How to Minimize the Weighted Sum AoI in Multi-Source Status Update Systems: TDMA or NOMA?

Jixuan Wang and Deli Qiao, Member, IEEE

Abstract—In this paper, the minimization of the weighted sum average age of information (AoI) in a multi-source status update communication system is studied. Multiple independent sources send update packets to a common destination node in a time-slotted manner under the limit of maximum retransmission rounds. Different multiple access schemes, i.e., time-division multiple access (TDMA) and non-orthogonal multiple access (NOMA), are exploited here over a block-fading multiple access channel (MAC). Constrained Markov decision process (CMDP) problems are formulated to describe the AoI minimization problems considering both transmission schemes. The Lagrangian method is used to convert CMDP problems to unconstrained Markov decision process (MDP) problems, and corresponding algorithms are designed to derive the power allocation policies. Also, a suboptimal threshold-based policy is proposed. On the other hand, for the case of unknown environments, two online reinforcement learning approaches considering both multiple access schemes are proposed to achieve near-optimal age performance. Numerical simulations validate the improvement of the proposed policy in terms of weighted sum AoI compared to the fixed power transmission policy and illustrate that NOMA is more favorable in the case of larger packet sizes.

Index Terms—Age of information (AoI), constrained Markov decision process (CMDP), non-orthogonal multiple access (NOMA), power allocation, reinforcement learning.

I. INTRODUCTION

In many emerging real-time Internet of Things (IoT) applications [1], [2], strictly guaranteeing the timeliness of information updates is crucial, particularly for systems that deal with time-sensitive data, e.g., autonomous driving, intelligent traffic-monitoring networks, and disaster alerting systems, since outdated information might become worthless. From a system perspective, the knowledge of the status of a remote sensor or system requires to be as timely as possible, so the timeliness of state updates has evolved into a new field of network research [3]. To characterize such information timeliness and freshness, the metric termed age of information (AoI), typically defined as the time elapsed since the most recent successfully received system information was generated at the source, has been proposed [4].

Most of the earlier work on AoI in various networks mainly considered simple single-source and single-destination status update system models (see, e.g., [4], [5], [6], [7], [8]), while recent research related to AoI optimization has shifted to more practical multi-source and/or multi-destination systems, and most of them involve the orthogonal multiple access (OMA) technique [9], [10], [11], [12], [13], [14]. For instance, the authors in [9] considered a system model in which a central controller collects data from multiple sensors via wireless links, and the AoI optimization problem is subject to both bandwidth and power consumption constraints. Besides, in [9], a truncated scheduling policy was also proposed to satisfy the hard bandwidth constraint. The work in [10] presented two multi-source information update problems in a practical IoT system, called Aol-aware Multi-Source Information Updating (Aol-MSIU) and AoI-Reduction-aware Multi-Source Information Updating (AoIR-MSIU) problems, respectively. The authors in [11] investigated and compared the AoI performances of two basic multiple access schemes, time-division multiple access (TDMA) and frequency-division multiple access (FDMA). Also, in [12], the authors considered the AoI minimization problems in the uplink of an energy harvesting (EH) wireless sensor network, where TDMA and FDMA were considered as possible multiple access schemes. In multi-user multi-channel systems [13], the authors proposed two asymptotic regimes to investigate how to exploit multi-channel flexibility to improve the age performance. In [14], to minimize the expected weighted sum AoI subject to minimum throughput requirements, the authors developed four low-complexity scheduling policies and then compared their performance to the optimal policy.

Even though the literature mentioned above is all about the OMA technique, non-orthogonal multiple access (NOMA) is a promising technique to reduce the average AoI [15], [16], [17], [18], [19] by using successive interference cancellation (SIC), which will greatly improve spectral efficiency compared to OMA in a large-scale system. In [15], the authors analyzed the performance of NOMA in minimizing the AoI of a two-node uplink network for the first time, and the results showed that OMA and NOMA can outperform each other in different configurations. A hybrid NOMA/OMA scheme was proposed in [16], in which the BS can adaptively switch between NOMA and OMA for the downlink transmission to minimize the AoI.
and a suboptimal policy called action elimination with lower computation complexity was also proposed. An uplink NOMA system combining physical-layer network coding (PNC) and multiuser decoding (MUD) was considered in [18]. In [19], the authors proposed an adaptive AoI-aware buffer-aided transmission scheme (ABTS) to improve the AoI performance in the downlink NOMA system. However, these articles do not consider the retransmission mechanism, which is essential to guarantee a certain degree of reliability of the received update packets.

Furthermore, dealing with the age-optimal scheduling problems using reinforcement learning approaches in an unknown environment has recently drawn great attention [6], [20], [21], [22], [23], [24], [25], [26], [27]. And to the best of our knowledge, the first application of RL approaches to the problem with a minimum AoI criterion appeared in [6], which employed the average-cost SARSA with softmax algorithm to learn the system parameters and the transmission policy under hybrid ARQ (HARQ) protocols. As an extension of [6], in [20], the age-optimal problem was extended to a multi-user setting in the downlink system with orthogonal transmissions, and three different RL methods were proposed to provide near-optimal performance. The work in [21] proposed one off-line power control policy and two online RL algorithms to investigate the joint optimization considering both AoI and total energy consumption in the fading channel, and now we extend this work to multi-source scenarios. The Policy Gradients and Deep Q-learning (DQN) methods were introduced in [22] to address a multi-queue AoI-optimal scheduling problem. The authors in [23] proposed two algorithms, i.e., a model-based VIA relying on dynamic programming and a model-free Q-learning method, to minimize the on-demand AoI in an IoT sensing network consisting of multiple users, multiple energy harvesting sensors, and a wireless edge node. A standard model-free Q-learning algorithm to obtain the optimal policy that minimizes the long-term average AoI over an uplink mmWave channel was proposed in [24], and Q-learning was utilized to improve system AoI performance in train-to-train communications [25]. An AoI-based trajectory planning (A-TP) algorithm employing deep reinforcement learning (DRL) technique was proposed in [26] to solve the online AoI-based trajectory planning problem in UAV-assisted IoT networks. In [27], the authors utilized the DRL method to develop single-agent and cooperative multi-agent virtual network function (VNF) placements for minimizing VNF placement cost, scheduling cost, and average AoI in the industrial Internet of Things (IIoT). However, none of the multi-user system work considers the NOMA transmission scheme when using RL to solve the AoI minimization problems in an unknown environment. Motivated by this, this paper makes the first attempt, to the best of our knowledge, to propose two RL algorithms considering both TDMA and NOMA schemes in a multi-source wireless uplink status update system with an unknown environment and achieve near-optimal age performance compared to the known environment.

In this paper, we investigate the weighted sum average AoI minimization problems for both TDMA and NOMA schemes, where we consider not only the retransmission mechanism on the uplink communication system with a limit on the maximum number of retransmission rounds, but also the fact that each source is subject to an individual power constraint. On the one hand, when the channel distribution information (CDI) is available in advance, we formulate the weighted sum average AoI minimization problems as CMDP problems. Through the Lagrangian method, we relax the CMDP problems into equivalent MDP problems and accordingly obtain the off-line algorithms to derive optimal policies in TDMA and NOMA, respectively. On the other hand, we consider transmitting update packets in an unknown environment. We propose two RL-based algorithms to find the optimal power allocation policy. The main contributions of this article can be summarized as follows:

- Unlike most previous works where transmit power is fixed, we assume different transmit power levels in this paper, varying with the instantaneous AoI and transmission rounds. As a result, power allocation policies are proposed to minimize the weighted sum average AoI.
- When the CDI is available in advance, we first propose two off-line value iteration algorithms (VIAs) to solve the Bellman optimality equations for both TDMA and NOMA. Then, with the minimization of value function, the optimal policy can be derived to achieve AoI-optimal performance under the average power constraints. Besides, a suboptimal threshold-based policy is derived in closed-form.
- When such environment is not known a priori, we design two online Q-learning algorithms with $\epsilon$-greedy exploration to find the optimal power allocation policy on both TDMA and NOMA schemes. And another RL algorithm termed State-Action-Reward-State-Action (SARSA) is also adopted as a benchmark scheme.
- Numerical results show the comparison of AoI performance in TDMA and NOMA and verify that the proposed optimal policy reduces the weighted sum average AoI significantly compared to the fixed power policy. Besides, RL-based approaches are shown to achieve performance close to the optimal policy and achieve faster convergence by normalization.

The remainder of this paper is organized as follows. In Section II, we discuss the preliminaries related to the system model and AoI. Section III formulates the CMDP optimization problems. Section IV presents the details of the solutions to the optimization problems, and Section V describes reinforcement learning approaches. Numerical results are provided in Section VI. Finally, Section VII concludes the paper.

II. PRELIMINARIES

A. System Model

In this paper, we consider a multi-source wireless uplink system as depicted in Fig. 1, where $N$ sources send their status update packets to a common receiver. Time is divided into slots of period $T$. A so-called generate-at-will model is adopted, where each sensor can generate a new status update at the beginning of any time slot. Additionally, $N$ independent error-free and delay-free feedback channels from destination to each source are considered. Here we assume that an update packet of $R$ bits is assigned to be transmitted in each slot. If the sent package is successfully decoded, the receiver will
send an ACK as feedback to the source so that the source is able to generate a new package that contains the latest status information with a time stamp. In the case of decoding failure, a NACK from receiver to source will be sent to request another round of transmission of the same packet. It is worth noting that one update packet can be transmitted up to $M$ times to ensure a certain degree of reliability. Namely, when reaching the maximum number of transmissions $M$, this packet should be discarded and a new packet will be generated and sent to the destination. For simplicity, $T$ is assumed to be 1 in the paper.

In this model, we consider a block-fading MAC where the channel gain remains constant in each slot and varies independently over different time slots. We use $h(n)$ to denote the channel coefficient of the signal channel between the source $n$ and the receiver. The mean of the channel gain is denoted as $E[h^2(n)] = z(n)$, where $E[.]$ denotes the expectation. We assume that each channel is independently and identically distributed (i.i.d) and omit $(n)$ from the notation in the following for simplicity. We assume that the CSI of all links is known at the receiver, and depending on the knowledge of the CDI at the transmitters, we have two different cases, which will be detailed later. Due to the block-fading assumption, each channel is modeled as a quantized channel, where the channel gain set is denoted as $F = \{z_0, z_1, \ldots, z_K\}$ in increasing order with $z_0 = 0$ and $z_K = \infty$. $K$ is the quantization level. The channel is assumed to be in state $i$ when channel gain $z \in [z_i, z_{i+1})$. Then, the probability of channel state transition from state $j$ to state $i$ is given by

$$
\Pr\{z_i \mid z_j\} = \Pr\{z_i\} = \int_{z_i}^{z_{i+1}} p_z(z) \, dz = \psi_i, \quad (1)
$$

where $p_z(z)$ is the probability density function of the channel gain $z$ and $\psi_i$ is the probability of the channel state $i$.

We would like to note that the traditional orthogonal multiple access technologies in current wireless communication systems are often broadly classified into two types: TDMA and FDMA. It is interesting that TDMA can provide a much lower average AoI than FDMA [11]. So, in this paper, we mainly consider TDMA. Some details about TDMA and NOMA are as follows:

1) **Time-division multiple access (TDMA):** We assume that each source is assigned a single orthogonal block that is free of any interference from the others. Assume that the $n$-th ($n \in [N] \triangleq \{1, 2, \ldots, N\}$) source occupies a $\rho_n$ fraction of one time slot to transmit an update packet. We should have $\sum_{n=1}^{N} \rho_n \leq 1$ and $\rho_n \geq 0, \forall n \in [N]$. As a consequence, by dividing the time slot, every source has the opportunity to transmit their update packets without interfering with each other during a time slot. The received signal in TDMA at the receiver is given by

$$
y^T = \sum_{n=1}^{N} \sqrt{P^T_{n,i}} h_{n,i} x_n + w, \quad i = 1, 2, \ldots, K \quad (2)
$$

where $P^T_{n,i}$ is the $i$-th transmit power level of source $n$ in TDMA, $x_n$ denotes the information transmitted by source $n$, and $w$ is the additive white Gaussian noise (AWGN).

2) **Non-orthogonal multiple access (NOMA):** On the other hand, NOMA allows every source to transmit update packets simultaneously to a common receiver with non-orthogonal signaling. It is clear that some signal interference exists between the wireless communication links. When NOMA is conducted in time slot $t$, for distinguishing between the sources, the sources are allocated different power levels based on the assigned decoding order. Subsequently, their corresponding information at the receiver can be correctly recovered in one time slot through SIC. So, with the decoding order $D(\cdot)$, the received signal in NOMA at the receiver is given by

$$
y^N = \sum_{n=1}^{N} \sqrt{p^{N,D}_{D(n),i}} h_{D(n),i} x_{D(n)} + w, \quad i = 1, 2, \ldots, K \quad (3)
$$

where $p^{N,D}_{D(n),i}$ denotes the transmit power of source $D(n)$ in state $i$ on the NOMA scheme when the decoding order is $D(\cdot)$.

### B. Age of Information (AoI)

Age of information (AoI) characterizes the timeliness and freshness of the information on the status update system. It is defined as the time elapsed since the most recent successfully received update was generated at the source. Suppose that at any time $t$, the last successfully received packet has a time stamp $U(t)$ representing its generation time. Then the age can be defined as

$$
\Delta(t) = t - U(t), \quad (4)
$$

The evolution process of the AoI $\Delta(t)$ for a slotted status update system is illustrated in Fig. 2, where we assume that a transmission decision is made at the beginning of each time slot based on the feedback received and the number of transmission rounds. In the figure, we can see that the AoI increases by one in the case of transmission failure (i.e., when receiving a NACK feedback), while it drops to the number of transmissions and a new packet will be generated when the packet is successfully decoded (i.e., when receiving an ACK feedback). What’s more, we have assumed that one update packet can be transmitted up to $M$ times. For example, at the seventh time slot in the figure, since the maximum number of transmissions has been reached (we assume $M = 4$ here), the old update packet is discarded, and a new one is generated for transmission. Meanwhile, the number of transmission rounds drops to one while the AoI continues to...
The age of information for a single source is given by
\[
\Delta_n = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau \Delta_n(t)dt, \quad n = 1, 2, \ldots, N
\]  
(5)

where the optimal values of the transmit power level of source \(n\) in TDMA can be calculated from (1) with the given values of \(\psi = (\psi_0, \psi_1, \ldots, \psi_{K-1})\) accordingly. For an arbitrary source, when the action \(a(m_n, \Delta_n) = P^n_{m,i}\) is taken, according to the independent block-fading assumption, the probability of erroneous transmission in TDMA can be calculated as follows:

\[
\epsilon^{T}_{zn,i} = \Pr\{R^{T}_{n,i} < \overline{R}^{T}\} = \Pr\{z < z_i\} = \sum_{m=0}^{i-1} \Pr\{z_m \leq z < z_{m+1}\} = \sum_{m=0}^{i-1} \psi_m,
\]  
(7)

where \(\overline{R}^{T}\) is the minimum preset rate required to meet the transmission requirements, which is set to be the same for all sources.

2) Actions in NOMA: In the case of NOMA, the signals from different sources will interfere with each other, resulting in a high probability of erroneous transmission. With the aim of decoding the corresponding information correctly at the receiver so as to minimize the weighted sum average AoI, the SIC technique is adopted here. When \(N\) sources transmit update packets, they can be decoded in \(N!\) different orders. Let \(\mathcal{D} = \{D_1(\cdot), D_2(\cdot), \ldots, D_{N}(\cdot)\}\) denote the decoding order set, where \(D_k(\cdot)\) represents the original index of the source that ranks \(k\)-th in the decoding order \(D_r\). Also, the optimal decoding order \(D^*\) is scheduled and decided at the receiver. Consequently, for each system state, the sources can choose different transmit power values, and the receiver can choose different decoding orders.

Taking the decoding order \(D_r\) as an example, the message from source \(D_r(1)\) is scheduled to be decoded and recovered first when the decoding order is \(D_r\) and is assigned the strongest transmit power, while other sources are allocated the weaker transmit power. Thus, the information from other sources is buried in the received signal and would be treated as interference noise by the common receiver. Based on the channel statistics, the transmit power set can be given by \(P^{N,D_r}\), consisting of \(N \times K\) elements, where the \((n \times k)\)-th element of \(P^{N,D_r}\), i.e., \(P^{N,D_r}_{n,k}\), represents the transmit power of source \(n\) in state \(k\) when decoding order is \(D_r\) considering the NOMA case. With the action set \(P^{N,D_r}\), the information rate for source \(D_r(n)\), denoted by \(R^{N,D_r}_{D_r(n)}\), is given by

\[
R^{N,D_r}_{D_r(n)} = \log_2 \left(1 + \frac{P^{N,D_r}_{D_r(n)} z_i}{1 + \sum_{m>n}^{N} P^{N,D_r}_{D_r(m)} z_{D_r(m)}}\right),
\]  
(8)

\(\rho_n\log_2 \left(1 + \frac{P^{N,D_r}_{D_r(n)} z_i}{P^{N,D_r}_{D_r(n)} z_{D_r(n)}}\right)\) with boosted transmit power \(P^{N,D_r}_{D_r(n)}\). And the optimal values of \(\rho_n\) are decided at the common receiver. So, the \(i\)-th transmit power level of source \(n\) in TDMA \(P^{T}_{D_r(n)}\) can be specified by the following formula:

\[
P^{T}_{D_r(n)} = \rho_n \frac{2^{\frac{\epsilon^{T}_{zn,i}}{z_i}} - 1}{z_i}, \quad i = 1, 2, \ldots, K, \forall n \in [N]
\]  
(6)

where \(R^{T}_{n,i}\) represents the data rate or packet size equivalently of source \(n\) following the TDMA transmission scheme, and the values of \(\{z_i\}\) can be calculated from (1) with the given values of \(\psi = (\psi_0, \psi_1, \ldots, \psi_{K-1})\) accordingly. For an arbitrary source, when the action \(a(m_n, \Delta_n) = P^n_{m,i}\) is taken, according to the independent block-fading assumption, the probability of erroneous transmission in TDMA can be calculated as follows:

\[
\epsilon^{T}_{zn,i} = \Pr\{R^{T}_{n,i} < \overline{R}^{T}\} = \Pr\{z < z_i\} = \sum_{m=0}^{i-1} \Pr\{z_m \leq z < z_{m+1}\} = \sum_{m=0}^{i-1} \psi_m,
\]  
(7)

where \(\overline{R}^{T}\) is the minimum preset rate required to meet the transmission requirements, which is set to be the same for all sources.
where $z^{D_i(n)}$ represents the channel gain of source $D_i(n)$. Under the condition that the other $N-1$ sources’ update packets have been correctly decoded at the receiver, the information rate for source $D_i(N)$ is given by

$$R_{N,D_i}^{N,D_i} = \log_2 \left( 1 + P_{N,D_i}^{N,D_i} z^{D_i(N)} \right).$$

(9)

Note that an outage occurs for source $D_i(n)$ if any message from source $D_i(m)$ with $m < n$ is decoded incorrectly or the message from source $D_i(n)$ is decoded erroneously while all messages from source $D_i(m)$ with $m < n$ are transmitted successfully. Similar to the TDMA scenario, let $R_N^N$ be the minimum preset rate considering the NOMA case, which is also same for all the sources. The error transmission probability of source $D_i(n)$, for all $n = 1, 2, \ldots, N$, is given by

$$
\epsilon_{N,D_i}^{N,D_i} = \Pr \left\{ R_{N,D_i}^{N,D_i} < R_N^{N} \right\} + \left( 1 - \Pr \left\{ R_{N,D_i}^{N,D_i} < R_N^{N} \right\} \right) \Pr \left\{ R_{N,D_i(n)}^{N,D_i(n)} < R_N^{N} \right\} + \ldots \\
+ \prod_{m=1}^{n-1} \left( 1 - \Pr \left\{ R_{N,D_i(m)}^{N,D_i(m)} < R_N^{N} \right\} \right) \Pr \left\{ R_{N,D_i(n)}^{N,D_i(n)} < R_N^{N} \right\}.
$$

(10)

Likewise, if the decoding order is $D_j, j \in \{1, 2, \ldots, N! \} \setminus i$, the corresponding rate and the error transmission probability of each source (i.e., $R_N^{N,D_i}$ and $\epsilon_{N,D_i}^{N,D_i}$) can be obtained from a similar slight modification of the above expressions.

We assume that power is not a continuously changeable parameter, and we quantize the transmit power as $A = \{P_{n,1}, P_{n,2}, \ldots, P_{n,K}\}$ based on the quantized channel gains. To sum up, depending on the current system state, different transmit power levels are adopted, and the power adaptation decision is decided at the transmitters individually. Hence, in both TDMA and NOMA cases, the transmit power $P_{n,i}$ of source $n$ belongs to the action space $A = \{P_{n,1}, P_{n,2}, \ldots, P_{n,K}\}$. Specifically, when $i = K$, $P_{n,i} = 0$, i.e., indicating idle action. It can be seen that the action $a(m_n, \Delta_n) = P_{n,i}$, channel state gain $z_i$, and error transmission probability $\epsilon_{n,i}$ are interrelated.

### C. Transition Probabilities

Assuming that an update packet from source $n$ fails to be decoded by the receiver with the action $a(m_n, \Delta_n) = P_{n,i}$ which occurs with probability $\epsilon_{n,i}$ as defined above, the system enters state $(m_n + 1, \Delta_n + 1)$ in the case of $m_n < M$. When the transmission round $m_n$ reaches the upper bound $M$, this packet is regarded as out of date, and a new packet should be generated, which causes the system to enter state $(1, \Delta_n + 1)$. On the other hand, if the transmission is successful, the state will change to $(1, m_n)$.

Considering that the decoding error probability of a source depends on whether the message decoded earlier is correct or not, the state transitions for $N > 2$ sources are more complicated, and it is difficult if not intractable to express all state transitions. For instance, if the message for $D_k(j)$ is decoded incorrectly, messages for all sources $D_k(i)$ with $i > j$ will be decoded in error definitely, i.e., the state transitions from $(m_1, \Delta_1; m_2, \Delta_2; \ldots; m_N, \Delta_N)$ to $(1, m_1; \ldots; 1, m_{j-1}; m_j + 1, \Delta_j + 1; \ldots; m_N + 1, \Delta_N + 1)$ if $m_i < M, \forall i \geq j$. Therefore, we mainly concentrate on the two-source scenario. To summarize, we list the transition probabilities of the CMDP problems in a two-source scenario using TDMA given by

$$
\begin{align*}
\Pr(m_1 + 1, \Delta_1 + 1; m_2, \Delta_2 | S, P^T_{1,i}) &= \epsilon_{1,i}, m_1 < M \\
\Pr(m_1, \Delta_1; m_2 + 1, \Delta_2 + 1 | S, P^T_{2,j}) &= \epsilon_{2,j}, m_2 < M \\
\Pr(1, \Delta_1 + 1; m_2, \Delta_2 | S, P^T_{1,i}) &= \epsilon_{1,i}, m_1 = M \\
\Pr(m_1, \Delta_1; 1, \Delta_2 + 1 | S, P^T_{2,j}) &= \epsilon_{2,j}, m_2 = M \\
\Pr(m_1, m_2; 1, \Delta_2 | S, P^T_{2,j}) &= 1 - \epsilon_{2,j}, (12)
\end{align*}
$$

using NOMA:

$$
\begin{align*}
\Pr(m_1 + 1, \Delta_1 + 1; m_2, \Delta_2 | S, P^N_{1,i}, P^N_{2,j}) &= \epsilon_{1,i}^{N} (1 - \epsilon_{2,j}^{N}), m_1 < M, \forall m_2 \\
\Pr(m_1, \Delta_1; m_2 + 1, \Delta_2 + 1 | S, P^N_{1,i}, P^N_{2,j}) &= (1 - \epsilon_{1,i}^{N}) \epsilon_{2,j}^{N}, m_2 < M, \forall m_1 \\
\Pr(1, \Delta_1 + 1; m_2, \Delta_2 | S, P^N_{1,i}, P^N_{2,j}) &= \epsilon_{1,i}^{N} (1 - \epsilon_{2,j}^{N}), m_1 = M, \forall m_2 \\
\Pr(m_1, \Delta_1; 1, \Delta_2 + 1 | S, P^N_{1,i}, P^N_{2,j}) &= (1 - \epsilon_{1,i}^{N}) \epsilon_{2,j}^{N}, m_2 = M, \forall m_1 \\
\Pr(m_1 + 1, \Delta_1 + 1; m_2 + 1, \Delta_2 + 1 | S, P^N_{1,i}, P^N_{2,j}) &= \epsilon_{1,i}^{N} \epsilon_{2,j}^{N}, m_1 < M, m_2 < M \\
\Pr(m_1 + 1, \Delta_1 + 1; 1, \Delta_2 + 1 | S, P^N_{1,i}, P^N_{2,j}) &= \epsilon_{1,i}^{N} \epsilon_{2,j}^{N}, m_1 = M, m_2 < M \\
\Pr(m_1 + 1, \Delta_1 + 1; m_2 + 1, \Delta_2 + 1 | S, P^N_{1,i}, P^N_{2,j}) &= \epsilon_{1,i}^{N} \epsilon_{2,j}^{N}, m_1 < M, m_2 = M \\
\Pr(m_1, \Delta_1 + 1; m_2 + 1, \Delta_2 + 1 | S, P^N_{1,i}, P^N_{2,j}) &= \epsilon_{1,i}^{N} \epsilon_{2,j}^{N}, m_1 = M, m_2 = M \\
\Pr(1, \Delta_1 + 1; 1, \Delta_2 + 1 | S, P^N_{1,i}, P^N_{2,j}) &= \epsilon_{1,i}^{N} \epsilon_{2,j}^{N}, m_1 = M, m_2 = M \\
\Pr(1, \Delta_1 + 1; 1, \Delta_2 + 1 | S, P^N_{1,i}, P^N_{2,j}) &= \epsilon_{1,i}^{N} \epsilon_{2,j}^{N}, m_1 = M, m_2 = M \\
\text{and otherwise} \quad \Pr(m_1', \Delta_1'; m_2', \Delta_2' | S, P^N_{1,i}, P^N_{2,j}) &= 0, (13)
\end{align*}
$$

where $S$ represents the current state $(m_1, \Delta_1; m_2, \Delta_2)$.

Note that our state space is countably infinite, since the value of AoI can be arbitrarily large. In practice, however, we can use a large but finite space to approximate the countably infinite
state space by setting an upper bound on the age (which will be denoted by $\Delta_{\text{max}}$). Regarding the storage efficiency, we note that the storage space requirement is given by $M \times \Delta_{\text{max}} \times K$, and the deciding parameter is the $\Delta_{\text{max}}$ applied. Fortunately, we note that $\Delta_{\text{max}}$ in the considered settings is limited, and hence the upper bound on the storage space is fixed and affordable. Besides, we assume that whenever the AoI exceeds $\Delta_{\text{max}}$, we set it to be one. Clearly, when $\Delta_{\text{max}}$ is close to infinity, the optimal policy for the finite state space will converge to that of the original problem.

D. Rewards, Costs and Problem Formulation

For source $n$, let us define the decision $\mu_{n,t}(s)$ as a function that maps the system state $s_t = (m_1, \Delta_{1:t}; m_2, \Delta_{2:t}; \ldots; m_{N,t}, \Delta_{N,t})$ to the action $a(m_n, \Delta_n)$ to be taken in time block $t$. And the policy $\mu_n = \{\mu_{n,1}, \mu_{n,2}, \ldots\}$ is called stationary if the actions are independent of time slot $t$. Therefore, we denote $s_{t,n}^\mu$ as sequences of states induced by policy $\mu_n$. For a stationary policy $\mu_n$, we define the reward function of the CMDP problems (i.e., the long-term weighted sum average AoI) in both TDMA and NOMA environments as

$$\bar{\Delta}(s, \mu) = (1 - \lambda) \sum_{n=1}^{N} \sum_{t=1}^{\infty} \lambda^{t-1} E[\omega_n \Delta^{n}_{n,t}],$$

where $\lambda \in (0, 1)$ is a discount factor, $\omega_n > 0$ is the weight coefficient representing the priority of source $n$, and $E[\cdot]$ is the expectation operator, which is taken over the policy $\mu_n$. Then the cost function (i.e., the long-term average power consumption) for source $n$ on the TDMA scheme is given by

$$\bar{P}^T_n(s, \mu_n) = (1 - \lambda) \sum_{t=1}^{\infty} \lambda^{t-1} E[P^T_{n,t}]$$

and the cost function for source $n$ when NOMA is employed can be written as

$$\bar{P}^N_n(s, \mu_n) = (1 - \lambda) \sum_{t=1}^{\infty} \lambda^{t-1} E[P^N_{n,t}]$$

and please note that the error transmission probability $\epsilon_{n,i}$ is only available when CDI is known at the transmitter. However, when the environment is not known a priori, such information is not available to the transmitters, and the CMDP optimization problems mentioned above are no longer applicable, as will be discussed in Section V.

IV. OPTIMAL POLICY WITH KNOWN CDI

In this section, we assume that the CDI of each link is known at the associated transmitter. And we adopt the Lagrangian relaxation method to solve the problems mentioned above. In this way, a CMDP problem can be converted into an equivalent unconstrained MDP problem by introducing a non-negative multiplier $\beta$, and the Lagrangian reward for source $n$ is defined as

$$r_n(s, \mu_n; \beta_n) = \omega_n \Delta_n(s, \mu_n) + \beta_n p_n(s, \mu_n),$$

where $p_n(s, \mu_n)$ denotes the power consumption in state $s$ with policy $\mu_n$ for the TDMA or NOMA scheme. Since expression (20) is valid for both transmission modes, we omit the superscripts from the notation for simplicity here. And then, accordingly, there exists a value function $V_{n,\beta_n}(s)$, satisfying

$$V_{n,\beta_n}(s) = \min_{a \in A} \left\{(1 - \lambda)r_n + \lambda \sum_{s' \in S} \Pr\{s' \mid s, a\} V_{n,\beta_n}(s')\right\},$$

called Bellman optimality equation for source $n$ in all states $s \in S$, where $s'$ denotes the next state obtained from state $s$ after taking action $a$. Hence, the policy for source $n$ can be indicated as follows:

$$\mu_{n,\beta_n}^*(s) = \arg \min_{a \in A} V_{n,\beta_n}(s).$$

A. Value Iteration Algorithm

A popular and effective approach, called value iteration algorithm (VIA), is adopted here to solve (21) for any given Lagrangian multiplier $\beta$. With any initialization of $V_{n,\beta_n}^{0}(s), \forall s, V_{n,\beta_n}^{k+1}(s)$ will be updated in each iteration satisfying

$$V_{n,\beta_n}^{k+1}(s) = \min_{a \in A} \left\{(1 - \lambda)r_n + \lambda \sum_{s' \in S} \Pr\{s' \mid s, a\} V_{n,\beta_n}^{k}(s')\right\},$$

until convergence.

Based on [28], [29], [30], there exist optimal stationary policies for the formulated CMDP problems, which are also optimal for the corresponding unconstrained problems considered in (20) for some $\beta_n = \beta^*_n$. To simplify the notation, the dependence on $n$ is suppressed in the following. For an arbitrary source, this optimal policy $\mu^*$ is a probabilistic mixture of two deterministic policies $\mu_{\beta^+}$ and $\mu_{\beta^-}$. To be more precise, there exists $\xi \in [0, 1]$ such that the optimal policy $\mu^*(s)$, in any state $s$, selects the action $\mu_{\beta^+}$ with probability $\xi$ and the action $\mu_{\beta^-}$ with probability $1 - \xi$. Let $\bar{P}_\beta$ be the average power consumption associated with the policy $\mu_{\beta^+}(s)$. According to the fact that $\bar{P}_\beta$ is monotonically
Algorithm 1: Proposed Optimal Policy in TDMA.

1: Input: $\mathcal{F}, K, M, \Delta_{max}, R, N$;
2: Initialization: $\beta_n^0 = 0, v_n^0, v_{n,\beta_n} = 0$, where $n \in [N]$.
3: while \(\max_n |\beta_n^+ - \beta_n^-| > \Gamma_\beta\) do
4: Let $\beta_n = \beta_n^0 = (\beta_n^- + \beta_n^+)/2, i = i + 1, k = 0$;
5: while $\max_k \{v^k_{n,\beta_n} - v_n^k(s)\} > \Gamma_V$ do
6: for $n = 1$ to $N$ do
7: $V_{n,\beta_n}^{k+1}(s) = \min_{a \in A_n}\{(1 - \lambda)r_n + \lambda\mathbb{E}[V_k] + \beta_n\}$.
8: $\pi_n^k = +\lambda\sum_{s' \in S} \Pr[s' | s, a]v_{n,\beta_n}(s'), \forall s$;
9: end for
10: $V_{n,\beta_n}^{k+1}(s) = \sum_{n=1}^{N} v_{n,\beta_n}^{k+1}(s)$;
11: $k = k + 1$;
12: end while
13: Derive the corresponding policy $\mu_n^\beta(s)$ from (22) and optimal values of $\rho_n$;
14: Compute the corresponding steady state distribution $\pi_n^\beta(s)$ and cost in TDMA:
$\hat{P}_n^\beta = \sum_{s \in A_n} \beta_n^s \pi_n^\beta(s)$;
15: if $\hat{P}_n^\beta > \bar{P}_n$ then
16: $\beta_n^- = \beta_n^+$;
17: else
18: $\beta_n^+ = \beta_n^-$;
19: end if
20: end while
21: Compute $\Delta_n, \bar{\Delta}_n, \bar{\Delta}_n, \bar{\Delta}_n$, $\hat{P}_n^\beta$, $\hat{P}_n^\beta$, and $\xi_n^\beta$ using (25)-(27);
22: Compute $\Delta_n = \xi_n^\beta \Delta_n + (1 - \xi_n^\beta) \bar{\Delta}_n$;
23: The optimal weighted sum average AoI is derived:
24: $\Delta = \sum_{n=1}^{N} \omega_n \Delta_n$.

Algorithm 2: Proposed Optimal Policy in NOMA.

1: Input: $\mathcal{F}, K, M, \Delta_{max}, R, N$;
2: Initialization: $\beta_n^0 = 0, v_n^0, v_{n,\beta_n} = 0$, where $n \in [N]$.
3: while \(\max_n |\beta_n^+ - \beta_n^-| > \Gamma_\beta\) do
4: Let $\beta_n = \beta_n^0 = (\beta_n^- + \beta_n^+)/2, i = i + 1, k = 0$;
5: while $\max_k \{v^k_{n,\beta_n} - v_n^k(s)\} > \Gamma_V$ do
6: Select the optimal decoding order $D^*$ that minimizes the value function $\hat{P}_n^\beta$.
7: for $n = 1$ to $N$ do
8: $V_{D^*(n),\beta_n}^{k+1}(s) = \min_{a \in A_n}\{(1 - \lambda)r_n + \lambda\mathbb{E}[V_k] + \beta_n\}$;
9: $\rho_n = +\lambda\sum_{s' \in S} \Pr[s' | s, a]v_{D^*(n),\beta_n}(s'), \forall s$;
10: end for
11: $V_{D^*(n),\beta_n}^{k+1}(s) = \sum_{n=1}^{N} v_{D^*(n),\beta_n}^{k+1}(s)$;
12: $k = k + 1$;
13: end while
14: Derive the corresponding policy $\mu_n^\beta(s)$ from (22);
15: Compute the steady state distribution $\pi_n^\beta(s)$ and cost in NOMA:
$\Delta_n = \sum_{n=1}^{N} \omega_n \Delta_n$;
16: if $\Delta_n > \bar{P}_n$ then
17: $\beta_n^- = \beta_n^+$;
18: else
19: $\beta_n^+ = \beta_n^-$;
20: end if
21: end while
22: Compute $\Delta_n, \bar{\Delta}_n, \bar{\Delta}_n, \bar{\Delta}_n$, $\hat{P}_n^\beta$, $\hat{P}_n^\beta$, and $\xi_n^\beta$ using (25), (26), and (27);
23: Compute $\Delta_n = \xi_n^\beta \Delta_n + (1 - \xi_n^\beta) \bar{\Delta}_n$;
24: The optimal weight sum average AoI is derived:
25: $\Delta = \sum_{n=1}^{N} \omega_n \Delta_n$.

respectively.

To sum up, the pseudocode of the VIA in TDMA is given in Algorithm 1 and that in NOMA is provided in Algorithm 2. We would like to note that the algorithms can be trained off-line, for which the complexity will not cause problems to the online implementation.

B. Threshold-Based Policy

In this part, a threshold-based policy is proposed, in which each source node will generate and transmit a new update packet with a fixed power level $P_0$ once the instantaneous age $\Delta$ exceeds a threshold value $\gamma \geq 1$, while keeping silent if $\Delta < \gamma$.

Accordingly, we have the following state transition probability matrix $A_{n,\gamma}$ for source $n$: \[
A_{n,\gamma} = \begin{bmatrix}
0 & 1 & 0 & \cdots & 0 & 0 \\
0 & 0 & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
1 - \epsilon_{n,\gamma} & 0 & 0 & \cdots & 0 & 0 \\
\epsilon_{n,\gamma} & 0 & 0 & \cdots & 0 & 0 \\
\end{bmatrix},
\]
where $\epsilon_{n,\gamma}$ is the error transmission probability for source $n$ when using the fixed power $P_n$.

Then, we have the following result.

**Proposition 1:** Given the value of $\gamma$ for source $n$, the average AoI and power consumption in the threshold-based policy for each source can be expressed as follows:

$$\sum_{n,\gamma} = \frac{1 - \epsilon_{n,\gamma}}{(\gamma - 1)(1 - \epsilon_{n,\gamma}) + 1} \times \left( \frac{\gamma(\gamma - 1)}{2} + \frac{\gamma}{1 - \epsilon_{n,\gamma}} + \frac{\epsilon_{n,\gamma}}{(1 - \epsilon_{n,\gamma})^2} \right),$$

$$\hat{P}_{n,\gamma} = \hat{P}_{n} (1 + (\gamma - 1)(1 - \epsilon_{n,\gamma})).$$

**Proof:** For each source, we combine (24) and (29) and then can obtain

$$\pi_{\Delta} = \frac{1 - \epsilon_{\gamma}}{(\gamma - 1)(1 - \epsilon_{\gamma}) + 1}, \quad \forall \Delta \leq \gamma,$$

$$\pi_{\Delta} = \epsilon_{\gamma} \pi_{\Delta - 1}, \quad \forall \Delta > \gamma.$$  \hspace{1cm} (32)

Subsequently, substituting (32) and (33) into (25) and (26), we can obtain (30) and (31), respectively.

Then, similar to the discussions in [21], for each $\gamma$, we can find a unique value of $P_{n,\gamma}$ that satisfies $\hat{P}_{n} = \hat{P}_{n} \rho_{n}$ or $\hat{P}_{n} = \hat{P}_{n} P_{n,\gamma}$, $\forall n$. Meanwhile, when $\gamma = 0$, the threshold-based policy reduces to the fixed power policy. Therefore, there should be some optimal threshold value $\gamma^*$ that minimizes the weighted sum average AoI, which can be determined numerically.

V. REINFORCEMENT LEARNING ALGORITHM IN AN UNKNOWN ENVIRONMENT

In the above section, it is assumed that CDI and error transmission probability are known in advance. However, in most practical scenarios, CDI is not available or may continuously change over time. So, in this section, we assume the sources do not have a priori information about CDI and have to learn it. As a result, we adopt online reinforcement learning (RL) approaches to minimize weighted sum AoI in TDMA and NOMA without degrading the performance significantly compared with VITA proposed above. In RL, an agent constantly interacts with the environment to learn a good scheduling policy without prior information. Then, at each iteration, the agent observes state $S^i$ and takes an action $a^i$. After performing the action $a^i$, the current state will transmit to $S^{i+1}$, and the agent will receive corresponding reward $r^{i+1}$ and the cost of this step $p^i$. Above, $i$ is the index of iterations.

Q-learning is one of the most popular reinforcement learning algorithms, in which the agent obtains a corresponding reward by interacting with the environment and gradually fills the Q-table by calculating the discounted accumulative rewards termed the state-action function, i.e., $Q(S, a)$. At each iteration, the Q-values are updated and then stored in the Q-table again. After learning a sufficient number of iterations or enough time, each $Q$-value will converge to a certain fixed value, and the agent will find the best policy for our problems based on the minimum non-zero value in the Q-table. Moreover, Q-learning is an off-policy learning algorithm, which means that the process of estimating the optimal state-action function $Q^*$ from the current state-action function $Q$ is independent of the exploration policy followed. The update process for the Q-function is shown as follows:

$$Q_n(S^i, a^i_n) \leftarrow Q_n(S^i, a^i_n) + \alpha_n[r^{i+1} + \lambda \min_{a_n} Q_n(S^{i+1}, a^i_n) - Q_n(S^i, a^i_n)].$$ \hspace{1cm} (34)

Nevertheless, there always exist some problems in Q-learning, such as curse of dimensionality and difficulty in convergence [31], [32]. What’s more, as discussed in Section III, there are $K$ possible power values for the agent to choose from in this work. So how to balance exploration and exploitation in the learning process is extremely important, as this determines how fast the algorithm converges and how long this learning process takes. Otherwise, adopting inappropriate policies will undermine our learning process. Here we adopt Q-learning with the $\epsilon$-greedy exploration algorithm to find the optimal policy for both TDMA and NOMA schemes. This means that, at the beginning of each training session, the agent selects a random action with some probability $\epsilon$ and executes a greedy policy (i.e., selects a historically optimal action according to the Q-table) with probability $1 - \epsilon$, as below:

$$\mu(s) = \Pr(a|s) = \begin{cases} \frac{r}{K} + 1 - \epsilon & \text{if } a^* = \arg \min_{a \in A} Q(S, a) \\ \frac{1}{K} & \text{else} \end{cases}$$

where $\epsilon$ is the exploration probability, which usually diminishes with the iterative process of the algorithm and eventually tends to zero. In this way, during the early iterative period, we encourage exploration, whereas after a sufficient number of iterations, we have enough exploration, the policy asymptotically becomes conservative and greedy so that the algorithm can converge stably.

For the purpose of comparing AoI performance with the known environment, for the source $n$, we employ the same transmit power set (also known as action space) $A = \{P_{n,1}, P_{n,2}, \ldots, P_{n,K}\}$ discussed in Section III-B and select action according to $\epsilon$–greedy exploration algorithm at the beginning. To speed up the convergence of Q-learning, we normalize the transmit power set using the average value of the power set and the average transmitted power constraint, i.e., $\|A^T\| = \{\frac{K P_{n,k}}{\sum_{k=1}^{K} P_{n,k}} \hat{P}_{n}\}$ and $\|A^N\| = \{\frac{K P_{n,k}}{\sum_{k=1}^{K} P_{n,k}} \hat{P}_{n}\}$, $k = 1, 2, \ldots, K$, for TDMA and NOMA, respectively. Numerical simulations show a considerable saving in training time by normalization without any significant change in performance. The details of Q-learning in TDMA and NOMA are given in Algorithms 3 and 4, respectively. It is worth noting that Q-learning is naturally implemented in an online and incremental manner, which gives it a clear advantage over other RL methods such as the Monte Carlo method. Consequently, this method enjoys a significant advantage in terms of time and space complexity.

In Algorithms 3 and 4, $a^i_n$ is the update parameter (or learning rate equivalently) in the $i$-th iteration for source $n$. However, in practice, it is hard to find the optimal Lagrange multiplier $\beta^*$ that satisfies $\hat{P}_{\beta} = \hat{P}$ or $\hat{P} = \hat{P}$. As a result, we have the following heuristic as in [6]: To find the desired non-negative optimal $\beta^*$,
Algorithm 3: Q-Learning in TDMA.
1: Input: $T$, $K, M$, $\Delta_{\max}$, $R$, $\lambda$, $\beta$, $I$, $|A^T|$, $N$;
2: Initialization: $Q^0 = 0$, where $n \in [N]$, $i \leftarrow 0$, $S^0 = (s^0_1, s^0_2, \ldots, s^0_N)$;
3: for $n = 1$ to $N$ do
4: while $i \leq I$ do
5: $a^i_n \leftarrow$ choose an action from $A$ based on $\epsilon$-greedy algorithm for $S^i$;
6: Observe next state $S^{i+1}$, and calculate corresponding reward function $r^{i+1}$;
7: Get average age $\tilde{\Delta}_n$ and cost function $p^i_n$;
8: Update
9: $n^i_\alpha = \frac{1}{\sqrt{n}}$; /* update parameter*/
10: $Q_n(S^t, a^i_n) \leftarrow Q_n(S^t, a^i_n) + a^i_n[r^{i+1} + \lambda \min_n Q_n(S^{i+1}, a^i_n) - Q_n(S^i, a^i_n)]$;
11: $i \leftarrow i + 1$; /*increase the iteration*/
12: end while
13: end for
14: /*Derive weighted sum average AoI*/
15: $\Delta = \sum_{n=1}^{N} \omega_n \tilde{\Delta}_n$.

we run the iterative algorithms via stochastic gradient descent with an initialized parameter $\beta^0$, which are given by

$$\beta^{i+1} = \max \left\{ 0, \beta^i + \zeta^i \left( \tilde{P}^T - \frac{\tilde{P}^T}{\beta^i} \right) \right\}, \quad (36)$$

$$\beta^{i+1} = \max \left\{ 0, \beta^i + \zeta^i \left( \tilde{P}^N - \frac{\tilde{P}^N}{\beta^i} \right) \right\}, \quad (37)$$

for TDMA and NOMA, respectively. $\zeta^i$ is a positive and decreasing sequence that satisfies the following conditions:

$$\sum_{i=1}^{\infty} \zeta^i = \infty, \quad \sum_{i=1}^{\infty} (\zeta^i)^2 = 0. \quad (38)$$

That is to say, we update the Lagrange multipliers based on the empirical power consumption in order to solve the proposed AoI optimization problem. In each iteration, we keep track of a value $\beta$ which drives the transmission cost close to the power constraint, and then we can derive the optimal policy. Please note that, in Algorithm 4, we must ensure that the gap between the weighted sum average AoI obtained from two consecutive training processes is smaller than a given threshold (i.e., $|\Delta - \tilde{\Delta}| < \Gamma_{\Delta}$) to guarantee the accuracy and reliability of the algorithm. This is due to the fact that NOMA requires to update the Q-values of both sources and their corresponding power consumption simultaneously, which makes it quite problematic to achieve convergence of the algorithm to obtain the desired results. By contrast, we have no such issues to worry about in TDMA’s Q-learning, where multiple sources are learning to update separately.

Algorithm 4: Q-Learning in NOMA.
1: Input: $T$, $K, M$, $\Delta_{\max}$, $R$, $\lambda$, $\beta$, $I$, $|A^N|$;
2: Initialization: $Q^0 = 0$, where $n \in [N]$, $i \leftarrow 0$, $S^0 = (s^0_1, s^0_2, \ldots, s^0_N)$ and specify $\Gamma_{\Delta}$;
3: while $|\Delta - \tilde{\Delta}| < \Gamma_{\Delta}$ do
4: while $i \leq I$ do
5: for $k = 1$ to $N$ do
6: Set decoding order $D = D_k$:
7: Execute Algorithm 3 (5–7) to get $a^i_{D_k(n)}, r^{i+1}_{D_k(n)}$, $\tilde{\Delta}^i_{D_k(n)}$ and $p^i_{D_k(n)}, n \in [N]$;
8: end for
9: Find optimal decoding order $D^*$ according to the reward function:
10: $a^i_n = a^i_{D^*}, \tilde{\Delta}^i = \tilde{\Delta}^i_{D^*}$ and $p^i_n = p^i_{D^*}, n \in [N]$;
11: Update
12: $\alpha^i = \alpha^i = \ldots = \alpha^i = \frac{1}{\sqrt{n}}$; /* update parameter*/
13: for $n = 1$ to $N$ do
14: $Q_n(S^t, a^i_n) \leftarrow Q_n(S^t, a^i_n) + a^i_n[r^{i+1} + \lambda \min_n Q_n(S^{i+1}, a^i_n) - Q_n(S^i, a^i_n)]$;
15: end for
16: $i \leftarrow i + 1$; /*increase the iteration*/
17: end while
18: /* Derive weighted sum average AoI*/
19: $\Delta = \sum_{n=1}^{N} \omega_n \tilde{\Delta}_n$.

VI. NUMERICAL RESULTS

In this section, we focus mainly on the two-source case and present some numerical results to validate our theoretical analyses. Unless otherwise specified, we assume that $M = 4$, $\Delta_{\max} = 100$ to approximate the countably infinite state space. Assuming that $R^T = R^N = R = 1.7 \text{ bps/Hz}$, $\lambda = 0.99$, $K = 128$, and $\omega_1 = \omega_2 = 1$ in the following. Rayleigh fading channels with unit mean are considered here, and in the off-line value iteration algorithms, we further assume that the channel state probabilities are $\psi_0 = \psi_1 = \ldots = \psi_{K-1} = \frac{1}{K}$. In the reinforcement learning algorithms, $I = 10000$ as the maximum number of iterations for each episode.

A. Performance in a Known Environment

First, in Fig. 3, we plot the weighted sum average AoI on TDMA and NOMA schemes as a function of $P_T$, where $P_T = P_T_1 = P_T_2$ such that the subscript $n$ is neglected for simplification. “Fixed power” refers to the policy that sets the transmit power to a fixed level, and “theoretical” stands for the statistical values derived directly from the algorithms. Moreover, we simulate the corresponding policies over 100 to approximate the countably infinite state space. Assuming that $\sum_{i=0}^{\infty} (s^i_1, s^i_2, \ldots, s^i_N)$ and specify $\Gamma_{\Delta}$;
the simulation results on both TDMA and NOMA schemes, which confirms the validity of the proposed optimal policy. Here we can see that NOMA always outperforms TDMA when $R = 1.7$ bps/Hz, which is reflected in a lower AoI. Compared with the “fixed power” policy, the proposed optimal policy reduces age significantly, especially in the low-power regime. Moreover, the suboptimal policy, i.e., threshold-based policy, can obtain a lower weighted sum average AoI than fixed power policy through a simple closed-form expression of the average AoI. As $P$ increases successively, each policy further converges to a similar performance.

Next, based on Algorithm 1, we put the focus on finding the optimal $\rho$ (in the two-source case, we set $\rho_1 = \rho$ and $\rho_2 = 1 - \rho$) that minimizes the AoI. In Fig. 4, we plot the weighted sum average AoI as a function of $\rho$ with the assumptions of $P = 0$ dBW. Generally speaking, the average AoI reaches the minimum when $\rho$ is around 0.5 (i.e., the update packet sent by source 1 occupies half of one time slot) since the initial conditions of the two sources are set to be the same. However, when the average power values change, the value of the optimal $\rho$ will change accordingly, e.g., as shown in Fig. 5(a).

We first fix the power constraint $P_1$ for source 1 and then adjust the power constraint $P_2$ for source 2 by the proportion $\alpha$ (i.e., $P_1 = P$, $P_2 = \alpha P$). In Fig. 5(a), we plot the derived optimal $\rho$ as the coefficient $\alpha$ varies from 0.1 to 10 under different power constraints, indicating a proportional reduction or amplification of $P_2$. Obviously, the optimal $\rho$ increases with $\alpha$ in all cases, indicating that source 1 can occupy more time fraction for status updates if the average power of source 2 gets larger. Moreover, we can see from the figure that the difference of the transmission time between the sources becomes smaller if the average power levels are low. This is due to the fact that in the low-power regime, more time is needed to successfully send update packets for each source, and hence a wise choice would be to compromise between the sources.

Likewise, in Fig. 5(b), we investigate the weighted sum average AoI as a function of $\alpha$ for different multiple access schemes when $P = -2$ dBW and 0 dBW. We notice that the weighted sum average AoI will increase significantly when $\alpha$ is less than 1, especially in the TDMA case. In other words, reducing one source’s power constraint will obviously lead to poor performance. On the other hand, when $\alpha > 1$, the age of information does not decrease dramatically with increases in $\alpha$ due to the fact that the power constraint is large enough that a larger $P_2$ will not lead to an obvious performance improvement, especially for a relatively large power constraint, as can be seen from the comparison of $P = 0$ dBW and $P = -2$ dBW in the figure.

Fig. 6 shows the weighted sum average AoI performance under different schemes in NOMA. Here we assume $R = 1$ bps/Hz. “Fixed $D$” is the policy that arranges to send the update.
packet with a fixed decoding order in NOMA, which is a newly proposed scheme to investigate the impact of decoding order \( D \) on AoI performance. This result is somewhat interesting since the optimal policy and the suboptimal policy with fixed decoding order achieve similar performance. This is because the decoding order is not the deciding parameter for the performance deterioration, whereas the power allocation policy considered in this paper is more prominent. Nevertheless, suboptimal decoding order will result in the performance loss that will not diminish.

In Fig. 7, we plot the optimal decoding order and associated power allocation policies obtained from Algorithm 2 with \( P = 5 \text{ dBW} \), respectively. The state space \( S \) is denoted as \( (s_1; s_2) \) where the value of state \( s_n \) is mapped to \( M \times (\Delta_n - 1) + m_n \), \( \forall n \). Here we truncate a portion of the complete state space to facilitate the analysis, and some illogical points have been removed. In Fig. 7(a), the distribution of decoding order is symmetrical diagonally, except that there are some reasonable fluctuations in the diagonal. Here we can see that when \( s_1 \) is large and \( s_2 \) is small, i.e., in the lower right area of the figure, optimal policy prefers to choose decoding order \( D_2 \), which means that the receiver first decodes the message from source 2 such that the message from source 1 sees no interference and attains successful transmissions with a higher probability to lower the AoI of source 1, and hence the weighted sum AoI is accordingly reduced.

In Fig. 8, we investigate the impact of maximum transmission rounds \( M \) on the AoI performance when \( P = 0 \text{ dB} \) and 5 dB. Obviously, as \( M \) increases, the weighted sum average AoI increases.
In Fig. 9, we plot AoI performance versus packet size $R$ for TDMA and NOMA. Compared to TDMA, NOMA has better performance when the packet size $R$ is larger. In other words, NOMA is better suited for transmitting large update packets, while TDMA is reasonable for sending small data packets. It is mainly attributed to the features of NOMA, which allows a large number of accesses and enables resource reuse at the expense of high device complexity. Besides, the large difference in transmit power between sources is key to the fundamental property of the SIC technique, since a larger value of $R$ implies a greater effect on the power values of the action set and therefore a larger range of power variations available to the sources, which can be evidenced in the power expressions (8) and (9).

**B. Performance in an Unknown Environment**

For the sake of performance comparison with the known environment, we set $R = 1.5$ bps/Hz in the following according to Fig. 9, in which scenario they have similar age performance. And to demonstrate the accuracy of the results obtained by the RL algorithm, we adopt another RL algorithm, SARSA, as a benchmark scheme. Q-learning is an off-policy learning algorithm, which means that the process of estimating the optimal state-action function $Q^*$ from the current state-action function $Q$ is independent of the exploration policy followed. However, SARSA is a related on-policy algorithm that learns the Q-value function for the policy the agent is actually executing. The update process of Q-function using SARSA is given by

$$Q_n(S^i, a^i_n) \leftarrow Q_n(S^i, a^i_n) + \alpha_n \left[ r_n + \gamma Q_n(S^{i+1}, a^i_{n+1}) - Q_n(S^i, a^i_n) \right],$$

(39)

where the action $a^i_{n+1}$ is the action that is actually executed by the current policy for state $S^{i+1}$. Note that the min-operator in Q-learning is replaced by the estimate of the value of the next action according to the policy.

In Fig. 10, we plot the weighted sum average AoI under different schemes as the average transmission power increases. We notice that the weighted sum AoI derived from the RL algorithms can achieve near-optimal performance for both TDMA and NOMA, and there is no significant deterioration in performance compared to the VIA (i.e., “Optimal” in the figure), indicating the validity of the reinforcement learning algorithm. On the other hand, the RL-based algorithms more markedly reduce the age of information compared to the policies with fixed power, even if no priori information is provided, confirming the advantage of the power control policy. Moreover, comparing the age performance obtained by two different RL algorithms, i.e., Q-learning and SARSA, we can find that the two learning algorithms can achieve similar age performance, especially under the TDMA scheme. However, Q-learning in NOMA can yield a lower weighted sum average AoI since Q-learning is a more courageous learning algorithm that can better deal with the large state space and complicated state transitions in NOMA.

In Fig. 11, we investigate the evolution of the average age during the learning process when $P = 0$ dBW. (a) Learning process in TDMA. (b) Learning process in NOMA.
perform almost identically, except for some differences at the beginning, when they are in the exploration period. While in the NOMA case, some undesired outcomes will occasionally occur depending on the channel realizations just like the second run in the Fig. 11(b). Consequently, in order to guarantee the reliability of the algorithm, we need to perform several runs of the simulation and ensure that the difference between the weighted sum average AoI $\Delta$ obtained from the two consecutive training processes is less than a given threshold, as described in Algorithm 4.

In Fig. 12, we plot the action probabilistic selection under different average power constraints using the NOMA scheme. We set the selection probability of all actions at the beginning to zero, as in Fig. 12(a). And we can see that, except that agents prefer to try some new and potentially good actions at early states, agents will always select the unique and optimal action after enough learning and exploration, confirming the convergence of our algorithms. By comparing Fig. 12(c) and (d), when the sources are equipped with less power, i.e., $-1$ dBW, the policy is going to visit more states because of the higher error transmission probability due to the restricted power set.

In Fig. 13, we show the convergence paths to find the optimal Lagrange multipliers $\beta_1$ and $\beta_2$ in Q-learning, and each episode consists of $10^4$ time steps. The gray dotted lines represent the end of the learning process, indicating that the learning process is complete and the optimal $\beta^*$ is derived. It is apparent from this figure that through normalization, we speed up the convergence of the algorithm and save a great deal of time, while the performance does not change significantly in the process of seeking Lagrange multipliers.
VII. CONCLUSION

In this paper, we have analyzed the AoI performance optimization on different multiple access schemes (i.e., TDMA and NOMA), where different transmit power levels are selected according to the channel state. We have assumed that multiple independent power-constrained sources send update packets to a common receiver, and the maximum number of retransmission rounds cannot exceed $M$ times. Thus, considering the two transmission schemes, we have formulated the CMDP problems to minimize the weighted sum average AoI under the average power constraint in TDMA and NOMA, respectively. We have resorted to the Lagrangian method to convert CMDP problems to equivalent MDP problems and obtained the corresponding value iteration algorithms to derive the power allocation policy. We have also proposed a threshold-based policy, where we can derive the closed-form expressions of the average AoI and power consumption. However, when the environment information is not known a priori, these analyses are no longer applicable, and we have proposed online Q-Learning RL algorithms. And SARSA is also proposed as a benchmark scheme. Through numerical results, we have verified that the proposed optimal policies reduce the weighted sum average AoI more significantly compared to the fixed power policy and demonstrated that NOMA is more suitable for transmitting large update packets due to the higher spectral efficiency introduced.

What’s more, Q-learning with $\epsilon$-greedy exploration algorithm achieves near-optimal age performance and saves a lot of time by normalization.

We would like to note that the learning methods proposed in this paper are model free. Other model-based reinforcement learning strategies may be possible extensions to achieve better time efficiency. Also, extensions of this work to other multiple access protocols, see, e.g., [33], [34], [35] and references therein, may give us interesting insights, and are left for future studies.

REFERENCES

[1] L. D. Xu, W. He, and S. Li, “Internet of Things in industries: A survey,” IEEE Trans. Ind. Informat., vol. 10, no. 4, pp. 2233–2243, Nov. 2014.

[2] T. Qiu, N. Chen, K. Li, M. Aiquzzaman, and W. Zhao, “How can heterogeneous Internet of Things build our future: A survey,” IEEE Commun. Surv. Tuts., vol. 20, no. 3, pp. 2011–2027, Thirdquarter, 2018.

[3] R. D. Yates, Y. Sun, D. R. Brown, S. K. Kaul, E. Modiano, and S. Ulukus, “Age of information: An introduction and survey,” IEEE J. Sel. Areas Commun., vol. 39, no. 5, pp. 1183–1210, May 2021.

[4] S. Kaul, R. Yates, and M. Gruteser, “Real-time status: How often should one update?,” in Proc. IEEE Conf. Comput. Commun., 2012, pp. 2731–2735.

[5] Q. Wang, H. Chen, Y. Gu, Y. Li, and B. Vucetic, “Minimizing the age of information of cognitive radio-based IoT systems under a collision constraint,” IEEE Trans. Wireless Commun., vol. 19, no. 12, pp. 8054–8067, Dec. 2020.

[6] E. T. Ceran, D. Gündüz, and A. Gürüy, “Average age of information with hybrid ARQ under a resource constraint,” IEEE Trans. Wireless Commun., vol. 18, no. 3, pp. 1900–1913, Mar. 2019.

[7] D. Qiao and M. C. Gursoy, “Age minimization for status update systems with packet based transmissions over fading channels,” in Proc. IEEE 11th Int. Conf. Wireless Commun. Signal Process., 2019, pp. 7–12.

[8] Y. Sun, E. Uysal-Biyikoglu, R. D. Yates, C. E. Koksal, and N. B. Shroff, “Update or wait: How to keep your data fresh,” IEEE Trans. Inf. Theory, vol. 63, no. 11, pp. 7492–7500, Nov. 2017.

[9] H. Tang, J. Wang, L. Song, and J. Song, “Minimizing age of information with power constraints: Multi-user opportunistic scheduling in multi-state time-varying channels,” IEEE J. Sel. Areas Commun., vol. 38, no. 5, pp. 854–868, May 2020.

[10] W. Pan, Z. Deng, X. Li, P. Zhou, and W. Wu, “Optimizing the age of information for multi-source information update in Internet of Things,” IEEE Trans. Netw. Sci. Eng., vol. 9, no. 2, pp. 904–917, Mar./Apr. 2022.

[11] H. Pan and S. C. Liew, “Information update: TDMA or FDMA?”, IEEE Wireless Commun. Lett., vol. 9, no. 6, pp. 856–860, Jun. 2020.

[12] N. Hiroswawa, H. Immori, K. Ishibashi, and G. T. F. Abreu, “Minimizing aging of information in energy harvesting wireless sensor networks,” IEEE Access, vol. 8, pp. 219934–219945, 2020.

[13] Z. Qian, F. Wu, J. Pan, K. Srinivasan, and N. B. Shroff, “Minimizing age of information in multi-channel time-sensitive information update systems,” in Proc. IEEE Conf. Comput. Commun., 2020, pp. 446–455.

[14] I. Kadota, A. Sinha, and E. Modiano, “Scheduling algorithms for optimizing age of information in wireless networks with throughput constraints,” IEEE/ACM Trans. Netw., vol. 27, no. 4, pp. 1359–1372, Aug. 2019.

[15] A. Maatouk, M. Assaad, and A. Ephremides, “Minimizing the age of information: NOMA or OMA?,” in Proc. IEEE Conf. Comput. Commun. Workshops, 2019, pp. 102–108.

[16] Q. Wang, H. Chen, C. Zhao, Y. Li, P. Popovski, and B. Vucetic, “Optimizing information freshness via multiser serving with adaptive NOMA/OMA,” IEEE Trans. Wireless Commun., vol. 21, no. 3, pp. 1766–1778, Mar. 2022.

[17] R. V. Bhat, R. Vaze, and M. Motani, “Minimization of age of information in fadind multiple access channels,” IEEE J. Sel. Areas Commun., vol. 39, no. 5, pp. 1471–1484, May 2021.

[18] H. Pan, J. Liang, S. C. Liew, V. C. M. Leung, and J. Li, “Timely information update with nonorthogonal multiple access,” IEEE Trans. Ind. Informat., vol. 17, no. 6, pp. 4096–4106, Jun. 2021.

[19] J. Chen et al., “On the adaptive Aol-aware buffer-aided transmission scheme for NOMA networks,” in Proc. IEEE Wireless Commun. Netw. Conf., 2021, pp. 1–6.

[20] E. T. Ceran, D. Gündüz, and A. Gürüy, “A reinforcement learning approach to age of information in multi-access networks with HARQ,” IEEE J. Sel. Areas Commun., vol. 39, no. 5, pp. 1412–1426, May 2021.

[21] H. Huang, D. Qiao, and M. C. Gursoy, “Age-energy tradeoff optimization for packet delivery in fading channels,” IEEE Trans. Wireless Commun., vol. 21, no. 1, pp. 179–190, Jan. 2022.

[22] H. B. Beytur and E. Uysal, “Age minimization of multiple flows using reinforcement learning,” in Proc. Int. Conf. Comput. Netw. Commun., 2019, pp. 339–343.

[23] M. Hatami, M. Leinonen, and M. Codreanu, “AoI minimization in status update control with energy harvesting sensors,” IEEE Trans. Commun., vol. 69, no. 12, pp. 8335–8351, Dec. 2021.

[24] L. Zhang and K. -W. Chin, “A distributed device selection method to minimize AoI in RF-charging networks,” IEEE Commun. Lett., vol. 25, no. 11, pp. 3733–3737, Nov. 2021.

[25] X. Wang, L. Liu, L. Zhu, and T. Tang, “Train-centric CBTC meets age of information in train-to-train communications,” IEEE Trans. Intell. Transp. Syst., vol. 21, no. 10, pp. 4072–4085, Oct. 2020.

[26] C. Zhouet al., “Deep RL-based trajectory planning for AoI minimization in UAV-assisted IoT,” in Proc. IEEE 11th Int. Conf. Wireless Commun. Signal Process., 2019, pp. 1–6.

[27] M. Akbari, M. R. Abedi, R. Joda, M. Pourghasemian, N. Mokari, and M. Erol-Kantarci, “Age of information aware VNF scheduling in industrial IoT-based systems,” IEEE Trans. Signal Process., vol. 69, no. 12, pp. 8335–8351, Dec. 2021.

[28] E. Altman, Constrained Markov Decision Processes, Boca Raton, FL, USA: CRC Press, 1995.

[29] M. H. Ngo and V. Krishnamurthy, “Monotonicity of constrained optimal transmission policies in correlated fading channels with ARQ,” IEEE Trans. Signal Process., vol. 58, no. 1, pp. 438–451, Jan. 2010.

[30] I. Moody and M. Safaee, “Learning to trade via direct reinforcement,” IEEE Trans. Neural Netw., vol. 21, no. 4, pp. 69–83, Jan. 1993.

[31] L. Baird and A. Moore, “Gradient descent for general reinforcement learning,” in Proc. Adv. Neural Inf. Process Syst., 1999, pp. 968–974.

[32] Z. Fang, J. Wang, Y. Ren, Z. Han, V. H. Poor, and L. Hanzo, “Age of information in energy harvesting aided massive multiple access networks,” IEEE J. Sel. Areas Commun., vol. 40, no. 4, pp. 1441–1456, May 2022.

[33] P. Zeng, G. Chen, Q. Wu, and D. Qiao, “Delay-aware multiple access design for intelligent reflecting surface aided uplink transmission,” in Proc. IEEE Globecom Workshops, 2022, pp. 1004–1009.
[35] Z. Ding, D. Xu, R. Schober, and H. V. Poor, “Hybrid NOMA offloading in multi-user MEC networks,” IEEE Trans. Wireless Commun., vol. 21, no. 7, pp. 5377-5391, Jul. 2022.

Jixuan Wang received the B.S. degree from Hangzhou Dianzi University, Hangzhou, China, in 2020, and the M. S. degree from East China Normal University, Shanghai, China, in 2023. His research focuses on AoI-based design for wireless communications.

Deli Qiao (Member, IEEE) received the B.E. degree in electrical engineering from the Harbin Institute of Technology, Harbin, China, in 2007, and the Ph.D. degree in electrical engineering from the University of Nebraska-Lincoln, Lincoln, NE, USA, in 2012. He is currently a Professor with East China Normal University, Shanghai, China. From 2012 to 2014, he was with Huawei Technologies Inc., as a Research Engineer, focusing on 5G techniques especially massive MIMO. His research interests include the general areas of wireless communications, information theory, and signal processing. He was the Editor of China Communications from 2016 to 2023, and has been an Associate Editor for IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY since 2022. He was the recipient of the Technical Committee on Green Communications and Computing (TCGCC) Best Journal Paper Award in 2017, and the Best Editor Award of China Communications 2019 and 2020.