Reinforcement Learning to Minimize Age of Information with an Energy Harvesting Sensor with HARQ and Sensing Cost

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*—The time average expected age of information (AoI) is studied for status updates sent from an energy-harvesting transmitter with a finite-capacity battery. The optimal scheduling policy is first studied under different feedback mechanisms when the channel and energy harvesting statistics are known. For the case of unknown environments, an average-cost reinforcement learning algorithm is proposed that learns the system parameters and the status update policy in real time. The effectiveness of the proposed methods is verified through numerical results.**

I. Introduction

There has been a growing interest in minimizing the age of information (AoI) of energy harvesting (EH) communication systems [1]–[8]. The AoI quantifies the staleness of the information at the receiver, and is defined as the time elapsed since the generation time of the most recent status update successfully received at the receiver.

Prior works have investigated online [1], [3], [6] and offline [1], [4] methods for different scenarios in order to optimize the timeliness of information under the energy causality constraints in EH systems. It is shown in [3], [6], [8] that the optimal policy is of a threshold type for a finite-size battery when the cost of sensing (monitoring) the status of a process is not considered. Until recently, prior literature in the AoI framework assumed that the cost of sensing (monitoring) the status of a process is negligible compared to the cost of transmitting the status update. However, in most practical sensing systems acquiring a new sample of the underlying process of interest also has a considerable energy cost. The sampling/sensing cost has been taken into account in [9] and [10], where a status update system with ARQ and an unlimited energy source is considered. In [9], closed form expressions are presented for the energy consumption and average AoI.

In this paper, similarly to [9], we study a status update system considering both the sensing and transmission energy costs. We consider an EH transmitter, which uses the energy harvested from the environment to power the sensing and communication operations. Moreover, we consider a hybrid automatic repeat request (HARQ) protocol, where the partial information obtained from previous unsuccessful transmission attempts is combined to increase the decoding probability.

In our previous work, we studied status-update systems under a transmission-rate constraint [11]–[13]. Here we consider the intermittent availability of energy and find the online status updating policy to minimize the average AoI at the receiver, subject to the energy causality constraints at the transmitter. However, in many practical scenarios the statistical information about either the energy arrival process or the channel conditions are not available or may change over time [14]. Previous work on EH communication systems without a-priori information on random processes governing the system exploited reinforcement learning (RL) methods in order to maximize throughput or minimize delay [15], [16].

To adapt the status-update scheme to the unknown energy arrival process and channel statistics, we propose a learning theoretic approach using RL algorithms. In particular, we consider a value-based RL algorithm, *GR-learning* [17], and a policy-based RL algorithm, *finite-difference policy gradient (PG)* [18], and compare their performances with the relative value iteration (RVI) algorithm which assumes a-priori knowledge on the system characteristics. We propose a suboptimal threshold policy and demonstrate that policy gradient algorithm exploiting the structural characteristics of a threshold policy outperforms GR-learning algorithm. We investigate the effects of the EH process on the average AoI, and we show by simulations that temporal correlations in EH increase the average AoI significantly. We compare the average AoI with EH with the average AoI under an average transmission constraint [11] and demonstrate that the performance of EH transmitter approximates to the one with average transmission constraint for a battery with unlimited capacity and zero sampling/sensing cost.

II. System Model

We consider a time-slotted status update system over an error-prone wireless communication link (see Figure 1). The transmitter (TX) can sense the underlying time-varying process and generate a status update at each time slot at a certain energy cost. Status updates are communicated to the receiver (RX) over a time-varying

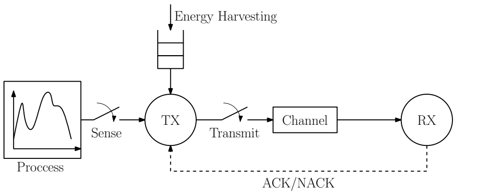


Figure 1: An EH status update system over an error-prone link in the presence of ACK/NACK feedback.

wireless channel. Each transmission attempt of a status update takes constant time, which is assumed to be equal to the duration of one time slot.

The AoI measures the timeliness of the status information at the receiver, and is defined at any time slot t as the number of time slots elapsed since the generation time $U(t)$ of the most up-to-date packet successfully decoded at the receiver. Formally, the AoI at the receiver at time t is defined as $\Delta_t^{rx} \triangleq \min(t - U(t), \Delta_{max})$, where a maximum value Δ_{max} on the AoI is imposed to limit the impact of the AoI on the performance after some level of staleness is reached.

We assume that the channel changes randomly from one time slot to the next in an independent and identically distributed (i.i.d.) fashion, and the instantaneous channel state information is available only at the receiver. We further assume the availability of an error- and delay-free feedback from the receiver to the transmitter for each transmission attempt. Successful reception of the status update at the end of time slot t is acknowledged by an ACK signal (denoted by $K_t = 1$), while a NACK signal is sent in case of a failure ($K_t = 0$).

There are three possible actions A_t the transmitter can take at each time slot t: it can either sample and transmit a new status update ($A_t = \mathrm{n}$), remain idle ($A_t = \mathrm{i}$) or retransmit the last transmitted status update ($A_t = \mathrm{x}$). If an ACK is received at the transmitter, we can restrict the action space to $\{\mathrm{i}, \mathrm{n}\}$ as retransmitting an already decoded status update is strictly suboptimal.

We consider the HARQ protocol: that is, the received signals from previous transmission attempts for the same packet are combined for decoding. The probability of error using r retransmissions, denoted by $g(r) < 1$, depends on r and the particular HARQ scheme used for combining multiple transmission attempts (an empirical method to estimate $g(r)$ is presented in [19]). As in any reasonable HARQ strategy, we assume that $g(r)$ is non-increasing in the number of retransmissions r. Standard HARQ methods only combine information from a finite maximum number of retransmissions [20]. Accordingly, we consider a truncated retransmission count of a status update, denoted by R_t for the status update transmitted at time t, where $R_t \in \{0, \ldots, R_{max}\}$; that is, the receiver can combine information from the last R_{max} retransmissions at most. We also assume that $R_0 = 0$ so that there is no previously transmitted packet at $t = 0$.

At the end of each time slot t, a random amount of energy is harvested and stored in a rechargeable battery at the transmitter, denoted by $E_t \in \mathcal{E} \triangleq \{0, 1, \ldots, E_{max}\}$, following a first-order discrete-time Markov model, characterized by stationary probabilities $p_E(e_1|e_2)$, defined as $p_E(e_1|e_2) \triangleq Pr(E_{t+1} = e_2|E_t = e_1)$, $\forall t$. It is also assumed that $p_E(0|e) > 0$, $\forall e \in \mathcal{E}$. Harvested energy is first stored in a rechargeable battery with a limited capacity of B_{max} energy units and the energy harvested when the battery is full is lost. The energy consumption for status sensing is denoted by $E^s \in \mathbb{Z}^+$, while the energy consumption for a transmission attempt is denoted by $E^{tx} \in \mathbb{Z}^+$.

The battery state at time t, denoted by B_t, and the energy causality constraints can be written as follows:

$$B_{t+1} = \min(B_t + E_t - E^s \mathbb{1}_{A_t = \mathrm{n}} - E^{tx} \mathbb{1}_{A_t \in \{\mathrm{x},\mathrm{n}\}}, B_{max}), \tag{1}$$

$$B_t \geq E^s \mathbb{1}_{A_t = \mathrm{n}} - E^{tx} \mathbb{1}_{A_t \in \{\mathrm{x},\mathrm{n}\}}, \tag{2}$$

where the indicator function $\mathbb{1}_C$ is equal to 1 if event C holds, and zero otherwise. Eqn. (1) implies that the battery overflows if energy is harvested when the battery is full, while Eqn. (2) imposes that sensing or transmission operations are not allowed if enough energy is not available in the battery.

The age Δ_t^{tx} of the most recently generated status update at the transmitter at the beginning of time slot t resets to 1 if a new status update is generated at time slot $t - 1$, and increases up to Δ_{max} otherwise, i.e.,

$$\Delta_{t+1}^{tx} = \begin{cases} 1 & \text{if } A_t = \mathrm{n}; \\ \min(\Delta_t^{tx} + 1, \Delta_{max}) & \text{otherwise.} \end{cases}$$

The AoI of the most recent successfully decoded packet at the receiver at time t, Δ_t^{rx}, evolves as follows:

$$\Delta_{t+1}^{rx} = \begin{cases} \min(\Delta_t^{rx} + 1, \Delta_{max}) & \text{if } A_t = \mathrm{i} \text{ or } K_t = 0; \\ 1 & \text{if } A_t = \mathrm{n} \text{ and } K_t = 1; \\ \min(\Delta_t^{tx} + 1, \Delta_{max}) & \text{if } A_t = \mathrm{x} \text{ and } K_t = 1. \end{cases}$$

We note that Δ_t^{tx} refers to the number of time slots elapsed since the generation of the most recently sensed status update at the transmitter side, while Δ_t^{rx} denotes the AoI of the most recent status update at the receiver. The system model also implies that whenever a new status update packet is generated, the previous packet at the transmitter is dropped and cannot be retransmitted. The number of retransmissions is zero for a newly sensed and generated status update and increases up to R_{max} as we keep retransmitting the same packet.

$$R_{t+1} = \begin{cases} 0 & \text{if } K_t = 1; \\ 1 & \text{if } A_t = \mathrm{n} \text{ and } K_t = 0; \\ R_t & \text{if } A_t = \mathrm{i}; \\ \min(R_t + 1, R_{max}) & \text{if } A_t = \mathrm{x} \text{ and } K_t = 0. \end{cases}$$

The state of the system is formed by five components $S_t = (E_t, B_t, \Delta_t^{rx}, \Delta_t^{tx}, R_t)$. At each time slot, the transmitter knows the state of the system and the goal is to find a policy π which minimizes the expected average AoI at the receiver over an infinite time horizon:

$$J^* \triangleq \min_{\pi} \lim_{T\to\infty} \frac{1}{T+1}\mathbb{E}\left[\sum_{t=0}^{T} \Delta_t^{rx}\right] \tag{3}$$

subject to (1) and (2).

III. Markov Decision Process (MDP) and RVI

An average-cost finite-state MDP provides the necessary framework for modeling and solving the AoI minimization problem in (3). An MDP is defined by the quadruple $\big(\mathcal{S}, \mathcal{A}, \mathrm{P}, c\big)$ [21]: The finite set of states $(E_t, B_t, \Delta_t^{rx}, \Delta_t^{tx}, R_t)$ is $\mathcal{S} = \mathcal{E} \times \{0, \ldots, B_{max}\} \times \{1, \ldots, \Delta_{max}\}^2 \times \{0, \ldots, R_{max}\}$ and the finite set of actions $\mathcal{A} = \{\mathrm{i}, \mathrm{n}, \mathrm{x}\}$ are already defined. P refers to the transition probabilities, where $P(s'|s,a) = \Pr(S_{t+1} = s' \mid S_t = s, A_t = a)$ is the probability that action a in state s at time t will lead to state s' at time $t+1$, which is characterized by the EH statistics and channel error probabilities. The cost function $c : \mathcal{S} \times \mathcal{A} \to \mathbb{Z}$, is the AoI at the receiver, and is defined as $c(s,a) = \Delta_t^{rx}$ for any $s \in \mathcal{S}$, $a \in \mathcal{A}$, independent of the action a.

We note that there exists an optimal stationary deterministic policy, $\pi : \mathcal{S} \to \mathcal{A}$, for this problem[1] [21]. In particular, there exists a function $h(s)$, called the *differential cost function* for all $s = (e, b, \delta^{rx}, \delta^{tx}, r) \in \mathcal{S}$, satisfying the following *Bellman optimality equations* for the average-cost finite-state finite-action MDP [21]:

$$h(s) + J^* = \min_{a\in\{\mathrm{i},\mathrm{n},\mathrm{x}\}} \big(\delta^{rx} + \mathbb{E}\left[h(s')|a\right]\big), \tag{4}$$

where $s' \triangleq (e', b', \delta^{rx\prime}, \delta^{tx\prime}, r')$ is the next state obtained from s after taking action a, and J^* represents the optimal achievable average AoI under policy π^*. We also introduce the *state-action cost function*:

$$Q(s,a) \triangleq \delta^{rx} + \mathbb{E}\left[h(s')|a\right] \ . \tag{5}$$

Then an optimal policy, for any $(e, b, \delta^{rx}, \delta^{tx}, r) \in \mathcal{S}$, takes the action achieving the minimum in (5):

$$\pi^*(s) \in \operatorname*{arg\,min}_{a\in\{\mathrm{i},\mathrm{n},\mathrm{x}\}} \big(Q(s,a)\big) \ . \tag{6}$$

An optimal policy solving (4), (5) and (6) can be found by relative value iteration (RVI) for finite-state finite-action average-cost MDPs from Section 8.5.5 of [21].

IV. A Reinforcement Learning Approach

In this section, we assume that the transmitter does not know the system characteristics *a-priori*, and has to learn them. We employ two different online learning algorithms. First, we employ a value-based RL algorithm, GR-learning, which converges to an optimal policy; then, we consider a structured policy search algorithm, finite-difference PG, which does not necessarily find the optimal policy but performs very well in practice, as demonstrated through simulations in Section V. We also note that GR-learning learns from a single trajectory generated during learning steps while PG uses Monte-Carlo roll-outs for each policy update. Thus, GR-learning is more applicable to real-time systems.

A. GR-Learning with Softmax

The literature for average-cost RL is quite limited compared to discounted cost problems [22]. For the average AoI minimization problem in (3), we employ a modified version of the *GR-learning* algorithm proposed in [17] with *Boltzmann* (*softmax*) exploration. The resulting algorithm is called *GR-learning with softmax*.

Notice that, by only knowing $Q(s,a)$, one can find the optimal policy π^* using (6) without knowing the transition probabilities P characterized by $g(r)$ and p_E. Thus, *GR-learning* starts with an initial estimation of $Q_0(s,a)$ and finds the optimal policy by estimating state-action values in a recursive manner. In the n^{th} iteration, after taking action A_n, the transmitter observes the next state S_{n+1}, and the instantaneous cost value Δ_n^{rx}. Based on this, the estimate of $Q_{n+1}(s,a)$ is updated by a weighted average of the previous estimate $Q_n(s,a)$ and the estimated expected value of the current policy in the next state S_{n+1}. Moreover, we update the gain J_n at every time slot based on the empirical average of AoI.

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 Δ_n^{rx},
- updates its estimate of $Q(S_n, A_n)$ using the current estimate of J_n by

$$Q_{n+1}(S_n, A_n) \leftarrow Q_n(S_n, A_n) + \alpha(m(S_n, A_n, n)) \\ [\Delta_n^{rx} - J_n + Q_n(S_{n+1}, A_{n+1}) - Q_n(S_n, A_n)], \tag{7}$$

  where $\alpha(m(S_n, A_n, n))$ is the update parameter (learning rate) in the n^{th} iteration, and depends on the function $m(S_n, A_n, n)$, which is the number of times the state–action pair (S_n, A_n) was visited till the n^{th} iteration.
- updates its estimate of J_n based on the empirical average as follows:

$$J_{n+1} \leftarrow J_n + \beta(n)\left[\frac{nJ_n + \Delta_n^{rx}}{n+1} - J_n\right] \tag{8}$$

  where $\beta(n)$ is the learning rate in the n^{th} iteration.

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* (*softmax*) action selection method, which chooses each action randomly relative to its expected cost, is used in this paper as follows: $\pi(a|S_n) =$

[1] For Markov chains corresponding to every stationary policy, there is only one recurrent class as the state $(0, 0, \Delta_{max}, \Delta_{max}, 0)$ is reachable from all other states (e.g., every transmission is successful but no EH is harvested for a period of $\max(\Delta_{max}, B_{max})$ time slots) from Theorem 8.4.3 of [21].

$\frac{\exp(-Q(S_n,a)/\tau_n)}{\sum_{a'\in\mathcal{A}}\exp(-Q(S_n,a')/\tau_n)}$, where τ is called the temperature parameter and decays exponentially with decay parameter γ. High τ corresponds to more uniform action selection (exploration) whereas low τ is biased toward the best action (exploitation). According to Theorem 2 of [17], if α, β satisfy $\sum_{m=1}^{\infty}\alpha(m), \sum_{m=1}^{\infty}\beta(m) \to \infty$, $\sum_{m=1}^{\infty}\alpha^2(m), \sum_{m=1}^{\infty}\beta^2(m) < \infty$, $\lim_{x\to\infty}\frac{\beta(m)}{\alpha(m)} \to 0$, *GR*-Learning converges to an optimal policy.

B. Finite-Difference Policy Gradient

GR-learning in Section IV-A is a value-based tabular RL method, which learns the state-action value function for each state-action pair [22]. In practice, Δ_{max} can be large, which might slow down the convergence of GR-learning due to a large state-space. In this section, we are going to simplify the problem and obtain a structured possibly sub-optimal policy, which can be learned via the PG method [18]. We make two assumptions on the policy space in order to obtain a more efficient RL:

- We assume that a packet is retransmitted until it is successfully decoded, provided that there is enough energy in the battery.
- The solution to the simplified problem is threshold-type, that is,
$$A_t = \begin{cases} \text{i} & \text{if } \Delta_t^{rx} < \mathcal{T}(e,b,\delta^{tx},r) \\ \text{n} & \text{if } \Delta_t^{rx} \geq \mathcal{T}(e,b,\delta^{tx},r) \text{ and } r=0 \\ \text{x} & \text{if } \Delta_t^{rx} \geq \mathcal{T}(e,b,\delta^{tx},r) \text{ and } r\neq 0 \end{cases} \quad (9)$$
for some $\mathcal{T}(e,b,\delta^{tx},r)$.

Note that $A_t = \text{i}$ if $b < E^{tx}$ ($b < E^{tx} + E^s$) for $r > 1$ ($r = 1$); that is, $\mathcal{T}(e,b,\delta^{tx},r) = \Delta_{max} + 1$. This ensures that energy causality constraints in (2) hold. Other thresholds will be determined using PG.

In order to employ PG method, we approximate the policy by a parameterized smooth function with parameters $\theta(e,b,\delta^{tx},r)$, and convert the discrete policy search problem into estimating the optimal values of some continuous parameters, which can be numerically solved by stochastic approximation algorithms [23].

In particular, with a slight abuse of notation, we let $\pi_\theta(e,b,\delta^{rx},\delta^{tx},r)$ denote the probability of taking action $A_t = \text{n}$ ($A_t = \text{x}$) if $r = 0$ ($r \neq 0$), and consider the parameterized sigmoid function:

$$\pi_\theta(e,b,\delta^{rx},\delta^{tx},r) \triangleq \frac{1}{1+e^{-\frac{\delta^{rx}-\theta(e,b,\delta^{tx},r)}{\tau}}}. \quad (10)$$

We note that $\pi_\theta(e,b,\delta^{rx},\delta^{tx},r) \to \{0,1\}$ and $\theta(e,b,\delta^{tx},r) \to \mathcal{T}(e,b,\delta^{tx},r)$ as $\tau \to 0$. Therefore, in order to converge to a deterministic policy π, $\tau > 0$ can be taken as a sufficiently small constant, or can be decreased gradually to zero. The total number of parameters to be estimated is $|\mathcal{E}| \times B_{max} \times \Delta_{max} \times R_{max} + 1$ minus the parameters corresponding to $b < E^{tx}$ ($b < E^{tx} + E^s$) for $r > 0$ ($r = 0$) due to energy causality constraints as stated previously.

With a slight abuse of notation, we map the parameters $\theta(e,b,\delta^{tx},r)$ to a vector $\overline{\theta}$ of size $d \triangleq |\mathcal{E}| \times B_{max} \times \Delta_{max} \times R_{max} + 1$. Starting with some initial estimates of $\overline{\theta}_0$, the parameters can be updated in each iteration n using the gradients as follows:

$$\overline{\theta}_{n+1} = \overline{\theta}_n - \gamma(n)\ \partial J/\partial\overline{\theta}_n, \quad (11)$$

where the step size parameter $\gamma(n)$ is a positive decreasing sequence and satisfies the first two convergence properties given at the end of Section IV-A.

Computing the gradient of the average AoI directly is not possible; however, several methods exist in the literature to estimate the gradient [23]. In particular, we employ the finite-difference PG [18] method. In this method, the gradient is estimated by estimating J at slightly perturbed parameter values. First, a random perturbation vector D_n of size d is generated according to a predefined probability distribution, e.g., each component of D_n is an independent Bernoulli random variable with parameter $q \in (0,1)$. The thresholds are perturbed with a small amount $\sigma > 0$ in the directions defined by D_n to obtain $\overline{\theta}_n^{\pm}(e,b,\delta^{tx},r) \triangleq \overline{\theta}_n(e,b,\delta^{tx},r) \pm \sigma D_n$. Then, empirical estimates $\hat{J}^{\pm}$ of the average AoI corresponding to the perturbed parameters $\overline{\theta}_n^{\pm}$, obtained from Monte-Carlo rollouts, are used to estimate the gradient:

$$\partial J/\partial\overline{\theta}_n \approx (D_n^\intercal D_n)^{-1} D_n^\intercal (\hat{J}^+ - \hat{J}^-)/(2\sigma). \quad (12)$$

where $D_n^\intercal$ denotes the transpose of vector D_n.

V. Simulation Results

In this section, we provide numerical results for the proposed algorithms, and compare the achieved average AoI. Motivated by previous research on HARQ [19], [20], we assume that the decoding error reduces exponentially with the number of retransmissions, 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. The exact value of λ depends on the particular HARQ protocol and the channel model. Following the *IEEE 802.16* standard [20], R_{max} is set to 3, and λ and p_0 are set to 0.5. E^{tx} and E^s are both assumed to be constant and equal to 1 unit of energy unless otherwise stated. Δ_{max} is set to 40. We choose the exact step sizes for the learning algorithms by fine-tuning. In particular, we use step size parameters of $\alpha(m), \beta(m), \gamma(m) = y/(m+1)^z$, where $0.5 < z \leq 1$ and $y > 0$ (which satisfy the convergence conditions) and choose y and z such that the oscillations are low and the convergence rate is high.

A. Uncorrelated EH

We first investigate the average AoI with HARQ when the EH process, $E_t \in \mathcal{E} = \{0,1\}$, is i.i.d. over time with probability distribution $Pr(E_t = 1) = p_e$, $\forall t$. The RVI algorithm is employed, and the effects of the battery capacity B_{max}, energy consumption of sensing E^s, and p_e on the average AoI are shown in Figure 2.

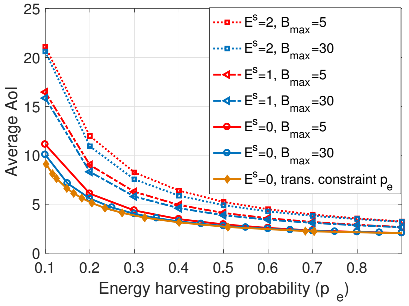


Figure 2: Average AoI for different B_{max}, E^s and p_e values when EH is i.i.d. and $E^{tx} = 1$.

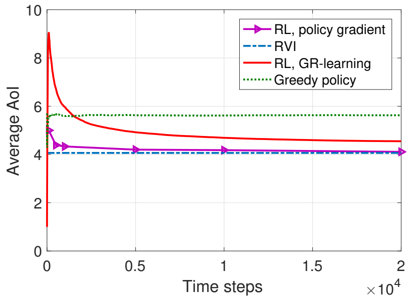
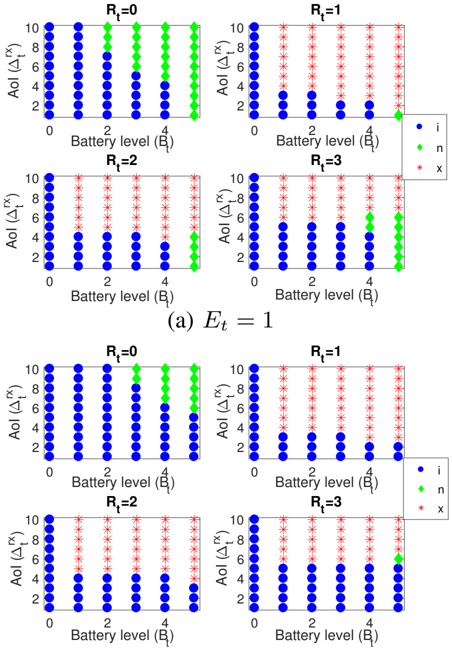


Figure 3: Performance of RL algorithms when $B_{max} = 5$, $E^s, E^{tx} = 1$, and $p_e = 0.5$.

As expected, the average AoI increases with decreasing B_{max}, decreasing p_e and increasing E^s. We note that, when $E^s\!=\!0$ and $B_{max}\!=\!\infty$, the problem defined in (3) corresponds to minimizing the average AoI under an average transmission rate constraint p_e, studied in [11], [13], which is also shown in Figure 2.

Figure 3 shows the evolution of the average AoI over time when the average-cost RL algorithms are employed. As a baseline, we have also included the performance of a greedy policy, which sends a new status update whenever there is sufficient energy for both sensing and transmission. It retransmits the last transmitted status update when the energy in the battery is sufficient only for transmission, and it remains idle otherwise; that is, $A_t\!=\!\text{n}$ if $B_t \geq E^{tx}\!+\!E^s$, $A_t\!=\!\text{x}$ if $E^{tx} \leq B_t < E^{tx}+E^s$ and $A_t = \text{i}$ if $B_t < E^{tx}$. It can be observed that the average AoI achieved by the proposed RL algorithms, converge to values close to the one obtained from the RVI algorithm, which has *a priori* knowledge of $g(r)$ and p_e, while the AoI of the greedy algorithm is significantly higher. Although the PG algorithm based on threshold policy does not allow preemption of an undecoded status update, it performs better than GR-learning since it tries to learn significantly smaller number of threshold values (i.e., $\Delta_{max}\!\times\!B_{max}\!\times\!R_{max}\!+\!1$) than GR-learning which learns one value for each state-action pair (i.e., $\Delta^2_{max}\!\times\!B_{max}\!\times\!(R_{max}\!+\!1)\!\times\!|\mathcal{A}|$).



(a) $E_t = 1$

(b) $E_t = 0$

Figure 4: Optimal policy for $B_{max} = 5$, $R_{max} = 3$, $p_E(1,1)$, $p_E(0,0)\!=\!0.7$, $E^s, E^{tx}\!=\!1$ and $\Delta^{tx}_t = R_t\!+\!1$.

B. Temporally Correlated EH

Next, we investigate the performance when the EH process has temporal correlations. A symmetric two-state Markovian EH process is assumed, such that $\mathcal{E} = \{0, 1\}$ and $Pr(E_{t+1} = 1|E_t = 0) = Pr(E_{t+1} = 0|E_t = 1) = 0.3$. That is, if the transmitter is in harvesting state, it is more likely to continue harvesting energy, and vice versa for the non-harvesting state.

Figure 4 illustrates the policy obtained by RVI. As it can be seen from the figure, the resulting policy is less likely to transmit if the battery level or the AoI is low. Moreover, the policy tends to retransmit the previous update rather than sensing a new update when the battery level is low and the AoI is high. When the system is in the non-harvesting state (i.e., $E_t = 0$), the transmitter is more conservative in transmitting the status updates compared to the case $E_t\!=\!1$, e.g., it might not transmit even if the battery is full depending on the AoI level.

Figure 5 shows the evolution of the average AoI over time when the average-cost RL algorithms are employed. It can be observed again that the average AoI achieved by the learned threshold parameters in Section IV-B, denoted by policy gradient in the figure, performs very close to the one obtained from the RVI algorithm, which has *a priori* knowledge of $g(r)$ and p_e. GR-learning, on the other hand, outperforms the greedy policy but converges to the optimal policy much more slowly, and the gap between the two RL algorithms is even longer compared to the i.i.d. case since the state space becomes larger with the addition of the EH state.

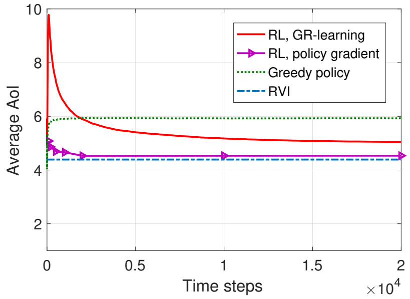


Figure 5: The performance of RL algorithms when $B_{max} = 5$, $p_E(1,1)$, $p_E(0,0) = 0.7$ and $E^s, E^{tx} = 1$.

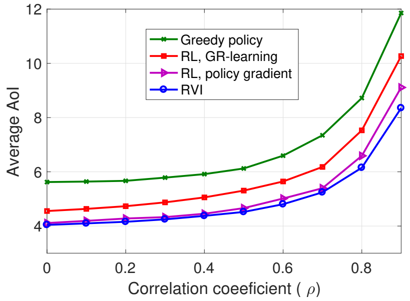


Figure 6: The performance of RL algorithms obtained after $2 \cdot 10^4$ time steps and averaged over 1000 runs for different temporal correlation coefficients.

Next, we investigate the impact of the burstiness of the EH process, measured by the correlation coefficient between E_t and E_{t+1}. Figure 6 illustrates the performance of the proposed RL algorithms for different correlation coefficients, which can be computed easily for the 2-state symmetric Markov chain; that is, $\rho \triangleq (2p_E(1,1) - 1)$. Note that $\rho = 0$ corresponds to the i.i.d. EH with $p_e = 1/2$. We note that the average AoI is minimized by transmitting new packets successfully at regular intervals, which has been well investigated in previous works [1], [2], [11]. Intuitively, for highly correlated EH, there are either successive transmissions or successive idle time slots, which increases the average AoI. Hence, the AoI is higher for higher values of ρ. Figure 6 also shows that both RL algorithms result in much lower average AoI than the greedy policy and policy gradient RL outperforms GR-learning since it benefits from the structural characteristics of a threshold policy.

VI. Conclusions

We have considered an EH system with a finite size battery and investigated scheduling policies transmitting time-sensitive data over a noisy channel with the average AoI as the performance measure. In addition to identifying a RVI solution for the optimal policy when the system characteristics are known, efficient RL algorithms are also presented for practical applications when the system characteristics are not known in advance. The algorithms studied in this paper are relevant to other systems concerning the timeliness of information or those powered by renewable energy sources.

References

[1] 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.

[2] R. D. Yates, “Lazy is timely: Status updates by an energy harvesting source,” in *IEEE Int’l Symposium on Information Theory (ISIT)*, 2015, pp. 3008–3012.

[3] B. T. Bacinoglu and E. Uysal-Biyikoglu, “Scheduling status updates to minimize age of information with an energy harvesting sensor,” *CoRR*, vol. abs/1701.08354, 2017.

[4] A. Arafa and S. Ulukus, “Age minimization in energy harvesting communications: Energy-controlled delays,” *CoRR*, vol. abs/1712.03945, 2017.

[5] X. Wu, J. Yang, and J. Wu, “Optimal status update for age of information minimization with an energy harvesting source,” *IEEE Tran. on Green Comms. and Netw.*, vol. 2, pp. 193–204, March 2018.

[6] B. T. Bacinoglu, Y. Sun, E. Uysal-Biyikoglu, and V. Mutlu, “Achieving the age-energy tradeoff with a finite-battery energy harvesting source,” *CoRR*, vol. abs/1802.04724, 2018.

[7] S. Feng and J. Yang, “Age of information minimization for an energy harvesting source with updating erasures: With and without feedback,” *CoRR*, 2018.

[8] A. Arafa, J. Yang, S. Ulukus, and H. V. Poor, “Age-minimal online policies for energy harvesting sensors with incremental battery recharges,” *Inf. Theory and Apps. Workshop*, Feb 2018.

[9] J. Gong, X. Chen, and X. Ma, “Energy-age tradeoff in status update communication systems with retransmission,” *CoRR*, vol. abs/1808.01720, 2018.

[10] B. Zhou and W. Saad, “Joint status sampling and updating for minimizing age of information in the internet of things,” *CoRR*, vol. abs/1807.04356, 2018.

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

[12] ——, “A reinforcement learning approach to age of information in multi-user networks,” in *IEEE Int’l Symp. on Personal, Indoor and Mobile Radio Comms. (PIMRC)*, Sep. 2018, pp. 1967–1971.

[13] E. T. Ceran, D. Gündüz, and A. György, “Average age of information with hybrid ARQ under a resource constraint,” *IEEE Transactions on Wireless Communications*, 2019.

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

[15] P. Blasco, D. Gunduz, and M. Dohler, “A learning theoretic approach to energy harvesting communication system optimization,” *IEEE Trans. on Wireless Comms.*, vol. 12, no. 4, pp. 1872–1882, April 2013.

[16] A. Ortiz, H. Al-Shatri, X. Li, T. Weber, and A. Klein, “Reinforcement learning for energy harvesting point-to-point communications,” in *IEEE Int’l Conf. on Comms. (ICC)*, 2016, pp. 1–6.

[17] A. Gosavi, “Reinforcement learning for long-run average cost,” *European Journal of Op. Res.*, vol. 155, pp. 654 – 674, 2004.

[18] J. Peters and S. Schaal, “Policy gradient methods for robotics,” in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2006, pp. 2219–2225.

[19] 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.

[20] “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.

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

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

[23] J. C. Spall, *Introduction to Stochastic Search and Optimization*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2003.