DRL-Aided Joint Resource Block and Beamforming Management for Cellular-Connected UAVs

Dr. Yuanjian Li, Prof. Mathini Sellathurai, Dr. Zheng Chu, Prof Pei Xiao, and Prof. Hamid Aghvami *Fellow, IEEE*

***Research Fellow* @ Nanyang Technological University (NTU), Singapore**

December 6, 2023

Outline

1. Brief Overview
2. Main Contribution
3. System Model
4. Problem Formulation
5. The Proposed DRL Algorithm
6. Simulation Results

Brief Overview

In this paper, we investigate a cellular-connected unmanned aerial vehicle (UAV) network, where multiple UAVs receive messages from base stations (BSs) in the down-link, and in the meantime, BSs serve their paired ground user equipments (UEs). To effectively manage inter-cell interferences (ICIs) among UEs due to intense reuse of time-frequency resource block (RB) resource, a first p-tier based RB coordination criterion is adopted. Then, to enhance wireless transmission quality for UAVs while protecting terrestrial UEs from being interfered by ground-to-air (G2A) transmissions, a radio resource management (RRM) problem of joint dynamic RB coordination and time-varying beamforming design is formulated to minimize UAV's ergodic outage duration (EOD).

Brief Overview

To cope with conventional optimization techniques' inefficiency in solving the formulated RRM problem, a deep reinforcement learning (DRL)-aided solution is proposed, where deep double duelling Q network (D3QN) and twin delayed deep deterministic policy gradient (TD3) are invoked to deal with RB coordination in the discrete action domain and beamforming design in the continuous action regime, respectively. Numerical results illustrate the effectiveness of the proposed hybrid D3QN-TD3 algorithm, compared to representative baselines.

Main Contribution

Radio resource management issue on interference coordination and beamforming design in down-link cellular-connected UAV networks is investigated in this paper. It is worth noting that the terrestrial transmissions between GUEs and BSs are protected to be not contaminated by the down-link G2A channels. The main contributions of this paper can be concluded as follows.

- A joint time-frequency resource block (RB) allocation and beamforming design optimization problem is formulated to minimize the ergodic outage duration (EOD) of UAV, for arbitrary given trajectory. Specifically, the RB allocation is utilized to assign proper RB resource to UAVs while insuring that the terrestrial transmissions are not violated by the potential co-channel interferences generated from BSs aligned to serve UAVs. To enhance the quality of received signals at UAVs after RB allocation, transmit beamforming design is invoked in the presence of imperfect G2A channel estimation.

Main Contribution

- To deal with the difficulty of traditional optimization-based methods solving the proposed EOD minimization problem, a deep reinforcement learning (DRL) aided solution is proposed via mapping the proposed EOD minimization problem into an outer Markov decision process (MDP) and an inner MDP. The outer MDP reflects the dynamic RB possession environment at BSs, while the inner MDP tracks the corresponding small-scale fading characteristics. The outer MDP contains discrete action space (i.e., RB indices), which is tackled by invoking deep double duelling Q network (D3QN), while the continuous action space (i.e., beamforming vectors) in the inner MDP is dealt with twin delayed deep deterministic policy gradient (TD3) approach. The proposed hybrid D3QN-TD3 algorithm can optimize EOD performance for UAVs via interactively interacting with the outer and inner environments, of which the D3QN and TD3 agents are able to offer independent EOD performance gains.

Main Contribution

- In contrast to the majority of related literature adopting statistical G2A channel model (e.g., probabilistic G2A channel model), LoS/NLoS G2A pathloss is determined via checking potential blockages between UAV and BSs in this paper, according to one realization of local building distribution suggested by the International Telecommunication Union (ITU). The considered G2A channel model is more practical than its statistical counterpart which can only reflect average pathloss gain over large number of similar building distribution realizations because building distribution in local area remains unchanged in practice.

System Model

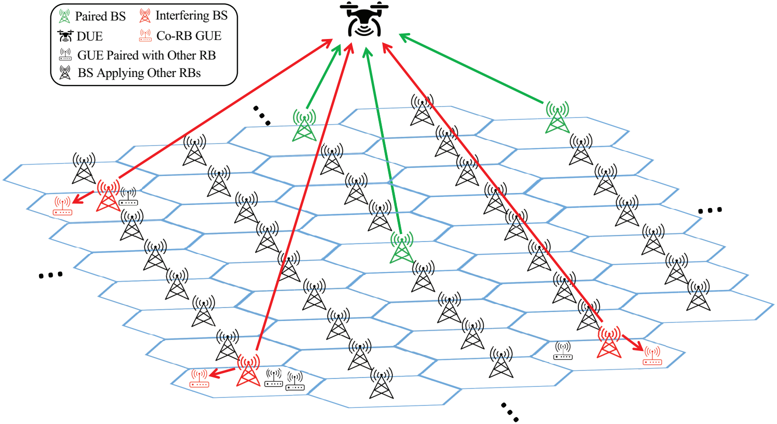
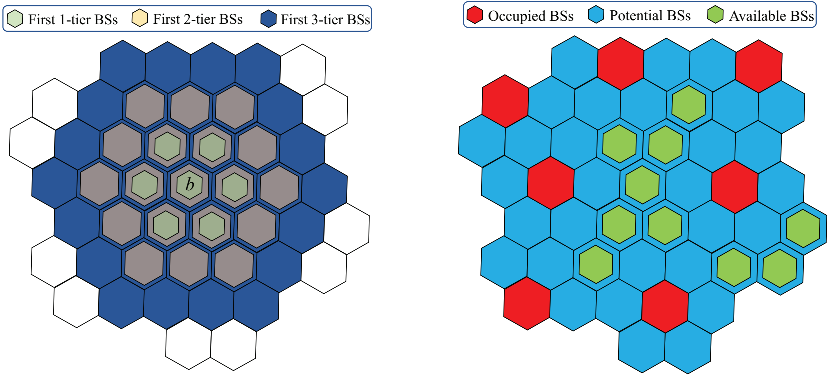


Figure 2: System model

BS Distribution

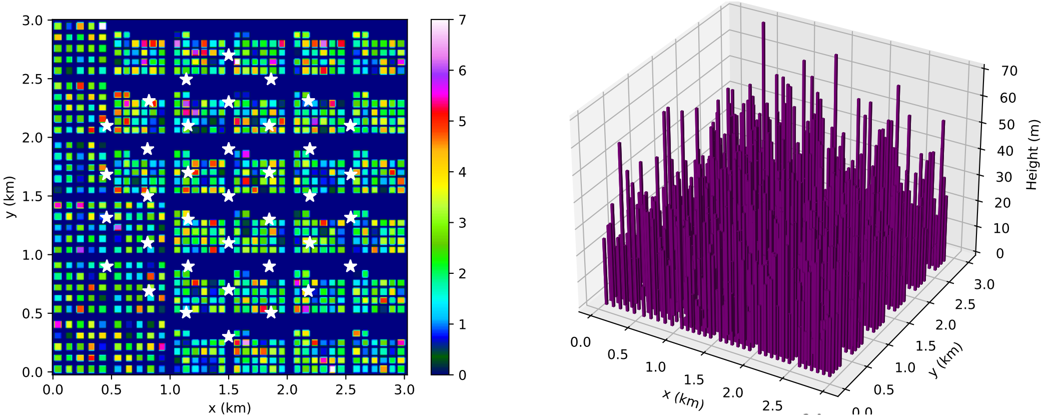


(a) First p-tier BSs

(b) An example of BS grouping

Figure 3: Illustrations of system model, the first p-tier set of BS b and an instance of BS grouping in the case of $p = 1$, where $card(\mathcal{TI}_b(1)) = 7$, $card(\mathcal{TI}_b(2)) = 19$ and $card(\mathcal{TI}_b(3)) = 37$

Building Distribution



(a) 2D distribution of local buildings and BSs

(b) 3D view of local building distribution

Figure 4: The considered building distribution

Channel Model

$$\mathrm{PL}_l = \begin{cases} 28.0 + 22\log_{10}\left(d_{bu}\right) + 20\log_{10}\left(f_c\right), & l = \mathrm{LoS} \\ -17.5 + \left[46 - 7\log_{10}\left(h\right)\right] \times \log_{10}\left(d_{bu}\right) + 20\log_{10}\left(\frac{40\pi f_c}{3}\right), & l = \mathrm{NLoS} \end{cases}, \quad (1)$$

in which $r_{bu} = \sqrt{d_{bu}^2 - h^2}$, $\varepsilon_1 = \max\{460\log_{10}(h) - 700, 18\}$, $\varepsilon_2 = 4300\log_{10}(h) - 3800$, f_c represents the carrier frequency and $d_{bu} = ||\vec{q}_u - \vec{q}_b||_2$ calculates the Euclidean distance between DUE u and ground BS b. Since the proposed design on beamforming vectors aims to be adaptive to arbitrary small-scale fading environment, we denote $\vec{h}_{bu} \in \mathbb{C}^{1\times M}, \forall b \in \mathcal{B}, u \in \mathcal{U}$ as the small-scale fading component for B2D channels and pose no assumptions on its modelling, while an example of specific B2D small-scale fading model will be discussed in numerical result section.

The Received Signal

The received signal of DUE u over RB k at time t can be given by

$$y_u^k(t) = C_u^k(t)\left[\sum_{b\in\breve{\mathcal{B}}_o^k}\sqrt{10^{\frac{-\mathrm{PL}_l}{10}}}\,\vec{h}_{bu}\vec{w}_{bu}x_u(t) + \sum_{b\in\mathcal{B}_o^k}\sqrt{10^{\frac{-\mathrm{PL}_l}{10}}}\,\vec{h}_{bu}\vec{w}_{bg}x_{bg}(t) + n_u^k\right], \quad (2)$$

where $\vec{w}_{bu} \in \mathbb{C}^{M\times 1}$ indicates the transmit beamforming vector at BS $b \in \breve{\mathcal{B}}_o^k$ for DUE u, $\vec{w}_{bg} \in \mathbb{C}^{M\times 1}$ represents the transmit beamforming vector at BS $b \in \mathcal{B}_o^k$ for corresponding GUEs, $x_u(t) \sim \mathcal{CN}(0, P)$ is the intended message from BS b to DUE u, $x_{bg}(t) \sim \mathcal{CN}(0, P)$ implies the signal for GUEs, and $n_u^k \sim \mathcal{CN}(0, \sigma^2)$ denotes the received additive complex Gaussian noise (AWGN) at DUE u.

SINR

Denote $C_u^k(n) \in \{0,1\}$ as the RB association indicator which means that DUE u is occupying RB k over time slot n when $C_u^k(n) = 1$, and $C_u^k(n) = 0$ otherwise. Each DUE is assumed to occupy at most one single RB each time, then we have $\sum_{k=1}^{K} C_u^k(n) \leq 1$. Then, instantaneous SINR of DUE u at time t can be calculated as

$$\Gamma_u(t) = \sum_{k=1}^{K} \frac{C_u^k(t) \left[\sum_{b \in \breve{\mathcal{B}}_o^k} \sqrt{P 10^{\frac{-\mathrm{PL}_l}{10}}} |\vec{h}_{bu} \vec{w}_{bu}|\right]^2}{I_u^k(t) + \sigma^2}, \tag{3}$$

where $I_u^k(t) = \sum_{b \in \mathcal{B}_o^k} P 10^{\frac{-\mathrm{PL}_l}{10}} |\vec{h}_{bu} \vec{w}_{bg}|^2$ means ICIs introduced by co-channel BSs from $\mathcal{B}_o^k$.

This paper focuses on minimizing ergodic outage duration (EOD) suffered from DUE's flight, approximately given by

$$\text{(P1)}: \min_{\substack{C_u^k(n),\\ \vec{w}_{bu}(n,i)}} \sum_{n=1}^{N_u} \sum_{i=1}^{\varsigma} \frac{\delta_u}{\varsigma} ITOP_u\{C_u^k(n),\vec{w}_{bu}(n,i);\hat{h}(n,i)\}, \tag{4a}$$

$$\text{s.t.} \Sigma_{k=1}^{K} C_u^k(n) \leq 1, \forall n \in [1, N_u], \tag{4b}$$

$$C_u^k(n) \in \{0,1\}, \forall k \in \mathcal{K}, \forall n \in [1, N_u]. \tag{4c}$$

$$||\vec{w}_{bu}(n,i)||^2 = 1, \forall b \in \breve{\mathcal{B}}_o^k, \forall n \in [1, N_u], \tag{4d}$$

where constraint (4d) ensures that the transmit power of each available BS $b \in \breve{\mathcal{B}}_o^k$ equals P. Denote Γ_{th} as the SINR threshold, indicator of transmission outage probability (TOP) is defined as $ITOP_u\{C_u^k(n), \vec{w}_{bu}(n,i); \hat{h}(n,i)\} = 1$ when $\Gamma_u(n,i) < \Gamma_{th}$, and $ITOP_u\{C_u^k(n), \vec{w}_{bu}(n,i); \hat{h}(n,i)\} = 0$ otherwise, where $\hat{h}(n,i)$ and $\vec{w}_{bu}(n,i)$ indicate one realization of fast fadings and corresponding beamforming vectors in time slot n, respectively. Then, to calculate the corresponding TOP, $\varsigma \gg 1$ times of SINR measurement are performed. The length of time slot is defined as δ_u and thus the number of time slots equals $N_u = T_u/\delta_u$ for the DUE u.

The Proposed DRL Solution

To derive a flexible solution solving (4) in a dynamic RB possession (RBP) and time-varying small-scale fading scenario, a deep reinforcement learning (DRL)-aided solution is proposed after mapping the considered optimization problem into two Markov decision processes (MDPs). Specifically, the deep double duelling netwrok (D3QN) maps the outer state and the RB selection into Q values, while actor of twin delayed deep deterministic policy gradient network (TD3) agent transforms the inner state to beamforming vector and TD3's critic network evaluates the corresponding Q values.

CNN-Attached Duelling DQN

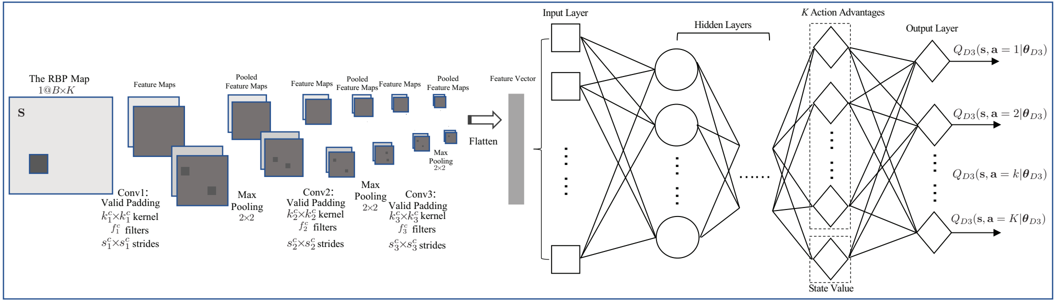
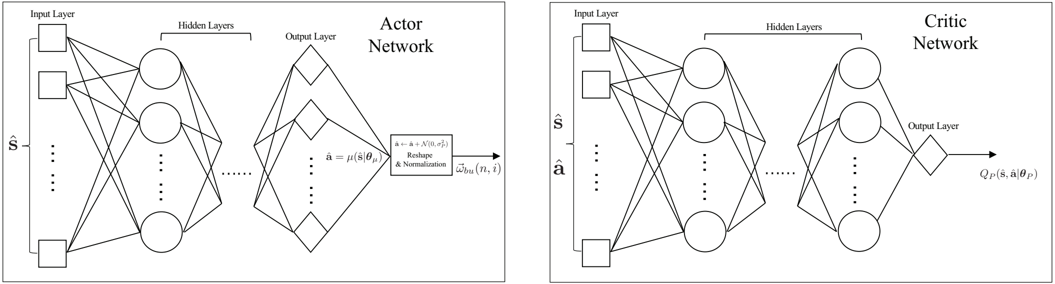

The RBP Map
$1@B\times K$
$\mathbf{s}$
Feature Maps
Pooled Feature Maps
Feature Maps
Pooled Feature Maps
Feature Maps
Pooled Feature Maps
Feature Vector
Conv1:
Valid Padding
$k_1^c\times k_1^c$ kernel
f_1^c filters
$s_1^c\times s_1^c$ strides
Max Pooling 2×2
Conv2:
Valid Padding
$k_2^c\times k_2^c$ kernel
f_2^c filters
$s_2^c\times s_2^c$ strides
Max Pooling 2×2
Conv3:
Valid Padding
$k_3^c\times k_3^c$ kernel
f_3^c filters
$s_3^c\times s_3^c$ strides
Max Pooling 2×2
Flatten
Input Layer
Hidden Layers
K Action Advantages
Output Layer
State Value
$Q_{D3}(\mathbf{s},\mathbf{a}=1|\boldsymbol{\theta}_{D3})$
$Q_{D3}(\mathbf{s},\mathbf{a}=2|\boldsymbol{\theta}_{D3})$
$Q_{D3}(\mathbf{s},\mathbf{a}=k|\boldsymbol{\theta}_{D3})$
$Q_{D3}(\mathbf{s},\mathbf{a}=K|\boldsymbol{\theta}_{D3})$

Figure 5: Architecture of CNN-attached duelling DQN

Architectures of TD3's Actor and Critic

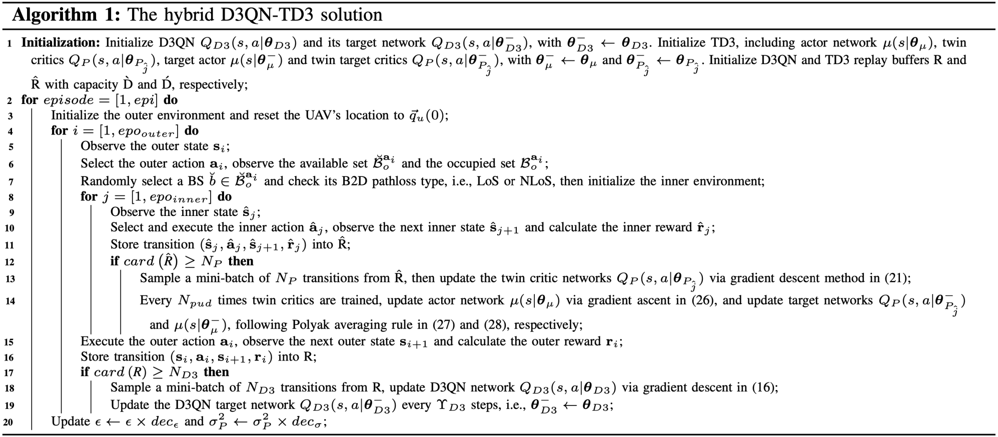

Input Layer
Hidden Layers
Output Layer
Actor
Network
$\hat{\mathbf{s}}$
......
$\hat{\mathbf{a}} = \mu(\hat{\mathbf{s}}|\boldsymbol{\theta}_\mu)$
$\hat{\mathbf{a}} \leftarrow \hat{\mathbf{a}} + \mathcal{N}(0, \sigma_P^2)$
Reshape
& Normalization
$\vec{\omega}_{bu}(n, i)$
Input Layer
Hidden Layers
Critic
Network
$\hat{\mathbf{s}}$
$\hat{\mathbf{a}}$
......
Output Layer
$Q_P(\hat{\mathbf{s}}, \hat{\mathbf{a}}|\boldsymbol{\theta}_P)$

(a) Layout of the actor network

(b) Layout of the critic network

Figure 6: Architecture of TD3 network

Pseudocode of the Proposed Hybrid D3QN-TD3 Solution



Algorithm 1: The hybrid D3QN-TD3 solution

```
Initialization: Initialize D3QN $Q_{D3}(s,a|\boldsymbol{\theta}_{D3})$ and its target network $Q_{D3}(s,a|\boldsymbol{\theta}_{D3}^{-})$, with $\boldsymbol{\theta}_{D3}^{-} \leftarrow \boldsymbol{\theta}_{D3}$. Initialize TD3, including actor network $\mu(s|\boldsymbol{\theta}_{\mu})$, twin critics $Q_P(s,a|\boldsymbol{\theta}_{P_{\hat{j}}})$, target actor $\mu(s|\boldsymbol{\theta}_{\mu}^{-})$ and twin target critics $Q_P(s,a|\boldsymbol{\theta}_{P_{\hat{j}}}^{-})$, with $\boldsymbol{\theta}_{\mu}^{-} \leftarrow \boldsymbol{\theta}_{\mu}$ and $\boldsymbol{\theta}_{P_{\hat{j}}}^{-} \leftarrow \boldsymbol{\theta}_{P_{\hat{j}}}$. Initialize D3QN and TD3 replay buffers R and R̂ with capacity D̀ and D́, respectively;
for $episode = [1, epi]$ do
    Initialize the outer environment and reset the UAV's location to $\vec{q}_u(0)$;
    for $i = [1, epo_{outer}]$ do
        Observe the outer state $\mathbf{s}_i$;
        Select the outer action $\mathbf{a}_i$, observe the available set $\breve{\mathcal{B}}_o^{\mathbf{a}_i}$ and the occupied set $\mathcal{B}_o^{\mathbf{a}_i}$;
        Randomly select a BS $\breve{b} \in \breve{\mathcal{B}}_o^{\mathbf{a}_i}$ and check its B2D pathloss type, i.e., LoS or NLoS, then initialize the inner environment;
        for $j = [1, epo_{inner}]$ do
            Observe the inner state $\hat{\mathbf{s}}_j$;
            Select and execute the inner action $\hat{\mathbf{a}}_j$, observe the next inner state $\hat{\mathbf{s}}_{j+1}$ and calculate the inner reward $\hat{\mathbf{r}}_j$;
            Store transition $(\hat{\mathbf{s}}_j, \hat{\mathbf{a}}_j, \hat{\mathbf{s}}_{j+1}, \hat{\mathbf{r}}_j)$ into R̂;
            if $card\ (\hat{R}) \geq N_P$ then
                Sample a mini-batch of $N_P$ transitions from R̂, then update the twin critic networks $Q_P(s,a|\boldsymbol{\theta}_{P_{\hat{j}}})$ via gradient descent method in (21);
                Every $N_{pud}$ times twin critics are trained, update actor network $\mu(s|\boldsymbol{\theta}_{\mu})$ via gradient ascent in (26), and update target networks $Q_P(s,a|\boldsymbol{\theta}_{P_{\hat{j}}}^{-})$ and $\mu(s|\boldsymbol{\theta}_{\mu}^{-})$, following Polyak averaging rule in (27) and (28), respectively;
        Execute the outer action $\mathbf{a}_i$, observe the next outer state $\mathbf{s}_{i+1}$ and calculate the outer reward $\mathbf{r}_i$;
        Store transition $(\mathbf{s}_i, \mathbf{a}_i, \mathbf{s}_{i+1}, \mathbf{r}_i)$ into R;
        if $card\ (R) \geq N_{D3}$ then
            Sample a mini-batch of $N_{D3}$ transitions from R, update D3QN network $Q_{D3}(s,a|\boldsymbol{\theta}_{D3})$ via gradient descent in (16);
            Update the D3QN target network $Q_{D3}(s,a|\boldsymbol{\theta}_{D3}^{-})$ every $\Upsilon_{D3}$ steps, i.e., $\boldsymbol{\theta}_{D3}^{-} \leftarrow \boldsymbol{\theta}_{D3}$;
    Update $\epsilon \leftarrow \epsilon \times dec_{\epsilon}$ and $\sigma_P^2 \leftarrow \sigma_P^2 \times dec_{\sigma}$;
```

Figure 7: Pseudocode of the proposed hybrid D3QN-TD3 solution

Workflow Diagram of the Proposed Hybrid D3QN-TD3 Solution

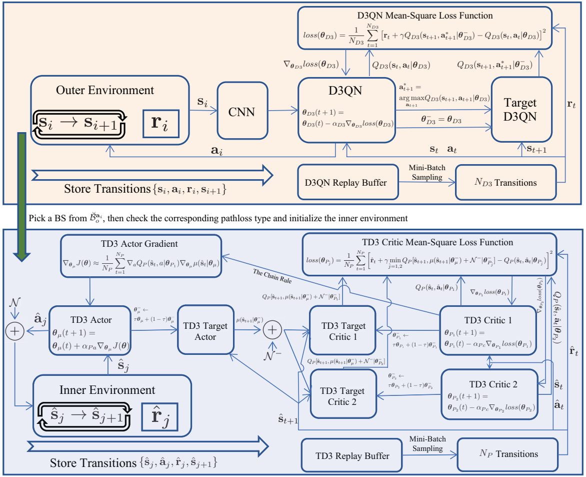

D3QN Mean-Square Loss Function
$loss(\boldsymbol{\theta}_{D3}) = \frac{1}{N_{D3}} \sum_{t=1}^{N_{D3}} [\mathbf{r}_t + \gamma Q_{D3}(\mathbf{s}_{t+1}, \mathbf{a}^*_{t+1}|\boldsymbol{\theta}^-_{D3}) - Q_{D3}(\mathbf{s}_t, \mathbf{a}_t|\boldsymbol{\theta}_{D3})]^2$
$\nabla_{\boldsymbol{\theta}_{D3}} loss(\boldsymbol{\theta}_{D3})$
$Q_{D3}(\mathbf{s}_t, \mathbf{a}_t|\boldsymbol{\theta}_{D3})$
$Q_{D3}(\mathbf{s}_{t+1}, \mathbf{a}^*_{t+1}|\boldsymbol{\theta}^-_{D3})$
Outer Environment
$\mathbf{s}_i \rightarrow \mathbf{s}_{i+1}$
$\mathbf{r}_i$
$\mathbf{s}_i$
CNN
D3QN
$\boldsymbol{\theta}_{D3}(t+1) =$
$\boldsymbol{\theta}_{D3}(t) - \alpha_{D3} \nabla_{\boldsymbol{\theta}_{D3}} loss(\boldsymbol{\theta}_{D3})$
$\mathbf{a}^*_{t+1} =$
$\arg\max_{\mathbf{a}_{t+1}} Q_{D3}(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}|\boldsymbol{\theta}_{D3})$
$\boldsymbol{\theta}^-_{D3} = \boldsymbol{\theta}_{D3}$
Target D3QN
$\mathbf{r}_t$
$\mathbf{a}_i$
$\mathbf{s}_t$ $\mathbf{a}_t$
$\mathbf{s}_{t+1}$
Store Transitions $\{\mathbf{s}_i, \mathbf{a}_i, \mathbf{r}_i, \mathbf{s}_{i+1}\}$
D3QN Replay Buffer
Mini-Batch Sampling
N_{D3} Transitions
Pick a BS from $\mathcal{B}_o^{\mathbf{a}_i}$, then check the corresponding pathloss type and initialize the inner environment
TD3 Actor Gradient
$\nabla_{\boldsymbol{\theta}_\mu} J(\boldsymbol{\theta}) \approx \frac{1}{N_P} \sum_{t=1}^{N_P} \nabla_a Q_P(\hat{\mathbf{s}}_t, a|\boldsymbol{\theta}_{P_1}) \nabla_{\boldsymbol{\theta}_\mu} \mu(\hat{\mathbf{s}}_t|\boldsymbol{\theta}_\mu)$
TD3 Critic Mean-Square Loss Function
$loss(\boldsymbol{\theta}_{P_j}) = \frac{1}{N_P} \sum_{t=1}^{N_P} \left[\hat{\mathbf{r}}_t + \gamma \min_{\hat{j}=1,2} Q_P[\hat{\mathbf{s}}_{t+1}, \mu(\hat{\mathbf{s}}_{t+1}|\boldsymbol{\theta}^-_\mu) + \mathcal{N}^-|\boldsymbol{\theta}^-_{P_{\hat{j}}}] - Q_P(\hat{\mathbf{s}}_t, \hat{\mathbf{a}}_t|\boldsymbol{\theta}_{P_j})\right]^2$
The Chain Rule
$\mathcal{N}$
$\hat{\mathbf{a}}_j$
TD3 Actor
$\boldsymbol{\theta}_\mu(t+1) =$
$\boldsymbol{\theta}_\mu(t) + \alpha_{Pa} \nabla_{\boldsymbol{\theta}_\mu} J(\boldsymbol{\theta})$
$\boldsymbol{\theta}^-_\mu \leftarrow$
$\tau\boldsymbol{\theta}_\mu + (1-\tau)\boldsymbol{\theta}^-_\mu$
TD3 Target Actor
$\mu(\hat{\mathbf{s}}_{t+1}|\boldsymbol{\theta}^-_\mu)$
$\mathcal{N}^-$
$Q_P[\hat{\mathbf{s}}_{t+1}, \mu(\hat{\mathbf{s}}_{t+1}|\boldsymbol{\theta}^-_\mu) + \mathcal{N}^-|\boldsymbol{\theta}^-_{P_1}]$
$Q_P(\hat{\mathbf{s}}_t, \hat{\mathbf{a}}_t|\boldsymbol{\theta}_{P_1})$
$\nabla_{\boldsymbol{\theta}_{P_1}} loss(\boldsymbol{\theta}_{P_1})$
$\nabla_{\boldsymbol{\theta}_{P_2}} loss(\boldsymbol{\theta}_{P_2})$
$Q_P(\hat{\mathbf{s}}_t, \hat{\mathbf{a}}_t|\boldsymbol{\theta}_{P_2})$
TD3 Target Critic 1
TD3 Critic 1
$\boldsymbol{\theta}_{P_1}(t+1) =$
$\boldsymbol{\theta}_{P_1}(t) - \alpha_{Pc} \nabla_{\boldsymbol{\theta}_{P_1}} loss(\boldsymbol{\theta}_{P_1})$
$\boldsymbol{\theta}^-_{P_1} \leftarrow$
$\tau\boldsymbol{\theta}_{P_1} + (1-\tau)\boldsymbol{\theta}^-_{P_1}$
$\hat{\mathbf{r}}_t$
$\hat{\mathbf{s}}_j$
$Q_P[\hat{\mathbf{s}}_{t+1}, \mu(\hat{\mathbf{s}}_{t+1}|\boldsymbol{\theta}^-_\mu) + \mathcal{N}^-|\boldsymbol{\theta}^-_{P_2}]$
Inner Environment
$\hat{\mathbf{s}}_j \rightarrow \hat{\mathbf{s}}_{j+1}$
$\hat{\mathbf{r}}_j$
TD3 Target Critic 2
$\boldsymbol{\theta}^-_{P_2} \leftarrow$
$\tau\boldsymbol{\theta}_{P_2} + (1-\tau)\boldsymbol{\theta}^-_{P_2}$
TD3 Critic 2
$\boldsymbol{\theta}_{P_2}(t+1) =$
$\boldsymbol{\theta}_{P_2}(t) - \alpha_{Pc} \nabla_{\boldsymbol{\theta}_{P_2}} loss(\boldsymbol{\theta}_{P_2})$
$\hat{\mathbf{s}}_t$
$\hat{\mathbf{a}}_t$
$\hat{\mathbf{s}}_{t+1}$
Store Transitions $\{\hat{\mathbf{s}}_j, \hat{\mathbf{a}}_j, \hat{\mathbf{r}}_j, \hat{\mathbf{s}}_{j+1}\}$
TD3 Replay Buffer
Mini-Batch Sampling
N_P Transitions

Figure 8: Workflow diagram

Example of Offline Exploitation

Note that it is conducted on a central coordinator (C2) that controls BSs via high-speed fronthaul links and is genuinely an online-centralized-learning-and-offline-decentralized-execution algorithm, for realizing its efficient implementation without introducing heavy burden of computations or excessive delays or overheads of information transfer, e.g., imperfect CSIs and designed beamforming vectors between the C2 and available BSs, during its exploitation for RRM.

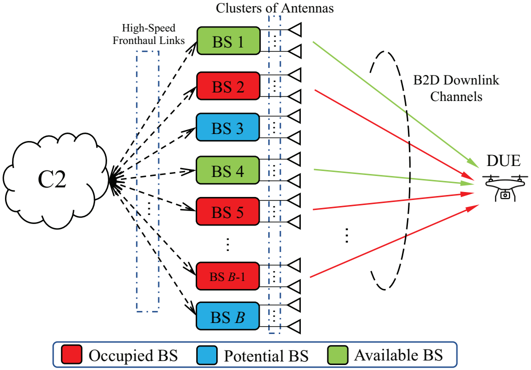
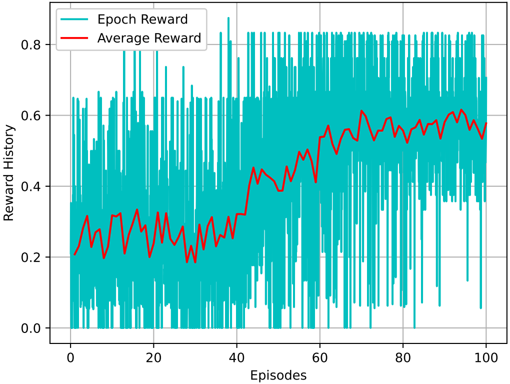
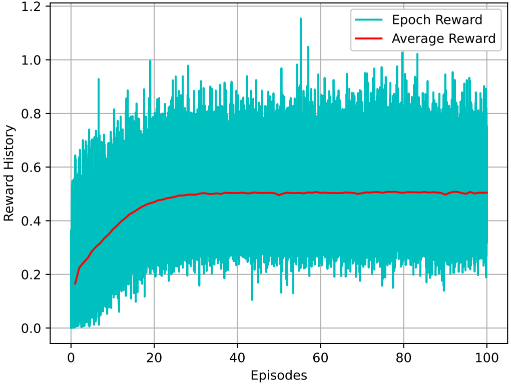


Figure 9: Offline exploitation

Simulation Results



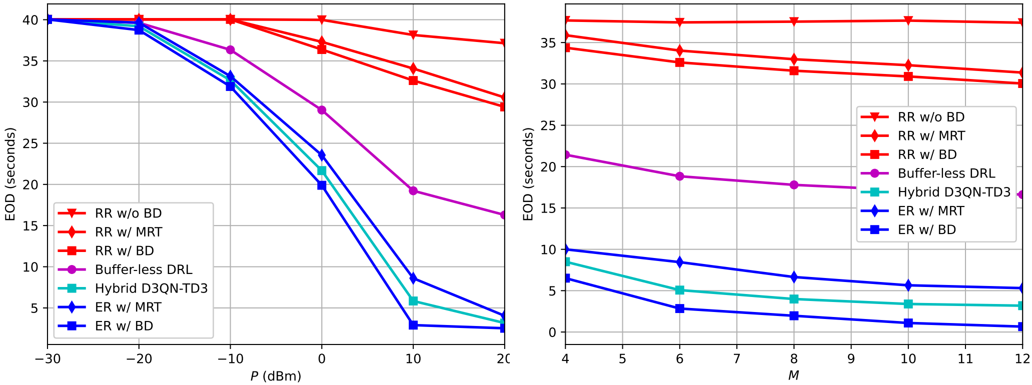
(a) Reward history of D3QN



(b) Reward history of TD3

Figure 10: Reward history

Simulation Results



(a) Performance comparison vs P

(b) Performance comparison vs M

Figure 11: Performance comparison

The End

Thanks for your attentions

This is the end of today's demonstration