Deep Reinforcement Learning-Assisted Age-optimal Transmission Policy for HARQ-aided NOMA Networks

Kunpeng Liu, Aimin Li, *Graduate Student Member, IEEE,* Shaohua Wu, *Member, IEEE*

Abstract—The recent interweaving of AI-6G technologies has sparked extensive research interest in further enhancing reliable and timely communications. *Age of Information* (AoI), as a novel and integrated metric implying the intricate trade-offs among reliability, latency, and update frequency, has been well-researched since its conception. This paper contributes new results in this area by employing a Deep Reinforcement Learning (DRL) approach to intelligently decide how to allocate power resources and when to retransmit in a *freshness-sensitive* downlink multi-user Hybrid Automatic Repeat reQuest with Chase Combining (HARQ-CC) aided Non-Orthogonal Multiple Access (NOMA) network. Specifically, an AoI minimization problem is formulated as a Markov Decision Process (MDP) problem. Then, to achieve deterministic, age-optimal, and intelligent power allocations and retransmission decisions, the Double-Dueling-Deep Q Network (DQN) is adopted. Furthermore, a more flexible retransmission scheme, referred to as Retransmit-At-Will scheme, is proposed to further facilitate the timeliness of the HARQ-aided NOMA network. Simulation results verify the superiority of the proposed intelligent scheme and demonstrate the threshold structure of the retransmission policy. Also, answers to whether user pairing is necessary are discussed by extensive simulation results.

Index Terms—Age of information, HARQ, NOMA, reinforcement learning, power allocation.

I. INTRODUCTION

With the rapid development of the 5th Generation Mobile Communication (5G), more and more Internet of Things (IoT) devices are accessed to the networks [1]. In the IoT application scenarios, such as automatic driving, smart factory, and smart healthcare, etc., the timeliness of information plays a significant role in enabling timely, accurate, and effective decision-making. In this regard, it is imperative to deliver *fresh* status updates in a timely and reliable manner, since the *stale* one usually contains little value.

To measure the timeliness of information, *Age of Information* (AoI) is first proposed in [2] for automated driving applications. AoI is defined as the time elapsed since the generation of the latest status which is received successfully. Different from the traditional metric in the 5G ultra-reliable-low-latency (URLLC), such as latency and throughput, AoI contemplates an intricate trade-off among reliability, latency, and update frequency. As such, AoI is also known as a type of semantic metrics, and has attracted massive interest from government agencies, industry, and academia.

With the massive access of IoT devices, the research of high timeliness access technique becomes a hot topic. Non-orthogonal multiple access (NOMA) is considered as a promising access technique for the future network due to its high spectral efficiency [3]. In [4] and [5], the authors have shed light on the potential superior AoI performance in the high signal-to-noise ratio (SNR) regime. However, in the low SNR regime, the age performance of NOMA system inexorably faces a bottleneck.

To this end, HARQ is introduced in [6] to ensure reliability in the low SNR regime by achieving retransmission-driven time diversity. Up to this point, the HARQ-aided NOMA network has demonstrated its superiority in both throughput and latency [7] and [8], while the age performance of HARQ-aided NOMA network is seldom investigated. An exception is our previous work in [9], which analyzes the average blocklength error ratio (BLER) of users in the HARQ-aided NOMA network, and provides an optimal transmission policy by formulating an MDP problem. However, the theory and analysis in [9] limited the number of users and the maximum transmission round to 2 only, which hinders flexible and timely system design.

In the multi-user NOMA network, user paring is a commonly-implemented technique to simplify the transmission mode. [10] has demonstrated that in terms of throughput performance, the optimal pairing policy manifests a strongest-weakest symmetrical structure, i.e., the user with the best channel condition is paired with the user with the worst channel condition sequentially, so as to accomplish the pairing of all users. Nevertheless, the insight on whether user pairing will benefit the timeliness performance of the NOMA-based network remains an open issue.

Besides, DRL algorithms are widely used in wireless communication to solve high-complexity problems. Reinforcement on Federated (RoF) scheme, based on deep multi-agent reinforcement learning is proposed in [11]. In [12], authors adopt TD3 algorithm to allocate the available resources for IoT.

Motivated by the above, we aim to investigate the optimal policy to minimize the average AoI in an HARQ-aided NOMA network with multiple users and unrestricted transmission rounds. Specifically, this paper achieves three-fold contribu-
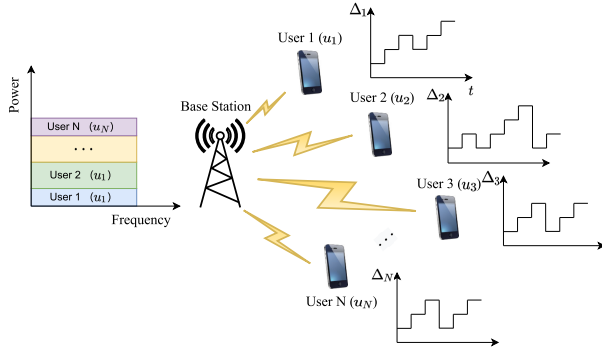
Fig. 1. HARQ-aided NOMA Network Model.

tions:

- We optimize the average AoI of an HARQ-aided NOMA network with multiple users and unrestricted transmission rounds. A Retransmit-At-Will policy that allows BS to retransmit flexibly is proposed, which manifests superiority in terms of AoI compared to existing works.
- We reformulate the age-optimal problem as an MDP. A Double-Dueling-DQN is trained to intelligently allocate power resources and retransmit at will. The solutions reveal the threshold structure of the retransmission policy.
- We investigate the user pairing issue in terms of AoI under the HARQ-aided NOMA network. The simulation results illustrate that user pairing can only enhance the AoI at low SNR.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. Transmission Modle of HAQR-aided NOMA Network

We consider a NOMA downlink network with a BS and N users $(u_i, i \in \{1, \cdots, N\})$ as Fig. 1, the distances from users to the BS are ordered as $d_1 < d_2 < d_3 < \cdots < d_N$. The BS sends freshness critical control commands or status updates to the users on the same shared channel. In a NOMA downlink network, the BS sends out multiple user signals by superposition coding (SC) technique in the power domain, and HARQ-CC is applied as a key technique to enhance transmission reliability by multiple retransmissions. Then, the signal sent by the BS in the $k-$th transmission can be expressed as

$$s_k = \sum_{i=1}^{N} \sqrt{\alpha_{i,k} P} x_{i,k},$$ (1)

where P denotes the total power of a single transmission, $\alpha_{i,k}$ and $x_{i,k}$ denotes the power allocation coefficient and the unit energy signal of user i in the $k-$th transmission. Since the receiver end of the NOMA network needs to decode superposed signals through successive interference cancellation (SIC), that is, a strong user with high channel quality needs to decode and eliminate weak users' signals iteratively, and then decodes its own signal. Generally, in order to improve the reliability of SIC, the power allocation coefficients satisfy $\alpha_{1,k} < \alpha_{2,k} < \cdots < \alpha_{N,k}$, $\sum_{i=1}^{N} \alpha_{i,k} = 1$, which indicates that weaker users are allocated with higher power.

We consider a wireless fading channel where the signal will experience quasi-static Rayleigh fading. The perfect channel state information (CSI) is assumed to be shared with the BS. Then the channel coefficient in the $k-$th transmission from user u_i to BS can be expressed as

$$h_{i,k} = \sqrt{d_i^{-\tau}} g,$$ (2)

where $g \sim \mathcal{CN}(0,1)$ denotes the unit Rayleigh channel coefficient and τ denotes the path loss exponent. In such a case, the signal received by u_i in the $k-$th transmission is given by:

$$y_{i,k} = h_{i,k} \sum_{i=1}^{N} \sqrt{\alpha_{i,k} P} x_{i,k} + n_{i,k},$$ (3)

where $n_{i,k}$ denotes the unit complex additive white Gaussian noise (AWGN) with zero mean and variance $\sigma^2 = 1$.

B. AoI Evolution of HARQ-aided NOMA Network

Noticing that as the number of NOMA users increases, SIC decoding becomes more complex, which makes it harder for users to decode successfully. To this end, we combine NOMA with HARQ-CC to ensure reliable reception of updates through retransmitting the former package. In HARQ-CC protocol, if u_i decodes its message successfully, a positive acknowledgment (ACK) will be sent to the BS, otherwise, a negative acknowledgment (NACK) will be sent. Different from [9], where i) the BS can choose whether retransmit or not only when it receives an NACK; ii) the BS can only transmit a new package if it receives an ACK. We propose a Retransmit-At-Will policy, which further relaxes these constraints, that is the BS can choose whether retransmit or not no matter what it receives.

We consider a time-slotted system where the continuous time is discretized into time slots[1]. The BS serves all the users through the *generate-at-will* model, which means the BS can generate packages for all the users at the beginning of every time slot. We adopt AoI as a metric to measure the timeliness of the information received by users. Let $\Delta_i(t)$ denote the instantaneous AoI of u_i at the time slot t, then if u_i receives a status update successfully, $\Delta_i(t)$ will decrease to the instantaneous AoI of this package, while $\Delta_i(t)$ will increase by 1 if u_i fails to decode its message. We denote $t_i, t_i^{'} \in \{1, 2, \cdots, T_{max}\}$ as the transmission round of u_i's package at the time slot $t, t+1$ respectively, which can also represent the instantaneous AoI of u_i's package, and T_{max} represents the maximum transmission rounds. Then the AoI evolution of u_i between two adjacent time slots can be expressed as

$$\Delta_i(t+1) = \begin{cases} \Delta_i(t) + 1, & w.p. \ \varepsilon_i \\ t_i^{'}, & w.p. \ 1 - \varepsilon_i \end{cases},$$ (4)

where ε_i denotes the BLER of u_i.

Noticing that if the BS retransmits the u_i's package, the transmission round t_i will increase by 1, and if the BS

[1]This is an extensively-implemented operation, which is also conducted in [13], [14]
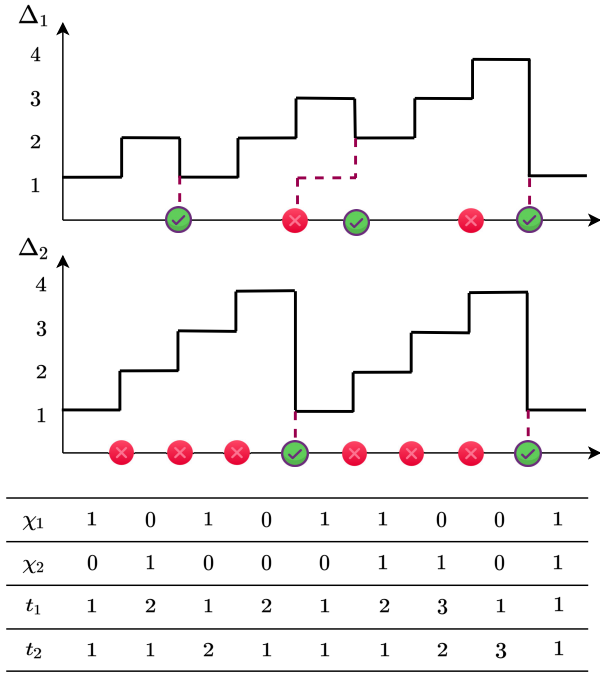
Fig. 2. A two-user example of instantaneous AoI evolution.

transmits a new package for u_i, t_i will be set as 1. Therefore, we can obtain:

$$t_i' = \begin{cases} t_i + 1, & \chi_i = 1 \\ 1, & \chi_i = 0 \end{cases}, \tag{5}$$

where $\chi_i = 1$ represents retransmitting the former u_i's package, and $\chi_i = 0$ represents transmitting a new package for u_i.

Fig. 2 gives the AoI evolution of the HARQ-aided NOMA network with two users as an example.

C. BLER analysis

In the $k-$th transmission, only the weakest user u_N can decode its message $x_{N,k}$ from $y_{N,k}$ directly. Other users $u_i(i \neq N)$ must use SIC technique to decode their message, which means u_i should decode and eliminate $x_{j,k}(i \leqslant j \leqslant N)$ from $y_{i,k}$ first, and then decode its own message $x_{i,k}$. In this case, the signal-to-interference plus noise ratio (SINR) at u_i for decoding x_j in the $k-$th transmission can be given as:

$$\gamma_{ij}^{(k)} = \frac{\alpha_{j,k} |h_{i,k}|^2}{\sum_{n=1}^{j-1} \alpha_{n,k} |h_{i,k}|^2 + 1/\rho} \ (i \leqslant j \leqslant N), \tag{6}$$

where $\rho = P/\sigma^2$ denotes the SNR.

In the HARQ-aided NOMA network, users can store the transmitted packages. When these packages are retransmitted, the stored information can be used to achieve maximal ratio combining (MRC) decoding to improve transmission reliability. By this means, the SINR of u_i for decoding the message of u_j in the t_j-th transmission round can be given as:

$$\gamma_{ij}(t_j) = \sum_{k=1}^{t_j} \gamma_{ij}^{(k)}, \tag{7}$$

where $t_j \in \{1, 2, \cdots, T_{max}\}$.

In [15], the BLER of u_i for decoding u_j's message which has been transmitted for t_j times can be approximated as:

$$\varepsilon_{ij}(t_j) \approx Q\left(\frac{C(\gamma_{ij}(t_j)) - \frac{N}{m}}{\sqrt{V(\gamma_{ij}(t_j))/m}} \right), \tag{8}$$

where N denotes the length of message sent to u_j, m denotes the number of encoded symbols, $C(\gamma_{ij}(t_j)) = \log_2(1 + \rho)$ is the channel capacity, and $V(\gamma_{ij}(t_j))$ is the channel dispersion, given as $V(\gamma_{ij}(t_j)) = (1 - 1/(1 + \gamma_{ij}(t_j))^2)(\log_2 e)^2$.

Denote a vector $\widetilde{\boldsymbol{\alpha_k}} \triangleq (\alpha_{1,k}, \alpha_{2,k}, \cdots, \alpha_{N,k})$ as the allocated power coefficients for N users at the $k-$th transmission. According to equation (6)(7)(8), when given a fixed N and m, $\varepsilon_{ij}(t_j)$ is equivalent to a function of t_j vectors. In this case, $\varepsilon_{ij}(t_j)$ can be expressed as:

$$\varepsilon_{ij}(t_j) \approx \psi_{ij}\left(\widetilde{\boldsymbol{\alpha_1}}, \widetilde{\boldsymbol{\alpha_2}}, \cdots, \widetilde{\boldsymbol{\alpha_{t_j}}} \right), \tag{9}$$

According to the principle of SIC, only when the user successfully decodes and eliminates other interference signals, and then successfully decodes its own signal, it is a successful decoding process. In this case, the BLER of u_i can be given as:

$$\varepsilon_i = 1 - \prod_{j=i}^{N} (1 - \varepsilon_{ij}(t_j)), \tag{10}$$

which means the BLER of u_i is associated with the transmission round of u_j's signal $t_j(i \leqslant j \leqslant N)$ and the corresponding power allocation coefficient vector $\widetilde{\boldsymbol{\alpha_k}}$.

D. Problem Formulation

In the HARQ-aided NOMA network, retransmissions can enhance the reliability but cause information aging. Besides, how to allocate power resources has a significant impact on the system timeliness. Therefore, it's necessary for the BS to adjust transmission policy π according to the state of the network to enhance the timeliness.

We use the expected weighted sum AoI of all the users to measure the AoI performance of the HARQ-aided NOMA network. To minimize the expected weighted sum AoI, the problem can be formulated as follows:

$$\arg\min_{\pi} \bar{\Delta} = \lim_{T_{slot} \to \infty} \frac{1}{T_{slot}} \sum_{t=1}^{T_{slot}} \sum_{i=1}^{N} \mathbb{E}[\omega_i \Delta_i^{\pi}(t)], \tag{11}$$

where ω_i denotes the weight of u_i. The transmission policy π contains two aspects: i) the power allocation scheme $\widetilde{\boldsymbol{\alpha}}$ at every time slot; ii) when and whether to retransmit the former package.

III. MDP FORMULATION

To obtain the age-optimal transmission policy for the HARQ-aided NOMA network, we reformulate the problem (11) as an MDP problem, which is characterized by a quaternary tuple $\{S, A, P, r\}$, where S, A, P, r denotes the state space, action space, state-action transition probability matrix, and the reward respectively. The elements of the MDP for HARQ-aided NOMA network are described specifically as follows:

- **State Space:** We denote s_t as the state of the network at the time slot t, which is characterized by a tuple $s_t \triangleq \left(\widetilde{\alpha_1}, \widetilde{\alpha_2}, \cdots, \widetilde{\alpha_{T_{max}-1}}, \mathcal{T}, \Delta \right)$, where
  1) The coefficient vector $\widetilde{\alpha_i} \triangleq (\alpha_{1,i}, \alpha_{2,i}, \cdots, \alpha_{N,i})$ denotes the $i-$th power allocation coefficient for all the N users which is stored in the buffer. We discretize the power into M levels, and the power coefficient can only take value from a discrete set, that is $\alpha_{i,k} \in \{\frac{1}{M}, \frac{2}{M}, ..., \frac{M-1}{M}\}$. Due to the computational complexity, the case $\alpha_{i,k} = 1$ which means BS serves only one user is not taken into consideration.
  2) The vector $\mathcal{T} \triangleq (t_1, t_2, \cdots, t_N)$, where t_j denotes the transmission round of u_j's signal at the time slot t, with $t_j \in \{1, 2, \cdots, T_{max}\}$.
  3) The AoI vector $\Delta \triangleq (\Delta_1, \Delta_2, \cdots, \Delta_N)$, where Δ_i denotes the instantaneous AoI of u_i at the time slot t.

- **Action Space:** The action $a_t \in \mathcal{A}$ contains two aspects: i): the power allocation policy: denote $\widetilde{\alpha} \triangleq (\alpha_1, \alpha_2, \cdots, \alpha_N)$ as the power coefficient allocated by the BS to all the N users at the time slot t, where $\alpha_i \in \{\frac{1}{M}, \frac{2}{M}, ..., \frac{M-1}{M}\}$; ii): the retransmission policy χ_i. Noticing that when $t_i = T_{max}$, the BS will only transmit a new package for u_i, that is $\chi_i = 0$.

- **Reward:** We denote the instantaneous expected weighted sum AoI as the reward r_t at the time slot t, which can be expressed as $r_t = \sum_{i=1}^{N} \omega_i \Delta_i$.

- **Transition Probability:** We denote $P(s_{t+1} \mid s_t, a_t)$ as the transition probability from state s_t to state s_{t+1} when taking the action a_t. The transition probability can be expressed as the following three parts:
  1) Power allocation coefficient:
  $$\begin{cases} \widetilde{\alpha_i}' = \widetilde{\alpha_{i+1}}, 1 \leq i \leq T_{max} - 2 \\ \widetilde{\alpha_{T_{max}-1}}' = \widetilde{\alpha} \end{cases}, \quad (12)$$

  where $\widetilde{\alpha_i}'$ denotes the $i-$th power coefficient vector which is stored in the buffer at the time slot $t+1$. At each iteration, we store the BS's power allocation scheme $\widetilde{\alpha}$ in the buffer as $\widetilde{\alpha_{T_{max}-1}}$, and then update $\widetilde{\alpha_i}$ into $\widetilde{\alpha_{i+1}}$. In this way, the buffer can always maintain $T_{max}-1$ freshest power coefficient vectors for MRC decoding.
  2) AoI evolution: the transition of AoI is given as (4). To get ε_i at the time slot $t+1$, we should first obtain $\varepsilon_{ij}(t_j')$, according to (10). Noticing that the MRC decoding for a t_j' times transmitted signal will use the power coefficient scheme $\widetilde{\alpha}$ and the freshest $t_j' - 1$ power coefficient vectors in the buffer. Combined with (9), we can get:
  $$\varepsilon_{ij}\left(t_j'\right) = \begin{cases} \psi_{ij}\left(\widetilde{\alpha}\right), & t_j' = 1 \\ \psi_{ij}\left(\widetilde{\alpha_{T_{max}-1}}, \widetilde{\alpha}\right), & t_j' = 2 \\ \cdots\cdots \\ \psi_{ij}\left(\widetilde{\alpha_1}, \cdots, \widetilde{\alpha_{T_{max}-1}}, \widetilde{\alpha}\right), & t_j' = T_{max} \end{cases}. \quad (13)$$

  3) Transmission round: the transition of transmission round is given as (5).

IV. PROPOSED DOUBLE-DUELING-DQN ALGORITHM

In this section, we first describe the idea of DQN, and then utilize the Double-Dueling-DQN algorithm to solve the formulated MDP.

A. Preliminaries of DQN

The main idea of reinforcement learning (RL) is to find the optimal decision policy by maximizing the long-term discounted return. We introduce the action-value function (Q-function) to measure the value under policy π by taking action a_t in a given state s_t, which can be expressed as:

$$Q^\pi(s_t, a_t) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k \cdot r_{t+k} | \pi \right], \quad (14)$$

where $\gamma \in [0, 1]$ represents the discounted factor. Then, the value function which measures the value starting from the state s_t under policy π can be expressed as:

$$V^\pi(s_t) = \mathbb{E}_{a_t}[Q^\pi(s_t, a_t)]. \quad (15)$$

Noticing that the Q-function evaluates the value of all actions, if we obtain Q^* which is the Q-function under the optimal policy π^*, the optimal action can be selected based on the Q-value, therefore, π^* can be obtained by:

$$\pi^* = \arg\max_{a_t \in \mathcal{A}} Q^*(s_t, a_t). \quad (16)$$

As a commonly used RL algorithm, Q-learning is precisely designed to obtain Q^*. Q-learning updates Q^π iteratively in the following way until it converges to Q^*.

$$Q^\pi(s_t, a_t) = Q^\pi(s_t, a_t) + \lambda(y_t - Q^\pi(s_t, a_t)), \quad (17)$$

where $y_t = r_t + \gamma\max_{a_{t+1}}Q^\pi(s_{t+1}, a_{t+1})$ is the temporal difference target (TD target), and λ denotes the learning rate.

However, Q-learning is difficult to converge when solving high-complexity problems. To this end, DQN algorithm combines neural networks with Q-learning to solve the high computational complexity problem. DQN is updated in a similar way to Q-learning, except that a neural network $Q^\pi(s_t, a_t; \theta)$ is used to approximate the Q-function, and the network parameter θ is updated by stochastic gradient descent method for each update.

B. Double-Dueling-DQN Algorithm

We adopt the Double-Dueling-DQN algorithm which is an improved version of DQN to deal with the problem (11).

In an interaction, the agent observes the current transmission state s_t of the HARQ-aided NOMA network and selects an action a_t by ϵ-*Greedy Policy*. After the NOMA network takes action a_t, the users' AoI and the transmission state will change. Then, the NOMA network returns a reward r_t and the next state s_{t+1}, which are in terms of a_t. The ϵ-*Greedy Policy* is given as:

$$a_t = \begin{cases} random\ action, & w.p.\ \epsilon \\ \arg\max_a Q^\pi(s_t, a; \theta), & w.p.\ 1 - \epsilon \end{cases}, \quad (18)$$

where ϵ denotes the exploration probability. This policy makes the agent inclined to explore more actions in the early stage of the algorithm, which can avoid the algorithm from converging to a local optimum. With continuous iterations, we gradually set $\epsilon \to 1$ to obtain a stable optimal solution.

After each round of interaction, we put (s_t, a_t, r_t, s_{t+1}) into the experience replay buffer $\mathcal{D}$. If the size of experiences in $\mathcal{D}$ exceeds a certain amount σ, a mini-batch experiences $\mathcal{B} = \{(s_t, a_t, r_t, s_{t+1})\}$ will be sampled from $\mathcal{D}$ for updating the network parameter during the training stage. This technique is called experience replay, which can improve the utilization of training data and reduce the correlation of adjacent samples.

In the training stage, to solve the overestimation problem of traditional DQN, two networks are used for training in our algorithm. The main network parameter is θ, which is used to approximate the Q-function, and the target network parameter is θ^-, which is used to calculate the TD target to update the main network. The TD target can be expressed as:

$$y_t^{target} = r_t + \gamma \max_{a_{t+1}} Q^\pi(s_{t+1}, a_{t+1}; \theta^-). \tag{19}$$

The main network updates the parameters θ by minimizing the loss function as follows at each interaction:

$$L(\theta) = \frac{1}{|\mathcal{B}|} \sum_{(s_t, a_t, r_t, s_{t+1}) \in \mathcal{B}} \left[y_t^{target} - Q(s_t, a_t; \theta) \right]^2. \tag{20}$$

The target network parameter θ^- is synchronized to the main network parameter θ every l updates.

Besides, we use dueling network architecture to improve the algorithm performance, which adopts two networks to approximate the Q-function. We approximate the value function $V^\pi(s_t)$ and the advantage function $A^\pi(s_t, a_t)$ with the same parameter, where $A^\pi(s_t, a_t) = Q^\pi(s_t, a_t) - V^\pi(s_t)$. Then, The Q-function can be obtained by these two networks, which can be expressed as:

$$Q^\pi(s_t, a_t; \theta) = V^\pi(s_t; \theta) - \operatorname*{mean}_{a_t} A^\pi(s_t, a_t; \theta). \tag{21}$$

Details of the training process are given in Algorithm 1.

V. SIMULATION RESULTS

In this section, we take a 4-users ($N = 4$) scenario as an example to analyze the timeliness performance of the HARQ-aided NOMA network. We set the distance from the user to the BS as $d_1 = 1.5$, $d_2 = 2$, $d_3 = 2.5$, $d_4 = 3$, the path loss exponent as $\tau = 2$, the message length as $N = 160$, the number of symbols as $m = 200$, and the power level as $M = 20$. The training parameters are set as $E = 500$, $T_{slot} = 10^3$, $\epsilon = 0.1$, $\eta = 0.99$, $\sigma = 200$, $\lambda = 10^{-3}$, $k = 4$, $l = 100$.

Fig. 3 compares the Retransmit-At-Will policy with the policy in [9]. Also, a non-optimal policy is given as a benchmark where the BS only transmits new packages with a fixed power allocation scheme. Retransmit-At-Will policy achieves the best AoI performance at low SNR, while two optimal policies perform similarly at high SNR. This is because retransmission is permitted when the BS received an ACK in the Retansmit-At-Will policy, which can enhance the transmission reliability at low SNR. However, since transmission is reliable enough at high SNR, the system will not choose this retransmission because it will cause stale delivery.

Algorithm 1: Double-Dueling-DQN Algorithm

Input: $E, T_{slot}, \epsilon, \eta, \sigma, \lambda, k, l$;
Output: The optimal Q-network $Q^*(s, a; \theta)$;

1 Initialize the network parameter θ and $\theta-$;
2 Initialize the state of HARQ-aided NOMA system s_0;
3 **for** $episode = 1$ *to* E **do**
4   Reset the system state $s_t = s_0$;
5   **for** $t = 1$ *to* T_{slot} **do**
6   Observe the system state s_t and calculate the Q value of all actions $Q(s_t, \cdot; \theta)$;
7   Select a action a_t according to the $\epsilon-greedy$ *policy* (18);
8   Update $\epsilon = \eta\epsilon$;
9   The BS selects action a_t, and gets the next state s_{t+1} and the reward r_t;
10   Store (s_t, a_t, r_t, s_{t+1}) in the replay buffer $\mathcal{D}$;
11   **if** *size of experiences* $> \sigma$ *and* $t \% k = 0$ **then**
12   Sample a mini-batch $\mathcal{B}$ from $\mathcal{D}$ randomly;
13   Calculate the TD target according to (19);
14   Training the main network with stochastic gradient descent method, and update θ with learning rate λ;
15   **if** $t \% l = 0$ **then**
16   Update the target network as $\theta^- = \theta$;

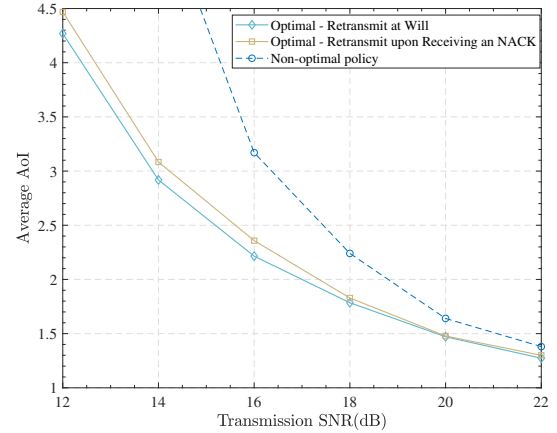17 **return** *The optimal Q-network* $Q^*(s, a; \theta)$

Fig. 3. The average AoI under different transmission policy with maximum retransmissions $T_{max} = 2$.

Fig. 4 illustrates the AoI performance under maximum retransmissions $T_{max} = 2$ and $T_{max} = 3$. The network realizes better AoI performance at low SNR when $T_{max} = 3$, but there is almost no difference under different T_{max} at high SNR. To explain this, we demonstrate the retransmission policy when $T_{max} = 3$ with the ratio of the retransmission rounds T. From the ratio, policy $T = 3$ accounts for a larger proportion at low SNR, which means the BS prefers to take more retransmissions to ensure transmission reliability. When the SNR exceeds 16 dB, the BS prefers fewer retransmissions to ensure the timeliness of transmission, and the policy struc-
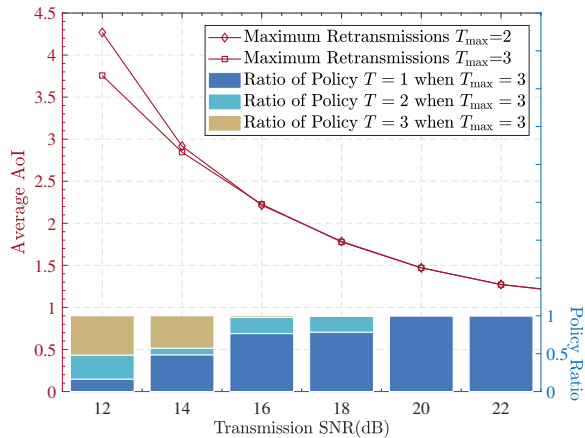
Fig. 4. The average AoI under different maximum retransmissions T_{max} and the policy structure when $T_{max} = 3$.

Fig. 5. The average AoI performance comparison of transmission with pairing and without pairing when $T_{max} = 2$.

ture of $T_{max} = 3$ is more like that of $T_{max} = 2$, so they perform similarly at high SNR. According to this, we can find that retransmission can only improve system timeliness at low SNR. When the SNR exceeds a certain threshold, the system prefers to always transmit new packages, and retransmission is meaningless for system timeliness.

To investigate the impact of user pairing on the system AoI, we set the optimal pairing scheme in [10] as a comparing benchmark. In this case, u_1 is paired with u_4, and u_2 is paired with u_3. The BS only transmits packages for one user pair at a time slot, and the other user pair can only wait at this time. Fig. 5 compares the performance of the Retransmit-At-Will policy with and without user pairing. We can find that transmission with pairing has a better AoI performance at low SNR, while transmission without pairing performs better at high SNR. That is because the user pairing technique reduces the users served simultaneously, which can reduce the complexity of SIC and ensure transmission reliability at low SNR. While reliability can be assured at high SNR, thus transmission without pairing performs better because it can serve more users simultaneously and save waiting time. As a result, user pairing can be used to enhance the timeliness at low SNR.

VI. CONCLUSION

In this paper, we optimize the AoI of the HARQ-aided NOMA network with multiple users and unrestricted retransmissions, which can intelligently decide how to allocate power resources and when to retransmit. First, a more flexible retransmission scheme with better AoI performance, referred to as Retransmit-At-Will is proposed. Then, we reformulate the age-optimal problem as an MDP, and the Double-Dueling-DQN is adopted to obtain the optimal policy. At last, a threshold structure of the retransmission policy is demonstrated, and user pairing is proved to be a technique enhancing the system freshness at low SNR.

REFERENCES

[1] T. Xia, M. M. Wang, J. Zhang, and L. Wang, "Maritime internet of things: Challenges and solutions," *IEEE Wireless Communications*, vol. 27, no. 2, pp. 188–196, 2020. I
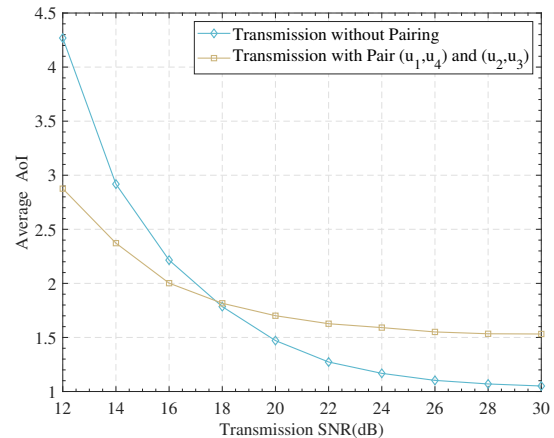
[2] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *2012 Proceedings IEEE INFOCOM*, 2012, pp. 2731–2735. I

[3] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, I. Chih-Lin, and H. V. Poor, "Application of Non-Orthogonal Multiple Access in LTE and 5G Networks," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 185–191, 2017. I

[4] J. Chen, Y. Liu, Q. Chen, X. Lan, G. Cheng, Y. Fu, and Z. Zhang, "On the Adaptive AoI-aware Buffer-aided Transmission Scheme for NOMA Networks," in *2021 IEEE Wireless Communications and Networking Conference (WCNC)*, 2021, pp. 1–6. I

[5] Q. Wang, H. Chen, Y. Li, and B. Vucetic, "Minimizing age of information via hybrid NOMA/OMA," in *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2020, pp. 1753–1758. I

[6] S. Lin, D. J. Costello, and M. J. Miller, "Automatic-repeat-request error-control schemes," *IEEE Communications Magazine*, vol. 22, no. 12, pp. 5–17, 1984. I

[7] F. Ghanami, G. A. Hodtani, B. Vucetic, and M. Shirvanimoghaddam, "Performance analysis and optimization of NOMA with HARQ for short packet communications in massive IoT," *IEEE Internet of Things Journal*, vol. 8, no. 6, pp. 4736–4748, 2020. I

[8] D. Marasinghe, N. Rajatheva, and M. Latva-Aho, "Block error performance of NOMA with HARQ-CC in finite blocklength," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2020, pp. 1–6. I

[9] S. Wu, Z. Deng, A. Li, J. Jiao, N. Zhang, and Q. Zhang, "Minimizing Age-of-Information in HARQ-CC Aided NOMA Systems," *IEEE Transactions on Wireless Communications*, 2022. I, II-B, V

[10] L. Zhu, J. Zhang, Z. Xiao, X. Cao, and D. O. Wu, "Optimal User Pairing for Downlink Non-Orthogonal Multiple Access (NOMA)," *IEEE Wireless Communications Letters*, vol. 8, no. 2, pp. 328–331, 2019. I, V

[11] W. Zhang, D. Yang, W. Wu, H. Peng, N. Zhang, H. Zhang, and X. Shen, "Optimizing Federated Learning in Distributed Industrial IoT: A Multi-Agent Approach," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 12, pp. 3688–3703, 2021. I

[12] W. Zhang, D. Yang, H. Peng, W. Wu, W. Quan, H. Zhang, and X. Shen, "Deep Reinforcement Learning Based Resource Management for DNN Inference in Industrial IoT," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 8, pp. 7605–7618, 2021. I

[13] J. P. Champati, M. H. Mamduhi, K. H. Johansson, and J. Gross, "Performance characterization using aoi in a single-loop networked control system," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2019, pp. 197–203. 1

[14] I. Krikidis, "Average age of information in wireless powered sensor networks," *IEEE Wireless Communications Letters*, vol. 8, no. 2, pp. 628–631, 2019. 1

[15] Y. Yu, H. Chen, Y. Li, Z. Ding, and B. Vucetic, "On the performance of non-orthogonal multiple access in short-packet communications," *IEEE Communications Letters*, vol. 22, no. 3, pp. 590–593, 2017. II-C