Average AoI Minimization in an HARQ-based Status Update System under Random Arrivals

Saeid Sadeghi Vilni, Mohammad Moltafet, and Markus Leinonen
Centre for Wireless Communications – Radio Technologies
University of Oulu, Finland
e-mail: {saeid.sadeghivilni, mohammad.moltafet, markus.leinonen
@oulu.fi}

Marian Codreanu
Department of Science and Technology
Linköping University, Sweden
e-mail: {marian.codreanu@liu.se}

***Abstract*—We consider a status update system consisting of one source, one buffer-aided transmitter, and one receiver. The source randomly generates status update packets and the transmitter sends the packets to the receiver over an unreliable channel using a hybrid automatic repeat request (HARQ) protocol. The system holds two packets: one packet in the buffer, which stores the last generated packet, and one packet currently under service in the transmitter. At each time slot, the transmitter decides whether to stay idle, transmit the last generated packet, or re-transmit the packet currently under service. We aim to find the optimal actions at each slot to minimize the average age of information (AoI) of the source under a constraint on the average number of transmissions. We model the problem as a constrained Markov decision process (CMDP) problem and solve it for the known and unknown learning environment as follows. First, we use the Lagrangian approach to transform the CMDP problem to an MDP problem which is solved with the relative value iteration (RVI) for the known environment and with deep Q-learning (DQL) algorithm for the unknown environment. Second, we use the Lyapunov method to transform the CMDP problem to an MDP problem which is solved with DQL algorithm for the unknown environment. Simulation results assess the effectiveness of the proposed approaches.**

***Index Terms*—Age of information (AoI), hybrid automatic repeat request (HARQ), policy, constrained Markov decision process (CMDP), Lyapunov method, deep Q-learning (DQL).**

I. Introduction

The number of time-critical cyber-physical applications is growing dramatically. For instance, in Internet of things and smart grid applications, delivery of fresh and accurate information is crucial to the decision making to obtain high system performance. One highly attracted metric to measure the information freshness is the age of information (AoI) [1], [2]. AoI is defined as the difference between the current time and the generation time of the last received packet at a destination. The destination can be kept updated about the status of a random process by assigning the source to send status update packets. Each status update packet contains a timestamp representing the time when the sample was generated and the measured value of the monitored process. At time instant t, by denoting the timestamp of the last received status update packet by U_t, the AoI, δ_t, is defined as $\delta_t = t - U_t$ [1]–[4].

For the data transmission systems with an unreliable communication channel, packet re-transmissions can considerably improve the system's reliability [5]. Automatic repeat request (ARQ) protocols are standard error control methods [5], where after each transmission, the transmitter receives a feedback about the reception of the packet as acknowledgement/negative-acknowledgement (ACK/NACK). Whenever the transmitter receives NACK, it keeps re-transmitting the previous packet until it receives ACK or reaches the maximum allowed number of re-transmissions. While the decoding in ARQ protocols utilizes only the last transmitted packet, the hybrid ARQ (HARQ) protocols use all the received versions of the packet, increasing the probability of successful decoding of the packet [5], [6].

We consider an HARQ-based status update system under random packet arrivals in which the buffer-aided transmitter communicates with a receiver over an unreliable channel. The size of the system is two packets: one packet in the buffer, which stores the last generated packet, and one packet currently under service in the transmitter. We construct a constrained Markov decision process (CMDP) problem to minimize the average AoI under a constraint on the average number of transmissions to find a policy that determines at each time slot the optimal action: stay idle, transmit the last generated packet, or re-transmit the packet currently under service. We solve the CMDP problem by relaxing it to an MDP problem via two approaches: I) a Lagrangian-based approach and II) a Lyapunov-based approach. In the Lagrangian approach, the MDP problem is solved under two learning scenarios: i) relative value iteration (RVI) for the known environment (i.e., the transmitter knows the packet arrival rate and the HARQ decoding function) and ii) deep Q-learning (DQL) algorithm [7] for the unknown environment. In the Lyapunov approach, the MDP problem is solved with DQL algorithm for the unknown environment. The numerical results illustrate the AoI performances of the proposed policies.

Related Works: AoI characterization from the perspective of the queueing theory has been extensively studied; see, e.g., [8]–[12] and the references therein. In particular, the first work that analyzed AoI under an HARQ protocol is [10]. The authors derived the closed-form expression of the

This research has been financially supported by the Infotech Oulu, the Academy of Finland (grant 323698), and Academy of Finland 6Genesis Flagship (grant 318927). The work of M. Leinonen has also been financially supported by the Academy of Finland (grant 340171).

average AoI for an HARQ-based M/G/1/1 queueing system. The authors of [11] considered the same queueing system as in [10] and used the closed-form expression of the average AoI derived in [10] as the objective function. They minimized the objective function under a constraint on the decoding error probability by finding the optimal number of symbols to send at each transmission attempt. The authors of [12] introduced a network-code-HARQ (NC-HARQ) protocol and derived the closed-form expression of AoI for the NC-HARQ-based M/1/1 queueing system.

Also, the AoI has been studied in HARQ-based systems from the perspective of sampling and transmission policies [13], [14]. The most related work to our paper is [13], where the authors considered a similar HARQ-based status update system to ours, yet with the following differences. The work [13] considers a *generate-at-will* model, i.e., the source can sample the process and thus generate a new packet any time; we consider random arrivals at the buffer-aided transmitter. Since the transmitter is unaware of the availability of fresh packets at the next slots, the system is more complicated. Moreover, while we use the Lagrangian-based approach similar to [13] for the known and unknown environments, we additionally propose a low-complexity Lyapunov-based approach, in conjunction with DQL, for the unknown environment. Recently, [13] was extended to a multi-user setup in [14].

II. System Model and Problem Formulation

We consider a status update system in which a transmitter (TX) communicates with a receiver (RX) through an error-prone wireless channel, as depicted in Fig. 1. We consider a time-slotted system with slots indexed by $t \in \{1, 2, ...\}$. The source generates packets according to the Bernoulli random process with parameter p_{b} at the beginning of slots. An arriving packet is stored in the buffer of TX, which keeps the last generated packet. We assume that a packet arriving at slot t is accessible to TX at the same slot t. TX retains the packet currently under service until it takes a packet from the buffer. Thus, the size of the system is two packets, i.e., one packet at the buffer and one packet currently under service.

The system employs the chase combining HARQ protocol, where upon a re-transmission, the source sends the entire previous packet [6]. After each transmission, RX sends a feedback to TX about the reception status of the packet as ACK (successful decoding) or NACK (unsuccessful decoding). We assume an error-free and zero-delay feedback channel. Transmission of each packet is completed in one time slot.

A. Status Update Procedure

At the beginning of slot t, TX takes one of the following actions from the action space $\mathcal{A} = \{0, 1, 2\}$: I) stay idle, $a_t = 0$, II) re-transmit the last sent packet, $a_t = 1$, and III) send the packet in the buffer, $a_t = 2$. The decision is influenced by 1) the number of (re-)transmissions of the packet currently under service, 2) the age of each packet in the system, and 3) the AoI of the source at RX. These will be elaborated next.

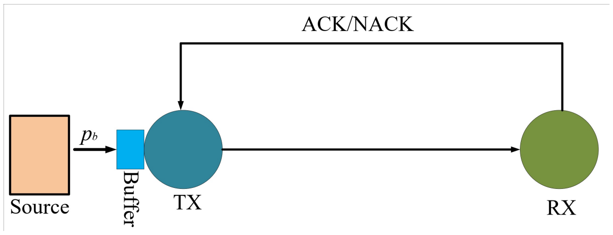


Figure 1. The considered status update system with an HARQ protocol. The transmitter (TX) has a buffer with capacity one which receives a new packet from the source according to the Bernoulli process with parameter p_{b}. The receiver (RX) sends a feedback as ACK for successful decoding and NACK otherwise.

Let x_t denote the number of transmissions of the packet currently under service at TX at time slot t. By utilizing the HARQ protocol, RX uses information from all transmission attempts (the first transmission and the re-transmissions) of the packet, which increases the probability of successful decoding. Thus, the probability of successful decoding of a packet is a non-decreasing function[1] of x_t, denoted as $f(x_t)$. The evolution of x_t is formulated as follows:

$$x_{t+1} = \begin{cases} 1 & a_t = 2 \\ x_t + 1 & a_t = 1 \\ x_t & a_t = 0. \end{cases} \tag{1}$$

Let δ_t be the AoI of the source at RX at the beginning of slot t, i.e., the number of time slots elapsed since the last received packet at RX was generated by the source. Let δ_t^{b} denote the age of the stored packet in the buffer at the beginning of time slot t, i.e., the number of time slots that the packet has stayed in the buffer up to slot t. Finally, let δ_t^{r} denote the age of the packet currently under service at TX at the beginning of time slot t. If at time slot t, TX transmits the packet with age δ_t^{r} and it is decoded successfully at RX, the AoI δ_t drops to $\delta_t^{\mathrm{r}} + 1$; otherwise, the AoI increases by one. Note that $\delta_t^{\mathrm{r}} = \delta_t^{\mathrm{b}}$ if TX transmits the packet in the buffer, i.e., $a_t = 2$. The evolution of the AoI, δ_t, can be formulated as

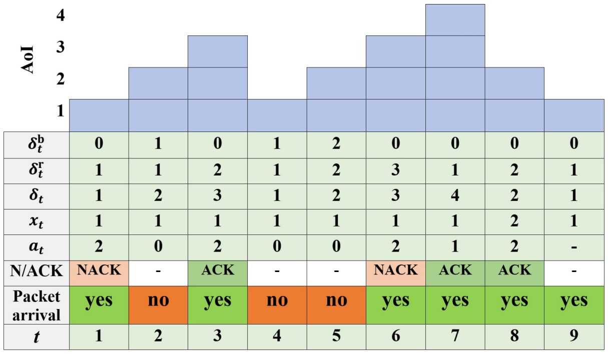
$$\delta_{t+1} = \begin{cases} \delta_t^{\mathrm{b}} + 1 & a_t = \{2\}, \text{ ACK} \\ \delta_t^{\mathrm{r}} + 1 & a_t = \{1\}, \text{ ACK} \\ \delta_t + 1 & a_t = \{1, 2\}, \text{ NACK} \\ \delta_t + 1 & a_t = 0. \end{cases} \tag{2}$$

We can conclude from (2) that $\delta_t \geq \delta_t^{\mathrm{r}}$. If TX transmits the packet in the buffer ($a_t = 2$), δ_t^{r} drops to δ_t^{b}, otherwise it increases by one. Thus, the evolution of δ_t^{r} is given by

$$\delta_{t+1}^{\mathrm{r}} = \begin{cases} \delta_t^{\mathrm{b}} + 1 & a_t = 2 \\ \delta_t^{\mathrm{r}} + 1 & \text{otherwise.} \end{cases} \tag{3}$$

From (2) and (3), we have $\delta_t \geq \delta_t^{\mathrm{r}} \geq \delta_t^{\mathrm{b}}$. If the buffer receives a new packet at slot $t + 1$, the age of the packet in the buffer

[1] The function $f(x_t)$ is a complicated function [15] and depends on several properties of the communication link such as the channel conditions and the channel coding methods. Thus, for the sake of simplicity, we consider a simple function in the numerical results.



δ_t^{b}	0	1	0	1	2	0	0	0	0
δ_t^{r}	1	1	2	1	2	3	1	2	1
δ_t	1	2	3	1	2	3	4	2	1
x_t	1	1	1	1	1	1	1	2	1
a_t	2	0	2	0	0	2	1	2	-
N/ACK	NACK	-	ACK	-	-	NACK	ACK	ACK	-
Packet arrival	yes	no	yes	no	no	yes	yes	yes	yes
t	1	2	3	4	5	6	7	8	9

Figure 2. Evolution of the age of the packet in the buffer (δ_t^{b}), the age of the packet currently under service at TX (δ_t^{r}), the AoI (δ_t), and the number of transmissions (x_t) for nine time slots. The actions (a_t), acknowledgements (ACK/NACK), and packet arrivals are also presented.

goes to zero, i.e., $\delta_{t+1}^{\mathrm{b}} = 0$; otherwise, $\delta_{t+1}^{\mathrm{b}} = \delta_t^{\mathrm{b}} + 1$. Thus, the evolution of δ_t^{b} can be formulated as

$$\delta_{t+1}^{\mathrm{b}} = \begin{cases} 0 & \text{a packet arrives at buffer at slot } t+1 \\ \delta_t^{\mathrm{b}} + 1 & \text{otherwise.} \end{cases} \tag{4}$$

The evolution of δ_t^{b}, δ_t^{r}, δ_t, and x_t are demonstrated in Fig. 2.

B. Problem Formulation

Our aim is to find the sequence of actions $\{a_t\}_{t=1,2,\dots}$ to minimize the expected long-term time average of the AoI while satisfying the constraint on the expected long-term time average number of transmissions. The expected long-term time average of the AoI, $\bar{\delta}$, is defined as

$$\bar{\delta} = \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\Big\{\sum_{t=1}^{T} \delta_t\Big\}, \tag{5}$$

where the expectation is taken over randomness of the system and the actions (this convention for $\mathbb{E}\{\cdot\}$ is used throughout the paper). The expected long-term time average number of transmissions, $\bar{P}$, is defined as

$$\bar{P} = \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\Big\{\sum_{t=1}^{T} P_t\Big\}, \tag{6}$$

where P_t indicates a transmission decision at slot t as

$$P_t = \begin{cases} 1 & a_t = 1,2 \\ 0 & a_t = 0. \end{cases} \tag{7}$$

Using (5) and (6), our considered AoI minimization problem is formulated as a stochastic optimization problem

$$\begin{aligned} \min_{\{a_t\}_{t=1,2,\dots}} \quad & \bar{\delta} \\ \text{s.t.} \quad & \bar{P} \le P^{\max}, \end{aligned} \tag{8}$$

where $P^{\max}$ is the maximum allowable average number of transmissions.

III. CMDP Problem and Proposed Solutions

In this section, we formulate the problem (8) as a CMDP problem and provide two approaches to solve it. In the first approach, we use the Lagrangian approach to transform the CMDP problem to an MDP problem. The MDP problem is then solved with RVI for the known environment and with DQL algorithm for the unknown environment. In the second approach, we use the Lyapunov method to transform the CMDP problem to an MDP problem and then solve the MDP problem via DQL.

The CMDP is defined by the five-tuple $(\mathcal{S}, \mathcal{A}, \mathrm{P}, C, D)$ [16, Sec. 2]. We define the state at time slot t as $s_t = (\delta_t^{\mathrm{b}}, \delta_t^{\mathrm{r}}, \delta_t, x_t) \in \mathcal{S}$, where $\mathcal{S}$ is the state space. The action space is $\mathcal{A} = \{0,1,2\}$, as defined in Section II. The state transition probability function $\mathrm{P}(s'|s,a) = \Pr(s' = s_{t+1}|s = s_t, a = a_t)$ gives the probability of moving from state s_t to state s_{t+1} when taking action a_t; P is specified below in (9). We define the objective cost as the instantaneous AoI, $C = \delta_t$. The transmission cost is defined as $D = D(a_t) = P_t$, which depends only on the current action.

Let us denote $p_{\mathrm{b}}' = 1 - p_{\mathrm{b}}$. The state transition probabilities of the CMDP at state $s_t = (\delta_t^{\mathrm{b}}, \delta_t^{\mathrm{r}}, \delta_t, x_t)$ are given as

$$\begin{aligned}
&\Pr\big((0, \delta_t^{\mathrm{r}} + 1, \delta_t{+}1, x_t)|s_t, 0\big) = p_{\mathrm{b}}, \\
&\Pr\big((\delta_t^{\mathrm{b}} + 1, \delta_t^{\mathrm{r}} + 1, \delta_t{+}1, x_t)|s_t, 0\big) = p_{\mathrm{b}}', \\
&\Pr\big((0, \delta_t^{\mathrm{r}} + 1, \delta_t^{\mathrm{r}}{+}1, x_t + 1)|s_t, 1\big) = p_{\mathrm{b}} f(x_t + 1), \\
&\Pr\big((0, \delta_t^{\mathrm{r}} + 1, \delta_t{+}1, x_t + 1)|s_t, 1\big) = p_{\mathrm{b}}\big(1 - f(x_t + 1)\big), \\
&\Pr\big((\delta_t^{\mathrm{b}} + 1, \delta_t^{\mathrm{r}} + 1, \delta_t^{\mathrm{r}}{+}1, x_t + 1)|s_t, 1\big) = p_{\mathrm{b}}' f(x_t + 1), \\
&\Pr\big((\delta_t^{\mathrm{b}}{+}1, \delta_t^{\mathrm{r}}{+}1, \delta_t{+}1, x_t{+}1)|s_t, 1\big) {=} p_{\mathrm{b}}'\big(1{-}f(x_t{+}1)\big), \\
&\Pr\big((0, \delta_t^{\mathrm{b}} + 1, \delta_t^{\mathrm{b}}{+}1, 1)|s_t, 2\big) = p_{\mathrm{b}} f(1), \\
&\Pr\big((0, \delta_t^{\mathrm{b}} + 1, \delta_t{+}1, 1)|s_t, 2\big) = p_{\mathrm{b}}\big(1 - f(1)\big), \\
&\Pr\big((\delta_t^{\mathrm{b}} + 1, \delta_t^{\mathrm{b}} + 1, \delta_t^{\mathrm{b}}{+}1, 1)|s_t, 2\big) = p_{\mathrm{b}}' f(1), \\
&\Pr\big((\delta_t^{\mathrm{b}} + 1, \delta_t^{\mathrm{b}} + 1, \delta_t{+}1, 1)|s_t, 2\big) = p_{\mathrm{b}}'\big(1 - f(1)\big),
\end{aligned} \tag{9}$$

for the other cases, the state transition probabilities are zero.

We define a policy $\pi : \mathcal{S} \to \mathcal{A}$ as a mapping from the states onto a distribution over the actions. Let $\bar{\delta}^{\pi}$ and $\bar{P}^{\pi}$ denote the average AoI in (5) and the average transmission cost in (6), obtained when following a policy π. Thus, our main problem in (8) is formulated as a CMDP problem

$$\begin{aligned} \min_{\pi} \quad & \bar{\delta}^{\pi} \\ \text{s.t.} \quad & \bar{P}^{\pi} \le P^{\max}. \end{aligned} \tag{10}$$

A. Lagrangian Approach

In this section, we utilize the Lagrangian approach to propose a policy that solves the CMDP problem (10). Based on [16, Theorem 4.4], if a CMDP is unichain, there is at least one stationary policy that solves the CMDP problem.

Theorem 1: The considered CMDP is unichain.

Proof: A CMDP is unichain, if under any policy, it induces a single recurrent class plus a possible empty set of transient states [17, pp. 78–80]. Consider now that our Markov chain is at state $s_t = (\delta_t^{\mathrm{b}}, \delta_t^{\mathrm{r}}, \delta_t, x_t)$, the source generates a new packet at slot t (i.e., $\delta_t^{\mathrm{b}} = 0$), TX takes action $a_t = 2$, and TX receives an ACK feedback. This means that the state moves to state $s_{t+1} = (\delta_{t+1}^{\mathrm{b}}, 1, 1, 0)$. Consider now that, in addition, the buffer receives a new packet at time slot $t+1$ (i.e., $\delta_{t+1}^{\mathrm{b}} = 0$). In other words, because the source keeps generating packets for any $p_{\mathrm{b}} \neq 0$, and RX is updated regularly by sending status updates, the system can go to the state $s = (0, 1, 1, 0)$ from every state with a non-zero probability. Therefore, the considered CMDP is a unichain CMDP. ■

We use a standard Lagrangian approach to relax the CMDP problem to an unconstrained MDP problem [16, Sec. 3.3]. The problem (10) is rewritten to its Lagrangian form for a Lagrangian multiplier λ as

$$\min_{\pi}. \quad \bar{L}(\pi, \lambda) = \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}\left\{ \sum_{t=1}^{T} \delta_t^{\pi} + \lambda \sum_{t=1}^{T} P_t^{\pi} \right\}. \tag{11}$$

The goal of the MDP problem (11) is to minimize the average Lagrangian cost $\bar{L}(\pi, \lambda)$ for a given λ. In the following subsections, we solve (11) under the known and unknown learning environment.

1) Known Learning Environment: Here, we consider the known environment where the transmitter knows the packet arrival rate and the HARQ decoding function. Based on [16, Theorem 4.3], an optimal policy π_λ^* for a given λ is obtained by solving the following Bellman equation:

$$\bar{L}^*(\lambda) + V(s_t, \lambda) = \min_{a_t \in \mathcal{A}} \left\{ L(\pi, \lambda) + \mathbb{E}\big\{V(s_{t+1}, \lambda)\big\} \right\}, \forall s_t \in \mathcal{S}, \tag{12}$$

where

$$L(\pi, \lambda) = \delta_t^{\pi} + \lambda P_t^{\pi}, \tag{13}$$

$$\bar{L}^*(\lambda) = \min_{\pi} \bar{L}(\pi, \lambda), \tag{14}$$

$$V(s_t, \lambda) = \mathbb{E}\Big\{ \sum_{t=1}^{\infty} \big(L(\pi, \lambda) - \bar{L}^*(\lambda) \big) | s_t \Big\}, \forall s_t \in \mathcal{S}, \tag{15}$$

where $L(\pi, \lambda)$ is the instantaneous Lagrangian cost, and $\bar{L}^*$ is the minimum average Lagrange cost achieved by the optimal policy π_λ^* for a given λ. $V(s_t, \lambda)$ is the value function for which the expectation in (12) is computed as

$$\mathbb{E}\big\{V(s_{t+1}, \lambda)\big\} = \sum_{s_{t+1} \in \mathcal{S}} \Pr(s_{t+1} | s_t, a_t) V(s_{t+1}, \lambda), \forall s_t \in \mathcal{S}. \tag{16}$$

Finally, the optimal policy π_λ^* for a given λ is obtained as

$$\pi_\lambda^*(s_t) = \operatorname*{arg\,min}_{a_t \in \mathcal{A}} \left\{ L(\pi, \lambda) + \mathbb{E}\big\{V(s_{t+1}, \lambda)\big\} \right\}, \forall s_t \in \mathcal{S}. \tag{17}$$

Let λ^* denote the optimal Lagrange multiplier with respect to CMDP problem (10). Since the CMDP problem (10) is unichain, according to [16, Theorem 4.4], there exists an optimal stationary policy π^*. In particular, π^* is a mixture of two deterministic policies π_1^* and π_2^* that are obtained by solving (11) given the optimal λ^* and they differ at most in one state. At that state, the policy π_1^* is selected with probability μ and the policy π_2^* is selected with probability $1 - \mu$, where $\mu \in [0, 1]$. However, finding λ^*, the optimal deterministic policies π_1^* and π_2^*, and the randomization factor μ is very difficult. Thus, similarly as the works [13], [17, p. 88], and [18], we use a heuristic method to find the two deterministic policies. Namely, we find two Lagrangian multipliers λ_1 and λ_2 so that 1) they both are very close to λ^*, 2) the obtained policy for λ_1 does not satisfy the transmission constraint, and 3) the obtained policy for λ_2 satisfies the constraint. Then, the heuristic policy $\bar{\pi}^*$ is derived by mixing these two policies such that the constraint is satisfied with equality, i.e., $\bar{P}^{\bar{\pi}^*} = P^{\max}$.

We search for λ^* by a gradient descent algorithm as

$$\lambda^{e+1} = \lambda^e + \sigma(\bar{P}^{\pi^*_{\lambda^e}} - P^{\max}), \tag{18}$$

where e is the iteration number and σ is the step size. In each iteration, an optimal policy $\pi^*_{\lambda^e}$ for the given λ^e is obtained by (17). Then, λ^e is updated based on (18). When λ^e converges, it is considered as λ^*. Then, we set $\lambda_1 = \lambda^* - \kappa$ and $\lambda_2 = \lambda^* + \kappa$, where the perturbation parameter κ is a sufficiently small constant and satisfies $\bar{P}^{\pi^*_{\lambda_1}} > P^{\max}$ and $\bar{P}^{\pi^*_{\lambda_2}} < P^{\max}$. By obtaining λ_1 and λ_2, we find the corresponding deterministic policies $\pi^*_{\lambda_1}$ and $\pi^*_{\lambda_2}$ as well as the average number of transmissions. The randomization factor is calculated as

$$\mu = \frac{P^{\max} - \bar{P}^{\pi^*_{\lambda_2}}}{\bar{P}^{\pi^*_{\lambda_1}} - \bar{P}^{\pi^*_{\lambda_2}}}. \tag{19}$$

To mix the obtained two deterministic policies $\pi^*_{\lambda_1}$ and $\pi^*_{\lambda_2}$, similarly as in [13], we consider the positive recurrent state $s = (0, 1, 1, 0)$ as the state in which we implement the randomization. In other words, whenever the chain reaches this state, the system chooses the policy $\pi^*_{\lambda_1}$ with probability μ, and $\pi^*_{\lambda_2}$ with probability $1 - \mu$.

To obtain the optimal policy π_λ^* for a given λ, we need to obtain $V(s, \lambda)$ (see (12)). To obtain $V(s, \lambda)$, we use the relative value iteration (RVI) [19]. To apply RVI, the state space needs to be finite. To this end, we assume an upper bound for the AoI δ_t as $\dot{\delta}$ [14], [20]. Accordingly, we redefine the AoI evolution (2) as

$$\delta_{t+1} = \begin{cases} \min\{\delta_t^{\mathrm{b}} + 1, \dot{\delta}\} & a_t = \{2\}, \text{ ACK} \\ \min\{\delta_t^{\mathrm{r}} + 1, \dot{\delta}\} & a_t = \{1\}, \text{ ACK} \\ \min\{\delta_t + 1, \dot{\delta}\} & a_t = \{1, 2\}, \text{ NACK} \\ \min\{\delta_t + 1, \dot{\delta}\} & a_t = 0. \end{cases}$$

We can conclude from (1)–(4) that δ_t is always greater than or equal to δ_t^{b}, δ_t^{r}, and x_t. Therefore, every state s_t is redefined as

$$s_t = \big(\min(\delta_t^{\mathrm{b}}, \dot{\delta}), \min(\delta_t^{\mathrm{r}}, \dot{\delta}), \min(\delta_t, \dot{\delta}), \min(x_t, \dot{\delta}) \big). \tag{20}$$

Now, the state space is finite, and we can use RVI to solve MDP problem (11).

2) Unknown Learning Environment: In the unknown environment, the decision-maker does not know the state transition probability function P. The solution steps are the same as in Section III-A1, however, instead of RVI, we use the DQL algorithm [7] to solve the MDP problem (11) for a given λ. After deriving λ_1 and λ_2, we mix their corresponding policies, $\pi^*_{\lambda_1}$ and $\pi^*_{\lambda_2}$, to derive the solution. In Section IV, we present the implementation of DQL and its parameters.

B. Lyapunov Approach

The major challenge in the CMDP problem (10) is to handle the average transmission constraint. As an alternative to the Lagrangian approach applied in Section III-A, we use here the Lyapunov method to transform the CMDP problem into an MDP problem [21]. Following the standard procedure of the Lyapunov method, we transform the constraint of the problem (10) into a virtual queue. Then, by stabilizing the virtual queue the constraint is satisfied [20], [21, Theorem 2.5]. Let Q_t denote the virtual queue associated with the constraint, which evolves as

$$Q_{t+1} = \max\{Q_t - P^{\max} + P_t, 0\}. \tag{21}$$

To enforce the stability of the virtual queue, we use the quadratic Lyapunov function defined as follows [21]

$$L\big(Q_t\big) = \frac{1}{2}Q_t^2. \tag{22}$$

If the Lyapunov function is small, then the queue is small, and if the Lyapunov function is large, then the queue is large. Thus, by minimizing the change of the Lyapunov function from one slot to the next slot, the virtual queue Q_t can be stabilized [21, Ch. 4].

The conditional Lyapunov drift $\Delta(\hat{s}_t)$, where $\hat{s}_t = \big(s_t, Q_t\big)$ is the current network state, is defined as the expected change in the Lyapunov function over one slot [21, Ch. 4]. Therefore, $\Delta(\hat{s}_t)$ is given by

$$\Delta(\hat{s}_t) = \mathbb{E}\big\{L\big(Q_{t+1}\big) - L\big(Q_t\big)|\hat{s}_t\big\}.$$

Using the drift-plus-penalty minimization method, the CMDP problem (10) is transformed to the following MDP problem [22]

$$\min_{\pi}. \quad \limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\big(\Delta(\hat{s}_t) + v\mathbb{E}\{\delta_t^\pi|\hat{s}_t\}\big), \tag{23}$$

where v is a positive parameter to adjust the trade-off between minimizing the objective function and constraint of the problem (10), and π is a stationary policy which configures the action a_t according to state s_t in each slot t.

The MDP problem in (23) is characterized by the four-tuple $(\hat{\mathcal{S}}, \mathcal{A}, \hat{\mathrm{P}}, \hat{C})$, where $\hat{\mathcal{S}}$ is the state space. The action space is $\mathcal{A} = \{0, 1, 2\}$, similarly as in Section III. The term $\hat{\mathrm{P}}$ denotes the state transition probability function. We define the objective cost as $\hat{C} = \hat{C}(\hat{s}_t, a_t) = L(Q_{t+1}) - L(Q_t) + v\delta_t$. We consider MDP problem (23) in the unknown learning environment in which we do not need the state transition probability function. Then, using the DQL algorithm presented in [7], we solve MDP problem (23).

Table I
SYSTEM PARAMETERS AND THEIR VALUES

Parameter	Symbol	Value
Maximum allowable average number of transmissions	$P^{\max}$	0.5
The parameter of function $f(\cdot)$	p_0	0.2
Maximum AoI	$\hat{\delta}$	20
The drift-plus-penalty method parameter	v	2.5
Step size	σ	0.5
The perturbation parameter	κ	0.05

IV. NUMERICAL RESULTS

In this section, we evaluate the performance of the proposed policies. In the spirit of [13], [14], we consider that the probability of successful decoding is given by $f(x_t) = 1 - p_0^{x_t}$, where $p_0 \in [0, 1]$. The main system parameters are provided in Table I.

Fig. 3 depicts the average AoI with respect to episodes for the proposed policies, i.e., the Lagrangian approach with RVI for the known environment, the Lagrangian approach with DQL algorithm for the unknown environment, and the Lyapunov method with DQL algorithm for the unknown environment. In both Lagrangian-based policies, we have $\lambda_1 = 1.6$ and $\lambda_2 = 1.7$. As can be seen, in the known environment, as the transmitter knows the state transition probability function, it results in the minimum average AoI among the proposed policies. In the unknown environment, the Lagrangian and Lyapunov approaches perform close to each other. Note that the Lyapunov approach needs only to select a reasonably high value for the parameter v, whereas the Lagrangian approach is subject to a tedious search for the Lagrangian multipliers and randomization factor. In addition, we compare the performance of the proposed algorithms to a fixed scheduling method representing a baseline policy. According to the baseline policy which satisfies the average number of transmissions constraint, TX transmits the last generated packet until the packet is delivered successfully or a new packet arrives, also RX uses all the received versions of a packet for decoding. As it can be seen, the proposed policies outperform the baseline policy.

In Fig. 4, we investigate the impact of the maximum allowable average number of transmissions and the packet arrival rates, p_{b}, on the average AoI performance. Here, we utilize the Lagrangian approach with RVI for the known environment. According to Fig. 4, as expected, the average AoI decreases when $P^{\max}$ increases, as TX can take more transmission attempts. In addition, the average AoI decreases when p_{b} increases, as TX will have fresh packets available more frequently. However, it can be seen from Fig. 4 that the average AoI values for the two different packet arrival rates are very close when $P^{\max} = 0.2$. This behavior is because the limitation on the number of transmissions is the dominant factor in the status updating related to the fresh packet availability.

V. CONCLUSION

We considered an HARQ-based status update system under random arrivals. We constructed a CMDP problem to find a

Table II
DQL ALGORITHM PARAMETERS AND THEIR VALUES

Parameter	Value
Learning rate	0.001
Hidden layers	(256,256)
Hidden activation	ReLU
Optimizer	Adam
Minibatch size	64
Training episodes	2100
Steps per episode	500
Replay memory size	100000
Discount factor	0.99

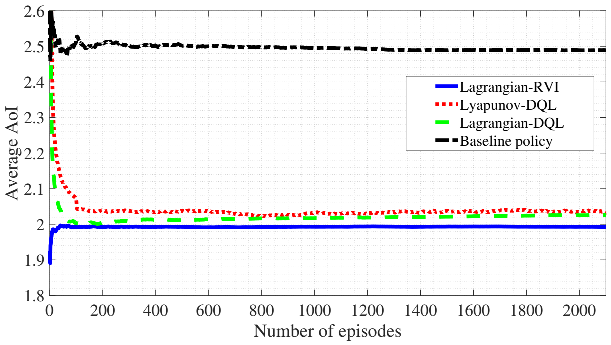


Figure 3. The average AoI over the number of episodes for the proposed policies where $\lambda_1 = 1.6$, $\lambda_2 = 1.7$, and $p_{\rm b} = 0.7$.

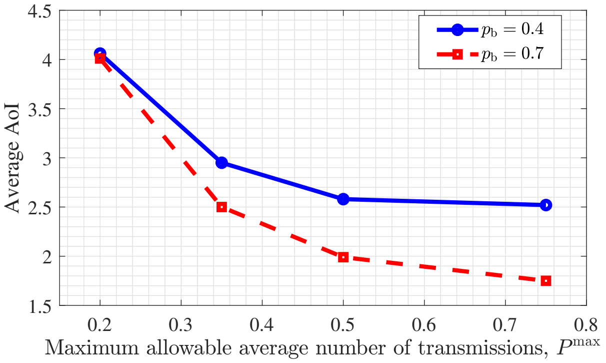


Figure 4. The average AoI for the Lagrangian approach with RVI versus the maximum allowable average number of transmissions, $P^{\max}$, for two different packet arrival rates, $p_{\rm b}$.

policy that minimizes the average AoI under the constraint on the average number of transmissions. We proposed two approaches to solve the problem for known and unknown environments. In the first approach, we used the Lagrangian approach to transform the CMDP problem to an MDP problem. We solved the MDP problem with the RVI algorithm for the known environment and the DQL algorithm for the unknown environment. In the second approach, we transformed the CMDP problem to an MDP problem with the Lyapunov drift-plus-penalty method and then solved the MDP problem with the DQL algorithm for the unknown environment. The numerical results showed that the proposed policies lead to significantly lower average AoI than that of the baseline policy. Moreover, the proposed DQL-based policies, which do not know the packet arrival rate and the probability of successful decoding, perform close to the RVI-based policy.

REFERENCES

[1] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *Proc. IEEE Int. Conf. on Computer Commun. (INFOCOM)*, Orlando, FL, USA, Mar. 25–30, 2012, pp. 2731–2735.

[2] S. K. Kaul, R. D. Yates, and M. Gruteser, "Status updates through queues," in *Proc. Conf. Inform. Sciences Syst. (CISS)*, Princeton, NJ, USA, Mar.21–23, 2012, pp. 1–6.

[3] R. D. Yates, "The age of information in networks: Moments, distributions, and sampling," *IEEE Trans. Inform. Theory*, vol. 66, no. 9, pp. 5712–5728, Sep. 2020.

[4] A. Kosta, N. Pappas, and V. Angelakis, "Age of information: A new concept, metric, and tool," *Found. Trends Netw.*, vol. 12, no. 3, pp. 162–259, Nov. 2017.

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

[6] "IEEE standard for air interface for broadband wireless access systems," *IEEE Std 802.16-2017 (Revision of IEEE Std 802.16-2012)*, pp. 1–2726, Mar. 2018.

[7] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[8] R. D. Yates and S. K. Kaul, "The age of information: Real-time status updating by multiple sources," *IEEE Trans. Inform. Theory*, vol. 65, no. 3, pp. 1807–1827, Mar. 2018.

[9] M. Moltafet, M. Leinonen, and M. Codreanu, "Average AoI in multi-source systems with source-aware packet management," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 1121–1133, Feb. 2021.

[10] E. Najm, R. Yates, and E. Soljanin, "Status updates through M/G/1/1 queues with HARQ," in *Proc. IEEE Int. Symp. Inform. Theory*, Aachen, Germany, Jun. 25–30, 2017, pp. 131–135.

[11] D. Li, S. Wu, Y. Wang, J. Jiao, and Q. Zhang, "Age-optimal HARQ design for freshness-critical satellite-IoT systems," *IEEE Internet of Things J.*, vol. 7, no. 3, pp. 2066–2076, Mar. 2020.

[12] S. Liu, J. Jiao, Z. Ni, S. Wu, and Q. Zhang, "Age-optimal NC-HARQ protocol for multi-hop satellite-based internet of things," in *Proc. IEEE Wireless Commun. and Networking Conf.*, Nanjing, China, Mar. 29–1, 2021, pp. 1–6.

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

[14] E. T. Ceran, D. Gündüz, and A. György, "A reinforcement learning approach to age of information in multi-user networks with HARQ," *IEEE J. Select. Areas Commun.*, vol. 39, no. 5, pp. 1412–1426, May 2021.

[15] X. Lagrange, "Throughput of HARQ protocols on a block fading channel," *IEEE Commun. Lett.*, vol. 14, no. 3, pp. 257–259, Mar. 2010.

[16] E. Altman, *Constrained Markov decision processes*. CRC Press, 1999.

[17] P. Agrawal, M. D. Andrews, P. J. Fleming, G. G. Yin, and L. Zhang, *Wireless communications*. Springer Science & Business Media, 2010.

[18] B. Zhou and W. Saad, "Joint status sampling and updating for minimizing age of information in the internet of things," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7468–7482, Nov. 2019.

[19] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.

[20] M. Moltafet, M. Leinonen, M. Codreanu, and N. Pappas, "Power minimization for age of information constrained dynamic control in wireless sensor networks," *arXiv preprint arXiv:2007.05364*, Jul. 2020.

[21] M. J. Neely, *Stochastic network optimization with application to communication and queueing systems*. Morgan & Claypool Publishers, 2010.

[22] W. Wu, P. Yang, W. Zhang, C. Zhou, and X. Shen, "Accuracy-guaranteed collaborative DNN inference in industrial IoT via deep reinforcement learning," *IEEE Trans. Ind. Informat.*, vol. 17, no. 7, pp. 4988–4998, Aug. 2021.