A Reinforcement Learning Approach to Age of Information in Multi-User Networks

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 time-sensitive data to multiple users over error-prone communication channels is studied with the goal of minimizing the long-term average *age of information (AoI)* at the users under a constraint on the average number of transmissions. The source can transmit only to a single user at each time slot, and after each transmission, it receives an instantaneous ACK/NACK feedback from the intended receiver, and decides on when and to which user to transmit the next update. 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. Then a *reinforcement learning* (RL) approach is introduced, which does not assume any *a priori* information on the random processes governing the channel states. Different RL methods are applied and compared through numerical simulations.**

I. Introduction

We consider a status-update system, where a source node communicates the state of a time-varying process to multiple users. The source can sample the process and schedule transmission to users over imperfect links. The timeliness of the information at each user is measured by the *age of information* (AoI), defined as the amount of time elapsed since the most recent status update received by that user was generated at the source. The goal of the update system is to minimize the *average* AoI across the users [1]–[3]. Most of the earlier work on AoI consider queue-based models, in which the status updates arrive at the source node randomly according to a Poisson process, and are stored in a buffer before being transmitted to the destination [2], [3]. Instead, in this paper we consider the so-called *generate-at-will* model, where the status updates of the underlying process of interest can be generated at any time [1], [4]–[7].

In this paper, we address the scheduling of status updates in a multi-user network under a transmission-rate constraint. This constraint is motivated by the fact that sensors sending status updates usually have limited energy supplies (e.g., are powered via energy harvesting [8]); hence, they cannot send an unlimited number of updates. We assume that the source can transmit to only a single user at each time slot, and the communication channels experience fading. While the transmitter does not have channel state information, we assume the presence of a single bit perfect feedback link from each user to the source terminal, across which the corresponding receiver can send ACK/NACK feedback after each transmission. We consider both the standard ARQ and the hybrid ARQ (HARQ) protocols. Note that, in the standard ARQ protocol, the same transmission is repeated until it is successfully received; however, in a status update system no retransmission takes place in this case, as it is always better to send a fresh status update. On the other hand, under HARQ, one may repeat previously sent packets as the probability of correct decoding increases with multiple transmissions. First, we assume that the success probability of each transmission attempt is known beforehand, in which case the source can judiciously decide when to transmit, or, in the case of HARQ, to retransmit or discard failed information and send a fresh update. Then, we consider scheduling status updates over unknown channels, in which case 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.

AoI in multi-user networks has been studied in [6], [7], [9], [10]. A source node sending time-sensitive information to a number of users through unreliable channels is considered in [7], where the problem is formulated as a multi-armed restless bandit. AoI in the presence of retransmissions has been considered in [9], [11]. In an earlier paper we studied a point-to-point status-update system under a transmission-rate constraint [12]. Here, we extend the results of [12] to the multi-user setting; in addition, more sophisticated reinforcement learning (RL) algorithms are used. We also demonstrate, through simulations, that the resulting update policy can perform very close the optimum.

II. System Model and Problem Formulation

We consider a slotted status update system. The source terminal monitors a time-varying process, for which it is able to generate a status update at the beginning of each time slot. It can only transmit the status update to a single user at each time slot. This can be either because of dedicated orthogonal links to the users, e.g., a wired network, or because the users are interested in distinct

This work was supported in part by the European Research Council (ERC) Starting Grant BEACON (grant agreement no. 725731).

$$
\mathcal{P}_{s,s'}(a) = \begin{cases} 1 & \text{if } a = \mathrm{i}, \delta_i' = \delta_i + 1, r_i' = r_i, \forall i; \\ 1 - g_j(0) & \text{if } a = \mathrm{n}_j, \delta_j' = 1, r_j' = 0, \delta_i' = \delta_i' + 1, r_i' = r_i, \forall i \neq j; \\ g_j(0) & \text{if } a = \mathrm{n}_j, \delta_j' = \delta_j{+}1, r_j'{=}1\delta_i' = \delta_i' + 1, r_i' = r_i, \forall i \neq j; \\ 1 - g_j(r_j) & \text{if } a = \mathrm{x}_j, \delta_j' = r_j, r_j' = 0, \delta_i' = \delta_i' + 1, r_i' = r_i, \forall i \neq j; \\ g_j(r_j) & \text{if } a = \mathrm{x}_j, \delta_j' = \delta_j + 1, r_j' = r_j' + 1, \delta_i' = \delta_i' + 1, r_i' = r_i, \forall i \neq j; \\ 0 & \text{otherwise.} \end{cases} \tag{1}
$$

processes. A transmission attempt of a status update to a single user takes constant time, which is assumed to be equal to the duration of one time slot.

We assume that the state of each of the channels 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 corresponding receivers. We assume the availability of an instantaneous error-free single-bit ACK/NACK feedback from each user to the source. Let M denote the number of users. Assume that the most up-to-date packet received by the the j^{th} user ($j \in \{1, \ldots, M\}$) before time slot t was generated in slot $U_j(t)$; then the AoI for user j at the beginning of time slot t is defined as $\delta_{j,t} \triangleq t - U_j(t) \in \mathbb{Z}^+$. Therefore, $\delta_{j,t}$ increases by one when the source chooses not to transmit to user j, or if the transmission fails, while it decreases to one or, in the case of HARQ, to the number of retransmissions, when a status update is successfully decoded.

In the standard ARQ, a packet is retransmitted after each NACK feedback, until it is successfully decoded. 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 with a fresh status update. Hence, the source always removes a failed status signal, and transmits a fresh update. On the other hand, in HARQ, signals from previous transmission attempts are combined; and therefore, the probability of error decreases with every retransmission [13].

For the j^{th} user, let $r_{j,t} \in \{0, \ldots, r_{max}\}$ denote the number of previous transmission attempts of the most recent packet (r_{max} may be infinite). Then, the state of the system can be described by the vector $s_t \triangleq (\delta_{1,t}, r_{1,t}, \ldots, \delta_{M,t}, r_{M,t})$, where s_t belongs to the set of possible states $\mathcal{S} \subset (\mathbb{Z}^+ \times \{1, \ldots, r_{max}\})^M$. At each time slot, the source node takes an action a from the set of actions $\mathcal{A} = \{\mathrm{i}, \mathrm{n}_1, \mathrm{x}_1, \ldots, \mathrm{n}_M, \mathrm{x}_M\}$: in particular,the source can i) remain idle ($a = \mathrm{i}$); ii) generate and transmit a new status update packet to the j^{th} user ($a = \mathrm{n}_j$); or, iii) retransmit the previously failed packet to the j^{th} user ($a = \mathrm{x}_j$, $j = 1, \ldots, M$).

For the j^{th} user, the probability of error after r retransmissions, denoted by $g_j(r)$, depends on r and the particular HARQ scheme used [13]. In any reasonable HARQ strategy, $g_j(r)$ is non-increasing in r, i.e., $g_j(r) \geq g_j(r')$ for all $r \leq r'$.

Note that, if no resource constraint is imposed on the source, remaining idle is clearly a suboptimal action. 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 $\lambda \in (0, 1]$.

This leads to the CMDP formulation, defined by the 5-tuple $\big(\mathcal{S}, \mathcal{A}, \mathcal{P}, c, d\big)$ [14]: The countable set of states $\mathcal{S}$ and the finite set of actions $\mathcal{A}$ have already been defined. $\mathcal{P}$ refers to the transition kernel, specified in (1), where $\mathcal{P}_{s,s'}(a) = \Pr(s_{t+1} = s' \mid s_t = s, a_t = a)$ is the probability that action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ at time t leads to state $s' \in \mathcal{S}$ at time $t{+}1$. The instantaneous cost function $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as the weighted sum of AoI for multiple users, independently of a. Formally, $c(s, a) = \Delta \triangleq w_1\delta_1 + \cdots + w_M\delta_M$, where the weight $w_j > 0$ represents priority of user j. The instantaneous transmission cost $d : \mathcal{A} \to \mathbb{R}$ is defined as $d(\mathrm{i}) = 0$ and $d(a) = 1$ if $a \neq \mathrm{i}$.

A stationary *policy* $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ maps each state $s \in \mathcal{S}$ to an action $a \in \mathcal{A}$ with probability $\pi(a|s)$ ($\pi(\cdot|s)$ is a distribution over $\mathcal{A}$). We use $s_t^\pi = (\delta_{1,t}^\pi, r_{1,t}^\pi, \ldots, \delta_{M,t}^\pi, r_{M,t}^\pi)$ and a_t^π to denote the sequences of states and actions, respectively, induced by policy π, while $\Delta_t^\pi \triangleq \sum_{j=1}^M w_j \delta_{j,t}^\pi$ denotes the instantaneous weighted cost.

The infinite horizon expected weighted average AoI for policy π starting from the initial state $s_0 \in \mathcal{S}$ is defined as

$$
J^\pi(s_0) \triangleq \limsup_{T \to \infty} \tfrac{1}{T} \mathbb{E}\left[\sum_{t=1}^T \Delta_t^\pi \middle| s_0\right]
$$

while the corresponding average number of transmissions is given by

$$
C^\pi(s_0) \triangleq \limsup_{T \to \infty} \tfrac{1}{T} \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}[a_t^\pi \neq \mathrm{i}] \middle| s_0\right] .
$$

We are interested in minimizing $J^\pi(s_0)$ given a constraint λ on the average number of transmissions $C^\pi(s_0)$, leading to the following CMDP optimization problem:

Problem 1. $\underset{\pi \in \Pi}{\text{Minimize}}$ $J^\pi(s_0)$ such that $C^\pi(s_0) \leq \lambda$.

It is possible to introduce some practical limitations on the set of policies that ensure that for any policy, the resulting Markov chain is *unichain* (details are given in the full version of the paper [15]). In the rest of the paper we will assume that such assumptions are in place, and hence the resulting CMDP is unichain. This implies that the process converges to a stationary distribution for any policy π, and the initial state s_0 has no effect on the asymptotic quantities $J^{\pi}(s_0)$ and $C^{\pi}(s_0)$. Therefore, s_0 will be omitted from the notation. A policy $\pi^* \in \Pi$ is called optimal if $J^* \triangleq J^{\pi^*} \leq J^{\pi}$ for all $\pi \in \Pi$.

III. Primal-Dual Algorithm to Minimize AoI

To solve the constrained MDP, we start by rewriting Problem 1 in its Lagrangian form. The average Lagrangian cost of a policy π with Lagrange multiplier $\eta \geq 0$, denoted by J^{π}_{η}, is defined as

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\Delta_t^{\pi}\right] - \eta\left(C_{max} - \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}[a_t^{\pi} \neq \mathrm{i}]\right]\right)$$

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

$$h_{\eta}(s) + J^*_{\eta} = \min_{a\in\mathcal{A}}\left(\Delta + \eta \cdot \mathbb{1}[a \neq \mathrm{i}] + \mathbb{E}\left[h_{\eta}(s')\right]\right) \quad (2)$$

for all $s \in \mathcal{S}$, where $s' \in \mathcal{S}$ is the next state obtained from s after taking action a. The optimal policy, for any s, is given by the action achieving the minimum in (2):

$$\pi^*_{\eta}(s) \in \arg\min_{a\in\mathcal{A}}\left(\Delta + \eta \cdot \mathbb{1}[a \neq \mathrm{i}] + \mathbb{E}\left[h_{\eta}(s')\right]\right). \quad (3)$$

We can solve (2) for any given η; and hence, find the optimal policy π^*_{η} using the relative value iteration (RVI) algorithm [16].

It is possible to characterize optimal policies for our CMDP problem using the deterministic policies π^*_{η}.: Specializing Theorem 4.4 of [14] to Problem 1 (since it has a single global constraint), 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. This actually means that, denoting by C_{η} the average resource consumption under the optimal policy π^*_{η}, the achievable optimal performance can be obtained as the lower convex hull of the points $\{(C_{\eta}, J^*_{\eta})\}_{\eta > 0}$.

Since C_{η} and J^*_{η} are obviously monotone functions of η, an approximate solution of finding a point on the lower convex hull with average transmission cost λ is to find $\eta_1 < \eta_2$ such that $C_{\eta_1} \geq \lambda \geq C_{\eta_2}$ and then combine the corresponding optimal policies $\pi^*_{\eta_1}$ and $\pi^*_{\eta_2}$ with a weight μ satisfying $\lambda = \mu C_{\eta_1} + (1-\mu)C_{\eta_2}$. A heuristic method to find such η_1 and η_2 is as follows: starting with an initial parameter η^0, we run an iterative algorithm updating η as $\eta^{m+1} = \eta^m + \alpha(C_{\eta^m} - \lambda)$ for some step size parameter $\alpha \triangleq 1/\sqrt{m}$. We continue this iteration until $|C_{\eta^m} - \lambda|$ becomes smaller than a given threshold. Denoting the resulting value as η^*, if $C_{\eta^*} \neq \lambda$, we can increase or decrease the η^* value until η^* and its modification satisfy the conditions (note that in case of a finite state space, which is an approximation we always use in computing an optimal policy numerically, π_{η}, and consequently C_{η} and J^*_{η}, are piecewise constant functions of η, and so η must be changed sufficiently to change the average transmission cost).

IV. AoI with Standard ARQ Protocol

Now, assume that the system adopts the standard ARQ protocol; that is, failed transmissions are discarded at the destination. Then the state space reduces to $(\delta_1, \delta_2, \ldots, \delta_M)$ as $r_{j,t} = 0,\ \forall j, t$, and the action space to $\mathcal{A} = \{\mathrm{i}, \mathrm{n}_1, \ldots, \mathrm{n}_M\}$. The probability of error of each status update is $p_j \triangleq g_j(0)$ for user j. State transitions in (1), Bellman optimality equations and the RVI algorithm can all be simplified accordingly.

A. Lower Bound

Thanks to these simplifications, we can derive a closed-form lower bound for the constrained MDP (the proof is provided in the full version of the paper [15]):

Theorem 1. *For a given network setup, we have $J_{LB} \leq J^{\pi}$, $\forall \pi \in \Pi$, where*

$$J_{LB} = \frac{1}{2\lambda}\left(\sum_{j=1}^{M}\sqrt{\frac{w_j}{1-p_j}}\right)^2 + \frac{\lambda w_{j^*} p_{j^*}}{2(1-p_{j^*})} + \frac{1}{2}\sum_{j=1}^{M} w_j,$$

and $j^* \triangleq \arg\min_{j} \dfrac{w_j p_j}{2(1-p_j)}$.

Previously, [7] proposed a universal lower bound on the average AoI for the broadcast channel with multiple users for the special case of $\lambda = 1$. Differently from [7], the lower bound derived in this paper shows the effect of the constraint (λ) and even for $\lambda = 1$, it is tighter than the lower bound provided in [7].

V. Learning in an unknown environment

In most practical scenarios, channel error probabilities for retransmissions may not be known at the time of deployment, or may change over time. We employ online learning algorithms to learn the error probabilities over time without degrading the performance significantly.

The Upper Confidence RL (UCRL2) algorithm [17] is a well-known RL algorithm for generic MDP problems which has strong theoretical performance guarantees. However, the computational complexity of the algorithm scales quadratically with the size of the state space, which makes the algorithm unsuitable for large state spaces. UCRL2 has been initially proposed for generic MDPs with unknown rewards and transition probabilities: thus, they need to be learned for each state-action pair. On the other hand, for the average AoI problem, the number of parameters to be learned can be reduced to the number of transmission error probabilities to each user; thus, the computational complexity can be reduced significantly. In addition, the constrained structure of the average AoI problem requires additional modifications to the UCRL2, which is achieved in this paper by updating the Lagrange multiplier according to the empirical resource consumption.

A. UCRL2 with HARQ

The details of the algorithm are given in Algorithm 1. UCRL2 exploits the optimistic MDP characterized by the optimistic estimation of error probabilities within a certain confidence interval, where $\hat{g}_j(r)$ and $\tilde{g}_j(r)$ represent the empirical and the optimistic estimates of the error probability for user j, after r retransmissions. In each episode, we keep track of a value η resulting in a transmission cost close to λ, and then find and apply a policy that is optimal for the optimistic MDP (i.e., the MDP with the smallest total cost from among all plausible ones given the observations so far) with Lagrangian cost. In contrast to the original UCRL2 algorithm, finding the optimistic MDP in our case is easy (choosing lower estimates of the error probabilities), and we can use standard value iteration (VI) to compute the optimal policy (instead of the much more complex extended VI used in UCRL2). The resulting algorithm will be called as *UCRL2-VI*.

Algorithm 1 UCRL2-VI

```
Input: Confidence parameter $\delta \in (0,1)$, $U$, update parameter $\alpha$, $\lambda$.
 1: $\eta = 0$, $t = 1$ and observe the initial state $s_1$.
 2: for episodes $k = 1, 2, \dots$ do
 3:     Set $t_k \triangleq t$.
 4:     for $j \in \{0, \dots, M\}$, $r \in \{1, \dots, r_{max}\}$ do
 5:         $N_k(j,r) \triangleq \#\{t < t_k : a_t = \mathrm{x}_j, r_{j,t} = r\}$.
 6:         $N_k(j,0) \triangleq \#\{t < t_k : a_t = \mathrm{n}_j\}$.
 7:         $E_k(j,r) \triangleq \#\{\tau < t_k : a_\tau = \mathrm{x}_j, r_{j,t} = r, failure\}$.
 8:         $E_k(j,0) \triangleq \#\{\tau < t_k : a_\tau = \mathrm{n}_i, failure\}$.
 9:         $\widehat{g_j}(r) \triangleq \frac{E_k(j,r)}{\max\{N_k(j,r),1\}}$.
10:     end for
11:     $C_k \triangleq \#\{\tau < t_k : a_\tau \neq \mathrm{i}\}$.
12:     $\eta \leftarrow \eta + \alpha(C_k/t_k - \lambda)$.
13:     Compute the optimistic error probability estimates
14:             $\tilde{g}_j(r) \triangleq \max\left\{0, \hat{g}_j(r) - \sqrt{\frac{U\log(SAt_k/\delta)}{max\{1,N_k(j,r)\}}}\right\}$.
15:     Use $\tilde{g}_j(r)$ and VI to find a policy $\tilde{\pi}_k$.
16:     Set $v_k(j,r) \leftarrow 0$.
17:     while $v_k(j_t,r) < N_k(j_t,r)$ do                /* run policy $\tilde{\pi}_k$ */
18:         Choose an action $a_t = \tilde{\pi}_k(s_t)$.
19:         If $a_t \neq \mathrm{i}$, set $j_t$ the target user, otherwise set $j_t = 0$.
20:         Obtain cost $\sum_{j=1}^{M} w_j\delta_j + \eta\mathbb{1}[a_t \neq \mathrm{i}]$ and observe $s_{t+1}$.
21:         Update $v_k(j_t,r) = v_k(j_t,r) + 1$.
22:         Set $t = t + 1$.
23:     end while
24: end for
```

B. UCRL2 with standard ARQ

In this section, we consider the RL algorithms for the standard ARQ with unknown error probabilities $p_j = g_j(0)$, and UCRL2-VI in Algorithm 1 can be simplified accordingly.

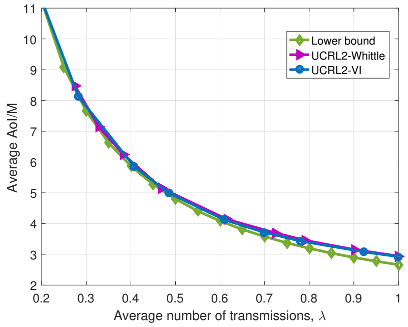

Figure 1. Average AoI with respect to λ for a 3-user network with $M = 3$ and error probabilities $p = [0.5\ 0.2\ 0.1]$, $w_j = 1$, $\forall j$. Time horizon is set to $T = 10^5$, and the results are averaged over 100 runs.

In order to reduce the computational complexity, we can replace the costly VI in the algorithm to find the $\tilde{\pi}_k$ with the suboptimal Whittle index policy proposed in [7]. The resulting algorithm, called *UCRL2-Whittle*, selects policy $\tilde{\pi}_k$ in step 15 as follows:

- Compute the index for each user (similarly to [7]),

$$I_j \triangleq w_j(1 - \tilde{p}(j))\delta_j \left(\delta_j + \frac{1 + \tilde{p}(j)}{1 - \tilde{p}(j)}\right). \quad (4)$$

- Compare the highest index with the Lagrange parameter η: if η is smaller then the source transmits to the user with the highest index, otherwise the source idles.

VI. Numerical Results

First, we analyze the average AoI in a multi-user setting with standard ARQ protocols. The asymptotic average AoI as a function of the resource constraint λ is shown in Figure 1 for a 3-user system with error probabilities $p = g(0) = [0.5\ 0.2\ 0.1]$. It can be seen from Figure 1 that both UCRL2-VI and UCRL2-Whittle perform very close to the lower bound, particularly when λ is small, i.e., the system is more constrained. Although UCRL2-Whittle has a significantly lower computational complexity, it performs very close to UCRL2-VI for all λ values.

Figure 2 illustrates the average AoI with standard ARQ with respect to the size of a network when there is no constraint on the average number of transmissions (i.e. $\lambda = 1$) and the performance of the UCRL2-Whittle is compared with the lower bound (UCRL2-VI is omitted since its performance is very similar to UCLR2-Whittle and has a much higher computational complexity, especially for large M). The performance of UCRL2-Whittle is close to the lower bound and very similar to that of the Whittle index policy [7] which requires the a priori knowledge of the error probabilities. Moreover, our algorithm outperforms the greedy benchmark policy which always transmits to the user with the highest age

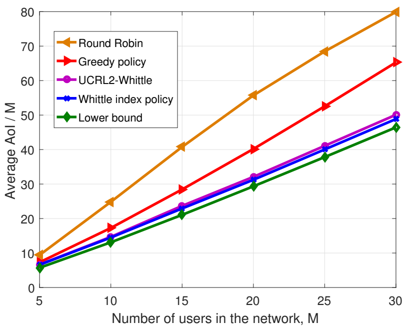

Figure 2. Average AoI for networks with different sizes where $p_j = j/M$, $\lambda = 1$ and $w_j = 1,\ \forall j$.The results are averaged over 100 runs.

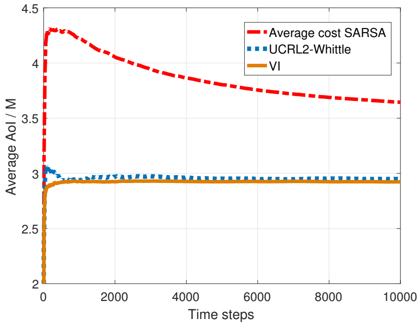

Figure 3. Average AoI for a 3-user ARQ network ($M = 3$) and error probabilities $p = [0.5\ 0.2\ 0.1]$ with $\lambda = 1$ and $w_j = 1,\ \forall j$. The simulation results are averaged over 100 runs.

and the Round Robin policy which transmits to each user in turns.

The performance of UCRL2-Whittle is compared with the performance of average cost SARSA, proposed in [12], in Figure 3. UCRL2-Whittle converges much faster compared to the standard average-cost SARSA, and it performs very close to the optimal algorithm computed by VI with known error probabilities. Figure 4 shows the performance of the learning algorithms for the HARQ protocol for a 2-user scenario. It is worth noting that although UCRL2-VI converges to the optimal policy in fewer iterations than average-cost SARSA, iterations in UCRL2-VI are computationally more demanding since the algorithm uses VI in each epoch. Therefore, UCRL2-VI is not practical for problems with large state spaces, in our case for large M. On the other hand, UCRL-Whittle can handle a large number of users since it is based on a simple index policy instead of VI.

VII. Conclusion

We considered scheduling the transmission of status updates to multiple destination nodes with the average AoI as the performance measure. Under a resource constraint, the problem is modeled as a CMDP considering both the ARQ and the HARQ. A lower bound on the average AoI was derived for the standard ARQ protocol. RL algorithms were presented for scenarios where the error probabilities may not be known in advance, and were numerically shown to provide near optimal performance in simple scenarios.

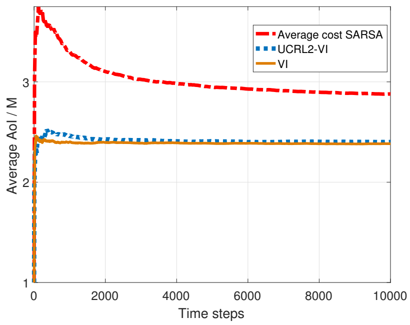

Figure 4. Average AoI for networks for a 2-user HARQ network with $M = 2$ and error probabilities $g_1(r_1) = 0.5 \cdot 2^{r_1}$ and $g_2(r_2) = 0.5 \cdot 2^{r_2}$ where $\lambda = 1$ and $w_j = 1,\ \forall j$. The simulation results are averaged over 100 runs.

References

[1] E. Altman, R. E. Azouzi, D. S. Menasché, and Y. Xu, "Forever young: Aging control in DTNs," *CoRR, 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.*, 2011.

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

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

[5] Y. Sun, E. Uysal-Biyikoglu, R. Yates, C. E. Koksal, and N. B. Shroff, "Update or wait: How to keep your data fresh," in *IEEE Int'l Conf. on Comp. Comms. (INFOCOM)*, April 2016, pp. 1–9.

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

[7] I. Kadota, E. Uysal-Biyikoglu, R. Singh, and E. Modiano, "Scheduling policies for minimizing age of information in broadcast wireless networks," *CoRR*, 2018.

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

[10] R. D. Yates and S. K. Kaul, "Status updates over unreliable multiaccess channels," in *IEEE Int'l Symp. on Inf. Theory (ISIT)*, June 2017, pp. 331–335.

[11] 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.

[12] E. T. Ceran, A. György, and D. Gündüz, "Average age of information with hybrid ARQ under a resource constraint," in *IEEE Wireless Comms. and Netw. Conf. (WCNC)*, April 2018.

[13] 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.

[14] E. Altman, *Constrained Markov Decision Processes*, ser. Stochastic modeling. Chapman & Hall/CRC, 1999.

[15] E. T. Ceran, A. György, and D. Gündüz, "A reinforcement learning approach to age of information in multi-user networks," *CoRR*, 2018.

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

[17] P. Auer, T. Jaksch, and R. Ortner, "Near-optimal regret bounds for reinforcement learning," in *Advances in Neural Inf. Processing Systems 21*. Curran Associates, Inc., 2009, pp. 89–96.