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* (AoI) 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. First, the optimal scheduling policy is 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. Then, for an unknown environment, an average-cost reinforcement learning (RL) algorithm is proposed that learns the system parameters and the transmission policy in real time. The effectiveness of the proposed methods are verified through numerical simulations.**

I. Introduction

We consider a source node which continually communicates the most up-to-date status packets to a destination (see Figure 1). In particular, we are interested in the *age of information (AoI)* [1], [2], [3] at the destination, for a system in which the source node samples an underlying time-varying process and sends the sample values over an imperfect link which introduces delays. The AoI at the destination at any point in time can simply be defined as the amount of time that elapsed since the most recent status update at the destination was generated. Our goal is to minimize the average AoI taking into account packet *retransmissions*. 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 latter, the receiver combines information from all previous transmission attempts of a packet in order to increase the success probability of decoding after each retransmission [4].

To address the trade-off between the success probability and the freshness of the status update to be transmitted, we develop scheduling policies to minimize the expected average AoI under an average transmission-rate constraint. This constraint is motivated by the fact that sensors sending status updates have usually limited energy supplies (e.g., are powered via energy harvesting); 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 depends on the number of previous unsuccessful transmissions); hence, the source can judiciously decide when to retransmit, and when to discard failed information and send a fresh update. Then, we consider transmitting status updates over an unknown channel, in which the success probabilities of transmission attempts are not known *a priori*, and must be learned in an online fashion using the ACK/NACK feedback signals.

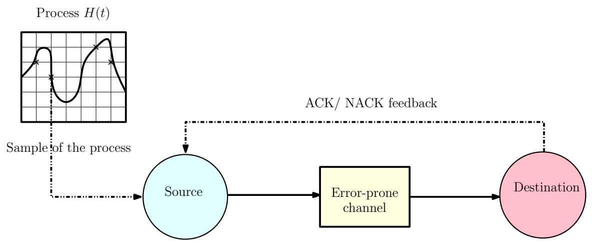


Figure 1. System model of a status update system over an error-prone point-to-point link in the presence of ACK/NACK feedback from the destination.

The main contributions of this paper are outlined next:

- 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.
- The optimal preemptive transmission policy for the standard ARQ protocol is shown to be threshold-type, and the optimal threshold value is derived in closed-form.
- An average-cost *reinforcement learning* (RL) algorithm is proposed to learn the optimal scheduling decisions when the transmission success probabilities are unknown.
- Extensive simulations are conducted in order to evaluate the impact of the resource constraint on the average AoI for the ARQ and HARQ protocols.

A. Related Work

Most of the earlier work on AoI considered 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 [2], [3]. Instead, in the so-called *generate-at-will* model, [1], [5], [6], [7], also considered in this paper, the status updates of the underlying process of interest can be sampled and generated at any time by the source node.

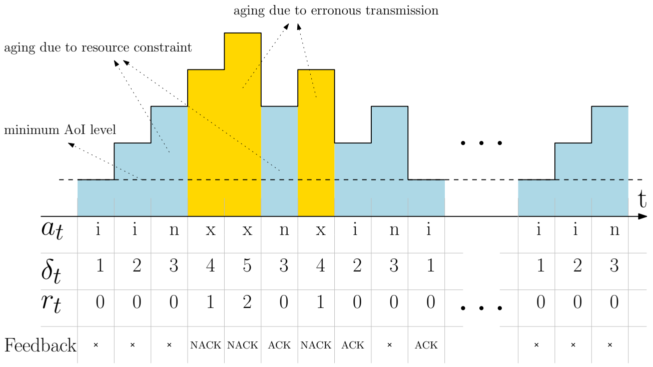
A constant packet failure probability for status update systems is investigated for the first time in [8], but this work focuses on an M/M/1 queuing model, and no feedback is considered for retransmissions. The paper [6] considers broadcasting of status updates to multiple clients over an unreliable broadcast channel. However, this paper only considers work-conserving policies, which update the information at every time slot, since no constraint is imposed on the number of updates. Optimizing the scheduling decisions in an AoI system multiple receivers is also investigated in [7], focusing on a perfect transmission medium, and an optimal scheduling algorithm for the MDP is shown to be threshold-type. The AoI in the presence of HARQ is modeled through an M/G/1/1 queue in [9]; however, no resource constraint is taken into account, and the status update arrivals are assumed to be memoryless and random, in contrast to our work, which considers general and controlled status update generation. To the best of our knowledge, this is the first work in the literature that addresses a status update system with HARQ in the presence of resource constraints.

The rest of the paper is organized as follows. In Section II, the system model is presented, and the problem of minimizing the average AoI with HARQ under a resource constraint is formulated as a *constrained Markov decision process* (CMDP). In Section III, a primal-dual algorithm is proposed to solve this CMDP. AoI under the standard ARQ protocol is investigated in Section IV, and a computationally efficient solution that minimizes the average AoI is proposed. Section V introduces an RL algorithm to minimize AoI in an unknown environment. Simulation results are presented in Section VI for both HARQ and ARQ protocols, and the paper is concluded in Section VII.

II. System Model and Problem Formulation

We consider a time-slotted status update system over an error-prone wireless link (see Figure 1 for an illustration). The source monitors an underlying time-varying process, and it is able to generate a status update at the beginning of each time slot; this is known as the *generate-at-will* model [6]. A transmission attempt of a status update takes constant time, which is assumed to be equal to the duration of one time slot.

We assume that the channel state changes randomly from one time slot to the next in an independent and identically distributed fashion. We further assume the availability of error- and delay-free single-bit feedback from the destination to the source node. Successful reception 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, and the received signal is discarded after each failed transmission attempt. In the considered AoI framework there is no point in retransmitting a failed out-of-date status packet if it has the same error probability with a fresh status update. Hence, the source always removes a failed status signal, and transmits a fresh status update (*preemption*). On the other hand, when the HARQ protocol is adopted, signals from all previous transmission attempts are combined for decoding; and therefore, the probability of error decreases with the number of retransmissions. In general, the error probability of each retransmission attempt depends on the particular combining technique used, as well as the channel conditions [4].



a_t	i	i	n	x	x	n	x	i	n	i		i	i	n
δ_t	1	2	3	4	5	3	4	2	3	1		1	2	3
r_t	0	0	0	1	2	0	1	0	0	0	. . .	0	0	0
Feedback	×	×	×	NACK	NACK	ACK	NACK	ACK	×	ACK		×	×	×

Figure 2. Illustration of the AoI in a slotted status update system with HARQ.

The AoI is defined as the time elapsed since the most up-to-date packet at the destination is generated. Assume that the most up-to-date packet at the destination at time t has a time stamp of generation $U(t)$, then the AoI at time t, denoted by δ_t, is defined as: $\delta_t \triangleq t - U(t)$. Therefore, the AoI increases by one when a transmission fails, while it decreases to one (or, to the number of retransmissions in the case of HARQ) when a status update is successfully decoded.

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 [10]). 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 $r_1 \leq r_2$. Standard HARQ methods only allow a finite maximum number of retransmissions r_{max} [11], which is also adopted here to simplify the analysis.

Let $\delta_t \in \mathbb{Z}^+$ denote the AoI at time slot t, and $r_t \in \{0, \dots, r_{max}\}$ denote the number of previous transmission attempts of the same packet. Then the state of the system can be described by the vector $s_t \triangleq (\delta_t, r_t)$. At each time slot, the source node takes one of the three possible actions, denoted by $a \in \mathcal{A}$, where $\mathcal{A} = \{\mathrm{i}, \mathrm{n}, \mathrm{x}\}$: (i) remain idle ($a = \mathrm{i}$); (ii) generate and transmit a new status update packet ($a = \mathrm{n}$); or (iii) retransmit the previously failed packet ($a = \mathrm{x}$). 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 a suboptimal action since it does not contribute to decreasing the AoI. However, continuous transmission is typically not possible in practice due to energy or interference constraints. To model these situations, we impose a constraint on the average number of transmissions, denoted by $C_{max} \in (0, 1]$.

This leads to the CMDP formulation, defined by the 5-tuple $(\mathcal{S}, \mathcal{A}, \mathrm{P}, c, d)$ [12]: The countable set of states $(\delta, r) \in \mathcal{S}$

and the finite action set $\mathcal{A} = \{\mathrm{i}, \mathrm{n}, \mathrm{x}\}$ are already 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 (1). The instantaneous cost function $c : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, which models the AoI at the destination, is defined as $c((\delta, r), a) = \delta$ for any $(\delta, r) \in \mathcal{S}$, $a \in \mathcal{A}$, independent of the action a. The instantaneous transmission cost related to the constraint, $d : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, is independent of the state and depends only on the action a, where $d = 0$ if $a = \mathrm{i}$, and $d = 1$, otherwise.

The transition probabilities of the CMDP are given as follows (with omitting the parenthesis from (δ, r)):

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

and $\mathrm{P}(\delta', r'|\delta, r, a) = 0$ otherwise.

A stationary *policy* is a decision rule represented by $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$, which maps the states $s \in \mathcal{S}$ into actions $a \in \mathcal{A}$ with some probability $\pi(a|s)$. We will 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 π. Let $J^\pi(s_0)$ denote the infinite horizon average age, and $C^\pi(s_0)$ denote the expected average number of transmissions, when policy π is employed with initial state s_0. Without loss of generality, we restrict our attention to stationary policies (see [12]), and denote the set of feasible policies by Π such that $C^\pi \leq C_{max}$ for all $\pi \in \Pi$.

We can state the CMDP optimization problem as follows, where $\mathbb{E}[\cdot]$ represents the expectation with respect to policy π and error probabilities $g(r)$:

Problem 1.

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

$$\text{s.t. } C^\pi(s_0) \triangleq \limsup_{T\rightarrow\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[a_t^\pi \neq \mathrm{i}] \middle| s_0\right] \leq C_{max}.$$

We define a policy π^* to be optimal if $J^* \triangleq J^{\pi^*} \leq J^\pi$ for all $\pi \in \Pi$. A policy is said to be deterministic if it chooses an action with probability one; with a slight abuse of notation, we will use $\pi(s)$ to denote the action taken with probability one in state s. Also, without loss of generality, we assume that the sender and the receiver are synchronized at the beginning, that is, $s_0 = (1, 0)$; and s_0 will be omitted from the notation for simplicity. We assume throughout this paper that the Markov decision process (MDP) is *unichain*: an MDP is said to be unichain if under any stationary policy the corresponding Markov chain contains a single (aperiodic) ergodic class [12]. The unichain assumption is not restrictive since any reasonable policy for Problem 1 will transmit a fresh update regularly, making the system to return to state $(1, 0)$, which results in an ergodic Markov chain.

A. Structure of the Optimal Policy

Note that for countable-state average-cost MDPs, an optimal deterministic stationary policy exists, and it can be found under certain conditions defined in [13], [14]. However, the optimal policies for constrained MDPs are no longer limited to the set of deterministic policies [12], [15], and randomized stationary policies should be considered. Since the CMDP defined in Problem 1 has a single global constraint, Corollary 1 below follows immediately from Theorem 4.4 of [12].

Corollary 1. *An optimal stationary policy for the CMDP in Problem 1 exists, and it is a mixture of two deterministic policies.*

The next proposition formalizes the simple observation that retransmitting a packet immediately after a failed attempt is better than remaining idle for some slots and then retransmitting (since remaining idle just increases the age, and the probability of successful retransmission will stay the same). The proof of the proposition is trivial, and hence omitted.

Proposition 1. *There exists an optimal policy for Problem 1 that takes a retransmission action only after a failed transmission event, that is $Pr(a_{t+1}^* = \mathrm{x}|a_t^* = \mathrm{i}) = 0$.*

Note that one can explicitly enforce the above property by slightly changing the transition kernel P by changing the first equation in (1) to $\mathrm{P}(\delta+1, 0)|\delta, r, \mathrm{i}) = 1$ (since retransmissions are not allowed in states $(\delta, 0)$).

III. Primal-Dual Algorithm to Minimize AoI

To solve the average cost CMDP in Problem 1, we adopt the Lagrangian primal-dual method [12], [15]. Similar methods have been adopted in other wireless network problems including [16], [17].

A. Relaxed Unconstrained MDP

Lagrangian relaxation of the constraint with non-negative multiplier η can be written as:

$$\begin{aligned} J_\eta^\pi = &\lim_{T\rightarrow\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \delta_t^\pi\right] \\ &- \eta\left(C_{max} - \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[a_t^\pi \neq \mathrm{i}]\right]\right), \end{aligned} \tag{3}$$

and the optimal J_η^* for a given η is defined as $J_\eta^* \triangleq \min_\pi J_\eta^\pi$. This formulation is equivalent to an unconstrained average-cost MDP, in which the instantaneous overall cost becomes $\delta_t + \eta\mathbb{1}[a_t^\pi \neq \mathrm{i}]$. It is well-known that there exits an optimal stationary deterministic policy for this problem. In particular, there exists a function $h_\eta(\delta, r)$, called the differential cost function, satisfying the so-called *Bellman optimality* equations

$$h_\eta(\delta, r) + J_\eta^* = \min_{a\in\{\mathrm{i},\mathrm{n},\mathrm{x}\}} \left(\delta + \eta \cdot \mathbb{1}[a \neq \mathrm{i}] + \mathbb{E}\left[h_\eta(\delta', r')\right]\right), \tag{4}$$

where (δ', r') is the next state obtained from (δ, r) after taking action a. We also introduce the state-action cost function:

$$Q_\eta(\delta, r, a) \triangleq \delta + \eta \cdot \mathbb{1}[a \neq \mathrm{i}] + \mathbb{E}\left[h_\eta(\delta', r')\right]. \tag{5}$$

Then the optimal policy, for any $(\delta, r) \in \mathcal{S}$, is given by the action achieving the minimum in (4):

$$\pi^*_\eta(\delta, r) \in \operatorname*{arg\,min}_{a \in \{\mathrm{i},\mathrm{n},\mathrm{x}\}} \left(Q_\eta(\delta, r, a)\right). \tag{6}$$

Note that the state space of our problem is possibly countably infinite, since the age can be arbitrarily large. However, in practice we can approximate the countable state space with a large finite space by setting a maximum 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 to N); this gives a finite state space approximation to the problem similarly to [1], [7]. Clearly, letting N go to infinity, the optimal policy for the restricted state space will converge to that of the original problem.

B. Relative Value Iteration (RVI)

The RVI algorithm can be employed to solve (4) for any given η; and hence, to find the optimal policy π^*_η [13]. Starting with an initialization of $h_0(\delta, r)$, $\forall(\delta, r)$, and setting an arbitrary but fixed reference state (δ^{ref}, r^{ref}), a single iteration for the RVI algorithm is given as follows:

$$Q_{n+1}(\delta, r, a) \leftarrow \delta + \eta \cdot \mathbb{1}[a^\pi \neq \mathrm{i}] + \mathbb{E}\left[h_n(\delta', r')\right], \tag{7}$$

$$V_{n+1}(\delta, r) \leftarrow \min_a (Q_{n+1}(\delta, r, a)), \tag{8}$$

$$h_{n+1}(\delta, r) \leftarrow V_{n+1}(\delta, r) - V_{n+1}(\delta^{ref}, r^{ref}), \tag{9}$$

where $Q_n(\delta, r, a)$, $V_n(\delta, r)$ and $h_n(\delta', r')$ denote the state action value function, value function and differential value function for iteration n, respectively.It can be shown that h_n converges to h_η, and $\pi^*_n(\delta, r) \triangleq \arg\min_a Q_n(\delta, r, a)$ converges to $\pi^*_\eta(\delta, r)$ [13]. Using the deterministic policies π^*_η, it is possible to characterize optimal policies for our CMDP problem according to Corollary 1.

C. Finding the Lagrange Parameter and Randomization

With the aim of finding a single η value such that $C_\eta \approx C_{max}$, starting with an initial parameter η^0, we run an iterative algorithm updating η as $\eta^{m+1} = \eta^m + \alpha(C_{\eta^m} - C_{max})$ for some step size parameter $\alpha \triangleq 1/\sqrt{m}$ (note that for each step we need to run the RVI algorithm to be able to determine C_{η^m}). We continue this iteration until $|\eta^{m+1} - \eta^m|$ is smaller than a given $\epsilon \in \mathbb{R}^+$, and denote the resulting value as η^*.

According to Corollary 1, one can think of the optimal policy as a randomized policy between two deterministic policies: in any state $s = (\delta, r)$, the optimal policy in the CMDP problem chooses action $\pi^*_{\eta_1}(s)$ with probability μ and $\pi^*_{\eta_2}(s)$ with probability $1-\mu$ independently for each time slot. For any η, let C_η denote the average resource consumption under the optimal policy π^*_η and J^*_η denote the average AoI for π^*_η (note that C_η and J^*_η can be computed directly through finding the stationary distribution of the chain, but can also be estimated empirically just by running the MDP with policy π^*_η). Obviously, C_η and J^*_η are monotone functions of η. Therefore, given η_1 and η_2, one can find the optimal weight, denoted by μ, by solving $\mu C_{\eta_1} + (1-\mu)C_{\eta_2} = C_{max}$, which has a solution $\mu \in [0, 1]$ if $C_{\eta_1} \geq C_{max} \geq C_{\eta_2}$. Next we approximate the values of η_1 and η_2 by $\eta^* \pm \xi$, where ξ is a small perturbation. Then the mixing coefficient can be obtained by setting the transmission rate $\mu C_{\eta_1} + (1-\mu)C_{\eta_2}$ of the mixture to C_{max}; that is,

$$\mu = \frac{C_{max} - C_{\eta_2}}{C_{\eta_1} - C_{\eta_2}}, \tag{10}$$

and the optimal policy is

$$\pi^*_{C_{max}} = \mu\pi^*_{\eta_1} + (1-\mu)\pi^*_{\eta_2}. \tag{11}$$

IV. AoI with Classical ARQ Protocol

Now, assume that the system adopts the classical ARQ protocol; that is, failed transmissions are discarded at the destination. In this case, there is no point in re-transmitting a failed packet since the successful transmission probabilities are the same for a retransmission and the transmission of a new update. The state space reduces to $\delta \in \{1, 2, \dots\}$ as $r_t = 0$, $\forall t$, and the action space to $\mathcal{A} \in \{\mathrm{i}, \mathrm{n}\}$. The probability of error of each status update is $p \triangleq g(0)$. State transitions in (1), Bellman optimality equations in (4), and the RVI algorithm can all be simplified accordingly. Thanks to these simplifications, we are able to provide a closed-form solution to the corresponding unconstrained MDP with Lagrange relaxation.

Lemma 1. *The optimal policy that minimizes J^π_η with the standard ARQ protocol is deterministic, and has a threshold structure:*

$$\pi^*(\delta) = \begin{cases} \mathrm{n} & \text{if } \delta \geq \Delta^*_\eta, \\ \mathrm{i} & \text{if } \delta < \Delta^*_\eta. \end{cases} \tag{12}$$

for some integer Δ^*_η that depends on η.

Proof. Proof is not included here due to space limitations, but will be provided in the extended version. □

Lemma 2. *Under the standard ARQ protocol, the optimal threshold value for the Lagrangian MDP with Lagrange multiplier η satisfies*

$$\Delta^*_\eta \in \left\{ \left\lfloor \frac{\sqrt{2\eta(1-p)+p} - p}{1-p} \right\rfloor, \left\lceil \frac{\sqrt{2\eta(1-p)+p} - p}{1-p} \right\rceil \right\}. \tag{13}$$

Proof. Proof will be provided in the extended version. □

The transmission cost (per time slot) of the threshold policy for any integer threshold Δ is given by

$$C^\Delta = \frac{1}{\Delta(1-p)+p}, \tag{14}$$

and the corresponding average AoI for the CMDP is

$$J^{\Delta} = \frac{(\Delta(1-p)+p)^2+p}{2(1-p)(\Delta(1-p)+p)} + \frac{1}{2}. \tag{15}$$

We note that, for all positive integers Δ, the points (C^{Δ}, J^{Δ}) lie on the lower convex hull of the $(C_\eta, J^*_\eta), \eta \geq 0$, and no other deterministic policy achieves the lower convex hull. Therefore, by (11), if $C_{max} \in (C^{\Delta}, C^{\Delta+1})$ for some Δ, then the optimal policy is a mixture of the threshold policies with thresholds Δ and $\Delta+1$. These threshold values can be found by inverting (14), and taking the closest integers to the resulting non-integer threshold value. Thus, we have obtained the following result which gives a closed form expression for the optimal policy under the ARQ protocol:

Theorem 1. *For any $C \in (0,1]$, let $\Delta_{C_{max}} = \frac{1/C_{max}-p}{1-p}$. Then the optimal policy for Problem 1 under the ARQ protocol is a mixture of two threshold policies with thresholds $\Delta_1 = \lfloor \Delta_{C_{max}} \rfloor$ and $\Delta_2 = \lceil \Delta_{C_{max}} \rceil$, respectively, with a mixture coefficient*

$$\mu = \frac{C_{max} - C^{\Delta_2}}{C^{\Delta_1} - C^{\Delta_2}} \ .$$

V. Learning to minimize AoI in an unknown environment

In most practical scenarios, channel error probabilities for all retransmissions may not be known at the time of deployment, or may change over time. In this section, we consider a practically motivated scenario, in which the source node does not have *a priori* information about 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 [18], [19]. For the average AoI minimization in Problem 1, an average cost version of the SARSA algorithm [19], as outlined in Algorithm 1, is employed with *Boltzmann* (*softmax*) exploration. Moreover, we update the gain J_η at every time slot based on the empirical average, instead of updating it at non-explored time slots and losing information. The resulting algorithm is called *average-cost SARSA with softmax*.

VI. Simulation Results

Decoding error probability is assumed to be given by $g(r) \triangleq p_0 2^{-r}$, where p_0 denotes the failure probability of the first transmission, and r is the retransmission count. The exponential behavior of the error probability follows from previous research on HARQ [4], [10]. In Figure 3, we illustrate the deterministic policies obtained by RVI and η search for given C_{max}, r_{max} and p_0 values. Final policies are generated by randomizing between $\pi_{\eta-\xi}$ and $\pi_{\eta+\xi}$. As it can be seen from the figure, the designed policy transmits less as η increases, and vice versa.

Figure 4 illustrates the performance of the proposed randomized HARQ policy with respect to C_{max} for different p_0 values. We also include the performance of deterministic and randomized threshold policies with ARQ. As expected, average AoI can be reduced by randomization between two deterministic threshold policies. We observe that the average AoI decreases exponentially with respect to C_{max}. The average AoI also decreases with decreasing p_0 and with increasing r_{max} as expected.

Figure 5 shows the evolution of the average AoI over time when the proposed average-cost SARSA learning algorithm is employed. It can be observed that the average AoI achieved by Algorithm 1 converges to the one obtained from the RVI algorithm which has *a priori* knowledge of $g(r)$. Average AoI achieved by the proposed online learning algorithm is

Algorithm 1: Average-cost modified SARSA with softmax

```
Input : Lagrange parameter $\eta$
$n \leftarrow 0$ /* time iteration */
$\tau \leftarrow 1$ /* softmax temperature parameter */
$Q_\eta^{N \times M \times 3} \leftarrow 0$ , $J^*_\eta \leftarrow 0$ /* initialization */
foreach $n$ do
    foreach $a \in \mathcal{A}$ do
        $\pi(a|s_n) = \frac{\exp(-Q_\eta(s_n,a)/\tau)}{\sum_{a' \in \mathcal{A}} \exp(-Q_\eta(s_n,a')/\tau)}$
    end
    Sample $a_n$ from $\pi(a|S_n)$, observe next state $s_{n+1}$ and cost $c_n = \delta_n + \eta \mathbb{1}[a_n \neq \mathrm{i}]$
    foreach $a \in \mathcal{A}$ do
        $\pi(a|s_{n+1}) = \frac{\exp(-Q_\eta(s_{n+1},a_{n+1})/\tau)}{\sum_{a'_{n+1} \in \mathcal{A}} \exp(-Q_\eta(s_{n+1},a'_{n+1})/\tau)}$
    end
    Sample $a_{n+1}$ from $\pi(a_{n+1}|s_{n+1})$
    Update $Q_\eta(s_n,a_n)$ as:
    $\alpha_n \leftarrow 1/\sqrt{n}$
    $Q_\eta(s_n,a_n) \leftarrow$
        $Q_\eta(s_n,a_n) + \alpha_n[c_n - J^*_\eta + Q_\eta(s_{n+1},a_{n+1}) - Q_\eta(s_n,a_n)]$
    /* update $J^*_\eta$ at every step */
    $J^*_\eta \leftarrow J^*_\eta + 1/n[c_n - J^*_\eta]$
    $n \leftarrow n+1$
end
```

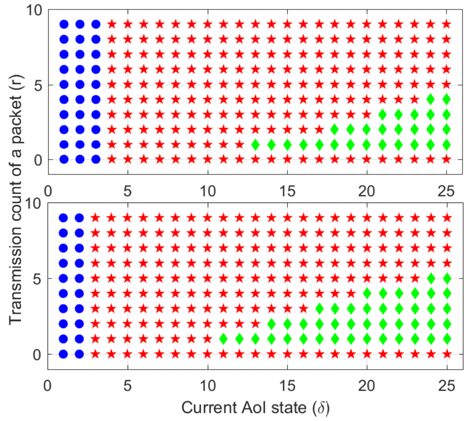


Figure 3. Deterministic policies $\pi_{\eta+\xi}$ (top) and $\pi_{\eta-\xi}$ (bottom) when $C_{max} = 0.4$, $p_0 = 0.3$, and $r_{max} = 9$. (Blue circles, red stars and green diamonds represent actions $\pi_\eta(\delta, r) = \mathrm{i}$, n and x, respectively.)

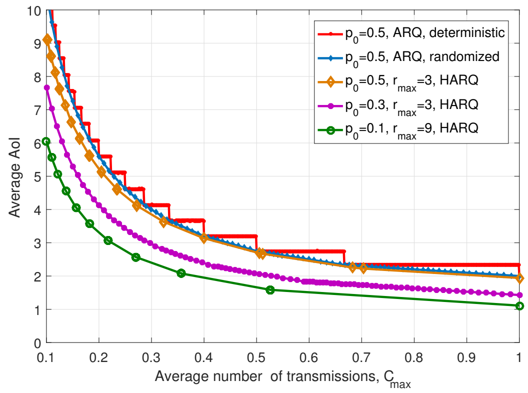


Figure 4. Expected average AoI J^{π} with respect to C_{max} for ARQ and HARQ protocols for different p_0 and r_{max} values. Time horizon is set to $T = 1000$, and the results are averaged over 1000 trials.

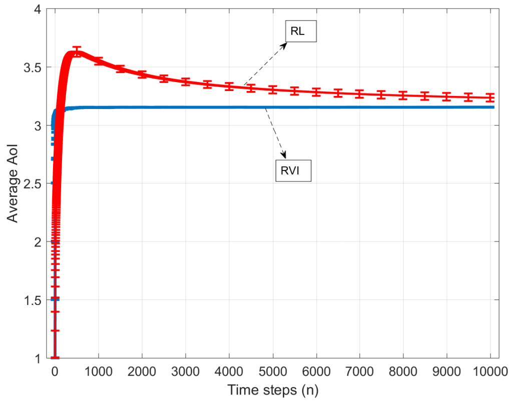


Figure 5. Performance of the average-cost SARSA for $r_{max} = 3$, $p_0 = 0.2$, $\eta = 10$ and $n \leq 10000$, averaged over 1000 trials (both the mean and the variance are shown).

presented in Figure 6 as a function of C_{max}, which shows that the performance is close to that of RVI.

VII. Conclusions

We have considered a wireless 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 destination. 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 proposed for practical deployments, when the system characteristics may not be known in advance. The algorithms designed in this paper are relevant to other systems concerning the timeliness of information, and the proposed methodology can be used in other CMDP problems. As a future work, the problem will be extended to time-correlated channel statistics in a multi-user setting.

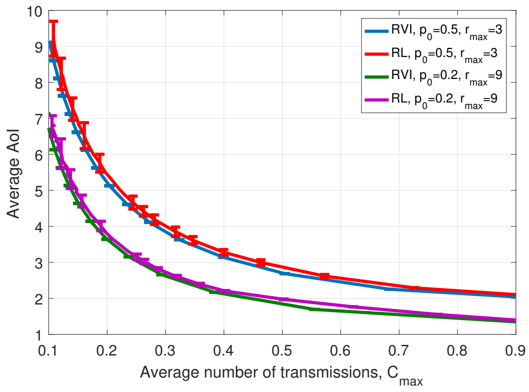


Figure 6. Performance of the RL (average-cost SARSA) with respect to RVI Algorithm for $n = 10000$ iterations, and values are averaged over 1000 trials for different p_0 and r_{max} values (both the mean and the variance are shown).

References

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

[2] 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.

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

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

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

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

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

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

[9] E. Najm, R. D. Yates, and E. Soljanin, "Status updates through M/G/1/1 queues with HARQ," *CoRR*, vol. abs/1704.03937, 2017.

[10] V. Tripathi, E. Visotsky, R. Peterson, and M. Honig, "Reliability-based type ii hybrid ARQ schemes," in *Communications, 2003. ICC '03. IEEE International Conference on*, vol. 4, May 2003, pp. 2899–2903 vol.4.

[11] "IEEE standard for local and metropolitan area networks-part 16: Air interface for fixed broadband wireless access systems," *IEEE Std P802.16/Cor1/D5*, 2005.

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

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

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

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

[16] M. H. Ngo and V. Krishnamurthy, "Monotonicity of constrained optimal transmission policies in correlated fading channels with ARQ," *IEEE Trans. on Signal Processing*, vol. 58, no. 1, pp. 438–451, Jan 2010.

[17] A. Roy and A. Karandikar, "Optimal radio access technology selection policy for lte-wifi network," in *Int'l Symp. Modeling and Opt. in Mobile, Ad Hoc, and Wireless Nets. (WiOpt)*, May 2015, pp. 291–298.

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

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