a_n

r_n

Environment

$\vec{q}_u(n)$

$\vec{q}_u(n+1)$

r_n

Store Transition $exp_n = \{\vec{q}_u(n), a_n, r_n, \vec{q}_u(n+1)\}$ into the Sliding Buffer Then Generate Multi-Step Transition and Record it into the QiER Buffer

Figure 3: Flow chart of the proposed DRL-QiER algorithm

Simulation Parameter Settings

Table 1: Parameter Settings for Simulation Environment

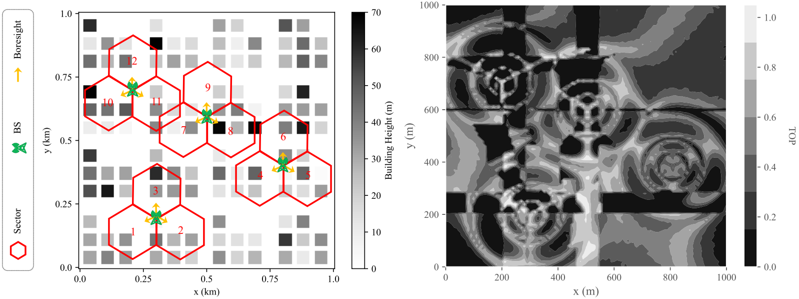
Parameters	Values	Parameters	Values
Amount of BSs B	4	Amount of sectors $3B$	12
Horizontal side-length of $\mathbb{A}$ D	1 km	Amount of each ULA's array elements M	8
Half-power beamwidth $\Theta_{\text{3dB}}/\Phi_{\text{3dB}}$	$65^\circ/65^\circ$	Speed of light c	3×10^8 m/s
Carrier frequency f_c	2 GHz	Wave length λ	15 cm
ULA's element spacing distance d_v	7.5 cm	ULA's electrically titled angle θ_{etilt}	100°
Antenna height of BS	25 m	Flying altitude of UAV	100 m
ITU building distribution parameter $\hat{\alpha}$	0.3	ITU building distribution parameter $\hat{\beta}$	118
ITU building distribution parameter $\hat{\gamma}$	25	Total amount of buildings $\hat{\beta}D^2$	118
Expected size of each building $\hat{\alpha}/\hat{\beta}$	0.0025 km^2	Maximum height of buildings	70 m
Transmit power of each sector P_i	20 dBm	Nakagami shape factor m for LoS/NLoS	3/1
Transmission outage threshold Γ_{th}	0 dB	Average power of AWGN σ^2	-90 dBm
Duration of time slot Δ_t	0.5 s	Velocity of the UAV V_u	30 m/s
Amount of signal Measurements L	1000	Weight balancing the minimization τ	50

Learning Parameter Settings

Table 2: Hyper-parameter Settings for Learning Process

Parameters	Values	Parameters	Values
Capacity of QiER buffer C	20000	Size of mini-batch N_{mb}	128
Initial ϵ-greedy factor ϵ	0.5	Annealing speed dec_{ϵ}	0.994/episode
Target D3QN update frequency Υ_{D3}	5	Length of sliding buffer N_{ms}	30
Positive special reward r_D	400	Negative special reward r_{ob}	-10000
Learning rate α_{lr}	Adam's default	Discount factor γ	1
Maximum training episodes $te_{\max}$	2000	Step threshold $N_{\max}$	400

Simulation Environment

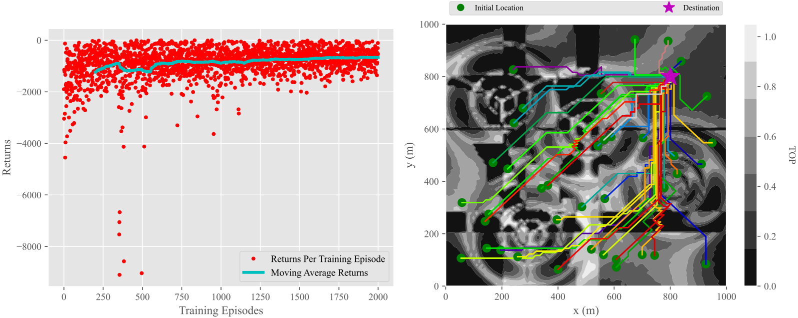


(a) The simulation environment

(b) The corresponding TOP distribution

Figure 4: Simulation environment and the corresponding preview on TOP distribution

Training results of the proposed DRL-QiER solution

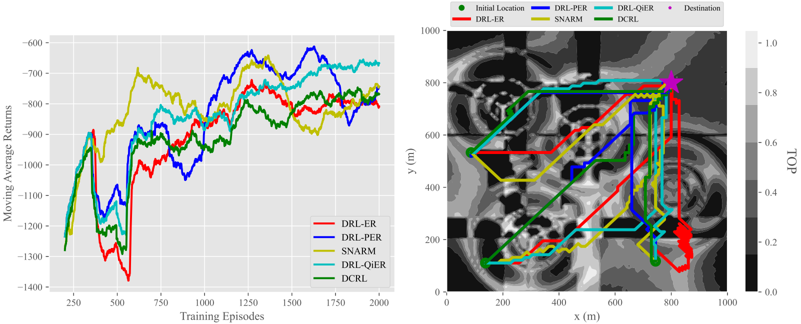


(a) Training return history

(b) The corresponding designed trajectories

Figure 5: Training results of the proposed DRL-QiER solution

Performance comparison on moving average returns and designed trajectories

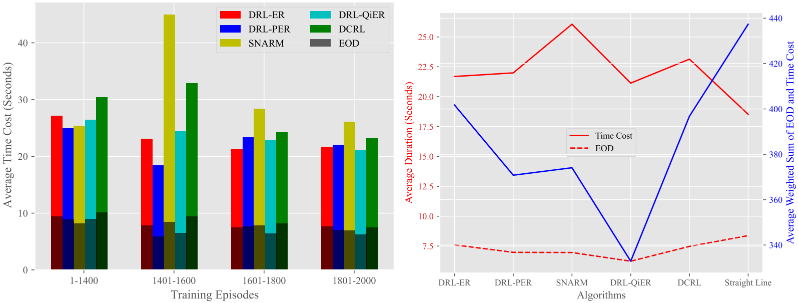


(a) Comparison on moving average returns

(b) Designed trajectories of trained agents

Figure 6: Performance comparison on moving average returns and designed trajectories

Performance comparison on average time costs and EOD



(a) Comparison on average time cost

(b) Comparison on average duration

Figure 7: Performance comparison on average time costs and EOD

Q & A

Thanks for your attentions

This is the end of this demonstration
Any Question?