Average Age of Information with Hybrid ARQ under a Resource Constraint

Elif Tuğçe Ceran, Deniz Gündüz, and András György
Department of Electrical and Electronic Engineering
Imperial College London
Email: {e.ceran14, d.gunduz, a.gyorgy}@imperial.ac.uk

Abstract

Scheduling the transmission of status updates over an error-prone communication channel is studied in order to minimize the long-term average *age of information* at the destination under a constraint on the average number of transmissions at the source node. After each transmission, the source receives an instantaneous ACK/NACK feedback, and decides on the next update without prior knowledge on the success of future transmissions. The optimal scheduling policy is first studied under different feedback mechanisms when the channel statistics are known; in particular, the standard automatic repeat request (ARQ) and hybrid ARQ (HARQ) protocols are considered. Structural results are derived for the optimal policy under HARQ, while the optimal policy is determined analytically for ARQ. For the case of unknown environments, an average-cost reinforcement learning algorithm is proposed that learns the system parameters and the transmission policy in real time. The effectiveness of the proposed methods is verified through numerical results.

Index Terms

Age of information, hybrid automatic repeat request (HARQ), constrained Markov decision process, reinforcement learning

I. Introduction

Motivated by the growing interest in timely delivery of information in status update systems, the *age of information (AoI)* has been introduced as a performance measure to quantify data staleness at the receiver [2]–[4]. Consider a source node that samples an underlying time-varying process and sends the sampled status of the process over an imperfect communication channel that introduces

Part of this work was presented at the IEEE Wireless Communications and Networking Conference, Barcelona, Spain, April 2018 [1].

delays. The AoI characterizes the data staleness (or tardiness) at the destination node, and it is defined as the time that has elapsed since the most recent status update available at the destination was generated. Different from classical performance measures, such as the delay or throughput, AoI jointly captures the latency in transmitting updates and the rate at which they are delivered.

Our goal in this paper is to minimize the average AoI at the destination taking into account *retransmissions* due to errors over the noisy communication channel. Retransmissions are essential for providing reliability of status updates over error-prone channels, particularly in wireless settings. Here, we analyze the AoI for both the standard ARQ and hybrid ARQ (HARQ) protocols.

In the HARQ protocol, the receiver combines information from all previous transmission attempts of the same packet in order to increase the success probability of decoding [5], [6], [7]. The exact relationship between the probability of error and the number of retransmission attempts varies depending on the channel conditions and the particular HARQ method employed [5], [6], [7]. In general, the probability of successful decoding increases with each transmission, but the AoI of the received packet also increases. Therefore, there is an inherent trade-off between retransmitting previously failed status information with a lower error probability, or sending a fresh status update with higher error probability. We address this trade-off between the success probability and the freshness of the status update to be transmitted, and develop scheduling policies to minimize the expected average AoI.

In the standard ARQ protocol, if a packet cannot be decoded, it is retransmitted until successful reception. Note, however, that, when optimizing for the AoI, there is no point in retransmitting the same packet, since a newer packet with more up-to-date information is available at the sender at the time of retransmission. Thus, after the reception of a NACK feedback, the actual packet is discarded, and the most recent status of the underlying process is transmitted (the exact timing of the transmission may depend on the feedback, i.e., on the success history of previous transmissions). A scheduling policy to decide whether to stay idle or transmit a status update should be designed considering a resource constraint on the average number of transmissions.

We develop scheduling policies for both the HARQ and the standard ARQ protocols to minimize the expected average AoI under a constraint on the average number of transmissions, which is motivated by the fact that sensors sending status updates have usually limited energy supplies (e.g., are powered via energy harvesting [8]); and hence, they cannot afford to send an unlimited number of updates, or increase the signal-to-noise-ratio in the transmission. First, we assume that the success probability before each transmission attempt is known (which, in the case of HARQ, depends on the number of previous unsuccessful transmission attempts); and therefore, the source node can

judiciously decide when to retransmit and when to discard a failed packet and send a fresh update. Then, we consider transmitting status updates over an unknown channel, in which case the success probabilities of transmission attempts are not known *a priori*, and must be learned in an online fashion. This latter scenario can model sensors embedded in unknown or time-varying environments. We employ reinforcement learning (RL) algorithms to balance exploitation and exploration in an unknown environment, so that the source node can quickly learn the environment based on the ACK/NACK feedback signals, and can adapt its scheduling policy accordingly, exploiting its limited resources in an efficient manner.

The main contributions of this paper are as follows:

- Average AoI is studied under a long-term average resource constraint imposed on the transmitter, which limits the average number of transmissions.
- Both retransmissions and pre-emption following a failed transmission are considered, corresponding, respectively, to the HARQ and ARQ protocols, and the structure of the optimal policy is determined in general.
- The optimal preemptive transmission policy for the standard ARQ protocol is shown to be a threshold-type randomized policy, and is derived in closed-form.
- An average-cost RL algorithm; in particular, *average-cost SARSA with softmax*, is proposed to learn the optimal scheduling decisions when the transmission success probabilities are unknown.
- Extensive numerical simulations are conducted in order to show the effect of feedback, resource constraint and ARQ or HARQ mechanisms on the freshness of the data.

A. Related Work

Most of the earlier work on AoI consider queue-based models, in which the status updates arrive at the source node randomly following a memoryless Poisson process, and are stored in a buffer before being transmitted to the destination over a noiseless channel [3], [4]. Instead, in the so-called *generate-at-will* model, [2], [9]–[12], also adopted in this paper, the status of the underlying process can be sampled at any time by the source node.

A constant packet failure probability for a status update system is investigated for the first time in [13], where status updates arrive according to a Poisson process, while the transmission time for each packet is exponentially distributed. Fast-come-first-served (FCFS) scheduling is analyzed and it is shown that packet loss and large queuing delay due to old packets in the queue result in an increase in the AoI. Different scheduling decisions at the source node are investigated; including the last-come-first-served (LCFS) principle, which always transmits the most up-to-date packet, and

retransmissions with preemptive priority, which preempts the current packet in service when a new packet arrives.

Broadcasting of status updates to multiple receivers over an unreliable broadcast channel is considered in [10]. A low complexity sub-optimal scheduling policy is proposed when the AoI at each receiver and the transmission error probabilities to all the receivers are known. However, only work-conserving policies are considered in [10], which update the information at every time slot, since no constraint is imposed on the number of updates. Optimizing the scheduling decisions with multiple receivers over a perfect channel is investigated in [11], and it is shown that there is an optimal scheduling algorithm that is of threshold-type. To our knowledge, the latter is the only prior work in the literature which applies RL in the AoI framework. However, their goal is to learn the data arrival statistics, and it does not consider either an unreliable communication link or HARQ. Moreover, we employ an average-cost RL method, which has significant advantages over discounted-cost methods, such as *Q-learning* [14].

The AoI in the presence of HARQ has been considered in [15], [16] and [17]. In [15] the affect of design decisions, such as the length of the transmitted codewords, on the average AoI is analyzed. The status update system is modeled as an M/G/1/1 queue in [16]; however, no resource constraint is considered, and the status update arrivals are assumed to be memoryless and random, in contrast to our work, which considers the *generate-at-will* model. Moreover, a specific coding scheme is assumed in [16], namely MDS (maximum distance separable) coding, which results in a particular formula for the successful decoding probabilities, whereas we allow general functions for these probabilities. From a queuing-systems perspective, our model can be considered as a G/G/1/1 queue with optimization of packet arrivals and pre-emption. In [17], HARQ is considered in a zero-wait system, where as soon as an update is successfully transmitted to the destination, the source starts transmitting a new status update, as no resource constraint or pre-emption is taken into account.

In [2] and [18], the receiver can choose to update its status information by downloading an update over one of the two available channels, a free yet unreliable channel, modeling a Wi-Fi connection, and a reliable channel with a cost, modeling a cellular connection. Although the Lagrangian formulation of our constrained optimization problem for the standard ARQ protocol is similar to the one considered in [2], our problem is more complicated due to several reasons: they have not considered the effect of retransmissions or any algorithm that learns the unknown system parameters, and even without these complications, we need to determine the Lagrange multiplier corresponding to the given constraints, while it is given in [2].

To the best of our knowledge, this is the first work in the literature that addresses a status update

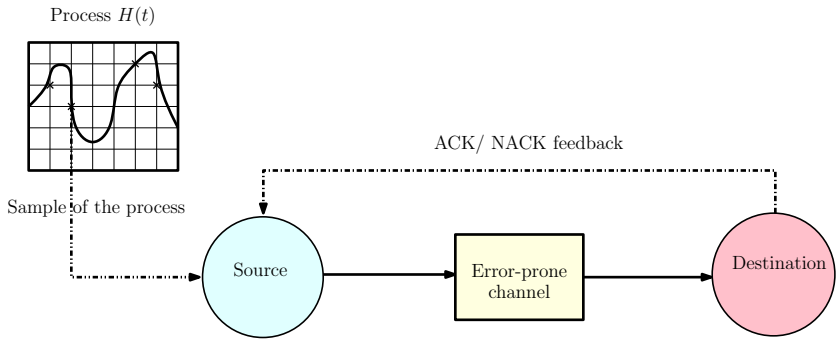


Figure 1. System model of a status update system over an error-prone point-to-point link in the presence of ACK/NACK feedback from the destination.

system with HARQ in the presence of resource constraints. In addition, no previous work has studied the average AoI over a channel with unknown error probabilities, and employed an average-cost RL algorithm.

II. System Model and Problem Formulation

We consider a time-slotted status update system over an error-prone communication link (see Figure 1). The source monitors an underlying time-varying process, and can generate a status update at any time slot; known as the *generate-at-will* model [12]. The status updates are communicated from the source node to the destination over a time-varying channel. Each transmission attempt of a status update takes constant time, which is assumed to be equal to the duration of one time slot. Throughout the paper, we will normalize all time durations by the duration of one time slot.

We assume that the channel changes randomly from one time slot to the next in an independent and identically distributed fashion, and the channel state information is available only at the destination node. We further assume the availability of an error- and delay-free single-bit feedback from the destination to the source node for each transmission attempt. Successful receipt of a status update is acknowledged by an ACK signal, while a NACK signal is sent in case of a failure. In the classical ARQ protocol, a packet is retransmitted after each NACK feedback, until it is successfully decoded (or a maximum number of allowed retransmissions is reached), and the received signal is discarded after each failed transmission attempt. Therefore, the probability of error is the same for all retransmissions. However, in the AoI framework there is no point in retransmitting a failed out-of-date status packet if it has the same error probability as that of a fresh update. Hence, we assume that if the ARQ protocol is adopted, the source always removes failed packets and transmits a fresh status update. On the other hand, if the HARQ protocol is used, the received signals from all previous transmission

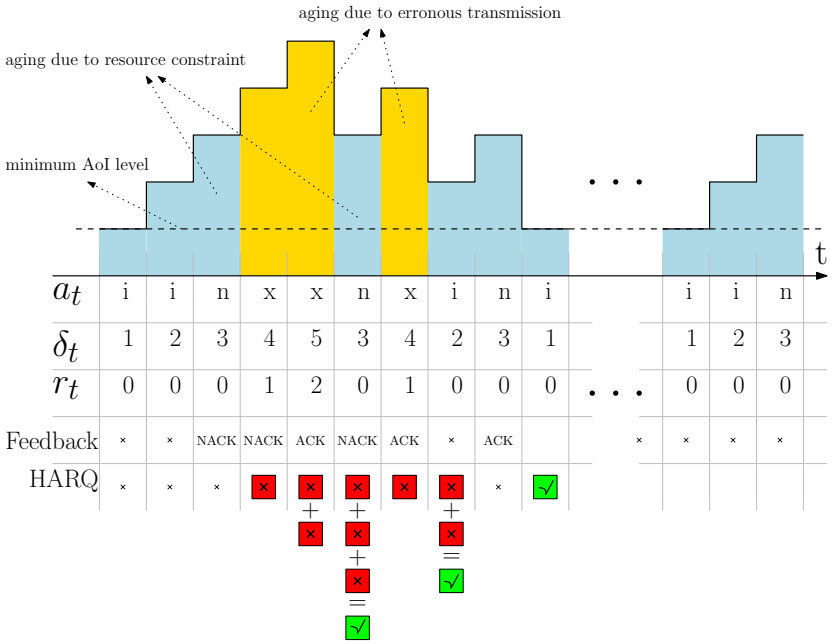


Figure 2. Illustration of the AoI in a slotted status update system with HARQ. (δ_t, r_t) represents the state of the system and the action is chosen based on the state (δ_t, r_t) and denoted by a_t. Packets with decoding errors (represented by red squares) are stored in the receiver and combined to decode the information successfully (represented by green squares).

attempts for the same packet are combined for decoding. Therefore, the probability of error decreases with every retransmission. In general, the error probability of each retransmission attempt depends on the particular combination technique used by the decoder, as well as on the channel conditions [5].

AoI measures the timeliness of the information at the receiver. It is defined as the number of time slots elapsed since the generation of the most up-to-date packet successfully decoded at the receiver. Formally, denoting the latter generation time for any time slot t by $U(t)$, the AoI, denoted by δ_t, is defined as

$$\delta_t \triangleq t - U(t). \tag{1}$$

We assume that a transmission decision is made at the beginning of each slot. The AoI increases by one when the transmission fails, while it decreases to one in the case of ARQ, or to the number of retransmissions plus one in the case of HARQ, when a status update is successfully decoded (the minimum age is set to 1 to reflect that the transmission is one slot long).

The probability of error after r retransmissions, denoted by $g(r)$, depends on r and the particular HARQ scheme used for combining multiple transmission attempts (an empirical method to estimate

$g(r)$ is presented in [6]). As in any reasonable HARQ strategy, we assume that $g(r)$ is non-increasing in the number of retransmissions r; that is, $g(r_1) \geq g(r_2)$ for all $r_1 \leq r_2$. For simplicity, we assume that $0 < g(0) < 1$, that is, the channel is noisy and there is a possibility that the first transmission is successful (if $g(0) = 0$, the problem becomes trivial, while $g(0) < 1$ can be easily relaxed to the condition that there exists an r such that $g(r) < 1$). Also, we will denote the maximum number of retransmissions by r_{max}, which may take the value ∞, unless otherwise stated. However, if $g(r) = 0$ for some r (i.e., a packet is always correctly decoded after r retransmissions), we set r_{max} to be the smallest such r. Finally note that standard HARQ methods only allow a finite maximum number of retransmissions [19].

For any time slot t, let $\delta_t \in \mathbb{Z}^+$ denote the AoI at the beginning of the time slot and $r_t \in \{0, \ldots, r_{max}\}$ denote the number of previous transmission attempts of the same packet. Then the state of the system can be described by $s_t \triangleq (\delta_t, r_t)$. At each time slot, the source node takes one of the three actions, denoted by $a \in \mathcal{A}$, where $\mathcal{A} = \{\mathrm{i}, \mathrm{n}, \mathrm{x}\}$: (i) remain idle ($a = \mathrm{i}$); (ii) transmit a new status update ($a = \mathrm{n}$); or (iii) retransmit the previously failed update ($a = \mathrm{x}$). The evolution of AoI for a slotted status update system is illustrated in Figure 2.

Note that if no resource constraint is imposed on the source, remaining idle is clearly suboptimal since it does not contribute to decreasing the AoI. However, continuous transmission is typically not possible in practice due to energy or interference constraints. Accordingly, we impose a constraint on the average number of transmissions, and we require that the long-term average number of transmission do not exceed $C_{max} \in (0, 1]$ (note that $C_{max} = 1$ corresponds to the case in which transmission is allowed in every slot).

This leads to the *constrained Markov decision process* (CMDP) formulation, defined by the 5-tuple $\big(\mathcal{S}, \mathcal{A}, \mathrm{P}, c, d\big)$ [20]: The countable set of states $(\delta, r) \in \mathcal{S}$ and the finite action set $\mathcal{A} = \{\mathrm{i}, \mathrm{n}, \mathrm{x}\}$ have already been defined. P refers to the transition function, where $\mathrm{P}(s'|s, a) = \Pr(s_{t+1} = s' \mid s_t = s, a_t = a)$ is the probability that action a in state s at time t will lead to state s' at time $t + 1$, which will be explicitly defined in (4). The cost function $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, is the AoI at the destination, and is defined as $c((\delta, r), a) = \delta$ for any $(\delta, r) \in \mathcal{S}$, $a \in \mathcal{A}$, independently of action a. The transmission cost $d : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is independent of the state and depends only on the action a, where $d = 0$ if $a = \mathrm{i}$, and $d = 1$ otherwise. Since our goal is to minimize the AoI subject to a constraint on the average transmission cost, the corresponding problem is a CMDP.

A policy is a sequence of decision rules $\pi_t : (\mathcal{S} \times \mathcal{A})^t \to [0, 1]$, which maps the past states and actions and the current state to a distribution over the actions; that is, after the state-action sequence $s_1, a_1, \ldots, s_{t-1}, a_{t-1}$, action a is selected in state s_t with probability $\pi_t(a_t|s_1, a_1, \ldots, s_{t-1}, a_{t-1}, s_t)$.

We use $s_t^\pi = (\delta_t^\pi, r_t^\pi)$ and a_t^π to denote the sequences of states and actions, respectively, induced by policy $\pi = \{\pi_t\}$. A policy $\pi = \{\pi_t\}$ is called *stationary* if the distribution of the next action is independent of the past states and actions given the current state, and time invariant; that is, with a slight abuse of notation, $\pi_t(a_t|s_1, a_1, \ldots, s_{t-1}, a_{t-1}, s_t) = \pi(a_t|s_t)$ for all t and $(s_i, a_i) \in \mathcal{S} \times \mathcal{A}$, $i = 1, \ldots, t$. Finally, a policy is said to be deterministic if it chooses an action with probability one; with a slight abuse of notation, we use $\pi(s)$ to denote the action taken with probability one in state s by a stationary deterministic policy.

Let $J^\pi(s_0)$ and $C^\pi(s_0)$ denote the infinite horizon average age and the average number of transmissions, respectively, when policy π is employed with initial state s_0. Then the CMDP optimization problem can be stated as follows:

Problem 1.

$$\text{Minimize } J^\pi(s_0) \triangleq \limsup_{T\to\infty} \frac{1}{T}\, \mathbb{E}\left[\sum_{t=1}^{T} \delta_t^\pi \Big| s_0\right], \tag{2}$$

$$\text{subject to } C^\pi(s_0) \triangleq \limsup_{T\to\infty} \frac{1}{T}\, \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[a_t^\pi \neq \mathrm{i}] \Big| s_0\right] \leq C_{max}. \tag{3}$$

A policy π that is a solution of the above minimization problem is called optimal, and we are interested in finding optimal policies. Without loss of generality, we assume that the sender and the receiver are synchronized at the beginning of the problem, that is, $s_0 = (1, 0)$; and we omit s_0 from the notation for simplicity.

Before formally defining the transition function P in our AoI problem, we present a simple observation that simplifies P: Retransmitting a packet immediately after a failed attempt is better than retransmitting it after waiting for some slots. This is true since waiting increases the age, without increasing the success probability. The difference in the waiting time is illustrated in Figure 3 for a simple scenario, where the first transmission of a status update results in a failure, while the retransmission is successful.

Proposition 1. *For any policy π there exists another policy π' (not necessarily distinct from π) such that $J^{\pi'}(s_0) \leq J^\pi(s_0)$, $C^{\pi'}(s_0) \leq C^\pi(s_0)$, and π' takes a retransmission action only following a failed transmission, that is, the probability $Pr(a_{t+1}^{\pi'} = \mathrm{x}|a_t^{\pi'} = \mathrm{i}) = 0$.*

The transition probabilities are given as follows (omitting the parenthesis from the state variables

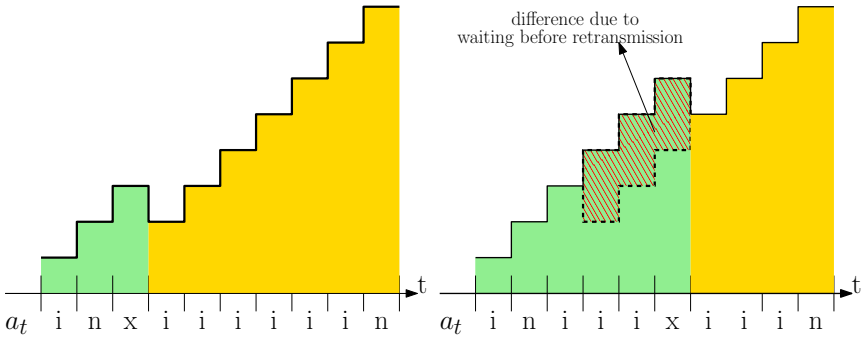


Figure 3. The difference of the AoI for policies without and with idle slots before retransmissions. The figure on the left shows the evolution of age (height of the bars) when retransmission occurs immediately after an error in transmission whereas the figure on the right represents the evolution of age when retransmission occurs after some idle slots.

(δ, r)):

$$\begin{aligned}
&\mathrm{P}(\delta+1, 0|\delta, r, \mathrm{i}) = 1,\\
&\mathrm{P}(\delta+1, 1|\delta, r, \mathrm{n}) = g(0),\\
&\mathrm{P}(1, 0|\delta, r, \mathrm{n}) = 1-g(0),\\
&\mathrm{P}(\delta+1, r+1|\delta, r, \mathrm{x}) = g(r),\\
&\mathrm{P}(r+1, 0|\delta, r, \mathrm{x}) = 1-g(r),
\end{aligned} \tag{4}$$

and $\mathrm{P}(\delta', r'|\delta, r, a) = 0$ otherwise. Note that the above equations set the retransmission count to 0 after each successful transmission, and it is not allowed to take a retransmission action in states where the transmission count is 0. Also, the property in Proposition 1 is enforced by the first equation in (4), that is, $\mathrm{P}(\delta+1, 0|\delta, r, \mathrm{i}) = 1$ (since retransmissions are not allowed in states $(\delta, 0)$). Since the starting state is $(1, 0)$, it also follows that the state set of our CMDP can be described as

$$\mathcal{S} = \{(\delta, r) : r < \min\{\delta, r_{max}+1\}, \delta, r \in \mathbb{N}\}\ . \tag{5}$$

III. Lagrangian Relaxation and the Structure of the Optimal Policy

In this section, we derive the structure of the optimal policy for Problem 1 based on [21]. A detailed treatment of finite state-finite action CMDPs is considered in [20], but here we need more general results that apply to countable state spaces. These results require certain technical conditions; roughly speaking, there must exist a deterministic policy that satisfies the transmission constraint while maintaining a finite average AoI, and any "reasonable" policy must induce a positive recurrent Markov chain. The precise formulation of the requirements is given in Appendix A, wherein Proposition 2

shows that the conditions of [21] are satisfied for Problem 1. Given this result, we follow [21] to characterize the optimal policy.

While there exists a stationary and deterministic optimal policy for countable-state finite-action average-cost MDPs [22]–[24], this is not necessarily true for CMDPs [20], [21]. To solve the CMDP, we start with rewriting the problem in its Lagrangian form. The average Lagrangian cost of a policy π with Lagrange multiplier $\eta \geq 0$ is defined as

$$L_\eta^\pi = \lim_{T\to\infty} \frac{1}{T}\left(\mathbb{E}\left[\sum_{t=1}^{T} \delta_t^\pi\right] + \eta \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[a_t^\pi \neq \mathrm{i}]\right]\right), \tag{6}$$

and, for any η, the optimal achievable cost L_η^* is defined as $L_\eta^* \triangleq \min_\pi L_\eta^\pi$. If the constraint on the transmission cost is less than one (i.e., $C_{max} < 1$), then we have $\eta > 0$, which will be assumed throughout the paper.[1] This formulation is equivalent to an unconstrained countable-state average-cost MDP in which the instantaneous overall cost is $\delta_t + \eta \mathbb{1}[a_t^\pi \neq \mathrm{i}]$. A policy π is called η-optimal if it achieves L_η^*. Since the assumptions of Proposition 3.2 of [21] are satisfied by Proposition 2 in Appendix A, the former implies that there exists a function $h_\eta(\delta, r)$, called the *differential cost function*, satisfying

$$h_\eta(\delta, r) + L_\eta^* = \min_{a\in\{\mathrm{i,n,x}\}} \big(\delta + \eta \cdot \mathbb{1}[a \neq \mathrm{i}] + \mathbb{E}\left[h_\eta(\delta', r')\right]\big), \tag{7}$$

called the *Bellman optimality equations*, for all states $(\delta, r) \in \mathcal{S}$, where (δ', r') is the next state obtained from (δ, r) after taking action a. Furthermore, Proposition 3.2 of [21] also implies that the function h_η satisfying (7) is unique up to an additive factor, and with selecting this additive factor properly, the *differential cost function* also satisfies

$$h_\eta(\delta, r) = \mathbb{E}\left[\sum_{t=0}^{\infty} (\delta_t + \eta \cdot \mathbb{1}[a \neq \mathrm{i}] - L_\eta^*) \big| \delta_0 = \delta, r_0 = r\right].$$

We also introduce the *state-action cost function* defined as

$$Q_\eta(\delta, r, a) \triangleq \delta + \eta \cdot \mathbb{1}[a \neq \mathrm{i}] + \mathbb{E}\left[h_\eta(\delta', r')\right] \tag{8}$$

for all $(\delta, r) \in \mathcal{S}, a \in \mathcal{A}$. Then, also implied by Proposition 3.2 of [21], the optimal deterministic policy for the Lagrangian problem with a given η takes, for any $(\delta, r) \in \mathcal{S}$, the action achieving the minimum in (23):

$$\pi_\eta^*(\delta, r) \in \arg\min_{a\in\{\mathrm{i,n,x}\}} Q_\eta(\delta, r, a)\ . \tag{9}$$

[1]If $C_{max} = 1$, a transmission (new update or retransmission) is allowed in every time slot, and instead of a CMDP we have an infinite state-space MDP with unbounded cost. Then it follows directly from part (ii) of the Theorem of [22] (whose conditions can be easily verified for our problem) that there exists an optimal stationary policy that satisfies the Bellman equations. In this paper we concentrate on the more interesting constrained case, while the derivation of this result is relegated to Appendix E.

Focusing on deterministic policies, it is possible to characterize optimal policies for our CMDP problem: Based on Theorem 2.5 of [21], we can prove the the following result:

Theorem 1. *There exists an optimal stationary policy for the CMDP in Problem 1 that is optimal for the unconstrained problem considered in* (6) *for some $\eta = \eta^*$, and randomizes in at most one state. This policy can be expressed as a mixture of two deterministic policies $\pi^*_{\eta^*,1}$ and $\pi^*_{\eta^*,2}$ that differ in at most a single state s, and are both optimal for the Lagrangian problem* (6) *with $\eta = \eta^*$. More precisely, there exists $\mu \in [0,1]$ such that the mixture policy $\pi^*_{\eta^*}$, which selects, in state s, $\pi^*_{\eta^*,1}(s)$ with probability μ and $\pi^*_{\eta^*,2}(s)$ with probability $1-\mu$, and otherwise follows these two policies (which agree in all other states) is optimal for Problem 1, and the constraint in* (3) *is satisfied with equality.*

Proof. By Proposition 2 in Appendix A, Theorem 2.5, Proposition 3.2, and Lemma 3.9 of [21] hold for Problem 1. By Theorem 2.5 of [21], there exists an optimal stationary policy that is a mixture of two deterministic policies, $\pi^*_{\eta^*,1}$ and $\pi^*_{\eta^*,2}$, which differ in at most one state and are η^*-optimal by Proposition 3.2 of [21] satisfying (7) and (23). From Lemma 3.9 of [21], the mixture policy π^*_μ, for any $\mu \in [0,1]$, also satisfies (7) and (23), and is optimal for the unconstrained problem in (6) with $\eta = \eta^*$. From the proof of Theorem 2.5 of [21], there exists a $\mu \in [0,1]$ such that $\pi^*_{\eta^*}$ satisfies the constraint in (3) with equality. This completes the proof of the theorem. □

Some other results in [21] will be useful in determining $\pi^*_{\eta^*}$. For any $\eta > 0$, let C_η and J_η denote the average number of transmissions and average AoI, respectively, for the optimal policy π^*_η. Note that these are multivalued functions since there might be more than one optimal policy for a given η. Note also that, C_η and J_η can be computed directly by finding the stationary distribution of the chain, or estimated empirically by running the MDP with policy π^*_η. From Lemma 3.4 of [21], L^*_η, C_η and J_η are monotone functions of η: if $\eta_1 < \eta_2$, we have $C_{\eta_1} \geq C_{\eta_2}$, $J_{\eta_1} \leq J_{\eta_2}$ and $L^*_{\eta_1} \leq L^*_{\eta_2}$. This statement is also intuitive since η effectively represents the cost of a single transmission in (7) and (23), as η increases, the average number of transmissions of the optimal policy cannot increase, and as a result, the AoI cannot decrease.

To determine the optimal policy, one needs to find η^*, the policies $\pi^*_{\eta^*,1}$ and $\pi^*_{\eta^*,2}$, and the weight μ. In fact, [21] shows that η^* is defined as

$$\eta^* \triangleq \inf\{\eta > 0 : C_\eta \leq C_{max}\}, \tag{10}$$

where the inequality $C_\eta \leq C_{max}$ is satisfied if it is satisfied for at least one value of (multivalued) C_η. By Lemma 3.12 of [21] and Proposition 2, η^* is finite, and $\eta^* > 0$ if $C_{max} < 1$.

If $C^{\pi^*_{\eta^*,i}} = C_{max}$ for $i = 1$ or $i = 2$, then it is the optimal policy, that is, $\pi^*_\mu = \pi^*_{\eta^*,i}$ and $\mu = 1$ if $i = 1$ and 0 if $i = 2$. Otherwise one needs to select μ such that $C^{\pi^*_\mu} = C_{max}$: that is, if $C^{\pi^*_{\eta^*,2}} < C_{max} < C^{\pi^*_{\eta^*,1}}$, then

$$\mu = \frac{C_{max} - C^{\pi^*_{\eta^*,2}}}{C^{\pi^*_{\eta^*,1}} - C^{\pi^*_{\eta^*,2}}}, \tag{11}$$

which results in an optimal policy.

In practice, finding both η^* and the policies $\pi^*_{\eta^*,1}$ and $\pi^*_{\eta^*,2}$ is hard. However, given two monotone sequences sequences $\eta_n \uparrow \eta^*$ and $\eta'_n \downarrow \eta^*$, there is a subsequence of η_n (resp., η'_n) such that the corresponding subsequence of the η_n-optimal policies $\pi^*_{\eta_n}$ (η'_n-optimal policies $\pi^*_{\eta'_n}$, resp.) satisfying the Bellman equation (7) converge. Then the limit points π and π'[2] are η^*-optimal by Lemma 3.7 (iii) of [21] and $C^\pi \geq C_{max} \geq C^{\pi'}$ by the monotonicity of C_η and the same Lemma 3.7. Although there is no guarantee that π and π' only differ in a single point, we can combine them to get an optimal randomized policy using μ defined in (11). In this case, Lemma 3.9 of [21] implies that the policy that first randomly selects if it should use π or π' (choosing π with probability μ) and then uses the selected policy forever is η^*-optimal. However, since $(1, 0)$ is a positive recurrent state of both policies and they have a single recurrent class by Proposition 3.2 of [21], we can do the random selection of between π and π' independently every time the system gets into state $(1, 0)$ without changing the long-term average or expected AoI and transmission cost (note that one cannot choose randomly between the two policies in, e.g., every step). Thus, the resulting randomized policy is η^*-optimal, and since μ is selected in such a way that the total transmission cost is C_{max}, it is also an optimal solution of Problem 1 by Lemma 3.10 of [21]. Note that to derive two η^*-optimal policies, which provably differ only in a single state, a much more elaborate construction is used in [21]. However, in practice, π and π' obtained above are often the same except for a single state. Furthermore, we can approximate π_1 and π_2 by $\pi^*_{\eta_n}$ and $\pi^*_{\eta'_n}$ for n large enough. This idea is explored in the next section.

Theorem 1 and the succeeding discussion present the general structure of the optimal policy. In Section IV, for practical implementation, a computationally efficient heuristic algorithm is proposed based upon the discussion in this section.

[2] $\pi_n \to \pi$ if for any state s, $\pi_n(s) = \pi(s)$ for n large enough.

IV. An Iterative Algorithm to Minimize the AoI under an Average Cost Constraint

While our state space is countably infinite, since the age can be arbitrarily large (r_{max} may also be infinite), in practice we can approximate the countable state space with a large but finite space by setting an upper bound on the age (which will be denoted by N), and by selecting a finite r_{max} (whenever the chain would leave this constrained state space, we truncate the value of the age and/or the retransmission number to N and r_{max}, respectively); this gives a finite state space approximation to the problem similarly to [2], [11]. Clearly, letting N and r_{max} go to infinity, the optimal policy for the restricted state space will converge to that of the original problem.

When we consider the finite state space approximation of our problem, we can employ the *relative value iteration* (RVI) [23] algorithm to solve (7) for any given η, and hence find (an approximation of) the optimal policy π_η^*. Note that the finite state space approximation is needed for the practical implementation of the RVI algorithm since each iteration in RVI requires the computation of the value function for each state-action pair (for the infinite state space we would need to use some sort of parametric approximation of the states or the value functions, which is out of the scope of this paper). The pseudocode of the RVI algorithm is given in Algorithm 1. To simplify the notation, the dependence on η is suppressed in the algorithm for h, V and Q.

After presenting an algorithm that can compute the optimal deterministic policy π_η^* for any given η (more precisely, an arbitrarily close approximation thereof in the finite approximate MDP), we need to find the particular Lagrange multiplier η^* as defined by (10). As the simplest solution, we would need to generate C_η for a reasonably large range of η values to determine η^*. This could be approximated by computing C_η for a fine grid of η values, but this approach might be computationally demanding (note that generating each point requires running an instance of RVI).

Instead, we can use the following heuristic: With the aim of finding a single η value with $C_\eta \approx C_{max}$, we start with an initial parameter η^0, and run an iterative algorithm updating η as $\eta^{m+1} = \eta^m + \alpha_m(C_{\eta^m} - C_{max})$ for a step size parameter α_m[3] (note that for each step we need to run RVI to be able to determine C_{η^m}). We continue this iteration until $|C_{\eta^m} - C_{max}|$ becomes smaller than a given threshold, and denote the resulting value by η^*. We can increase or decrease the η^* value until η^* and its modification satisfy the conditions (note that with a finite state space, which is an approximation always used when computing an optimal policy numerically, π_η, and consequently C_η

[3] α_m is a positive decreasing sequence and satisfies the following conditions: $\sum_m \alpha_m = \infty$ and $\sum_m \alpha_m^2 < \infty$ from the theory of stochastic approximation [25].

Algorithm 1: Relative value iteration (RVI) algorithm for a given η.

Input : Lagrange parameter η, error probability $g(r)$

(δ^{ref}, r^{ref}) /* choose an arbitrary but fixed reference state */
$n \leftarrow 0$ /* iteration counter */
$h_0^{N \times r_{max}} \leftarrow \mathbf{0}$ /* initialization */
while 1 /* until convergence */
do
  for *state* $s = (\delta, r) \in [1, \ldots, N] \times [1, \ldots, r_{max}]$ **do**
    for *action* $a \in \mathcal{A}$ **do**
      $Q_{n+1}(\delta, r, a) \leftarrow \delta + \eta \cdot \mathbb{1}[a^{\pi} \neq \mathrm{i}] + \mathbb{E}\left[h_n(\delta', r')\right]$
    end
    $V_{n+1}(\delta, r) \leftarrow \min_a(Q_{n+1}(\delta, r, a))$
    $h_{n+1}(\delta, r) \leftarrow V_{n+1}(\delta, r) - V_{n+1}(\delta^{ref}, r^{ref})$
  end
  if $|h_{n+1} - h_n| \leq \epsilon$ **then**
    /* compute the optimal policy */
    for $(\delta, r) \in [1, \ldots, N] \times [1, \ldots, r_{max}]$ **do**
      $\pi_{\eta}^{*}(\delta, r) \leftarrow \arg\min_a(Q(\delta, r, a))$
    end
    return π^*
  else
    increase the iteration counter: $n \leftarrow n + 1$
  end
end

and J_η, are piecewise constant functions of η, thus the step size must be chosen sufficiently large to change the average transmission cost).

In order to obtain two deterministic policies and the corresponding mixing coefficient, based on the discussion at the end of Section III, we want to find optimal policies for η values slightly smaller and larger than η^*, and so we compute the optimal policies (by RVI) for $\eta^* \pm \xi$ where ξ is a small perturbation and obtain a mixture coefficient according to (11) as

$$\mu = \frac{C_{max} - C_{\eta^*+\xi}}{C_{\eta^*-\xi} - C_{\eta^*+\xi}} \ . \tag{12}$$

If the optimal policies differ only in a single state, we can randomize in that state (by Theorem 1), while, if they are more different, we can randomly select between the policies (with probabilities μ

and $1-\mu$) every time after a successful transmission (i.e., when the system is in state $(1,0)$), as discussed at the end of Section III.

Numerical results obtained by implementing the above heuristics in order to minimize the average AoI with HARQ will be presented in Section VII. In the next section, we focus on the simpler scenario with the classical ARQ protocol.

V. AoI with Classical ARQ Protocol under an Average Cost Constraint

In the classical ARQ protocol, failed transmissions are discarded at the destination and the receiver tries to decode each retransmission as a new message. In the context of AoI, there is no point in retransmitting an undecoded packet since the probability of a successful transmission is the same for a retransmission and for the transmission of a new update. Hence, the state space reduces to $\delta \in \{1,2,\ldots\}$ as $r_t = 0$ for all t, and the action space reduces to $\mathcal{A} \in \{\mathrm{n},\mathrm{i}\}$, and the probability of error $p \triangleq g(0)$ is fixed for every transmission attempt.[4] State transitions in (4), Bellman optimality equations [23], [24] for the countable-state MDP in (7), and the RVI algorithm with the finite state approximation can all be simplified accordingly. We define

$$Q_\eta(\delta,\mathrm{i}) \triangleq \delta + h_\eta(\delta+1), \tag{13}$$

$$Q_\eta(\delta,\mathrm{n}) \triangleq \delta + \eta + p h_\eta(\delta+1) + (1-p)h_\eta(1), \tag{14}$$

where $h_\eta(\delta)$ is the optimal differential value function satisfying the Bellman optimality equation

$$h_\eta(\delta) + L_\eta^* \triangleq \min\left\{Q_\eta(\delta,\mathrm{i}), Q_\eta(\delta,\mathrm{n})\right\},\ \forall \delta \in \{1,2,\ldots\}. \tag{15}$$

Thanks to these simplifications, we are able to provide a closed-form solution to the corresponding Bellman equations in (13), (14) and (15).

Lemma 1. *The policy that satisfies the Bellman optimality equations for the standard ARQ protocol is deterministic and has a threshold structure:*

$$\pi^*(\delta) = \begin{cases} \mathrm{n} & \textit{if } \delta \geq \Delta_\eta, \\ \mathrm{i} & \textit{if } \delta < \Delta_\eta. \end{cases}$$

for some integer Δ_η that depends on η.

Proof. The proof is given in Appendix B. □

[4]This simplified model with classical ARQ protocol and Lagrangian relaxation is equivalent to the work in [2] when η is considered to be the cost of a single transmission and the assumption of a perfect transmission channel in [2] is ignored.

The next lemma characterizes the possible values of the threshold defined in Lemma 1.

Lemma 2. *Under the standard ARQ protocol, the η-optimal value of the threshold Δ_η can be found in closed-form:*

$$\Delta_\eta^* \in \left\{ \left\lfloor \frac{\sqrt{2\eta(1-p)+p}-p}{1-p} \right\rfloor, \left\lceil \frac{\sqrt{2\eta(1-p)+p}-p}{1-p} \right\rceil \right\}.$$

Proof. The proof is given in Appendix C. □

The main result of this section, given below, shows that the optimal policy for Problem 1 is a randomized threshold policy which randomizes over the above two thresholds for the optimal value of η^*. Let $\Delta_{C_{max}} \triangleq \frac{1/C_{max}-p}{1-p}$, $\Delta_1 \triangleq \lfloor \Delta_{C_{max}} \rfloor$ and $\Delta_2 \triangleq \lceil \Delta_{C_{max}} \rceil$, and consider the mixture of the threshold policies with thresholds Δ_1 and Δ_2, respectively, and mixture coefficient $\mu \in [0,1]$. The resulting policy $\pi^*_{C_{max},\mu}$ can be written in closed form: if $\Delta_{C_{max}}$ is an integer then $\pi^*_{C_{max},\mu}(\delta) = \mathrm{n}$ if $\delta \geq \Delta_{C_{max}}$ and i otherwise. If $\Delta_{C_{max}}$ is not an integer, then $\pi^*_{C_{max},\mu}(\delta) = \mathrm{n}$ if $\delta \geq \lceil \Delta_{C_{max}} \rceil$, $\pi^*_{C_{max},\mu}(\delta) = \mathrm{i}$ if $\delta < \lfloor \Delta_{C_{max}} \rfloor$, while $\pi^*_{C_{max},\mu}(\mathrm{n}|\delta) = \mu$ and $\pi^*_{C_{max},\mu}(\mathrm{i}|\delta) = 1-\mu$ for $\delta = \lfloor \Delta_{C_{max}} \rfloor$. The mixture coefficient μ is selected so that $C^{\pi^*_{C_{max},\mu}} = C_{max}$: From the proof of Lemma 2 one can easily deduce that the transmission cost (per time slot) of the threshold policy for any integer threshold Δ is given by

$$C^{\Delta} = \frac{1}{\Delta(1-p)+p}. \tag{16}$$

Hence, selecting $\mu^* = \frac{C_{max}-C^{\Delta_2}}{C^{\Delta_1}-C^{\Delta_2}}$, as described in (11), ensures $C^{\pi^*_{C_{max},\mu^*}} = C_{max}$. Denoting $\pi^*_{C_{max}} = \pi^*_{C_{max},\mu^*}$, we obtain the following theorem (the proof is given in Appendix D).

Theorem 2. *For any $C_{max} \in (0,1]$, the stationary policy $\pi^*_{C_{max},\mu^*}$ defined above is an optimal policy (i.e., a solution of Problem 1) under the ARQ protocol.*

Numerical results obtained for the above algorithm will be presented and compared with those from the HARQ protocol in Section VII.

VI. Learning to minimize AoI in an unknown environment

In the CMDP formulation presented in Sections IV and V, we have assumed that the channel error probabilities for all retransmissions are known in advance. However, in most practical scenarios, these error probabilities may not be known at the time of deployment, or may change over time. Therefore, in this section, we assume that the source node does not have *a priori* information about

Algorithm 2: Average-cost SARSA with softmax

Input : Lagrange parameter η /* error probability $g(r)$ is unknown */

$n \leftarrow 0$ /* time iteration */

$\tau \leftarrow 1$ /* softmax temperature parameter */

$Q_\eta^{N\times M\times 3} \leftarrow 0$ /* initialization of Q */

$L_\eta \leftarrow 0$ /* initialization of the gain */

for n **do**

  OBSERVE the current state s_n

  for $a \in \mathcal{A}$ **do**

    /* since it is a minimization problem, use minus Q function in softmax */

    $$\pi(a|s_n) = \frac{\exp(-Q_\eta(s_n,a)/\tau)}{\sum_{a'\in\mathcal{A}} \exp(-Q_\eta(s_n,a')/\tau)}$$

  end

  SAMPLE a_n from $\pi(a|S_n)$

  OBSERVE the next state s_{n+1} and cost $c_n = \delta_n + \eta 1_{\{a_n=1,2\}}$

  for $a \in \mathcal{A}$ **do**

    /* softmax is also used for the next state s_{n+1}, so that it is on-policy */

    $$\pi(a|s_{n+1}) = \frac{\exp(-Q_\eta(s_{n+1},a_{n+1})/\tau)}{\sum_{a'_{n+1}\in\mathcal{A}} \exp(-Q_\eta(s_{n+1},a'_{n+1})/\tau)}$$

  end

  SAMPLE a_{n+1} from $\pi(a_{n+1}|s_{n+1})$

  UPDATE

  $\alpha_n \leftarrow 1/\sqrt{n}$ /* update parameter */

  $Q_\eta(s_n,a_n) \leftarrow Q_\eta(s_n,a_n) + \alpha_n[\delta + \eta\cdot \mathbb{1}[a_n \neq \mathrm{i}] - J_\eta + Q_\eta(s_{n+1},a_{n+1}) - Q_\eta(s_n,a_n)]$

  $L_\eta \leftarrow L_\eta + 1/n[\delta + \eta\cdot \mathbb{1}[a_n \neq \mathrm{i}] - J_\eta]$ /* update J_η at every step */

  $n \leftarrow n+1$ /* increase the iteration */

end

the decoding error probabilities, and has to learn them. We employ an online learning algorithm to learn $g(r)$ over time without degrading the performance significantly.

The literature for average-cost RL is quite limited compared to discounted cost problems [14], [26]. SARSA [26] is a well-known RL algorithm, originally proposed for discounted MDPs, that learns the optimal policy for an MDP based on the action performed by the current policy in a recursive

manner. For average AoI minimization in Problem 1, an average cost version of the SARSA algorithm is employed with *Boltzmann* (*softmax*) exploration. The resulting algorithm is called *average-cost SARSA with softmax*.

As indicated by (7) and (23) in Section III, $Q_\eta(s_n, a_n)$ of the current state-action pair can be represented in terms of the immediate cost of the current state-action pair and the differential state-value function $h_\eta(s_{n+1})$ of the next state. Notice that, one can select the optimal actions by only knowing $Q_\eta(s, a)$ and choosing the action that will give the minimum expected cost as in (9). Thus, by only knowing $Q_\eta(s, a)$, one can find the optimal policy π^* without knowing the transition probabilities P characterized by $g(r)$ in (4).

Similarly to SARSA, *average-cost SARSA with softmax* starts with an initial estimation of $Q_\eta(s, a)$ and finds the optimal policy by estimating state-action values in a recursive manner. In the n^{th} time iteration, after taking action a_n, the source observes the next state s_{n+1}, and the instantaneous cost value c_n. Based on this, the estimate of $Q_\eta(s, a)$ is updated by weighing the previous estimate and the estimated expected value of the current policy in the next state s_{n+1}. Also note that, in general, c_n is not necessarily known before taking action a_n because it does not know the next state s_{n+1} in advance. In our problem, the instantaneous cost c_n is the sum of AoI at the destination and the cost of transmission, i.e. $\delta_n + \eta \cdot \mathbb{1}[a_n \neq \mathrm{i}]$; hence, it is readily known at the source node.

In each time slot, the learning algorithm

- observes the current state $s_n \in \mathcal{S}$,
- selects and performs an action $a_n \in \mathcal{A}$,
- observes the next state $s_{n+1} \in \mathcal{S}$ and the instantaneous cost c_n,
- updates its estimate of $Q_\eta(s_n, a_n)$ using the current estimate of $_\eta$ by

$$Q_\eta(s_n, a_n) \leftarrow Q_\eta(s_n, a_n) + \alpha_n[\delta + \eta \cdot \mathbb{1}[a_n \neq \mathrm{i}] - L_\eta + Q_\eta(s_{n+1}, a_{n+1}) - Q_\eta(s_n, a_n)], \quad (17)$$

  where α_n is the update parameter (learning rate) in the n^{th} iteration.
- updates its estimate of L_η based on empirical average.

The details of the algorithm are given in Algorithm 2. We update the gain L_η at every time slot based on the empirical average, instead of updating it at non-explored time slots.

As we discussed earlier, with the accurate estimate of $Q_\eta(s, a)$ at hand the transmitter can decide for the optimal actions for a given η as in (9). However, until the state-action cost function is accurately estimated, the transmitter action selection method should balance the *exploration* of new actions with the *exploitation* of actions known to perform well. In particular, the *Boltzmann* action

selection method, which chooses each action probabilistically relative to expected costs, is used in this paper. The source assigns a probability to each action for a given state s_n, denoted by $\pi(a|s_n)$:

$$\pi(a|s_n) \triangleq \frac{\exp(-Q_\eta(s_n,a)/\tau)}{\sum_{a'\in\mathcal{A}} \exp(-Q_\eta(s_n,a')/\tau)}, \tag{18}$$

where τ is called the temperature parameter such that high τ corresponds to more uniform action selection (exploration) whereas low τ is biased toward the best action (exploitation).

In addition, the constrained structure of the average AoI problem requires additional modifications to the algorithm, which is achieved in this paper by updating the Lagrange multiplier according to the empirical resource consumption. In each time slot, we keep track of a value η resulting in a transmission cost close to C_{max}, and then find and apply a policy that is optimal (given the observations so far) for the MDP with Lagrangian cost as in Algorithm 2.

The performance of *average-cost SARSA with softmax*, and its comparison with the RVI algorithm will be presented in the next section.

VII. Numerical Results

In this section, we provide numerical results for all the proposed algorithms, and compare the achieved average performances. For the simulations employing HARQ, motivated by previous research on HARQ [5], [6], [7], we assume that decoding error reduces exponentially with the number of retransmission, that is, $g(r) \triangleq p_0\lambda^r$ for some $\lambda \in (0,1)$, where p_0 denotes the error probability of the first transmission, and r is the retransmission count (set to 0 for the first transmission). The exact value of the rate λ depends on the particular HARQ protocol and the channel model. Note that ARQ corresponds to the case with $\lambda = 1$ and $r_{max} = 0$. Following the *IEEE 802.16* standard [19], the maximum number of retransmissions is set to $r_{max} = 3$; however, we will present results for other r_{max} values as well. We note that we have also run simulations for HARQ with relatively higher r_{max} values and $r_{max} = \infty$, and the improvement on the performance is not observable beyond $r_{max} = 3$. Numerical results for different p_0, λ and C_{max} values, corresponding to different channel conditions and HARQ schemes, will also be provided.

Figure 4 illustrates the deterministic policies obtained by RVI and the search for η^* for given C_{max} and p_0 values, while λ is set to 0.5. The final policies are generated by randomizing between $\pi^*_{\eta^*-\xi}$ and $\pi^*_{\eta^*+\xi}$; the approximate η^* values found for the settings in Figures 4(a) and 4(b) are 5 and 19, respectively, and ξ is set to 0.2. As it can be seen from the figures, the resulting policy transmits less as the average cost constraint becomes more limiting, i.e., as η increases. We also note that,

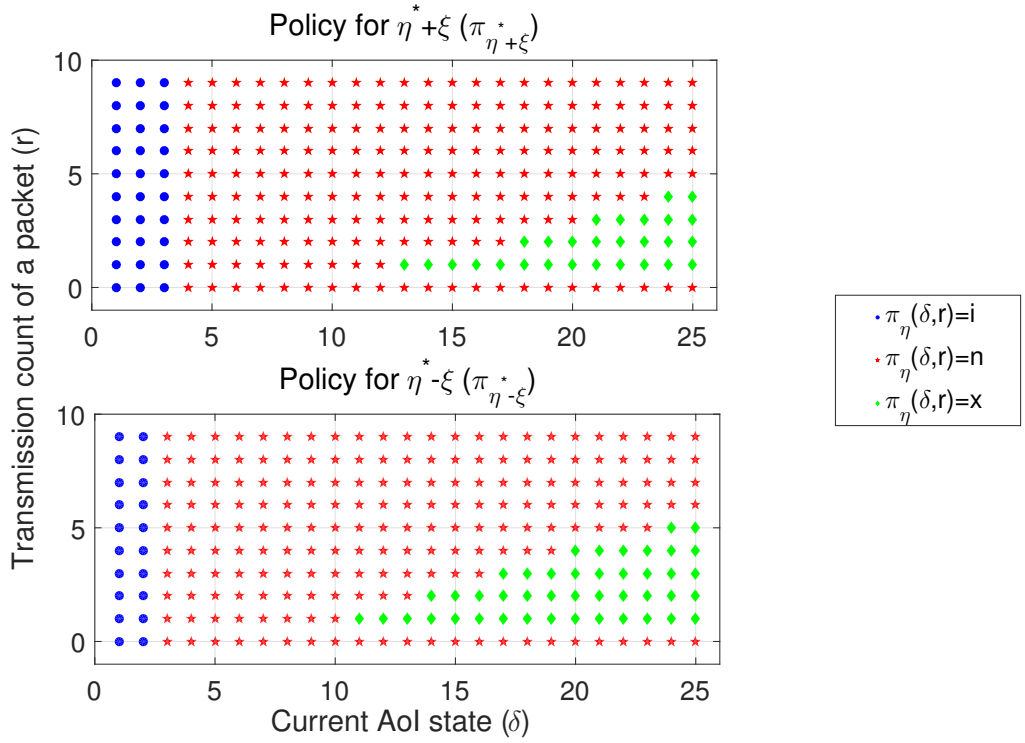
although the policies $\pi^*_{\eta^*-\xi}$ and $\pi^*_{\eta^*+\xi}$ are obtained for similar η^* values, and hence, have similar average number of transmissions, they may act quite differently especially for large C_{max} values.

Figure 5 illustrates the performance of the proposed randomized HARQ policy with respect to C_{max} for different p_0 values when λ is set to 0.5. We also include the performance of the optimal deterministic and randomized threshold policies with ARQ, derived in Section V, for $p_0 = 0.5$. For baseline, we use a simple no-feedback policy that periodically transmits a fresh status update with a period of $\lceil 1/C_{max} \rceil$, ensuring that the constraint on the average number of transmissions holds. The effect of feedback on the performance can be seen immediately: a single-bit ACK/NACK feedback, even with the ARQ protocol, decreases the average AoI considerably, although receiving feedback might be costly for some status update systems. The two curves for the ARQ policies demonstrate the effect of randomization: the curve corresponding to the randomized policy is the lower convex hull of the piecewise constant AoI curve for deterministic policies. For the same $p_0 = 0.5$, HARQ with $\lambda = 0.5$ improves only slightly over ARQ. Smaller p_0 results in a decrease in the average AoI as expected, and the gap between the AoIs for different p_0 values is almost constant for different C_{max} values.

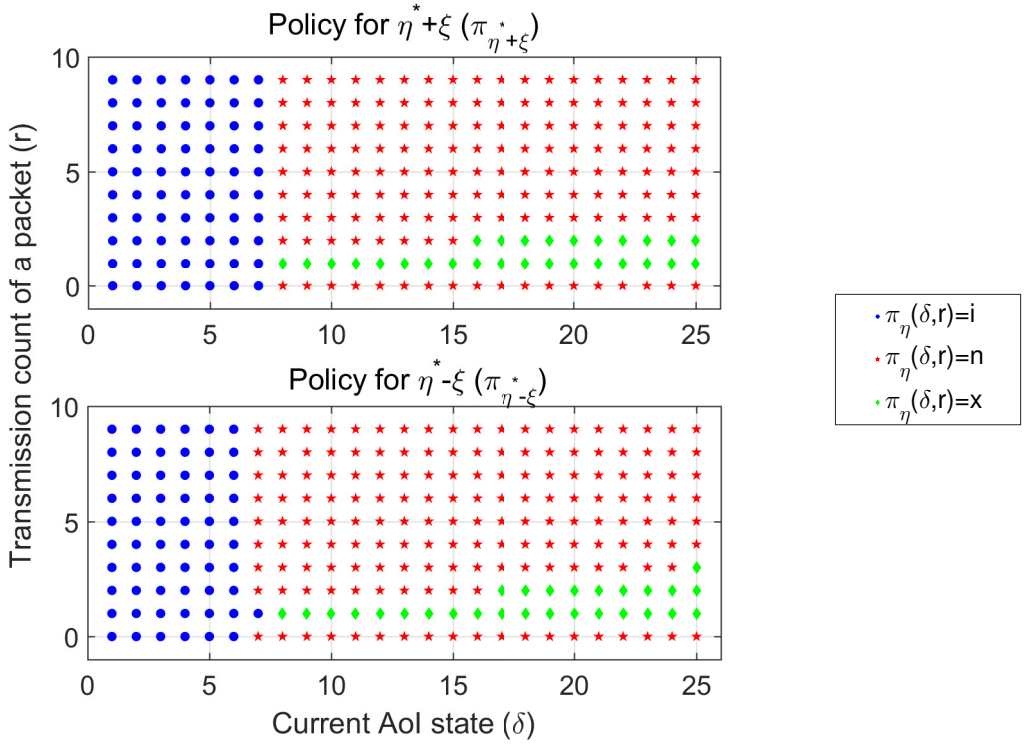
More significant gains can be achieved from HARQ when the error probability decreases faster with retransmissions (i.e., small λ), or more retransmissions are allowed. This is shown in Figure 6. On the other hand, the effect of retransmissions on the average AoI (with respect to ARQ) is more pronounced when p_0 is high and λ is low.

Figure 7 shows the average AoI achieved by the HARQ protocol with respect to different p_0 and λ values for $r_{max} = 3$. Similarly to Figure 5, the gap between the average AoI values is higher for unreliable environments with higher error probability, and the performance gap due to different λ values are not observable for relatively reliable environments, for example, when $p_0 = 0.3$. The performance difference for different λ values (with a fixed p_0) is more pronounced when the average number of transmissions, C_{max}, is low, since then less resources are available to correct an unsuccessful transmission.

Figure 8 shows the evolution of the average AoI over time when the average-cost SARSA learning algorithm is employed. It can be observed that the average AoI achieved by Algorithm 2, denoted by *RL* in the figure, converges to the one obtained from the RVI algorithm which has *a priori* knowledge of $g(r)$. We can observe from Figure 8 that the performance of SARSA achieves that of RVI in about 10000 iterations. Figure 9 shows the performance of the two algorithms (with again 10000 iterations in SARSA) as a function of C_{max} in two different setups. We can see that SARSA performs very close to RVI with a gap that is more or less constant for the whole range of C_{max} values. We can also



(a) $C_{max} = 0.4$, $p_0 = 0.3$



(b) $C_{max} = 0.2$, $p_0 = 0.4$

Figure 4. Deterministic policies $\pi_{\eta^*+\xi}$ (top) and $\pi_{\eta^*-\xi}$ (bottom) when $\lambda = 0.5$ and $r_{max} = 9$. (Blue circles, red stars, and green diamonds represent actions $\pi_\eta(\delta, r) =$ i, n and x, respectively.)

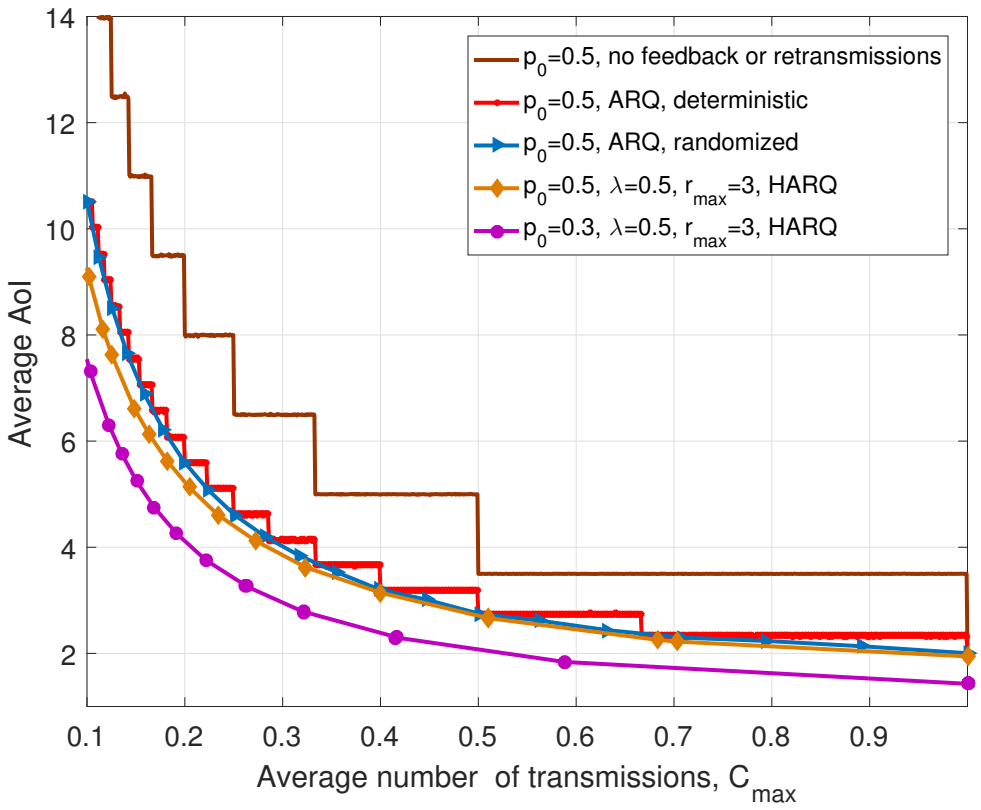


Figure 5. Expected average AoI as a function of C_{max} for ARQ and HARQ protocols for different p_0 values. Time horizon is set to $T = 10000$, and the results are averaged over 1000 runs.

observe that the variance of the average AoI achieved by SARSA is much larger when the number of transmissions is limited, which also limits the algorithm's learning capability.

VIII. Conclusions

We have considered a communication system transmitting time-sensitive data over an imperfect channel with the average AoI as the performance measure, which quantifies the timeliness of the data available at the receiver. Considering both the classical ARQ and the HARQ protocols, preemptive scheduling policies have been proposed by taking into account retransmissions under a resource constraint. In addition to identifying a randomized threshold structure for the optimal policy when the error probabilities are known, an efficient RL algorithm is also presented for practical applications when the system characteristics may not be known in advance. The effects of feedback and the HARQ structure on the average AoI are demonstrated through numerical simulations. The algorithms adopted in this paper are also relevant to different systems concerning the timeliness of information, and the

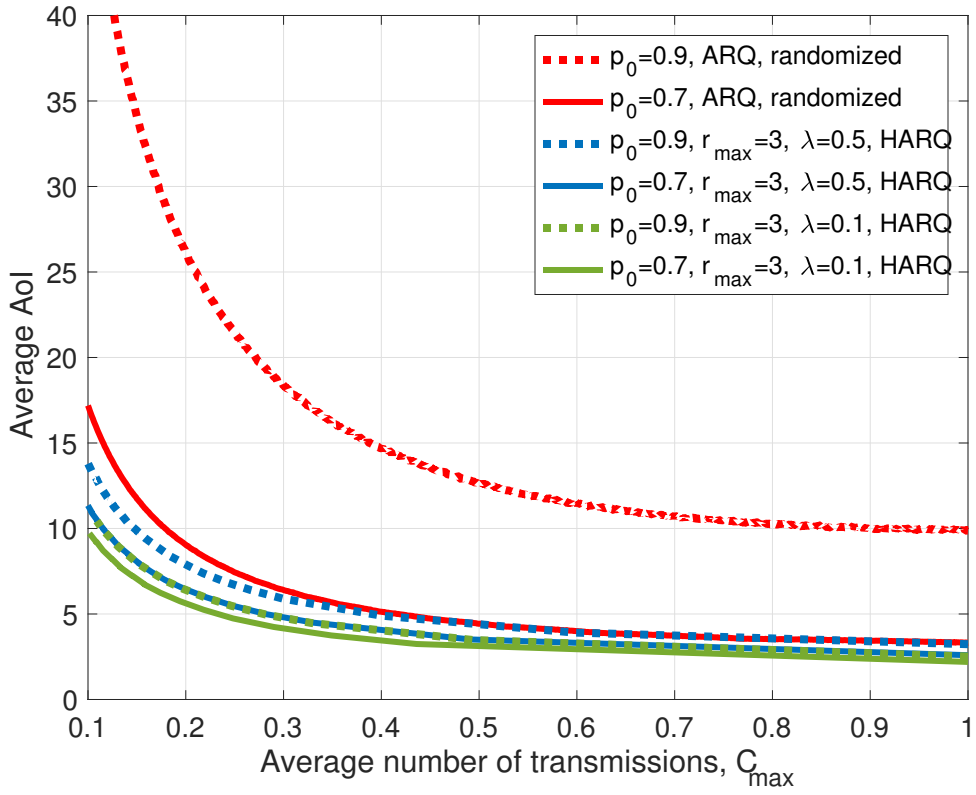


Figure 6. Expected average AoI with respect to C_{max} for ARQ and HARQ protocols for different p_0 and r_{max} values. Time horizon is set to $T = 10000$, and the results are averaged over 1000 runs.

proposed methodology can be used in other CMDP problems. As future work, the problem will be extended to time-correlated channel statistics in a multi-user setting.

APPENDIX

A. *Verifying the assumptions of [21]*

In this section, we show that the assumptions for the main results of [21] are satisfied for Problem 1. We start with a few standard definitions about Markov chains: In a Markov chain with a countable state space $\mathcal{S}$, a state $s \in \mathcal{S}$ is called *positive recurrent* if the expected number of transitions needed to return to state s given that the chain started in state s is finite. A communication class $Z \subset \mathcal{S}$ is defined as a subset of the state space $\mathcal{S}$ such that all states within it communicate; that is, for any $s, s' \in Z$, starting from state s the chain reaches state s' with some positive probability. A communication class is positive recurrent if and only if all states in a communication class are positive recurrent. [27]

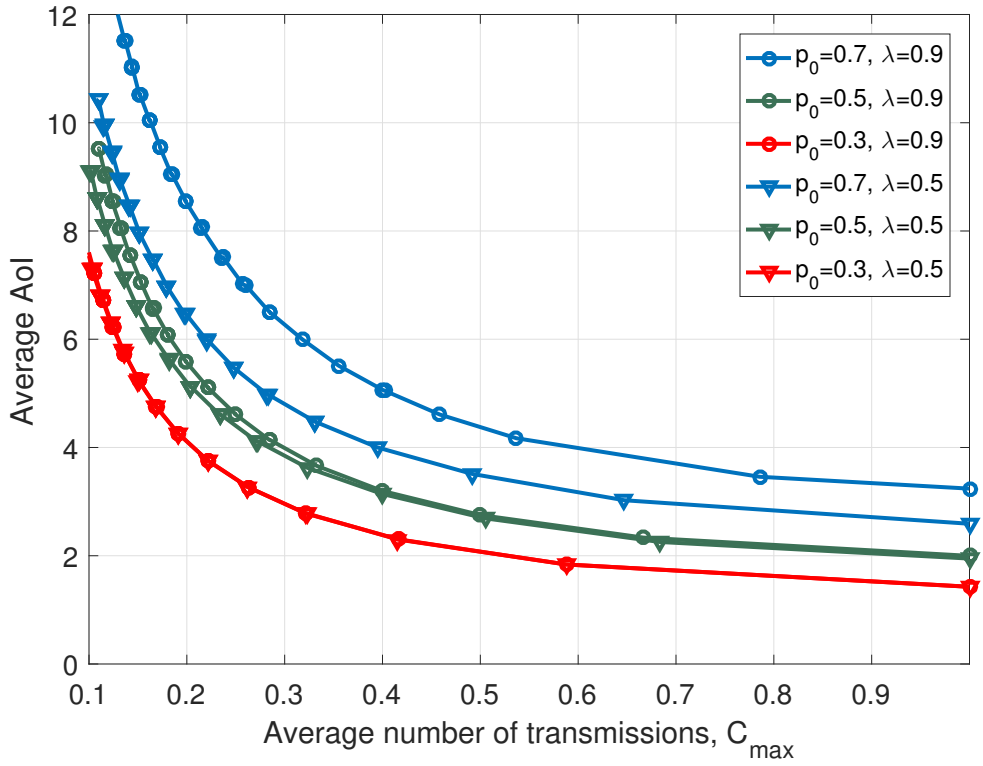


Figure 7. Expected average AoI with respect to C_{max} for HARQ protocols with different $g(r) = p_0\lambda^r$ values corresponding to different p_0 and λ values with $r_{max} = 3$. The time horizon is set to $T = 10000$, and the results are averaged over 1000 runs.

We continue with Definition 2.3 of [21]: Let $G \subset \mathcal{S}$ be a nonempty set of states of a CMDP. Given a state $s \in \mathcal{S}$, let $\mathcal{R}(s, G)$ be the class of policies such that $P^\pi(s_t \in G \text{ for some } t \geq 1 \ | s_0 = s) = 1$ and the expected time $m_{s,G}(\pi)$ of the first passage from s to G under π is finite. Let $\mathcal{R}^*(s, G)$ be the class of policies $\pi \in \mathcal{R}(s, G)$ such that, in addition, the expected average AoI $c_{s,G}(\pi)$ and the expected transmission cost $d_{s,G}(\pi)$ of a first passage from s to G are finite.

Proposition 2. *The following hold for Problem 1:*

(i) For all $b > 0$, the set $G(b) \triangleq \{s|$ there exists an action a such that $c(s, a) + d(s, a) \leq b\}$ is finite (Assumption 1 of [21]).

(ii) There exists a deterministic policy π that induces a Markov chain with the following properties: the state space $\mathcal{S}^\pi$ consists of a single (nonempty) positive recurrent class R and a set U of transient states such that $\pi \in R^*(s, R)$, for any $s \in U$, and both the average AoI J^π and the average transmission cost C^π on R are finite (Assumption 2 of [21]).

(iii) Given any two states $s, s' \in \mathcal{S}$, there exists a policy π (a function of s and s') such that

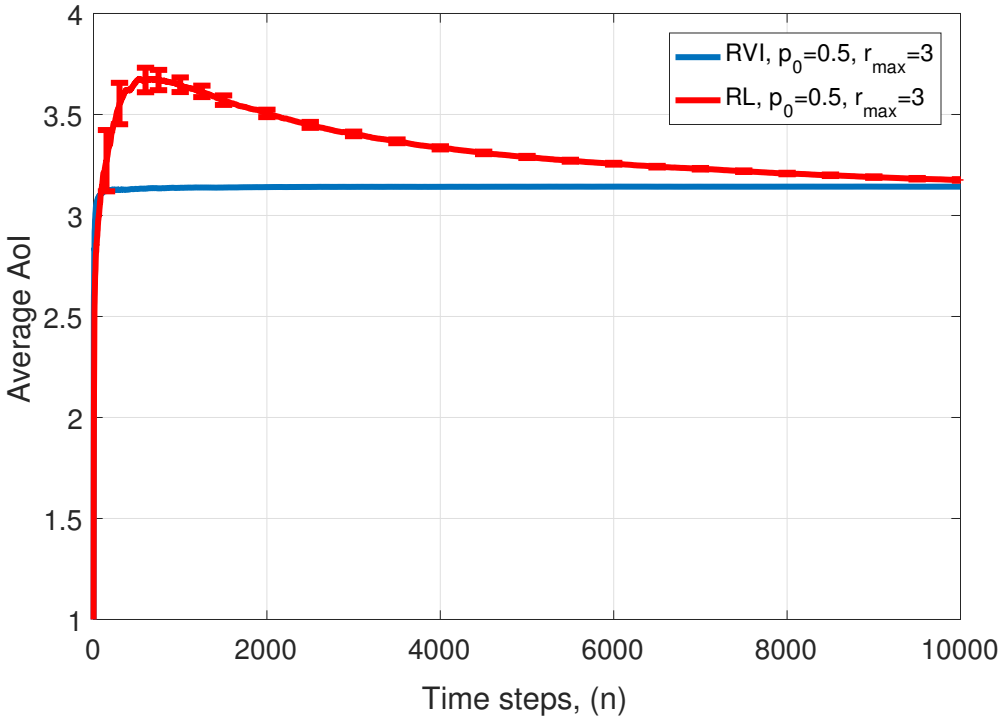


Figure 8. Performance of the average-cost SARSA for $r_{max} = 3$, $p_0 = 0.5$, $\lambda = 0.5$, $C_{max} = 0.4$ and $n = 10000$, averaged over 1000 runs (both the mean and the variance are shown).

$\pi \in \mathcal{R}^*(s, \{s'\})$ *(Assumption 3 of [21]).*

(iv) If a deterministic policy has at least one positive recurrent state, then it has a single positive recurrent class, and this class contains the state $(1, 0)$ *(Assumption 4 of [21]).*

(v) There exists a policy π *such that* $J^\pi < \infty$ *and* $C^\pi < C_{max}$ *(Assumption 5 of [21]).*

Proof. Note that (i)-(iv) are independent from the constraint (3), and the policies required in the proposition need not be deterministic unless specifically required.

First note that (i) holds trivially, since for any b, if state $(\delta, r) \in G(b)$ then $r < \delta \leq b$ by (5).

To prove (ii), consider the policy $\pi(\delta, r) = \mathrm{n}$ for all $(\delta, r) \in \mathcal{S}$. Since $0 < g(0) < 1$, $R = \{(1, 0)\} \cup \{(\delta, 1) : \delta = 1, 2, \ldots, \}$ is a recurrent class since from any state $(\delta, r) \in R$, the next state is either $(1, 0)$ or $(\delta + 1, 1)$, both belonging to R. Furthermore, the set of states $U = \mathcal{S} \setminus R$ is clearly transient: starting from any $s \in U$, the probability of not getting to state $(1, 0)$ (and hence to R) in at most k steps is $g(0)^k$. The latter also implies that $\pi \in R^*(s, R)$. Finally, $C^\pi = 1$, and $J^\pi = 1/(1 - g(0))$, proving (ii).

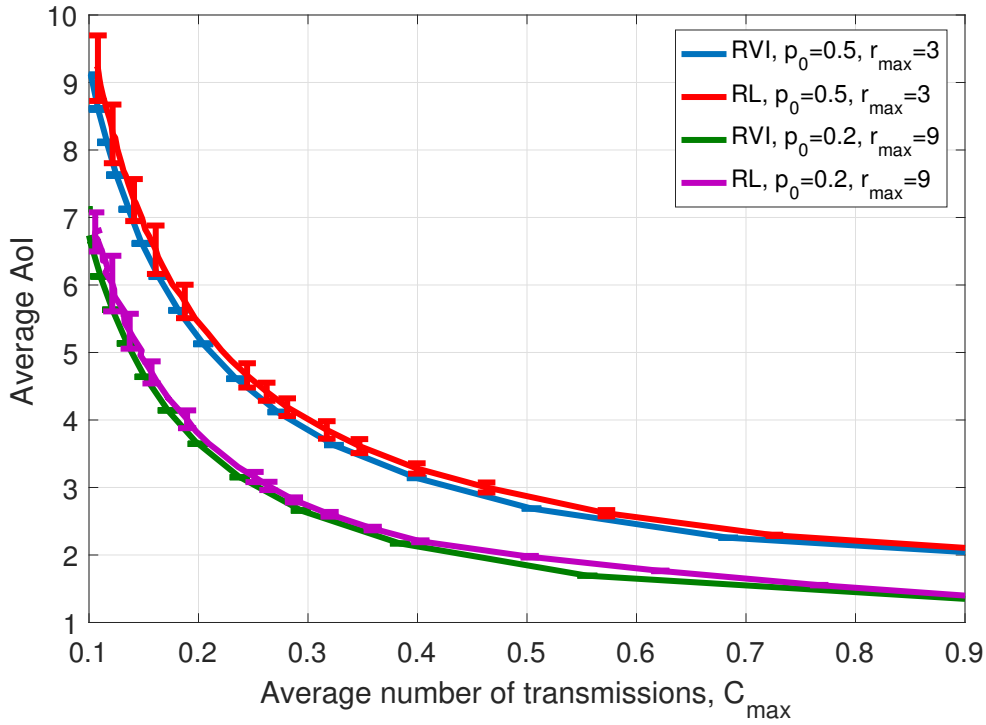


Figure 9. Performance of the proposed RL algorithm (average-cost SARSA) and its comparison with the RVI algorithm for $n = 10000$ iterations, and values are averaged over 1000 runs for different p_0 and r_{max} values when $\lambda = 0.5$ (both the mean and the variance are shown).

To prove (iii), let $s = (\delta, r)$ and $s' = (\delta', r')$. For any s, s', we construct the required policy. First note that from $(1, 0)$, we can govern the state with positive probability to any valid state (δ', r') by being idle in states $(\delta'', 0)$ for $\delta'' < \delta' - r'$, sending a new packet in state $(\delta' - r', 0)$, and retransmitting in states $(\delta' - r' + k, k)$ for $k = 1, \ldots, r' - 1$. Sending a new packet in any other state (δ'', r'') will send the chain to state $(1, 0)$ as quickly as possible, with the number of steps being exponentially distributed with parameter $g(0)$. It is trivial to see that the proposed policy belongs to $\mathcal{R}^*(s, \{s'\})$.

To see (iv), notice that the only way the AoI does not increase in one step is if there is a successful transmission, after which the chain returns to state $(1, 0)$. Thus, any (positive) recurrent class must contain the state $(1, 0)$; and hence, there can only be a single positive recurrent class.

Finally, it is easy to see that the policy π defined as $\pi(\delta, r) = \mathrm{n}$ if $\delta - 1$ is a multiple of $2\lceil 1/C_{max} \rceil$, and $\pi(\delta, r) = \mathrm{i}$ otherwise, satisfies the requirements of (v): $C^\pi = 1/(2\lceil 1/C_{max} \rceil) \leq C_{max}/2 < C_{max}$, and $J^\pi < \infty$ since $P^\pi(\delta > 2k\lceil 1/C_{max} \rceil) = g(0)^k$ for any $k \geq 0$. This completes the proof of the proposition. □

B. Proof of Lemma 1

We are going to show that the decision to transmit ($a = \mathrm{n}$) is monotone with respect to the age δ, that is if $a^*(\delta^1) = \mathrm{n}$, then $a^*(\delta^2) = \mathrm{n}$ for all $\delta^2 \geq \delta^1$. By (9), this holds if $Q_\eta(\delta, a)$ has a *sub-modular* structure [28]: that is, when the difference between the Q functions is monotone with respect to the state-action pair (δ, a). We have

$$Q_\eta(\delta^1, \mathrm{n}) - Q_\eta(\delta^1, \mathrm{i}) \geq Q_\eta(\delta^2, \mathrm{n}) - Q_\eta(\delta^2, \mathrm{i}), \tag{19}$$

for any $\delta^2 \geq \delta^1$. From (13) and (14), for any $\delta > 0$, we have

$$Q_\eta(\delta, \mathrm{n}) - Q_\eta(\delta, \mathrm{i}) = \eta + (1-p)h_\eta(1) - (1-p)h_\eta(\delta+1). \tag{20}$$

We can see that (19) holds if and only if $h_\eta(\delta)$ is a non-decreasing function of the age.

We compare the costs incurred by the systems starting in states δ^1 and δ^2 via coupling the stochastic processes governing the behavior of the system; that is, we assume that the realization of the channel behavior is the same for both systems over the time horizon (this is valid since channel states/errors are independent of the ages and the actions). Assume a sequence of actions $\{a_t^2\}_{t=1}^\infty$ corresponds to the optimal policy starting from age δ^2 for a particular realization of channel errors, and let $\{\delta_t^i\}$ denote the sequence of states obtained after following actions $\{a_t^2\}$ starting from state $\delta_1 = \delta^i$, $i = 1, 2$. Then, if $\delta^1 \leq \delta^2$, clearly $\delta_t^1 \leq \delta_2^t$ for all t. Furthermore, by the Bellman optimality equation (7),

$$\begin{aligned}
h_\eta(\delta^1) &\leq \delta_1^1 + \eta \cdot \mathbb{1}[a_1^2 \neq \mathrm{i}] - L_\eta^* + \mathbb{E}\left[h_\eta(\delta_2^1)\right] \\
&\leq \delta_1^1 + \eta \cdot \mathbb{1}[a_1^2 \neq \mathrm{i}] - L_\eta^* + \mathbb{E}\left[\delta_2^1 + \eta \cdot \mathbb{1}[a_2^2 \neq \mathrm{i}] - L_\eta^* + \mathbb{E}\left[h_\eta(\delta_3^1)\right]\right] \\
&\;\;\vdots \\
&\leq \mathbb{E}\left[\sum_{t=1}^{\infty} (\delta_t^1 + \eta \cdot \mathbb{1}[a_t^2 \neq \mathrm{i}] - L_\eta^*) \middle| \delta_1^1 = \delta^1\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{\infty} (\delta_t^2 + \eta \cdot \mathbb{1}[a_t^2 \neq \mathrm{i}] - L_\eta^*) \middle| \delta_1^1 = \delta^2\right] \\
&= h_\eta(\delta^2)\ .
\end{aligned}$$

This completes the proof of the lemma. □

C. Proof of Lemma 2

First we compute the steady state probabilities p_δ of the age δ for a given integer threshold Δ, for all $\delta = 1, 2, \ldots, N$. We have

$$p_\delta = \begin{cases} p_1 & \text{if } 1 \leq \delta \leq \Delta \\ p_{\delta-1}p = p_1 p^{\delta-\Delta} & \text{if } \delta \geq \Delta + 1 \ . \end{cases}$$

Since $\sum_{\delta=1}^{\infty} p_\delta = 1$, we can compute the p_δ in closed form when N goes to infinity:

$$p_\delta = \begin{cases} \frac{1}{\Delta + \frac{p}{1-p}} & \text{if } \delta \leq \Delta; \\ \frac{p^{\delta-\Delta}}{\Delta + \frac{p}{1-p}} & \text{otherwise.} \end{cases} \tag{21}$$

Then, the closed form of the expected Lagrangian cost function can be computed as:

$$\begin{aligned} L_\eta^\Delta &= \sum_{\delta=1}^{\infty} p_\delta(\delta + \eta \mathbb{1}[\delta \geq \Delta]) = p_1 \left(\sum_{\delta=1}^{\Delta-1} \delta + \sum_{\delta=\Delta}^{\infty} p^{\delta-\Delta}(\delta + \eta) \right) \\ &= p_1 \left(\frac{(\Delta-1)\Delta}{2} + \frac{\eta + \Delta}{1-p} + \frac{p}{(1-p)^2} \right). \end{aligned} \tag{22}$$

Substituting p_1 from (21) and minimizing over Δ (by setting the derivative $\partial L_\eta^\Delta / \partial \Delta$ to zero) yields that the optimal non-integer value of Δ is given by

$$\hat{\Delta}_\eta = \frac{\sqrt{2\eta(1-p) + p} - p}{1-p} \ .$$

Using that L_η^Δ is a convex function of Δ by (22), the optimal integer threshold Δ_η^* is either

$$\left\lfloor \frac{\sqrt{2\eta(1-p) + p} - p}{1-p} \right\rfloor \ \text{ or } \ \left\lceil \frac{\sqrt{2\eta(1-p) + p} - p}{1-p} \right\rceil \ .$$

Computing just the cost term from (22), we obtain the formula for C^Δ for any integer threshold Δ. □

D. Proof of Theorem 2

Let π_{η^*} denote the deterministic solution of the Bellman equation (7). If $C^{\pi_{\eta^*}} = C_{max}$ then π_{η^*} is the optimal solution to Problem 1 (by Proposition 3.2 and Lemma 3.10 of [21]), and since it is a threshold policy by Lemma 1 with threshold $\Delta_{C_{max}} = \Delta_1 = \Delta_2$ (as can be obtained by inverting equation (16)), the theorem holds.

For $C^{\pi_{\eta^*}} \neq C_{max}$, we first show that the optimal policy is a mixture of two threshold policies that differ in at most a single state, based on the construction used to prove Theorem 2.5 of [21]. Assume without loss of generality that $C^{\pi_{\eta^*}} > C_{max}$, and consider a sequence of Lagrange multipliers

$\eta_n \downarrow \eta^*$ such that the corresponding deterministic solutions $\pi^*_{\eta_n}$ of (7) (which are also η_n-optimal by Proposition 3.2 of [21]) converge to a policy π^*.[5] By Lemma 2, these are all threshold policies, and so π_{η^*} and π^* are both threshold policies. By Lemma 3.7 (iii) of [21], π^* is η^*-optimal, and $C^{\pi^*} \leq C_{max}$ by Lemma 3.4 of [21]. If $C^{\pi^*} = C_{max}$ then the proof can be completed as in the case of $C^{\pi_{\eta^*}} = C_{max}$. Thus, we are left with the case of $C^{\pi^*} < C_{max}$. Denoting by $(\mu, \pi_{\eta^*}, \pi^*)$ the randomized policy that selects π_{η^*} with probability $\mu = \frac{C_{max} - C^{\pi^*}}{C^{\pi_{\eta^*}} - C^{\pi^*}}$ and π^* with probability $1-\mu$ before the system starts and then uses the selected policy forever, it follows that $(\mu, \pi_{\eta^*}, \pi^*)$ has average transmission cost C_{max}, while it is also η^*-optimal by Lemma 3.9 of [21]. Therefore, by Lemma 3.10 of [21], $(\mu, \pi_{\eta^*}, \pi^*)$ is an optimal solution to Problem 1.

Next we show that the thresholds of the two policies are Δ_1 and Δ_2. From the proof of Lemma 2, one can easily deduce that the average AoI of a threshold policy for any integer threshold Δ is given by

$$J^{\Delta} = \frac{(\Delta(1-p)+p)^2 + p}{2(1-p)(\Delta(1-p)+p)} + \frac{1}{2}.$$

Expressing J^{Δ} as a function of C^{Δ} (given in (16)), and extending the definition of C^{Δ} and J^{Δ} to positive real values of Δ, one can see that

$$J^{\Delta} = \frac{1}{2(1-p)C^{\Delta}} + \frac{1}{2} + \frac{pC^{\Delta}}{2(1-p)}$$

is a convex function of C^{Δ}. Denoting the threshold of π_{η^*} and π^* by Δ_{η^*} and Δ^*, respectively, we obviously have that the expected average AoI of $(\mu, \pi_{\eta^*}, \pi^*)$ is $\mu J^{\Delta_{\eta^*}} + (1-\mu)J^{\pi^*}$, while the expected average transmission cost is C_{max}. By the convexity of J^{Δ}, and since $C^{\Delta^*} = C^{\pi^*} < C_{max} < C^{\pi_{\eta^*}} = C^{\Delta_{\eta^*}}$ it follows that the integer threshold values minimizing the AoI must be the closest integers (from above and below) to $\Delta_{C_{max}}$, the minimizer of J^{Δ} over the reals. That is, $\Delta_{\eta^*} = \Delta_1 = \lfloor \Delta_{C_{max}} \rfloor$ and $\Delta^* = \Delta_2 = \lceil \Delta_{C_{max}} \rceil$ (recall that the transmission cost C^{Δ} is a decreasing function of the threshold Δ). Note that this also implies that $\mu = \mu^*$ (recall that μ^* is specified in the statement of the theorem).

To complete the proof, define $(\pi_{\eta^*}, \pi^*) - \mu$ to be the policy that randomly selects between π_{η^*} and π^* every time state $(1,0)$ is reached (independently, and with probability μ and $1-\mu$, resp.) and follows the chosen policy until state $(1,0)$ is reached again. Since $(1,0)$ is a positive recurrent state of both π^* and π_{η^*}, the policy $(\pi_{\eta^*}, \pi^*) - \mu$ has the same expected AoI and transmission cost as $(\mu, \pi_{\eta^*}, \pi^*)$, which randomizes once at the beginning. Therefore, $(\pi_{\eta^*}, \pi^*) - \mu$ is optimal. Moreover, since π^* and π_{η^*} only differ in state $\delta = \Delta_2$, the randomization can be performed only in that state.

[5] If $C^{\pi_{\eta^*}} > C_{max}$, η_n should be increasing to η^*, and the rest of the proof follows the same lines as for the case of $C^{\pi_{\eta^*}} < C_{max}$.

Thus, since $\mu = \mu^*$, the policy $(\pi_{\eta^*}, \pi^*) - \mu$ is identical to $\pi^*_{C_{max},\mu^*}$, defined in the theorem, proving that $\pi^*_{C_{max},\mu^*}$ is optimal. □

E. Unconstrained case ($C_{max} = 1$)

Here we analyze the problem without a transmission constraint, that is, when $C_{max} = 1$. We show that the conditions of part (ii) of the Theorem in [22] hold, implying that there exists a deterministic optimal policy satisfying the Bellman equation (7) with $\eta = 0$ and a restricted to $\{\mathrm{n}, \mathrm{x}\}$, namely

$$h(\delta, r) + L^* = \min_{a \in \{\mathrm{n},\mathrm{x}\}} \left(\delta + \mathbb{E}\left[h(\delta', r')\right]\right) \tag{23}$$

for some function $h(\delta, r)$ and constant L^*.

For any $\alpha \in (0, 1)$, policy π (here a policy is an arbitrary, possibly randomized decision strategy that may depend on the whole history) and initial state s_0, let

$$J^\pi_\alpha(s_0) \triangleq \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t \delta^\pi_t \Big| s_0\right] ,$$

and $J_\alpha(s_0) = \inf_\pi J^\pi_\alpha(s_0)$.

Consider policy π_{n}, which transmits a new update in every step. One can verify (e.g., by induction) that the stationary distribution of the Markov chain induced by this policy is a geometric distribution over states $(\delta, 0)$ with parameter $1-p$, where $p = g(0)$: that is, the probability of being in state $(\delta, 0)$ is $(1-p)p^{\delta-1}$.

Next, we verify Assumption 1 of [22], which requires that for any α and state $s = (\delta_0, r)$, $J_\alpha(s)$ is finite. Note that given the first state is (δ_0, r), we have $\delta_t \le \delta_0 + t$. Therefore,

$$J_\alpha((\delta_0, r)) \le J^{\pi_{\mathrm{n}}}_\alpha((\delta_0, r)) \le \sum_{t=0}^{\infty} \alpha^t(\delta_0 + t) < \infty,$$

which is what we wanted to prove.

Let $h_\alpha(s) = J_\alpha(s) - J_\alpha(s_0)$, where $s_0 = (1, 0)$. In what follows, we give upper and lower bounds on h_α. Consider an arbitrary policy π starting from state $s = (\delta, r)$. Since in every time step a transmission is successful with probability at least $1-p$ (since the success probability cannot decrease with retransmissions), and if two successive transmissions are successful, the second must be a new update, in two steps the process returns to state s_0 with probability at least $(1-p)^2$. Thus, if the MDP is started from s and s_0, with probability at least $q = (1-p)^4$, they synchronize after two steps, after which the terms in the summations defining $J^\pi_\alpha(s)$ and $J^\pi_\alpha(s_0)$ become identical: denoting the AoI at time t by $\delta_t(s)$ and $\delta_t(s_0)$ for the processes started in state s and s_0, respectively, and by T the first time step they simultaneously reach the same state, we have $\delta_t(s) - \delta_t(s_0) \le \delta + t - 1$ for

$t < T$ (before the synchronization happens) and $\delta_t(s) = \delta_t(s_0)$ for $t \geq T$ (after synchronization). By our argument above, for any k, $\Pr(T \geq 2k) \leq (1-q)^k$, and so

$$\begin{aligned}
&\mathbb{E}\left[J_\alpha^\pi(s) - J_\alpha^\pi(s_0)\right] = \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t(\delta_t(s) - \delta_t(s_0))\right] \\
&= \sum_{t=0}^{\infty} \mathbb{E}\Big[\alpha^{2t}(\delta_{2t}(s) - \delta_{2t}(s_0)) + \alpha^{2t+1}(\delta_{2t+1}(s) - \delta_{2t+1}(s_0))\Big| 2t+1 < T\Big] \Pr(2t+1 < T) \\
&\leq \sum_{t=0}^{\infty} \left(\alpha^{2t}(\delta + 2t - 1) + \alpha^{2t+1}(\delta + 2t)\right)(1-q)^{t+1} \\
&< 2\delta \sum_{t=0}^{\infty} (1-q)^{t+1} + 4\sum_{t=0}^{\infty} t(1-q)^{t+1} \\
&= \frac{2(1-q)}{q}\delta + \frac{4(1-q)^2}{q^2}
\end{aligned}$$

Therefore, we have

$$h_\alpha(s) \leq \sup_\pi \mathbb{E}\left[J_\alpha^\pi(s) - J_\alpha^\pi(s_0)\right] < \frac{2(1-q)}{q}\delta + \frac{4(1-q)^2}{q^2} \triangleq M_\delta.$$

Similarly, since $\delta_t(s_0) - \delta_t(s) \leq t$, we can prove that $\mathbb{E}\left[J_\alpha(s_0) - J_\alpha(s)\right] < \frac{4(1-q)}{q^2}$, which implies $-\frac{4(1-q)}{q^2} \leq h_\alpha(s)$. The latter directly proves Assumption 2 of [22], which requires $h_\alpha(s)$ to be uniformly bounded from below by a nonpositive constant for all $\alpha \in (0,1)$ and state s.

Finally, for any starting state $s = (\delta, r)$ let $s' = (\delta', r')$ denote the next state following action a. Assumptions 3 and 3* of [22] are satisfied if $\mathbb{E}\left[M_{\delta'}|s,a\right] < \infty$ holds for all s, s' and a. Since for any s and a there can be only two states s' with non-zero probability and all $M_{\delta'}$ are finite, this is trivially satisfied.

Therefore, Assumptions 1–3* of Sennott [22] are satisfied, and hence part (ii) of her Theorem implies that there exists a deterministic, optimal policy satisfying the Bellman equation (23) (equivalently, equation (7) with $\eta = 0$ and a restricted to $\{\mathrm{n}, \mathrm{x}\}$).

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their careful reading of the paper and their insightful comments.

REFERENCES

[1] E. T. Ceran, D. Gündüz, and A. György, "Average age of information with hybrid ARQ under a resource constraint," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2018.

[2] E. Altman, R. E. Azouzi, D. S. Menasché, and Y. Xu, "Forever young: Aging control in DTNs," *CoRR, abs/1009.4733*, 2010.

[3] S. Kaul, M. Gruteser, V. Rai, and J. Kenney, "Minimizing age of information in vehicular networks," in *IEEE Coms. Society Conf. on Sensor, Mesh and Ad Hoc Coms. and Nets.*, June 2011, pp. 350–358.

[4] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *Proc. IEEE INFOCOM,*, March 2012, pp. 2731–2735.

[5] P. Frenger, S. Parkvall, and E. Dahlman, "Performance comparison of HARQ with chase combining and incremental redundancy for HSDPA," in *Proc. IEEE Vehicular Technology Conf.*, vol. 3, 2001, pp. 1829–1833.

[6] V. Tripathi, E. Visotsky, R. Peterson, and M. Honig, "Reliability-based type ii hybrid ARQ schemes," in *IEEE Int'l Conf. on Communications,*, vol. 4, May 2003, pp. 2899–2903 vol.4.

[7] X. Lagrange, "Throughput of HARQ protocols on a block fading channel," *IEEE Communications Letters*, vol. 14, no. 3, pp. 257–259, March 2010.

[8] D. Gunduz, K. Stamatiou, N. Michelusi, and M. Zorzi, "Designing intelligent energy harvesting communication systems," *IEEE Communications Magazine*, vol. 52, pp. 210–216, 2014.

[9] B. T. Bacinoglu, E. T. Ceran, and E. Uysal-Biyikoglu, "Age of information under energy replenishment constraints," in *2015 Information Theory and Applications Workshop (ITA)*, Feb 2015, pp. 25–31.

[10] I. Kadota, E. Uysal-Biyikoglu, R. Singh, and E. Modiano, "Minimizing age of information in broadcast wireless networks," in *Allerton Conf. On on Communication, Control, and Computing*, Sep. 2016.

[11] Y. P. Hsu, E. Modiano, and L. Duan, "Age of information: Design and analysis of optimal scheduling algorithms," in *IEEE Int'l Symposium on Information Theory (ISIT)*, June 2017, pp. 561–565.

[12] 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 Transactions on Information Theory*, vol. 63, no. 11, pp. 7492–7508, Nov 2017.

[13] K. Chen and L. Huang, "Age-of-information in the presence of error," in *IEEE Int'l Symposium on Information Theory (ISIT)*, July 2016, pp. 2579–2583.

[14] S. Mahadevan, "Average reward reinforcement learning: Foundations, algorithms, and empirical results," *Machine Learning*, vol. 22, no. 1, pp. 159–195, 1996.

[15] P. Parag, A. Taghavi, and J. F. Chamberland, "On real-time status updates over symbol erasure channels," in *IEEE Wireless Communications and Networking Conference (WCNC)*, March 2017, pp. 1–6.

[16] E. Najm, R. Yates, and E. Soljanin, "Status updates through M/G/1/1 queues with HARQ," in *IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 131–135.

[17] R. D. Yates, E. Najm, E. Soljanin, and J. Zhong, "Timely updates over an erasure channel," in *IEEE Int'l Symposium on Information Theory (ISIT)*, June 2017, pp. 316–320.

[18] M. R. el Fenni, R. El-Azouzi, D. S. Menasche, and Y. Xu, "Optimal sensing policies for smartphones in hybrid networks: A POMDP approach," in *Int'l ICST Conf. on Performance Evaluation Methodologies and Tools*, Oct 2012, pp. 89–98.

[19] "Approved draft IEEE standard for local and metropolitan area networks corrigendum to IEEE standard for local and metropolitan area networks-part 16: Air interface for fixed broadband wireless access systems (incorporated into IEEE std 802.16e-2005 and IEEE std 802.16-2004/cor 1-2005 e)," *IEEE Std P802.16/Cor1/D5*, 2005.

[20] E. Altman, *Constrained Markov Decision Processes*, ser. Stochastic modeling. Boca Raton, London: Chapman & Hall/CRC, 1999.

[21] L. I. Sennott, "Constrained average cost Markov decision chains," *Probability in Eng. and Informational Sciences*, vol. 7, no. 1, p. 6983, 1993.

[22] ——, "Average cost optimal stationary policies in infinite state Markov decision processes with unbounded costs," *Operations Research*, vol. 37, no. 4, pp. 626–633, 1989.

[23] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY, USA: John Wiley & Sons, 1994.

[24] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 2nd ed. Athena Scientific, 2000.

[25] H. J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications*, 1997.

[26] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*, 1st ed. Cambridge, MA, USA: MIT Press, 1998.

[27] S. M. Ross, *Introduction to Probability Models, Ninth Edition*. Orlando, FL, USA: Academic Press, Inc., 2006.

[28] D. M. Topkis, "Minimizing a submodular function on a lattice," *Oper. Res.*, vol. 26, no. 2, pp. 305–321, Apr. 1978.