A Reinforcement Learning Framework for Optimizing Age of Information in RF-powered Communication Systems

Mohamed A. Abd-Elmagid, Harpreet S. Dhillon, and Nikolaos Pappas

***Abstract*—In this paper, we study a real-time monitoring system in which multiple source nodes are responsible for sending update packets to a common destination node in order to maintain the freshness of information at the destination. Since it may not always be feasible to replace or recharge batteries in all source nodes, we consider that the nodes are powered through *wireless energy transfer* (WET) by the destination. For this system setup, we investigate the optimal online sampling policy (referred to as the *age-optimal policy*) that jointly optimizes WET and scheduling of update packet transmissions with the objective of minimizing the long-term average weighted sum of Age of Information (AoI) values for different physical processes (observed by the source nodes) at the destination node, referred to as the *sum-AoI*. To solve this optimization problem, we first model this setup as an average cost Markov decision process (MDP) with finite state and action spaces. Due to the extreme curse of dimensionality in the state space of the formulated MDP, classical reinforcement learning algorithms are no longer applicable to our problem even for reasonable-scale settings. Motivated by this, we propose a deep reinforcement learning (DRL) algorithm that can learn the age-optimal policy in a computationally-efficient manner. We further characterize the structural properties of the age-optimal policy analytically, and demonstrate that it has a threshold-based structure with respect to the AoI values for different processes. We extend our analysis to characterize the structural properties of the policy that maximizes average throughput for our system setup, referred to as the *throughput-optimal policy*. Afterwards, we analytically demonstrate that the structures of the age-optimal and throughput-optimal policies are different. We also numerically demonstrate these structures as well as the impact of system design parameters on the optimal achievable average weighted sum-AoI.**

***Index Terms*—Age of Information, RF energy harvesting, Markov Decision Process, Reinforcement learning.**

I. Introduction

A typical real-time monitoring system consists of source and destination nodes, where source nodes observe underlying stochastic processes while the destination nodes keep track of the status of these processes through status updates transmitted (often wirelessly) by the source nodes. Examples of the source nodes include Internet of Things (IoT) devices, aggregators and sensors, while of the destination nodes include cellular base stations (BSs) [2]. The performance of many such real-time systems and applications depends upon how *fresh* the status updates are when they reach the destination nodes. In practice, the timely delivery of the measurements to the destination nodes is greatly restricted by the limited energy budget of the source nodes and the pathloss of the wireless channel between the source and destination nodes. Specifically, this could result in the loss or out-of-order reception of the measurements at the destination nodes. Consequently, the staleness of information status at the destination nodes increases, which eventually degrades the performance of such real-time applications.

Since it is highly inefficient or even impractical to replace or recharge batteries in many source nodes, energy harvesting solutions have been considered to enable a self-perpetuating operation of communication networks by supplementing or even circumventing the use of replaceable batteries in the source nodes. Due to its ubiquity and cost efficient implementation, radio-frequency (RF) energy harvesting has quickly emerged as an appealing solution for charging low-power source nodes (especially the ones that are deployed at difficult-to-reach places) [3]. This necessitates designing efficient transmission policies for freshness-aware RF-powered communication systems, which is the main objective of this paper. Towards this objective, we use the concept of AoI to quantify the freshness of information at the destination nodes [4]. This raises the obvious question of optimally scheduling packet transmissions from these RF-powered source nodes with the objective of minimizing the average AoI at the destination nodes, subject to the energy causality constraints at the source nodes. To address this question, this paper makes the first attempt, to the best of our knowledge, to develop a reinforcement learning-based framework in which we: i) propose a computationally-efficient approach to characterize the age-optimal transmission policy numerically, ii) analytically derive the structural properties of the age-optimal policy, and iii) analytically characterize key differences in the structural properties of the age-optimal and throughout-optimal policies.

A. Related Work

First introduced in [4], AoI is a new metric that quantifies the freshness of information at a destination node due to the transmission of update packets by the source node. Formally, AoI is defined as the time passed since the latest successfully received update packet at the destination was generated at the source node. Under a simple queue-theoretic model in which randomly generated packets arrive at the source according to

M. A. Abd-Elmagid and H. S. Dhillon are with Wireless@VT, Department of ECE, Virginia Tech, Blacksburg, VA. Email: {maelaziz, hdhillon}@vt.edu. N. Pappas is with the Department of Science and Technology, Linköping University, SE-60174 Norrköping, Sweden. Email: nikolaos.pappas@liu.se. The support of the U.S. NSF (Grant CPS-1739642) is gratefully acknowledged. This paper was presented in part at the IEEE Globecom, 2019 [1].

a Poisson process and then are transmitted to the destination using a first-come-first-served (FCFS) discipline, the authors of [4] characterized the average AoI expression. Afterwards, a series of works [5]–[12] aimed at characterizing the average AoI and its variations (e.g., Peak Age-of-Information (PAoI) [8]–[10] and Value of Information of Update (VoIU) [11]) for adaptations of the queueing model studied in [4]. Another direction of research [13]–[33] focused on employing AoI as a performance metric for different communication systems that deal with time critical information while having limited resources, e.g., multi-server information-update systems [14], broadcast networks [15]–[17], multi-hop networks [18], cognitive networks [19], unmanned aerial vehicle (UAV)-assisted communication systems [20]–[22], IoT networks [2], [23], [24], ultra-reliable low-latency vehicular networks [25], multicast networks [26], decentralized random access schemes [32], and multi-state time-varying networks [33]. Particularly, the objective of this research direction was to characterize optimal policies that minimize average AoI, referred to as *age-optimal polices*, by applying different tools from optimization theory. Note that [13]–[33] did not consider energy harvesting as a powering source for the source nodes.

Different from [13]–[33], another line of research [34]–[48] focused on the class of problems in which the source node is powered by energy harvesting under various system settings. The objective of this line of research was to investigate age-optimal offline/online policies for update packet transmissions subject to the energy causality constraint at the source under various assumptions regarding the battery size, transmission time of update packets and channel modeling. Specifically, the infinite battery capacity case was studied in [34]–[37], [44] whereas [38]–[43], [45], [46] considered the case of finite battery capacity. Different from [36]–[41] where it was assumed that each update packet could be transmitted to the destination instantly subject to the energy causality constraint, [34], [43], [44] considered stochastic transmission time and [35], [45], [46] studied the non-zero fixed transmission time case. While [34]–[36], [38]–[42], [45] considered error-free channel models, i.e., every update packet transmission is successfully received at the destination, a noisy channel model was considered in [37], [43], [44], [46]. A common model of the energy harvesting process in [34]–[45] is an external point process (e.g., Poisson process) independent from all the system design parameters. In contrast, when the source node is powered by RF energy harvesting, as considered in this paper, the energy harvested at the source is a function of the temporal variation of the channel state information (CSI). This, in turn, means that the age-optimal polices studied in [34]–[44] are not directly applicable to this setting. In particular, one needs to incorporate CSI statistics in the process of decision-making, which adds another layer of complexity to the analysis of age-optimal policies for such settings.

Before going into more details about our contributions, it is instructive to note that the problem of age-optimal policy in wireless powered communications systems has been studied very recently in [47], [48] for a single source-destination pair model. However, neither of the policies proposed in [47], [48] took into account the evolution of the battery level at the source and the variation of CSI over time in the process of decision-making. It is also worth noting that [22], [46], [49]–[52] have recently applied reinforcement learning-based algorithms to characterize the age-optimal policy. However, none of these works applied a DRL-based algorithm to efficiently design freshness-aware RF-powered communication systems. Different from these, we consider a more general model in which multiple RF-powered source nodes are deployed to potentially sense different physical processes. For this setting, we provide a novel reinforcement learning framework in which we: 1) develop a DRL-based algorithm that characterizes the online age-optimal sampling policy while considering the dynamics of batteries, AoI values for different processes and CSI, and 2) analytically characterize key differences between the structures of the online age-optimal and throughput-optimal polices. More details on our contributions are provided next.

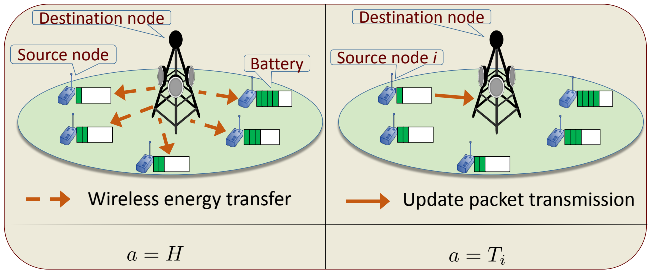

Fig. 1. An illustration of the system setup.

B. Contributions

This paper studies a real-time monitoring system in which multiple source nodes are supposed to keep the status of their observed physical processes fresh at a common destination node by transmitting update packets frequently over time. Furthermore, each source node is assumed to be powered by harvesting energy from RF signals broadcast by the destination node. For this setup, our main contributions are listed next.

A novel DRL algorithm for optimizing average weighted sum-AoI. Given an importance weight for each physical process at the destination node, we study the long-term average weighted sum-AoI (i.e., sum of AoI values for different processes at the destination node) minimization problem in which WET and scheduling of update packet transmissions from different source nodes are jointly optimized. To tackle this problem, we model it as an average cost MDP with finite state and action spaces. In particular, the MDP determines whether each time slot should be allocated for WET or an update packet transmission from one of the source nodes. This decision is based on the available energies at the source nodes (or their *battery levels*), the AoI values of different processes at the destination node, and the CSI. Due to the extreme curse of dimensionality in the state space of the formulated MDP, it is computationally infeasible to characterize the age-optimal policy using classical reinforcement learning algorithms [53], [54] such as relative value iteration algorithm (RVIA), value iteration algorithm (VIA) or policy iteration algorithm (PIA). To overcome this hurdle, we propose a novel DRL algorithm that can learn the age-optimal policy in a computationally-efficient manner.

Analytical characterization for the structural properties of the age-optimal policy. By analytically establishing the monotonicity property of the value function associated with the formulated MDP, we show that the age-optimal policy is a threshold-based policy with respect to each of the AoI values for different processes [1]. Moreover, for the single source-destination pair model (i.e., the case of having a single source node), our results demonstrate that the age-optimal policy is a threshold-based policy with respect to each of the system state variables, i.e., the battery level at the source, the AoI at the destination and the channel power gains. This result is of interest on its own because of the relevance of the source-destination pair model in plethora of applications, such as predicting and controlling forest fires, safety of an intelligent transportation system, and efficient energy utilization in future smart homes. Not surprisingly, this model has been of interest in a large proportion of the prior work on AoI. Furthermore, this result allows us to analytically demonstrate the key differences between the structures of the age-optimal and throughput optimal policies.

System design insights. Our results provide several useful system design insights. For instance, they show that the differences between the structures of the age-optimal and throughput-optimal policies in the single source-destination pair model mainly depend upon the AoI value of the observed process at the destination node. In particular, while the age-optimal and throughput-optimal policies have different structures when the AoI value is large, these differences start to vanish as the AoI value decreases. After showing the convergence of our proposed DRL algorithm, our numerical results also demonstrate the impact of system design parameters, such as the capacity of batteries and the size of update packets, on the achievable average weighted sum-AoI. Specifically, they reveal that the achievable average weighted sum-AoI by the DRL algorithm is monotonically decreasing (monotonically increasing) with the capacity of batteries (the size of update packets).

C. Organization

The rest of the paper is organized as follows. Section II presents our system model. The long-term weighted sum-AoI minimization problem is then formulated in Section III, where a DRL algorithm is proposed to obtain its solution. Afterwards, we present our analysis used to characterize the structural properties of the age-optimal policy in Section IV. Using the analytical results derived in Section IV, the key differences between the structural properties of the age-optimal and throughput-optimal policies in the single source-destination pair model are demonstrated in Section V. Section VI verifies our analytical findings from Sections IV and V as well as evaluates the performance of our proposed DRL algorithm numerically. Finally, Section VII concludes the paper.

II. System Model

A. Network Model

We study a real-time monitoring system in which a set $\mathcal{I}$ of N source nodes is deployed to observe potentially different physical processes, such as temperature or humidity. Each source node is supposed to keep the information status of its observed process at a destination node (for instance, a cellular BS) fresh by sending status update packets over time. In the context of IoT networks, the source node could refer to a single IoT device or an aggregator located near a group of IoT devices, which transmits update packets collected from them to the destination node. The destination node is assumed to have a stable energy source whereas each source node is equipped with an RF energy harvesting circuitry as its only source of energy. In particular, the source nodes harvest energy from the RF signals broadcast by the destination in the downlink such that the energy harvested at source node i is stored in a battery with finite capacity $B_{\max,i}$ Joules. The source and destination nodes are assumed to have a single antenna each and operate over the same frequency channel. Hence, at a given time instant, each source node cannot simultaneously harvest wireless energy in downlink and transmit data in uplink.

We consider a discrete time horizon composed of slots of unit length (without loss of generality) where slot $k = 0, 1, \ldots$ corresponds to the time duration $[k, k+1)$. Denote by $B_i(k)$ and $A_i(k)$ the amount of available energy at source node i and the AoI of its observed process i at the destination, respectively, at the beginning of time slot k. We assume that $A_i(k)$ is upper bounded by a finite value $A_{\max,i}$ which can be chosen to be arbitrarily large, i.e., $A_i(k) \in \{1, 2, \cdots, A_{\max,i}\}$. When $A_i(k)$ reaches $A_{\max,i}$, it means that the available information at the destination nodes about process i is too stale to be of any use. In addition, this assumption makes the AoI variable of each process only take finite number of values, i.e., the AoI state space of each process is finite. This will facilitate the solution of MDP, as will be clarified in the next section. Let $g_i(k)$ and $h_i(k)$ denote the downlink and uplink channel power gains between the destination and source node i over slot k, respectively. The downlink and uplink channels are assumed to be affected by quasi-static flat fading, i.e., they remain constant over a time slot but change independently from one slot to another. The locations of the source nodes are known *a priori*, and hence their average channel power gains are pre-estimated and known at the destination node. In particular, at the beginning of an arbitrary time slot, the destination node has perfect knowledge about the channel power gains in that slot, and only a statistical knowledge for future slots. This is a very reasonable assumption for many IoT applications.

B. State and Action Spaces

At the beginning of an arbitrary time slot k, the state $s_i(k)$ of a source node i is characterized by its battery level, the

[1] Note that constructing a threshold-based optimal policy under the analytical framework of MDPs is common in other research areas (such as power control and distributed detection) as well. However, the novelty of our MDP formulation lies in the use of the newly emerging concept of AoI in the objective function to quantify freshness of information, which has not been done in the other research areas. This process of decision-making is performed while accounting for various system design parameters (i.e., the battery levels, the AoI values at the destination node, and the CSI) as system state variables.

AoI of its observed process i at the destination, and its uplink and downlink channel power gains from the destination node, i.e., $s_i(k) \triangleq (B_i(k), A_i(k), g_i(k), h_i(k)) \in \mathcal{S}_i^a$. Note that $\mathcal{S}_i^a$ is the state space which contains all the combinations of $B_i(k), A_i(k), g_i(k)$ and $h_i(k)$, where the superscript a indicates that it is defined for the average AoI minimization problem. The state of the system at slot k is then given by $s(k) = \{s_i(k)\}_{i\in\mathcal{I}} \in \mathcal{S}^a$, where $\mathcal{S}^a$ is the system state space. Based on $s(k)$, the action taken at slot k is given by $a(k) \in \mathcal{A} \triangleq \{H, T_1, T_2, \cdots, T_N\}$, as illustrated in Fig. 1. When $a(k) = H$, slot k is dedicated for WET where the destination broadcasts RF energy signal in the downlink to charge the batteries at the source nodes. Particularly, the amount of energy harvested by an arbitrary source node i can be expressed as

$$E_i^{\mathrm{H}}(k) = \eta P g_i(k), \tag{1}$$

where η is the efficiency of the energy harvesting circuitry and P is the average transmit power by the destination. We assume that P is sufficiently large such that the energy harvested at each source node due to uplink data transmissions by the other source nodes is negligible. On the other hand, when $a(k) = T_i$, slot k is allocated for information transmission where source i sends an update packet about its observed process to the destination. We consider a *generate-at-will policy* [13], where the source scheduled for transmission generates an update packet at the beginning of the time slot whenever that slot is allocated for information transmission. According to Shannon's formula, when the energy consumed by source i to transmit an update packet of size S in slot k is $E_i^{\mathrm{T}}(k)$, its maximum reliable transmission rate is $\log_2\left(1 + \frac{h_i(k)E_i^{\mathrm{T}}(k)}{\sigma^2}\right)$ bits/Hz (recall that the slot length is unity), where σ^2 is the noise power at the destination. Hence, the action T_i can only be decided if the battery level at source i satisfies the following condition

$$B_i(k) \geq E_i^{\mathrm{T}}(k) = \frac{\sigma^2}{h_i(k)}\left(2^{\bar{S}} - 1\right). \tag{2}$$

In every time slot, the battery level at each source node and the AoI values for different processes at the destination are updated based on the action decided. Specifically, if $a(k) = T_i$, then the battery level at source i decreases by $E_i^{\mathrm{T}}(k)$, and the AoI value of its observed process i becomes one (recall that a generate-at-will policy is employed); if $a(k) = H$, then the battery level at source i increases by $E_i^{\mathrm{H}}(k)$ and the AoI value of process i increases by one; otherwise, the battery level at source i does not change and the AoI value of process i increases by one. Hence, the evolution of the battery level at source i and the AoI value of its observed process at the destination node can be expressed, respectively, by

$$B_i(k+1) = \begin{cases} B_i(k) - E_i^{\mathrm{T}}(k), \text{ if } a(k) = T_i, \\ \min\left\{B_{\max,i}, B_i(k) + E_i^{\mathrm{H}}(k)\right\}, \text{ if } a(k) = H, \\ B_i(k), \text{ otherwise.} \end{cases} \tag{3}$$

$$A_i(k+1) = \begin{cases} 1, & \text{if } a(k) = T_i, \\ \min\left\{A_{\max,i}, A_i(k) + 1\right\}, & \text{otherwise.} \end{cases} \tag{4}$$

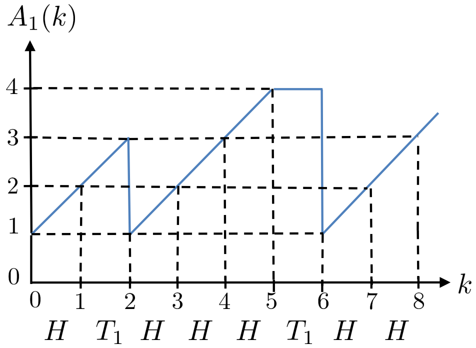

Fig. 2. AoI evolution vs. time when $N = 1$ and $A_{\max,1} = 4$.

To help visualize (4), Fig. 2 shows the AoI evolution for process 1 as a function of actions taken over time when $N = 1$ and $A_{\max,1} = 4$.

III. Problem Formulation and Proposed Solution

A. Problem Statement

Our objective is to obtain the optimal policy, which specifies the actions taken at different states of the system over time, achieving the minimum average weighted sum-AoI, i.e., sum of AoI values for different processes at the destination. Particularly, a policy $\pi = \{\pi_0, \pi_1, \cdots\}$ is a sequence of probability measures of actions over the state space. For instance, the probability measure π_k specifies the probability of taking action $a(k)$, conditioned on the sequence s^k which includes the past states and actions, and the current state, i.e., $s^k \triangleq \{s(0), a(0), \cdots, s(k-1), a(k-1), s(k)\}$. Formally, π_k specifies $\mathbb{P}(a(k) \mid s^k)$ such that $\sum_{a(k)\in\mathcal{A}(s(k))} \mathbb{P}(a(k) \mid s^k) = 1$, where $\mathcal{A}(s(k))$ is the set of possible actions at state $s(k) \in \mathcal{S}^a$. The policy π is said to be stationary when $\mathbb{P}(a(k) \mid s^k) = \mathbb{P}\left(a\left(k\right) \mid s\left(k\right)\right), \forall k$, and is called deterministic if $\mathbb{P}(a(k) \mid s^k) = 1$ for some $a(k) \in \mathcal{A}(s(k))$. Under a policy π, the long-term average AoI of process i at the destination starting from an initial state $s(0)$ can be expressed as

$$\bar{A}_i^{\pi} \triangleq \limsup_{K\to\infty} \frac{1}{K+1} \sum_{k=0}^{K} \mathbb{E}\left[A_i(k) \mid s(0)\right], \tag{5}$$

where the expectation is taken with respect to the channel conditions and the policy. Our goal is to find the optimal policy $\pi^\star$, referred to as the *age-optimal policy*, that minimizes the average weighted sum-AoI such that

$$\pi^\star = \arg\min_{\pi} \sum_{i\in\mathcal{I}} \theta_i \bar{A}_i^{\pi}, \tag{6}$$

where $\theta_i \geq 0$ and $\sum_{i=1}^{N} \theta_i = 1$. Here, θ_i is a weight accounting for the importance of process i at the destination node. Our intention behind using a weighted average cost function is to provide a generic problem formulation that can account for the potential differences between the observed physical processes by the source nodes in terms of the impact of the AoI value of each process on the optimal actions taken at the destination node. In particular, the weights can be chosen according to the importance of the AoI values of different processes at the destination node. For instance, if the destination node only cares about the AoI value of the process observed by source node i, then we can set $\theta_i = 1$ and $\theta_j = 0$ for all $j \neq i$. Clearly, the optimal strategy π^* in that case is to select whether each time slot is dedicated for WET $(a = H)$ or is allocated for an update packet transmission from source i $(a = T_i)$, depending upon the AoI value of process i, and the battery level and channel power gains at source i. Hence in this scenario, the achievable average AoI values for the other processes are given by $\bar{A}_j^{\pi^\star} = A_{\max,j}, \forall j \neq i$.

B. MDP Formulation

Due to the nature of evolution of the battery level at source i and the AoI value of process i at the destination (as described by (3) and (4), $\forall i \in \mathcal{I}$), and the independence of channel power gains over time, the problem can be modeled as an MDP. In particular, we denote by $b_i(k) \in \{0, 1, \cdots, b_{\max,i}\}$ the discrete battery level at source i at the beginning of slot k, where $b_{\max,i}$ represents the maximum amount of energy quanta that can be stored in the battery at source i such that each energy quantum contains $\frac{B_{\max,i}}{b_{\max,i}}$ Joules. In this case, the quantities $E_i^{\mathrm{T}}(k)$ and $E_i^{\mathrm{H}}(k)$ in (3) should be replaced by two integer variables expressed in terms of energy quanta. Therefore, by defining $e_i^{\mathrm{T}}(k) \triangleq \left\lceil \frac{b_{\max,i}}{B_{\max,i}} E_i^{\mathrm{T}}(k) \right\rceil$ and $e_i^{\mathrm{H}}(k) \triangleq \left\lfloor \frac{b_{\max,i}}{B_{\max,i}} E_i^{\mathrm{H}}(k) \right\rfloor$, the dynamics of the battery at source i for the discrete model can be expressed as

$$b_i(k+1) = \begin{cases} b_i(k) - e_i^{\mathrm{T}}(k), & \text{if } a(k) = T_i, \\ \min\left\{b_{\max,i}, b_i(k) + e_i^{\mathrm{H}}(k)\right\}, & \text{if } a(k) = H, \\ b_i(k), & \text{otherwise,} \end{cases} \tag{7}$$

where we used the ceiling and floor in the definitions of $e_i^{\mathrm{T}}(k)$ and $e_i^{\mathrm{H}}(k)$ to obtain a lower bound to the performance of the continuous system. Clearly, an upper bound to the performance of the continuous system can be obtained by reversing the use of the floor and ceiling in the definitions of $e_i^{\mathrm{T}}(k)$ and $e_i^{\mathrm{H}}(k)$. Similarly, if the channel power gains are modeled by continuous random variables, we divide their support into a finite number of intervals with the same probability according to the fading probability density function (PDF). In this sense, the problem is modeled as a *finite-state finite-action MDP* with state $s(k) \triangleq \{(b(k), A(k), g(k), h(k))\}_{i \in \mathcal{I}} \in \mathcal{S}_{\mathrm{d}}^{a}$ (the state space of the discrete model) and action $a(k) \in \mathcal{A}(s(k)) \subseteq \mathcal{A}$. Since there exists an optimal stationary deterministic policy for solving finite-state finite-action MDPs [53], we aim at investigating that age-optimal stationary deterministic policy in the sequel and omit the time index. Note that as the number of discrete levels for both batteries and channel power gains increase, the discrete model can be considered as a good approximation for the continuous one, but this comes at the expense of a high computational complexity to characterize $\pi^\star$.

Due to taking an action a, the transition probability of moving from state $s_i = (b_i, A_i, g_i, h_i)$ to state $s_i' = (b_i', A_i', g_i', h_i')$ at source node i is given by

$$\begin{aligned} \mathbb{P}\left(s_i' \mid s_i, a\right) &\triangleq \mathbb{P}\left(b_i', A_i', g_i', h_i' \mid b_i, A_i, g_i, h_i, a\right) \\ &\overset{(a)}{=} \mathbb{P}\left(b_i', A_i' \mid b_i, A_i, g_i, h_i, a\right) \mathbb{P}(g_i') \mathbb{P}(h_i') \\ &\overset{(b)}{=} \mathbb{P}\left(b_i' \mid b_i, g_i, h_i, a\right) \mathbb{P}\left(A_i' \mid A_i, a\right) \mathbb{P}(g_i') \mathbb{P}(h_i'), \end{aligned} \tag{8}$$

where step (a) follows from the independence of the channel power gains over time and from other random variables, where $\mathbb{P}(g_i')$ and $\mathbb{P}(h_i')$ denote the probability mass functions for the downlink and uplink channel power gains (after discretization if they were expressed originally by continuous random variables), respectively. Note that for the case of a Markovian fading channel model, the conditional probabilities $\mathbb{P}(g_i' \mid g_i)$ and $\mathbb{P}(h_i' \mid h_i)$ will replace $\mathbb{P}(g_i')$ and $\mathbb{P}(h_i')$, respectively. These conditional probabilities are determined according to the Markovian fading channel model considered in the problem. However, all our analytical results regarding the structures of the age-optimal and throughput-optimal policies (derived in Sections IV and V) will remain the same. Step (b) follows since given s_i and a, the next battery level b_i' and the value of AoI A_i' can be obtained deterministically, separately from each other. Specifically, b_i' only depends on the current battery level and channel power gains, i.e., (b_i, g_i, h_i), and A_i' only depends upon its current value A_i. Thus, from (4) and (7), b_i' and A_i' can be determined, respectively, as

$$\mathbb{P}(b_i' \mid b_i, g_i, h_i, a) = \begin{cases} \mathbb{1}\left(b_i' = b_i - e_i^{\mathrm{T}}\right), \text{ if } a = T_i, \\ \mathbb{1}\left(b_i' = \min\left\{b_{\max,i}, b_i + e_i^{\mathrm{H}}\right\}\right), \text{ if } a = H, \\ \mathbb{1}\left(b_i' = b_i\right), \text{ otherwise,} \end{cases} \tag{9}$$

$$\mathbb{P}(A_i' \mid A_i, a) = \begin{cases} \mathbb{1}\left(A_i' = 1\right), \text{ if } a = T_i, \\ \mathbb{1}\left(A_i' = \min\left\{A_{\max,i}, A_i + 1\right\}\right), \text{ otherwise,} \end{cases} \tag{10}$$

where $\mathbb{1}(\cdot)$ is the indicator function. Note that in the case of having an infinite state space for AoI (i.e., setting $A_{\max,i}$ to ∞), the term $\mathbb{1}\left(A_i' = \min\left\{A_{\max,i}, A_i + 1\right\}\right)$ in (10) reduces to $\mathbb{1}\left(A_i' = A_i + 1\right)$. The overall transition probability of moving from state $s = \{s_i\}_{i \in \mathcal{I}}$ to state $s' = \{s_i'\}_{i \in \mathcal{I}}$, after taking an action a, can then be expressed as

$$\mathbb{P}\left(s' \mid s, a\right) \overset{(a)}{=} \prod_{i \in \mathcal{I}} \mathbb{P}\left(s_i' \mid s_i, a\right), \tag{11}$$

where (a) follows from the fact that given action a, the state of each source node evolves separately from the other source nodes. The following Lemma characterizes the optimal policy $\pi^\star$ satisfying (6).

Lemma 1. *The optimal policy $\pi^\star$ can be evaluated by solving the following Bellman's equations for average cost MDPs [53]:*

$$\bar{A}^\star + V(s) = \min_{a \in \mathcal{A}(s)} Q(s, a), s \in \mathcal{S}_{\mathrm{d}}^{a}, \tag{12}$$

where $\bar{A}^\star$ is the achievable optimal average AoI under $\pi^\star$ which is independent of the initial state $s(0)$, $V(s)$ is the value function, and $Q(s,a)$ is the Q-function $\big($also referred to as the

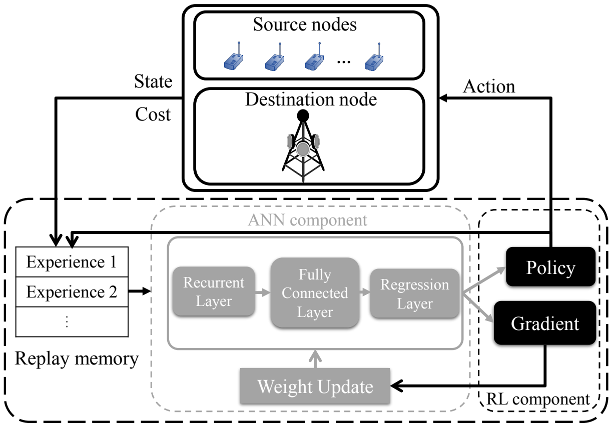

Fig. 3. The DRL architecture.

Q-factors, $\forall s \in \mathcal{S}_{\mathrm{d}}^{a}$ and $a \in \mathcal{A}(s)\big)$, which is the expected cost resulting from taking action a in state s, i.e.,

$$Q(s,a) = \sum_{i \in \mathcal{I}} \theta_i A_i + \sum_{s' \in \mathcal{S}_{\mathrm{d}}^{a}} \mathbb{P}(s' \mid s,a) V(s'), \qquad (13)$$

where $\mathbb{P}(s' \mid s,a)$ is evaluated using (11). In addition, the optimal action taken at state s is given by

$$\pi^{\star}(s) = \arg\min_{a \in \mathcal{A}(s)} Q(s,a). \qquad (14)$$

Since the weak accessibility condition holds for our problem, a solution for the Bellman's equations in Lemma 1 is guaranteed to exist [53]. Characterizing the optimal policy by solving Bellman's equations using classical reinforcement learning algorithms [53, Sec. 4.3 and Sec. 4.4], [54] (e.g., VIA, PIA or RVIA) requires to evaluate the policy improvement setup in (14) for each state at each iteration. Note that although all the environment parameters are known in our problem, the term “learning” refers to the process of learning the optimal policy in our case. Defining G_i and H_i as the number of discrete values that the state variables g_i and h_i can take, respectively, the number of states inside the state space $\mathcal{S}^a$ can then be computed as $|\mathcal{S}^a| = \prod_{i \in \mathcal{I}} \left(A_{\max,i} G_i H_i \left(b_{\max,i}+1\right)\right)$. Clearly, for a reasonable number of both the discrete values for each state variable (i.e., $A_{\max,i}, G_i, H_i$, and $b_{\max,i}+1$) and the source nodes deployed in the network (N), the state space will have a massive number of states. For instance, if we consider that each state variable can only take 10 values and there are three source nodes in the network, then the number of states becomes 10^{12}. As a result, it becomes computationally infeasible to obtain the optimal policy using classical reinforcement learning algorithms as the number of states increases (due to either increasing the number of discrete values for each state variable or the number of source nodes). This calls for investigating new approaches for characterizing the optimal policy in such large-scale setups. In order to overcome this problem, we propose a DRL algorithm to obtain the age-optimal policy numerically in the next subsection. We will also derive several key structural properties of the age-optimal policy analytically in Section IV.

C. Deep Reinforcement Learning for Optimizing AoI

DRL is suitable for our problem since it can reduce the dimensionality of the large state space while learning the optimal policy at the same time [55]. As shown in Fig. 3, the proposed DRL algorithm has two components: i) an artificial neural network (ANN), that reduces the dimension of the system state space by extracting its useful features, and ii) a reinforcement component, which is used to find the best policy based on the ANN's extracted features. Further, the ANN component has three layers: i) a recurrent layer consisting of long short term memory (LSTM) blocks [56], [57], ii) a fully connected (FC) layer in which the neurons have connections to all the outputs of the recurrent layer [57], and iii) a regression layer consisting of a single neuron whose output gives the Q value of the state-action pair input. The reason behind using a recurrent layer is its ability to store information for long periods of time, which allows it to learn long-term time correlations within a given sequence of inputs (i.e., it is useful for time series analysis) [56]–[58]. This perfectly complies with the nature of our problem in which we aim to extract useful features from actions and states of previous time slots such that the dimension of the system state space is implicitly reduced.

The reinforcement learning component is represented by the Q-learning algorithm [53], [54], [59]. As per the Q-learning algorithm, an update step for the Q-function value of the current state is performed at the beginning of each time slot, based on the action taken as well as the resulting next state. In particular, at the beginning of slot $k+1$, the update step of the Q-learning algorithm for our average cost MDP can be expressed as [53, Sec. 6.6.3]:

$$\begin{aligned} Q_{k+1}\left(s(k),a(k)\right) &= Q_k\left(s(k),a(k)\right) + \alpha(k)\Bigg(c(k) \\ &+ \min_{\bar{a} \in \mathcal{A}(s(k+1))} Q_k\left(s\left(k+1\right),\bar{a}\right) - \min_{\bar{a} \in \mathcal{A}(\bar{s})} Q_k\left(\bar{s},\bar{a}\right) \\ &- Q_k\left(s\left(k\right),a\left(k\right)\right)\Bigg), \end{aligned} \qquad (15)$$

where $c(k) = \sum_{i \in \mathcal{I}} \theta_i A_i(k)$ represents the resulting cost from taking action $a(k)$ in state $s(k)$ at slot k, $\alpha(k)$ is the learning rate at slot k, and $\bar{s}$ is the special state, which remains fixed over all the iterations and can be chosen arbitrarily. Note that (15) results from applying the Q-learning method to the relative value iteration of the Q-factors for average cost MDPs [53]. The sequence of values $\min_{\bar{a} \in \mathcal{A}(\bar{s})} Q_k\left(\bar{s},\bar{a}\right)$ is expected to converge to the optimal average AoI $\bar{A}^{\star}$ under the following conditions [59]: i) $\sum_{k=1}^{\infty} \alpha(k)$ is infinite and $\sum_{k=1}^{\infty} \left(\alpha(k)\right)^2$ is finite , ii) all potential state-action pairs are visited infinitely often, and iii) the state transition probability is stationary under the optimal stationary policy. By applying the update step in (15), the system can always *exploit* the learning process by taking the action which minimizes the long-term average cost, i.e., the action that minimizes the Q-function value of the current state. On the other hand, according to condition ii), the system has to *explore* all state-action pairs for the convergence of the algorithm. Thus, an ϵ-greedy policy has to be employed [55], where a random action is decided at the current state with probability $0 < \epsilon < 1$ with the objective of exploring the environment rather than exploiting the learning process. Meanwhile, the value of ϵ could be reduced to 0 as the learning goes in order to ensure that the learning process is exploited efficiently, i.e., not too much time is spent on exploring the environment.

Using the Q-learning algorithm (presented above) alone to characterize the optimal policy is efficient for cases where the system state space has a relatively small number of states. However, when the number of states is extremely large (which is the case in our problem), it becomes impractical to store the Q-function values for all state-action pairs (a massive memory is required for this) or even ensure that all state action-pairs will be visited so that the convergence can be achieved. Thus, as the cardinality of the (discrete) support set of each state variable and/or the number of source nodes increase in our problem, using Q-learning alone to characterize the optimal policy is not sufficient. In order to tackle this hurdle, we employ ANNs which are very effective at extracting features from data points and summarizing them in smaller dimensions. Specifically, a deep Q network approach [55] is used in which the learning steps are the same as in Q-learning, but the Q-function is approximated using an ANN $Q(s,a|\boldsymbol{\beta})$ (whose structure is constructed as explained above), where $\boldsymbol{\beta}$ is the vector containing the weights of the ANN. The objective is to find the optimal values for $\boldsymbol{\beta}$ such that the stored Q-function by the ANN becomes as close as possible to the optimal Q-function. To this end, we define a loss function for any combination of $(s(k),a(k),c(k),s(k+1))$, as follows:

$$\begin{aligned} L(\boldsymbol{\beta}_{k+1}) = \Bigg(&c(k) + \min_{\bar{a}\in\mathcal{A}(s(k+1))} Q_k\left(s\left(k+1\right),\bar{a}|\boldsymbol{\beta}_k\right) \\ &- \min_{\bar{a}\in\mathcal{A}(\bar{s})} Q_k\left(\bar{s},\bar{a}|\boldsymbol{\beta}_k\right) - Q_k\left(s(k),a(k)|\boldsymbol{\beta}_{k+1}\right)\Bigg)^2, \end{aligned} \tag{16}$$

where subscript $k+1$ is the time slot at which the weights are updated. Furthermore, a *replay memory* is used to save the evaluation of the state, action, and cost of past *experiences*, i.e., past state-action pairs and their resulting costs. In particular, after every time slot, we sample a random batch of a finite number of past experiences from the replay memory, and the gradient of the ANN's weights using this batch is evaluated as follows:

$$\begin{aligned} \nabla_{\boldsymbol{\beta}_{k+1}} L(\boldsymbol{\beta}_{k+1}) = \Bigg(&c(k) + \min_{\bar{a}\in\mathcal{A}(s(k+1))} Q_k\left(s\left(k+1\right),\bar{a}|\boldsymbol{\beta}_k\right) \\ &- \min_{\bar{a}\in\mathcal{A}(\bar{s})} Q_k\left(\bar{s},\bar{a}|\boldsymbol{\beta}_k\right) - Q_k\left(s(k),a(k)|\boldsymbol{\beta}_{k+1}\right)\Bigg) \\ &\times \nabla_{\boldsymbol{\beta}_{k+1}} Q_k(s(k),a(k)|\boldsymbol{\beta}_{k+1}). \end{aligned} \tag{17}$$

The weights of the ANN are then trained using this loss function. Note that it has been shown in [55] that using the batch method and replay memory improves the convergence of DRL. Algorithm 1 summarizes the steps of the proposed DRL algorithm.

So far, we have presented our proposed approach to obtain the optimal policy numerically. In the next section, we explore the structural properties of the age-optimal policy $\pi^\star$ analytically.

Algorithm 1 Deep reinforcement learning for average weighted sum-AoI minimization

```
Initialize a replay memory and an ANN $Q$ with a vector of weights $\boldsymbol{\beta}_0$.
Observe the initial state $s(0)$ and set $k = 0$.
Repeat:
  Select an action $a(k)$:
    select a random action $a(k) \in \mathcal{A}(s(k))$ with probability $\varepsilon$,
    otherwise select $a(k) = \arg\min_{\bar{a}} Q\left(s(k),\bar{a}|\boldsymbol{\beta}_k\right)$
  Perform action $a(k)$.
  Evaluate the cost $c(k)$ and observe the new state $s(k+1)$.
  Store experience $\{s(k),a(k),c(k),s(k+1)\}$ in the replay memory.
  Sample a random batch of experiences $\{\hat{s}(\zeta),\hat{a}(\zeta),\hat{c}(\zeta),\hat{s}(\zeta+1)\}$ from the replay memory.
  Calculate the set of target values $\{t(\zeta)\}$ corresponding to the experiences of the sampled
  batch:
    $t(\zeta) = \hat{c}(\zeta) + \min_{\bar{a}\in\mathcal{A}(\hat{s}(\zeta+1))} Q\left(\hat{s}\left(\zeta+1\right),\bar{a}|\boldsymbol{\beta}_k\right) - \min_{\bar{a}\in\mathcal{A}(\bar{s})} Q\left(\bar{s},\bar{a}|\boldsymbol{\beta}_k\right).$
  Train the network $Q$ using the gradient in (17).
  $k = k + 1$.
Until convergence to some value of average weighted sum-AoI.
```

IV. Structural Properties of the Age-optimal Policy

In this section, we derive the structural properties of the age-optimal policy $\pi^\star$ analytically using the VIA. Note that the obtained analytical results can be derived using the RVIA as well [53]. For completeness, we start this discussion by summarizing the VIA. According to the VIA, the value function $V(s)$ can be evaluated iteratively such that $V(s)$ at iteration m, $m = 1, 2, \cdots$, is computed as

$$\begin{aligned} V(s)^{(m)} &= \min_{a\in\mathcal{A}(s)} Q(s,a)^{(m-1)} \\ &= \min_{a\in\mathcal{A}(s)} \Bigg\{ \sum_{i\in\mathcal{I}} \theta_i A_i + \sum_{s'\in\mathcal{S}_{\mathrm{d}}^a} \mathbb{P}(s' \mid s,a) V(s')^{(m-1)} \Bigg\}, \end{aligned} \tag{18}$$

where $s \in \mathcal{S}_{\mathrm{d}}^a$. Hence, the optimal policy at iteration m is given by

$$\pi^{\star(m)}(s) = \arg\min_{a\in\mathcal{A}(s)} Q(s,a)^{(m-1)}. \tag{19}$$

As per the VIA, under any initialization of the value function $V(s)^{(0)}$, the sequence $\left\{V(s)^{(m)}\right\}$ converges to $V(s)$ which satisfies the Bellman's equation in (12), i.e.,

$$\lim_{m\to\infty} V(s)^{(m)} = V(s). \tag{20}$$

Based on the VIA, the following Lemma characterizes the monotonicity property of the value function with respect to the system state variables.

Lemma 2. *The value function $V(s)$, satisfying the Bellman's equation in (12) and corresponding to the age-optimal policy $\pi^\star$, is non-increasing with respect to the battery level b_j, the downlink channel power gain g_j and the uplink channel power gain $h_j, \forall j \in \mathcal{I}$. In contrast, $V(s)$ is non-decreasing with respect to the AoI $A_j, \forall j \in \mathcal{I}$.*

Proof: First, to prove that $V(s)$ is non-increasing with respect to b_j, let us define two states $s^1 = \left\{\left(b_i^1, A_i^1, g_i^1, h_i^1\right)\right\}_{i\in\mathcal{I}}$ and $s^2 = \left\{\left(b_i^2, A_i^2, g_i^2, h_i^2\right)\right\}_{i\in\mathcal{I}}$ where: i) $b_j^1 \leq b_j^2$, ii) $b_i^1 = b_i^2, \forall i \neq j$, and iii) $A_i^1 = A_i^2$, $g_i^1 = g_i^2$ and $h_i^1 = h_i^2, \forall i \in \mathcal{I}$. Hence, the objective is to show that $V(s^1) \geq V(s^2)$. According to (20), it is then sufficient to show that $V(s^1)^{(m)} \geq V(s^2)^{(m)}, \forall m$, which we prove using mathematical induction. Particularly, the relation holds by construction for $m = 0$ since it corresponds to the initial values for the value function which can be chosen arbitrarily. Now, we assume that $V(s^1)^{(m)} \geq V(s^2)^{(m)}$ holds for some m, and then show that it holds for $V(s^1)^{(m+1)} \geq V(s^2)^{(m+1)}$ as well. Let $C_1 = \sum\limits_{i\in\mathcal{I}} \theta_i A_i^2$ and $C_3 = \sum\limits_{i\in\mathcal{I}} \theta_i A_i^1$. According to (18) and (19), $V(s^2)^{(m+1)}$ and $V(s^1)^{(m+1)}$ can then be expressed, respectively, as

$$\begin{aligned} V(s^2)^{(m+1)} &= C_1 + \sum_{s^{2'} \in \mathcal{S}_{\mathrm{d}}^a} \mathbb{P}\left(s^{2'} \mid s^2, \pi^{\star(m)}\left(s^2\right)\right) V(s^{2'})^{(m)} \\ &\overset{(a)}{\leq} C_1 + \sum_{s^{2'} \in \mathcal{S}_{\mathrm{d}}^a} \mathbb{P}\left(s^{2'} \mid s^2, \pi^{\star(m)}\left(s^1\right)\right) V(s^{2'})^{(m)} \\ &\overset{(b)}{=} C_1 + C_0 \sum_{g_1'}\sum_{h_1'}\cdots\sum_{g_N'}\sum_{h_N'} V\left(\left\{b_i^{2'}, A_i', g_i', h_i'\right\}_{i\in\mathcal{I}}\right)^{(m)}, \end{aligned} \tag{21}$$

$$\begin{aligned} V(s^1)^{(m+1)} &= C_3 + \sum_{s^{1'} \in \mathcal{S}_{\mathrm{d}}^a} \mathbb{P}\left(s^{1'} \mid s^1, \pi^{\star(m)}\left(s^1\right)\right) V(s^{1'})^{(m)} \\ &= C_3 + C_0 \sum_{g_1'}\sum_{h_1'}\cdots\sum_{g_N'}\sum_{h_N'} V\left(\left\{b_i^{1'}, A_i', g_i', h_i'\right\}_{i\in\mathcal{I}}\right)^{(m)}, \end{aligned} \tag{22}$$

where $C_0 = \prod\limits_{i\in\mathcal{I}} \mathbb{P}(g_i')\mathbb{P}(h_i')$. Step (a) follows since it is not optimal to take action $\pi^{\star(m)}(s^1)$ in state s^2, and step (b) follows from (8)-(11) where, for a given $\pi^{\star(m)}(s^1)$, the set of values $\{A_i'\}_{i\in\mathcal{I}}$ can be evaluated based on (10), and the sets $\{b_i^{2'}\}_{i\in\mathcal{I}}$ and $\{b_i^{1'}\}_{i\in\mathcal{I}}$ are determined using (9). Note that since $b_i^1 = b_i^2, \forall i \neq j$, we have $b_i^{1'} = b_i^{2'}, \forall i \neq j$. On the other hand since $b_j^1 \leq b_j^2$, we can observe from (9) that $b_j^{1'} \leq b_j^{2'}$ for $\pi^{\star(m)}(s_1) \in \mathcal{A}$, and hence $V\left(\left\{b_i^{1'}, A_i', g_i', h_i'\right\}_{i\in\mathcal{I}}\right)^{(m)} \geq V\left(\left\{b_i^{2'}, A_i', g_i', h_i'\right\}_{i\in\mathcal{I}}\right)^{(m)}$. Therefore the expression in (21) is less than or equal to $V(s^2)^{(m+1)}$ which implies $V(s^1)^{(m+1)} \geq V(s^2)^{(m+1)}$ and indicates that the value function is non-increasing with respect to b_j. Note that increasing g_j (h_j) increases e_j^{H} (reduces e_j^{T}) which leads to a larger amount of energy in the battery at source j at the next time slot and hence a lower value function. This proves that $V(s)$ is non-increasing with respect to g_j and h_j, $\forall j \in \mathcal{I}$.

Next, using the same approach, we can show that $V(s)$ is non-decreasing with respect to A_j. Now, consider that the two states s^1 and s^2 are defined such that: i) $A_j^1 \geq A_j^2$, ii) $A_i^1 = A_i^2, \forall i \neq j$, and iii) $b_i^1 = b_i^2$, $g_i^1 = g_i^2$ and $h_i^1 = h_i^2, \forall i \in \mathcal{I}$. The goal is then to show that $V(s^1) \geq V(s^2)$. This can again be proven using mathematical induction by showing that $V(s^1)^{(m)} \geq V(s^2)^{(m)}, \forall m$. In particular, (21) and (22) can be rewritten for this case as

$$\begin{aligned} V(s^2)^{(m+1)} &\leq C_1 + \sum_{s^{2'} \in \mathcal{S}_{\mathrm{d}}^a} \mathbb{P}\left(s^{2'} \mid s^2, \pi^{\star(m)}\left(s^1\right)\right) V(s^{2'})^{(m)} \\ &= C_1 + \underbrace{C_0 \sum_{g_1'}\sum_{h_1'}\cdots\sum_{g_N'}\sum_{h_N'} V\left(\left\{b_i', A_i^{2'}, g_i', h_i'\right\}_{i\in\mathcal{I}}\right)^{(m)},}_{C_2} \end{aligned} \tag{23}$$

$$\begin{aligned} V(s^1)^{(m+1)} &= C_3 + \sum_{s^{1'} \in \mathcal{S}_{\mathrm{d}}^a} \mathbb{P}\left(s^{1'} \mid s^1, \pi^{\star(m)}\left(s^1\right)\right) V(s^{1'})^{(m)} \\ &= C_3 + \underbrace{C_0 \sum_{g_1'}\sum_{h_1'}\cdots\sum_{g_N'}\sum_{h_N'} V\left(\left\{b_i', A_i^{1'}, g_i', h_i'\right\}_{i\in\mathcal{I}}\right)^{(m)},}_{C_4} \end{aligned} \tag{24}$$

where $A_i^{2'} = A_i^{1'}, \forall i \neq j$ due to the fact that $A_i^1 = A_i^2, \forall i \neq j$. Note that we have $C_3 \geq C_1$ by construction since $A_j^1 \geq A_j^2$. It is then sufficient to show that $C_4 \geq C_2$ for all possible actions $\pi^{\star(m)}(s^1) \in \mathcal{A}(s^1)$. Specifically, there are two different cases: 1) $\pi^{\star(m)}(s^1) = T_j$, and 2) $\pi^{\star(m)}(s^1) \in \mathcal{A}(s^1)\backslash\{T_j\}$. Based on (10), we have $A_j^{1'} = A_j^{2'} = 1$ for the first case and hence $C_4 = C_2$. On the other hand, we have $A_j^{1'} \geq A_j^{2'}$ for the second case, which leads to $C_4 \geq C_2$. Consequently, $V(s^1)^{(m+1)} \geq V(s^2)^{(m+1)}, \forall \pi^{\star(m)}(s^1) \in \mathcal{A}(s^1)$ which proves that $V(s)$ is non-decreasing with respect to $A_j, \forall j \in \mathcal{I}$. ■

Based on Lemma 2, the following Theorem characterizes the structure of the age-optimal policy $\pi^\star$ with respect to the AoI values for different processes at the destination node.

Theorem 1. *Define two states $s^1 = \left\{\left(b_i^1, A_i^1, g_i^1, h_i^1\right)\right\}_{i\in\mathcal{I}}$ and $s^2 = \left\{\left(b_i^2, A_i^2, g_i^2, h_i^2\right)\right\}_{i\in\mathcal{I}}$ such that: i) $A_j^2 \geq A_j^1$, ii) $A_i^2 = A_i^1, \forall i \neq j$, and iii) $b_i^1 = b_i^2$, $g_i^1 = g_i^2$ and $h_i^1 = h_i^2, \forall i \in \mathcal{I}$. If $\pi^\star(s^1) = T_j$, then $\pi^\star(s^2) = T_j$.*

Proof: First, we observe that proving $\pi^\star(s^1) = \bar{a}$ implies $\pi^\star(s^2) = \bar{a}$ is equivalent to showing that

$$Q(s^2, \bar{a}) - Q(s^2, a') \leq Q(s^1, \bar{a}) - Q(s^1, a'), \forall a' \neq \bar{a}. \tag{25}$$

This is because if $\bar{a}$ is optimal in state s^1, then we have $Q(s^1, \bar{a}) - Q(s^1, a') \leq 0, \forall a' \neq \bar{a}$, which implies $Q(s^2, \bar{a}) \leq Q(s^2, a'), \forall a' \neq \bar{a}$, i.e., taking action $\bar{a}$ is optimal in state s_2. Hence, in order to complete the proof, we need to show that (25) holds for all possible choices of $a' \in \mathcal{A}(s^2) \setminus \{T_j\}$ when $\bar{a} = T_j$. To maintain generality, we consider the case where $\mathcal{A}(s^2) = \mathcal{A}$. Particularly, from (8)-(11) and (13), we have

$Q(s^n, a) =$

$$\sum_{i\in\mathcal{I}}\theta_i A_i^n + \underbrace{C_0\sum_{g_1'}\sum_{h_1'}\cdots\sum_{g_N'}\sum_{h_N'}V\left(\left\{b_i', A_i^{n'}, g_i', h_i'\right\}_{i\in\mathcal{I}}\right)}_{C(n,a)}, \tag{26}$$

where $n \in \{1,2\}$. According to (25), we first note that the term $\sum\limits_{i\in\mathcal{I}}\theta_i A_i^n$ is canceled out from all $Q(s^n, a)$, $n \in \{1,2\}$ and $a \in \{\bar{a}, a'\}$. When $a = T_j$, we have $A_j^{1'} = A_j^{2'} = 1$ from (10). This means $C(1, T_j)$ will equal $C(2, T_j)$ and (25) will hold if $C(2,a) \geq C(1,a), \forall a \in \mathcal{A}\setminus\{T_j\}$. For any $a \in \mathcal{A}\setminus\{T_j\}$, it follows that $A_j^{n'} = \min\left\{A_{\max,j}, A_j^n + 1\right\}$ from (10). Since $A_j^2 \geq A_j^1$ from i), we then have $A_j^{2'} \geq A_j^{1'}$. Now, based on Lemma 2 along with taking into account ii) and iii), it follows that $V\left(\left\{b_i', A_i^{2'}, g_i', h_i'\right\}_{i\in\mathcal{I}}\right) \geq V\left(\left\{b_i', A_i^{1'}, g_i', h_i'\right\}_{i\in\mathcal{I}}\right)$. Hence, we have $C(2,a) \geq C(1,a)$, which completes the proof. ■

Remark 1. *For the case of having multiple source nodes deployed in the network, i.e., $N > 1$, Theorem 1 indicates that the age-optimal policy $\pi^\star$ has a threshold-based structure with respect to each of the AoI state variables for different processes, i.e., $A_j, j \in \mathcal{I}$. For instance, for a fixed combination of state variables excluding A_j, if $A_{\mathrm{th},j}$ is the minimum AoI value of process j for which it is optimal to take an action $a = T_j$, then for all states with $A_j \geq A_{\mathrm{th},j}$, the optimal decision is T_j as well. This is also intuitive, since when the value of AoI for some process becomes large, it is optimal to update the status of information for that process at the destination by sending a new update packet.*

Note that by checking (25), one can show that $\pi^\star$ does not have a threshold-based structure with respect to the other system state variables, i.e., the levels of batteries and the channel power gains, for the case of $N > 1$. However, for the case of $N = 1$, the following Theorem provides more structural properties of the optimal policy $\pi^\star$ with respect to all system state variables.

Theorem 2. *Given $N = 1$, for any $s^1 = (b_1^1, A_1^1, g_1^1, h_1^1)$ and $s^2 = (b_1^2, A_1^2, g_1^2, h_1^2)$, the age-optimal policy $\pi^\star$ has the following structural properties:*
(i) When $s^1 \preceq s^2$ and $b_1^1 \geq \max\left\{b_{\max,1} - e_1^{\mathrm{H},1}, e_1^{\mathrm{T},1}\right\}$, if $\pi^\star(s^1) = T$, then $\pi^\star(s^2) = T$.
(ii) When $s^1 \succeq s^2$ and $b_1^2 \geq \max\left\{b_{\max,1} - e_1^{\mathrm{H},2}, e_1^{\mathrm{T},2}\right\}$, if $\pi^\star(s^1) = H$, then $\pi^\star(s^2) = H$.
Note that the symbols $\preceq$ and $\succeq$ represent the element-wise inequalities.

Proof: Since the action space becomes $\mathcal{A} \triangleq \{H, T_1\}$ for the case of $N = 1$, (i) is proven ((ii) is proven) if (25) holds for $\bar{a} = T_1$ and $a' = H$ ($\bar{a} = H$ and $a' = T_1$). Therefore, in the remaining, we focus on the proof of (i) while (ii) can be proven similarly. Particularly, from (8)-(10) and (13), we have

$$Q(s^n, T_1) = A_1^n + C_0\sum_{g_1'}\sum_{h_1'}V(b_1^n - e_1^{\mathrm{T},n}, 1, g_1', h_1'), \tag{27}$$

$$Q(s^n, H) = A_1^n + C_0\sum_{g_1'}\sum_{h_1'}V(b_{\max,1}, \min\{A_{\max,1}, A_1^n + 1\}, g_1', h_1'), \tag{28}$$

where $n \in \{1,2\}$ and the next battery level in (28) is equal to $b_{\max,1}$ since $b_1^1 + e_1^{\mathrm{H},1} \geq b_{\max,1}$ and $b_1^1 \leq b_1^2$. Since $s^1 \preceq s^2$ and based on Lemma 2, we have $V(b_1^1 - e_1^{\mathrm{T},1}, 1, g_1', h_1') \geq V(b_1^2 - e_1^{\mathrm{T},2}, 1, g_1', h_1')$ ($e_1^{\mathrm{T},1} \geq e_1^{\mathrm{T},2}$) and $V(b_{\max,1}, \min\{A_{\max,1}, A_1^2 + 1\}, g_1', h_1') \geq V(b_{\max,1}, \min\{A_{\max,1}, A_1^1 + 1\}, g_1', h_1')$. Hence, (25) holds for $\bar{a} = T_1$ and $a' = H$, which completes the proof of (i). ■

Remark 2. *Note that according to Theorem 2, the age-optimal policy $\pi^\star$ has a threshold-based structure over the set of states $\mathcal{S}_{\mathrm{d}}^{\mathrm{th},a} \triangleq \left\{s \in \mathcal{S}_{\mathrm{d}}^a : b_1 \geq \max\{b_{\max,1} - e_1^{\mathrm{H}}, e_1^{\mathrm{T}}\}\right\}$, for the case of $N = 1$. Particularly, $\pi^\star$ is a threshold-based policy with respect to each of the system state variables, i.e., b_1, A_1, g_1, and h_1. For instance, for a fixed (b_1, g_1, h_1), if $A_{\mathrm{th},1}$ is the minimum value of AoI for which it is optimal to take an action $a = T_1$, then for all states $s \in \mathcal{S}_{\mathrm{d}}^{\mathrm{th}}$ such that $A_1 \geq A_{\mathrm{th},1}$, the optimal decision is T_1 as well. In addition, if there exists a state $s^{\mathrm{th}} = (b_{\mathrm{th},1}, A_{\mathrm{th},1}, g_{\mathrm{th},1}, h_{\mathrm{th},1})$, where $b_{\mathrm{th},1}, g_{\mathrm{th},1}$, and $h_{\mathrm{th},1}$ are defined similar to $A_{\mathrm{th},1}$, then $\pi^\star(s) = T_1, s \in \mathcal{S}_{\mathrm{d}}^{\mathrm{th}}$, such that $s \succeq s^{\mathrm{th}}$.*

Based on Remark 2, the computational complexity of characterizing the age-optimal policy using standard classical reinforcement learning algorithms such as VIA or PIA can be significantly reduced. In particular, the threshold-based structure of the age-optimal policy with respect to the system state variables can be exploited to reduce the complexity of the policy improvement step. More specifically, the optimal actions at some states can now be directly determined based on the optimal actions taken at some other states (due to the threshold-based structure of the age-optimal policy), and hence the computational complexity of the policy improvement step can be greatly reduced. We refer the readers to [16], [24] for a detailed discussion on this matter. It is also worth noting that the case of $N = 1$ in our system setup refers to the classical single source-destination pair model studied in most prior works on AoI in the literature, e.g., [4], [6], [8]–[13]. Since the single source-destination pair model may actually be sufficient to study a diverse set of applications [4] (e.g., predicting and controlling forest fires, safety of an intelligent transportation system, and efficient energy utilization in future smart homes), the results obtained in Theorem 2 for $N = 1$ are of interest on their own in many applications. Furthermore, the results of Theorem 2 are very useful to investigate the differences between the structural properties of the age-optimal and throughput-optimal policies for the single source-destination pair model, as will be discussed in the next section.

V. Age-optimal Policy vs. Throughput-optimal Policy

In this section, we aim to analytically compare the structural properties of the age-optimal and the throughput-optimal policies. Due to its higher tractability (as demonstrated in the previous section), we will focus on the single source-destination pair model for this comparison. Specifically, we first formulate the average throughput maximization problem for the case of $N = 1$ in the system setup presented in Section II. Afterwards, we investigate some structural properties of the throughput-optimal policy from which we highlight the differences between the structures of the age-optimal and throughput-optimal polices.

A. Average Throughput Maximization Formulation and Proposed Solution

When the objective is to maximize the average throughput, the system state at slot k for the case of $N = 1$ is defined as $s(k) = \{b_1(k), g_1(k), h_1(k)\} \in \mathcal{S}_{\mathrm{d}}^{r}$, where $\mathcal{S}_{\mathrm{d}}^{r}$ is the state space of the discrete model for the throughput maximization problem, i.e., when the battery and channel power gain are discretized. Note that the AoI is not included now in the state of the system. For such single source-destination pair model, the action space is defined as $\mathcal{A} \triangleq \{H, T_1\}$, where the source node can either harvest energy or transmit a packet of size S at each time slot. The evolution of the battery is then given by (7). Hence, the average throughput maximization problem is modeled as a finite-state finite-action MDP for which there exists an optimal stationary deterministic policy [53]. Particularly, under a policy μ, the long-term average throughput is defined as

$$\bar{R}_1^{\mu} \triangleq \liminf_{K\to\infty} \frac{1}{K+1} \sum_{k=0}^{K} \mathbb{E}\left[\mathbb{1}\left(a(k) = T_1\right) S \mid s(0)\right], \quad (29)$$

where the system receives some reward equal to S in an arbitrary time slot only if this slot is allocated for data transmission to the destination node. This implies that the goal of this problem is to maximize the long-term average throughput resulting from transmitting update packets to the destination node. More specifically, we aim at characterizing the throughput-optimal policy μ^* such that

$$\mu^\star = \arg\max_{\mu} \bar{R}_1^{\mu}. \quad (30)$$

Under a stationary deterministic policy μ, the probability of moving from state s to state s' can be expressed as

$$\mathbb{P}\left(s' \mid s, \mu(s)\right) = \mathbb{P}\left(b_1' \mid b_1, g_1, h_1, \mu(s)\right) \mathbb{P}(g_1')\mathbb{P}(h_1'), \quad (31)$$

where $\mathbb{P}\left(b_1' \mid b_1, g_1, h_1, \mu(s)\right)$ can be expressed as in (9). The optimal policy $\mu^\star$ can then be obtained by solving the following Bellman's equation using the VIA (similar to (18) and (19))

$$\bar{R}^\star + V(s) = \max_{a\in\mathcal{A}(s)} Q(s,a), s \in \mathcal{S}_{\mathrm{d}}^{r}, \quad (32)$$

where $\bar{R}^\star$ is the optimal average throughput achievable by $\mu^\star$, and $Q(s,a)$ can be expressed (according to the definition of the Q-function in (13) and the expression of $\bar{R}_1^{\mu}$ in (29)) as

$$Q(s,a) = \mathbb{1}\left(a = T_1\right) S + \sum_{s'\in\mathcal{S}_{\mathrm{d}}^{r}} \mathbb{P}(s' \mid s,a) V(s'), \quad (33)$$

where $\mathbb{P}(s' \mid s,a)$ is computed by (31). Clearly, $Q(s,a)$ represents the expected reward resulting from taking action a in state s. In addition, $\mu^\star(s)$ is given by

$$\mu^\star(s) = \arg\max_{a\in\mathcal{A}(s)} Q(s,a). \quad (34)$$

B. Structural Properties of the Throughput-optimal Policy

Lemma 3. *The value function $V(b_1, g_1, h_1)$, corresponding to the throughput-optimal policy $\mu^\star$, is non-decreasing with respect to the battery level b_1, the downlink channel power gain g_1, and the uplink channel power gain h_1.*

Proof: By using (31), the result can be obtained using the same approach used in the proof of Lemma 2, i.e., by applying mathematical induction to the iterations of the VIA. ■

Using Lemma 3, some structural properties of the throughput-optimal policy are presented in the following Theorem.

Theorem 3. *For any $s^1 = (b_1^1, g_1^1, h_1^1)$ and $s^2 = (b_1^2, g_1^2, h_1^2)$, the throughput-optimal policy $\mu^\star$ has the following structural properties:*
(i) When $s^1 \preceq s^2$ and $b_1^1 \geq \max\left\{b_{\max,1} - e_1^{\mathrm{H},1}, e_1^{\mathrm{T},1}\right\}$, if $\mu^\star(s^1) = T_1$, then $\mu^\star(s^2) = T_1$.
(ii) When $s^1 \succeq s^2$ and $b_1^2 \geq \max\left\{b_{\max,1} - e_1^{\mathrm{H},2}, e_1^{\mathrm{T},2}\right\}$, if $\mu^\star(s^1) = H$, then $\mu^\star(s^2) = H$.

Proof: This result can be obtained using the same approach used in the proof of Theorem 2. Note that since this is a maximization problem, proving that $\mu^\star(s^1) = \bar{a}$ leads to $\mu^\star(s^2) = \bar{a}$ is now equivalent to showing that

$$Q(s^2,\bar{a}) - Q(s^2,a') \geq Q(s^1,\bar{a}) - Q(s^1,a'), \forall a' \neq \bar{a}. \quad (35)$$

■

Remark 3. *Similar to Remark 2, Theorem 3 shows that the throughput-optimal policy has a threshold-based structure over the set of states $\mathcal{S}_{\mathrm{d}}^{\mathrm{th},r} = \left\{s \in \mathcal{S}_{\mathrm{d}}^{r} : b_1 \geq \max\{b_{\max,1} - e_1^{\mathrm{H}}, e_1^{\mathrm{T}}\}\right\}$.*

Remark 4. *Our results in Theorems 2 and 3 clearly demonstrate that the structures of the age-optimal and throughput-optimal policies are different, which will also be verified in the numerical results section. Specifically, let us consider a state $\bar{s} = (\bar{b}_1, \bar{g}_1, \bar{h}_1) \in \mathcal{S}_{\mathrm{d}}^{\mathrm{th},r}$ such that $\mu^\star(\bar{s}) = T_1$. Note that the set of states $\bar{\mathcal{S}}_{\mathrm{d}}^{\mathrm{th},a} = \{(b_1, A_1, g_1, h_1) : (b_1, g_1, h_1) = \bar{s}, 1 \leq A_1 \leq A_{\max,1}\}$ belongs to $\mathcal{S}_{\mathrm{d}}^{\mathrm{th},a}$ since $s \in \mathcal{S}_{\mathrm{d}}^{\mathrm{th},r}$. Similar to the definition of $A_{\mathrm{th},1}$ in Remark 2, let us define $\bar{A}_{\mathrm{th},1} = \min\left(\left\{A_1 : \pi^\star(\bar{b}_1, A_1, \bar{g}_1, \bar{h}_1 = T_1)\right\}\right)$. Now, for a given state $s \in \bar{\mathcal{S}}_{\mathrm{d}}^{\mathrm{th},a}$ such that $A_1 < \bar{A}_{\mathrm{th},1}$, according to Lemma 3, we note that $\pi^\star(s) = H$. This indicates that $\mu^\star(\bar{s})$ and $\pi^\star(s)$ are different even though the states s and $\bar{s}$ have the same combination $(\bar{b}_1, \bar{g}_1, \bar{h}_1)$ which demonstrates the difference between the structures of the age-optimal and the throughput-optimal polices.*

VI. Numerical Results

In this section, we verify our analytical results derived in section IV, and show the performance of our proposed DRL algorithm in terms of the achievable average weighted sum-AoI as a function of system design parameters. The downlink and uplink channel power gains between the destination and source nodes are modeled as $g_i = h_i = \Gamma\psi^2 d_i^{-\nu}$; where Γ is the signal power gain at a reference distance of 1 meter,

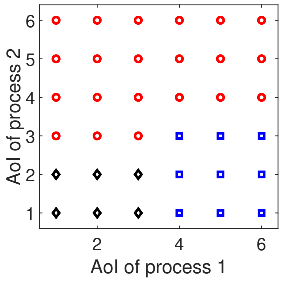

Fig. 4. Structure of the age-optimal policy when $N = 2$, $b_1 = b_2 = 1$ and $g_1 = g_2 = 6$. We use $d_1 = 25$ meters, $d_2 = 40$ meters, $B_{\max,1} = B_{\max,2} = 0.4$ mJoules, $b_{\max,1} = b_{\max,2} = 5$, $S = 15$ Mbits, and $A_{\max,i} = H_i = G_i = 6, i \in \{1, 2\}$.

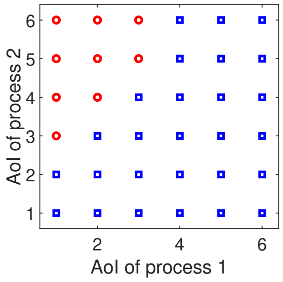

Fig. 5. Structure of the age-optimal policy when $N = 2$, $b_1 = b_2 = 5$ and $g_1 = g_2 = 2$. Other parameters are same as Fig. 4.

$\psi^2 \sim \exp(1)$ denotes the small-scale fading gain, and $d_i^{-\nu}$ represents standard power law path-loss with exponent ν. Recall that we denote the number of discrete values that the state variables g_i and h_i can take by G_i and H_i, respectively. In the following, we use $g_i = j$ ($h_i = j$) to refer to the value of the channel power gain at its j-th level where $j \in \{1, 2, \cdots, G_i\}$ ($j \in \{1, 2, \cdots, H_i\}$). Unless otherwise specified, we use the following values for different system parameters: $W = 1$ MHz, $P = 37$ dBm, $\eta = 0.5$, $\sigma^2 = -95$ dBm, $\Gamma = 0.2$, $\nu = 2$ and $\theta_i = \frac{1}{N}, i \in \mathcal{I}$.

A. Verification of Analytical Results

In Figs. 4 and 5 (Figs. 6 and 7), we present the structure of the age-optimal policy for the case of $N = 2$ ($N = 1$). Particularly, a point in each of these figures represents a potential state of the system where a blue square point (a red circle point) (a black diamond point) indicates that the optimal action at this state is T_1 (T_2) (H). In addition, for the single source-destination pair model in Figs. 6 and 7, the points located inside the solid polygon refer to the states for which it is possible to transmit an update packet (take T_1 action), i.e., for each of those states $b_1 \geq e_1^{\mathrm{T}}$. Furthermore, the points located inside the dotted polygon represent the set $\mathcal{S}_{\mathrm{d}}^{\mathrm{th},a}$ (defined in Remark 2), i.e., the set of states over which the age-optimal policy has a threshold-based structure. Note that the dotted polygon is the same as the solid one in Fig. 7. From these results, we can easily verify that the analytical structural properties of the age-optimal policy, derived in Theorems 1 and 2, are satisfied. For instance, in Fig. 4, since the optimal action at the point $(2, 3)$ is T_2, we observe that the optimal action at the points $(2, y)$, where $y > 3$, is T_2 as well (Theorem 1). In addition, in Fig. 7, the optimal action at the point $(1, 2)$ is T_1, and hence, we observe that it is optimal to take action T_1 at all the states (x, y) located inside the set $\mathcal{S}_{\mathrm{d}}^{\mathrm{th},a}$ (i.e., the solid polygon) such that $x \geq 1$ and $y \geq 2$ (Theorem 2, (i)). On the other hand, we observe that the optimality of taking action H at the point $(2, 1)$ implies that it is optimal to take action H at the point $(1, 1)$ as well (Theorem 2, (ii)).

B. Comparison of the Structures of the Age-optimal and Throughput-optimal Policies

The difference between the structures of the age-optimal and throughput-optimal polices can be understood by comparing Figs. 8 and 9. Specifically, according to the AoI value A_1, we have two different regimes: i) when A_1 is small (for instance, $A_1 = 1$ in our simulation setup), the destination node has a fresh information about process 1, and hence there is no

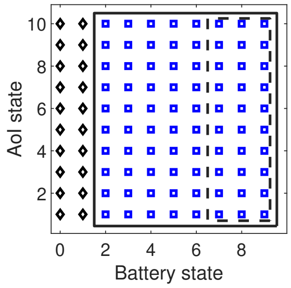

Fig. 6. Structure of the age-optimal policy when $N = 1$ and $g_1 = 2$. We use $d_1 = 35$ meters, $B_{\max,1} = 0.3$ mJoules, $S = 12$ Mbits, $A_{\max,1} = H_1 = G_1 = 10$ and $b_{\max,1} = 9$.

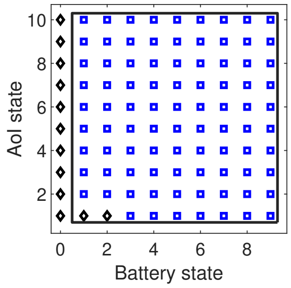

Fig. 7. Structure of the age-optimal policy when $N = 1$ and $g_1 \in \{5, 6, \cdots, 10\}$. Other parameters are same as Fig. 6.

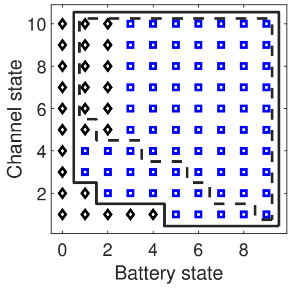

Fig. 8. Structure of the throughput-optimal policy as well as the age-optimal policy for $N = 1$ and $A_1 = 1$. We use the same simulation setup as in Fig 6.

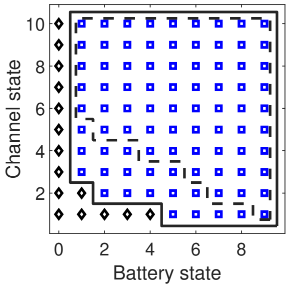

Fig. 9. Structure of the age-optimal policy for $N = 1$ and $A_1 \in \{2, 3, \cdots, 10\}$. Other parameters are same as Fig. 8.

urgency to transmit an update packet, because of which the structures of the age-optimal and throughput-optimal policies are similar (they are the same in our simulation setup when $A_1 = 1$, as shown in Fig. 8), and ii) when A_1 is large ($A_1 > 1$), different from the throughput-optimal policy, it is always optimal to take action T_1 regardless of the amount of available energy in the battery according to the age-optimal policy. This is intuitive since if the values of AoI and the battery state are small, it is wise to save the current energy in battery for future update packet transmissions when the AoI value becomes large.

Fig. 8 also verifies the analytical structural properties of the throughput-optimal policy, presented in Theorem 3. For instance, we observe that it is optimal to take action T_1 at all the states (x, y) located inside the set $\mathcal{S}_{\rm d}^{{\rm th},r}$ (i.e., the dotted polygon) such that $x \geq 4$ and $y \geq 4$, since the optimal action at the point $(4, 4)$ is T_1 (Theorem 3, (i)). Furthermore, since the optimal action at the point $(2, 10)$ is H, we observe that it is optimal to take action H as well at all states (x, y) located inside $\mathcal{S}_{\rm d}^{{\rm th},r}$ such that $x \leq 2$ and $y \leq 10$ (Theorem 3, (ii)).

C. Impact of System Design Parameters on Optimal Average Weighted Sum-AoI

Due to the curse of dimensionality in the state space of our formulated MDP, the age-optimal policy obtained by applying classical reinforcement learning algorithms [53], e.g., the RVIA, can only be evaluated numerically for small-scale settings (i.e., small values for both N and the cardinality of the discrete support set of each state variable). Therefore, we first consider the case of $N = 1$ in Fig. 10 to check the convergence of our proposed DRL algorithm while quantifying its performance in terms of the gap between its achievable average AoI and the optimal value obtained by the RVIA. Afterwards, we demonstrate the impact of system design parameters on the achievable average weighted sum-AoI for a larger value of N ($N = 3$) in Fig. 11, using the DRL algorithm. Clearly, Fig. 10 shows that our proposed reinforcement learning algorithm is able to learn the optimal policy quickly, and hence approaches the optimal average AoI. Note that the slight gap between the optimal value and the achievable average AoI by the DRL algorithm is due to using an ϵ-greedy policy in the DRL algorithm (required for exploring all the state-action pairs while learning the optimal policy, and hence guaranteeing the convergence of the algorithm). However, after the DRL algorithm converges to some value, one can check that the algorithm learns the optimal policy. Hence, the optimal value of average AoI can be achieved by reducing the value of ϵ to zero after the algorithm has converged (i.e., exploiting the learning process without the need of wasting time in exploring the environment anymore).

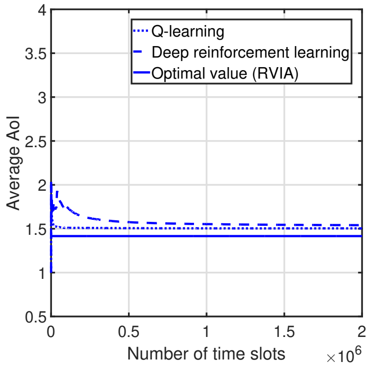

Fig. 10. Convergence of deep reinforcement learning algorithm when $N = 1$. We use $d_1 = 25$ meters, $B_{\max,1} = 0.3$ mJoules, $S = 12$ Mbits, $A_{\max,1} = H_1 = G_1 = 4$ and $b_{\max,1} = 3$.

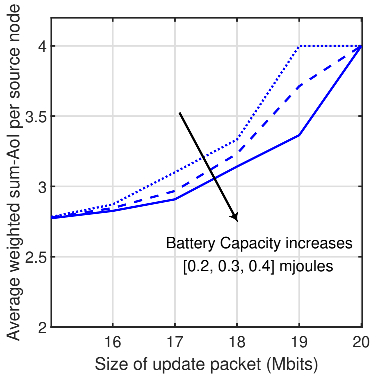

Fig. 11. Impact of size of update packets and capacity of batteries on the achievable average weighted sum-AoI by the deep reinforcement learning algorithm, for $N = 3$. We use $d_1 = 25$ meters, $d_2 = 40$ meters, $d_3 = 20$ meters, $A_{\max,i} = H_i = G_i = 4, i \in \{1, 2, 3\}$ and $b_{\max,1} = b_{\max,2} = b_{\max,3} = 3$. We also consider that $B_{\max,1} = B_{\max,2} = B_{\max,3}$.

Fig. 11 shows the impact of the capacity of batteries and size of update packets on the achievable optimal average weighted sum-AoI $\bar{A}^{\star}$, satisfying the Bellman's equations in (12). It is observed that the achievable average sum-AoI monotonically decreases as the size of update packets decreases and/or the capacity of batteries increases. This is due to the fact that decreasing the size of update packets reduces the amount of energy needed to transmit an update packet from each source node, and increasing the capacity of batteries allows to store more harvested energy inside the batteries. This, in turn, increases the likelihood that each source node will have enough energy required for an update packet transmission when the AoI value of its observed process is large, and hence the achievable average weighted sum-AoI is reduced.

VII. Conclusion

In this paper, we have proposed an implementable age-optimal sampling strategy for designing freshness-aware RF-powered communication systems. In particular, we studied a real-time monitoring system in which multiple RF-powered source nodes are sending update packets to a destination node with the objective of keeping its information status about their observed processes fresh. For this system setup, the long-term average weighted sum-AoI minimization problem was formulated, where the WET by the destination node and scheduling of update packet transmissions from the source nodes are jointly optimized. To obtain the age-optimal policy, the problem was modeled as an average cost MDP with finite state and action spaces. Since the state space in the formulated MDP is extremely large, we proposed a DRL algorithm that can learn the optimal policy efficiently. An analytical characterization for the structural properties of the age-optimal policy was also provided, where it was proven that the age-optimal policy has a threshold-based structure with respect to the AoI values for different processes. Moreover, it was demonstrated that the age-optimal policy has a threshold based structure with respect to all system state variables for the single-source destination pair model. We then extended our analysis to the average throughput maximization problem using which we mathematically characterized key differences in the structural properties of the age-optimal and throughput-optimal policies for our system setup.

Multiple system design insights were drawn from our numerical results. For instance, they showed that the structures of the age-optimal and throughput-optimal policies in the single source-destination pair model are similar when the AoI value is relatively small (i.e., there is no urgency to update the information status at the destination node). In contrast, the age-optimal and throughput-optimal polices have completely different structures when the AoI value grows. Our results also revealed that the optimal average weighted sum-AoI is a monotonically increasing (decreasing) function with respect to the size of update packets (capacity of batteries at the source nodes).

References

[1] M. A. Abd-Elmagid, H. S. Dhillon, and N. Pappas, “Online age-minimal sampling policy for RF-powered IoT networks,” *Proc., IEEE Globecom*, Dec. 2019.

[2] M. A. Abd-Elmagid, N. Pappas, and H. S. Dhillon, “On the role of age of information in the internet of things,” *IEEE Commun. Magazine*, vol. 57, no. 12, pp. 72–77, 2019.

[3] M. A. Abd-Elmagid, M. A. Kishk, and H. S. Dhillon, “Joint energy and SINR coverage in spatially clustered RF-powered IoT network,” *IEEE Trans. on Green Commun. and Networking*, vol. 3, no. 1, pp. 132–146, March 2019.

[4] S. Kaul, R. Yates, and M. Gruteser, “Real-time status: How often should one update?” in *Proc., IEEE INFOCOM*, 2012.

[5] R. D. Yates and S. Kaul, “Real-time status updating: Multiple sources,” in *Proc., IEEE Intl. Symposium on Information Theory*, 2012.

[6] C. Kam, S. Kompella, and A. Ephremides, “Age of information under random updates,” in *Proc., IEEE Intl. Symposium on Information Theory*, 2013.

[7] L. Huang and E. Modiano, “Optimizing age-of-information in a multi-class queueing system,” in *Proc., IEEE Intl. Symposium on Information Theory*, 2015.

[8] M. Costa, M. Codreanu, and A. Ephremides, “On the age of information in status update systems with packet management,” *IEEE Trans. on Info. Theory*, vol. 62, no. 4, pp. 1897–1910, 2016.

[9] K. Chen and L. Huang, “Age-of-information in the presence of error,” in *Proc., IEEE Intl. Symposium on Information Theory*, 2016.

[10] B. Barakat, S. Keates, I. Wassell, and K. Arshad, “Is the zero-wait policy always optimum for information freshness (peak age) or throughput?” *IEEE Commun. Letters*, vol. 23, no. 6, pp. 987–990, June 2019.

[11] A. Kosta, N. Pappas, A. Ephremides, and V. Angelakis, “Age and value of information: Non-linear age case,” in *Proc., IEEE Intl. Symposium on Information Theory*, 2017.

[12] A. Javani and Z. Wang, “Age of information in multiple sensing of a single source,” 2019, available online: arxiv.org/abs/1902.01975.

[13] Y. Sun, E. Uysal-Biyikoglu, R. D. Yates, C. E. Koksal, and N. B. Shroff, “Update or wait: How to keep your data fresh,” *IEEE Trans. on Info. Theory*, vol. 63, no. 11, pp. 7492–7508, 2017.

[14] A. M. Bedewy, Y. Sun, and N. B. Shroff, “Optimizing data freshness, throughput, and delay in multi-server information-update systems,” in *Proc., IEEE Intl. Symposium on Information Theory*, 2016.

[15] I. Kadota, E. Uysal-Biyikoglu, R. Singh, and E. Modiano, “Minimizing the age of information in broadcast wireless networks,” in *Proc., Allerton Conf. on Commun., Control, and Computing*, 2016.

[16] Y.-P. Hsu, E. Modiano, and L. Duan, “Scheduling algorithms for minimizing age of information in wireless broadcast networks with random arrivals,” *IEEE Trans. on Mobile Computing*, to appear.

[17] X. Chen and S. S. Bidokhti, “Benefits of coding on age of information in broadcast networks,” 2019, available online: arxiv.org/abs/1904.10077.

[18] R. Talak, S. Karaman, and E. Modiano, “Minimizing age-of-information in multi-hop wireless networks,” in *Proc., Allerton Conf. on Commun., Control, and Computing*, 2017.

[19] A. Valehi and A. Razi, “Maximizing energy efficiency of cognitive wireless sensor networks with constrained age of information,” *IEEE Trans. on Cognitive Commun. and Networking*, vol. 3, no. 4, pp. 643–654, Dec 2017.

[20] M. A. Abd-Elmagid and H. S. Dhillon, “Average peak age-of-information minimization in UAV-assisted IoT networks,” *IEEE Trans. on Veh. Technology*, vol. 68, no. 2, pp. 2003–2008, Feb. 2019.

[21] J. Liu, X. Wang, B. Bai, and H. Dai, “Age-optimal trajectory planning for UAV-assisted data collection,” in *Proc., IEEE INFOCOM Workshops*, 2018.

[22] M. A. Abd-Elmagid, A. Ferdowsi, H. S. Dhillon, and W. Saad, “Deep reinforcement learning for minimizing age-of-information in UAV-assisted networks,” *Proc., IEEE Globecom*, Dec. 2019.

[23] Y. Gu, H. Chen, Y. Zhou, Y. Li, and B. Vucetic, “Timely status update in internet of things monitoring systems: An age-energy tradeoff,” *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5324–5335, June 2019.

[24] B. Zhou and W. Saad, “Joint status sampling and updating for minimizing age of information in the Internet of Things,” *IEEE Trans. on Commun.*, vol. 67, no. 11, pp. 7468–7482, 2019.

[25] M. K. Abdel-Aziz, C.-F. Liu, S. Samarakoon, M. Bennis, and W. Saad, “Ultra-reliable low-latency vehicular networks: Taming the age of information tail,” in *Proc., IEEE Globecom*, 2018.

[26] B. Buyukates, A. Soysal, and S. Ulukus, “Age of information in Two-hop multicast networks,” in *Proc., IEEE Asilomar*, 2018.

[27] R. D. Yates, E. Najm, E. Soljanin, and J. Zhong, “Timely updates over an erasure channel,” in *Proc., IEEE Intl. Symposium on Information Theory)*, 2017.

[28] A. Maatouk, M. Assaad, and A. Ephremides, “Energy efficient and throughput optimal CSMA scheme,” *IEEE/ACM Trans. on Networking*, vol. 27, no. 1, pp. 316–329, Feb 2019.

[29] M. Bastopcu and S. Ulukus, “Minimizing age of information with soft updates,” 2018, available online: arxiv.org/abs/1812.08148.

[30] N. Lu, B. Ji, and B. Li, “Age-based scheduling: Improving data freshness for wireless real-time traffic,” in *ACM Intl. Symposium on Mobile Ad Hoc Networking and Computing*, 2018, pp. 191–200.

[31] T. Z. Ornee and Y. Sun, “Sampling for remote estimation through queues: Age of information and beyond,” 2019, available online: arxiv.org/abs/1902.03552.

[32] J. Sun, Z. Jiang, B. Krishnamachari, S. Zhou, and Z. Niu, “Closed-form whittle’s index-enabled random access for timely status update,” *IEEE Trans. on Commun.*, vol. 68, no. 3, pp. 1538–1551, 2020.

[33] H. Tang, J. Wang, J. Wang, L. Song, J. Song, and J. Song, “Minimizing age of information with power constraints: Multi-user opportunistic scheduling in multi-state time-varying channels,” *IEEE Journal on Selected Areas in Commun.*, to appear.

[34] R. D. Yates, “Lazy is timely: Status updates by an energy harvesting source,” in *Proc., IEEE Intl. Symposium on Information Theory*, 2015.

[35] A. Arafa and S. Ulukus, "Timely updates in energy harvesting two-hop networks: Offline and online policies," *IEEE Trans. on Wireless Commun.*, vol. 18, no. 8, pp. 4017–4030, Aug. 2019.

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

[37] S. Feng and J. Yang, "Age of information minimization for an energy harvesting source with updating erasures: With and without feedback," 2018, available online: arxiv.org/abs/1808.05141.

[38] A. Arafa, J. Yang, S. Ulukus, and H. V. Poor, "Age-minimal transmission for energy harvesting sensors with finite batteries: Online policies," 2018, available online: arxiv.org/abs/1806.07271.

[39] B. T. Bacinoglu, Y. Sun, E. Uysal-Bivikoglu, and V. Mutlu, "Achieving the age-energy tradeoff with a finite-battery energy harvesting source," in *Proc., IEEE Intl. Symposium on Information Theory*, 2018.

[40] A. Baknina, S. Ulukus, O. Oze, J. Yang, and A. Yener, "Sening information through status updates," in *Proc., IEEE Intl. Symposium on Information Theory*, 2018, pp. 2271–2275.

[41] X. Wu, J. Yang, and J. Wu, "Optimal status update for age of information minimization with an energy harvesting source," *IEEE Trans. on Green Commun. and Networking*, vol. 2, no. 1, pp. 193–204, 2018.

[42] S. Farazi, A. G. Klein, and D. R. Brown, "Average age of information for status update systems with an energy harvesting server," in *Proc., IEEE INFOCOM Workshops*, 2018.

[43] G. Stamatakis, N. Pappas, and A. Traganitis, "Control of status updates for energy harvesting devices that monitor processes with alarms," 2019, available online: arxiv.org/abs/1907.03826.

[44] Z. Chen, N. Pappas, E. Björnson, and E. G. Larsson, "Age of information in a multiple access channel with heterogeneous traffic and an energy harvesting node," in *Proc., IEEE INFOCOM Workshops*, 2019.

[45] S. Leng and A. Yener, "Age of information minimization for an energy harvesting cognitive radio," *IEEE Trans. on Cognitive Commun. and Networking*, vol. 5, no. 2, pp. 427–439, 2019.

[46] E. T. Ceran, D. Gündüz, and A. György, "Reinforcement learning to minimize age of information with an energy harvesting sensor with harq and sensing cost," in *Proc., IEEE INFOCOM Workshops*, 2019.

[47] Y. Lu, K. Xiong, P. Fan, Z. Zhong, and K. B. Letaief, "Online transmission policy in wireless powered networks with urgency-aware age of information," in *Proc., Intl. Wireless Commun. Mobile Computing Conf.*, 2019.

[48] I. Krikidis, "Average age of information in wireless powered sensor networks," *IEEE Wireless Commun. Letters*, 2019.

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

[50] ——, "A reinforcement learning approach to age of information in multi-user networks," in *Proc., IEEE PIMRC*, 2018.

[51] H. B. Beytur and E. Uysal, "Age minimization of multiple flows using reinforcement learning," in *Intl. Conf. on Computing, Networking and Commun. (ICNC)*, 2019.

[52] E. Sert, C. Sönmez, S. Baghaee, and E. Uysal-Biyikoglu, "Optimizing age of information on real-life tcp/ip connections through reinforcement learning," in *Proc., Signal Processing and Commun. Applications Conf.*, 2018.

[53] D. P. Bertsekas, "Dynamic programming and optimal control 3rd edition, volume ii," *Belmont, MA: Athena Scientific*, 2011.

[54] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[55] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[56] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc., IEEE Intl. Conf. on Acoustics, Speech, and Sig. Proc.*, 2013.

[57] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: A tutorial," *IEEE Commun. Surveys & Tutorials*, vol. 21, no. 4, pp. 3039–3071, 2019.

[58] A. Ferdowsi, U. Challita, W. Saad, and N. B. Mandayam, "Robust deep reinforcement learning for security and safety in autonomous vehicle systems," in *Proc., IEEE Intl. Conf. on Intelligent Transportation Systems*, 2018.

[59] W. B. Powell, *Approximate Dynamic Programming: Solving the curses of dimensionality*. John Wiley & Sons, 2007, vol. 703.

Mohamed A. Abd-Elmagid is a Ph.D. Candidate in the Bradley Department of Electrical and Computer Engineering at Virginia Tech. He received his B.Sc. degree in Electronics and Electrical Communications Engineering from Cairo University, Egypt in 2014 and M.S. degree in Wireless Communications from Nile University, Egypt in 2017. He worked as an Exchange Research Assistant at Sabanci University, Turkey, from Oct. 2016 to Feb. 2017. Prior to that, he was a Research Assistant at Nile University, Egypt, from Sep. 2014 to Sep. 2016. His research interests include cyber-physical internet of things systems, age of information, energy harvesting, wireless networks, machine learning, optimization and stochastic geometry.

Harpreet S. Dhillon (S'11–M'13–SM'19) received the B.Tech. degree in electronics and communication engineering from IIT Guwahati in 2008, the M.S. degree in electrical engineering from Virginia Tech in 2010, and the Ph.D. degree in electrical engineering from the University of Texas at Austin in 2013.

After serving as a Viterbi Postdoctoral Fellow at the University of Southern California for a year, he joined Virginia Tech in 2014, where he is currently an Associate Professor of electrical and computer engineering and the Elizabeth and James E. Turner Jr. '56 Faculty Fellow. His research interests include communication theory, wireless networks, stochastic geometry, and machine learning. He is a Clarivate Analytics Highly Cited Researcher and has coauthored five best paper award recipients including the 2014 IEEE Leonard G. Abraham Prize, the 2015 IEEE ComSoc Young Author Best Paper Award, and the 2016 IEEE Heinrich Hertz Award. He was named the 2017 Outstanding New Assistant Professor, the 2018 Steven O. Lane Junior Faculty Fellow, the 2018 College of Engineering Faculty Fellow, and the recipient of the 2020 Dean's Award for Excellence in Research by Virginia Tech. His other academic honors include the 2008 Agilent Engineering and Technology Award, the UT Austin MCD Fellowship, and the 2013 UT Austin Wireless Networking and Communications Group leadership award. He currently serves as a Senior Editor for the IEEE WIRELESS COMMUNICATIONS LETTERS and an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING.

Nikolaos Pappas (S'07-M'13) received the B.Sc. degree in computer science, the B.Sc. degree in mathematics, the M.Sc. degree in computer science, and the Ph.D. degree in computer science from the University of Crete, Greece, in 2005, 2012, 2007, and 2012, respectively. From 2005 to 2012, he was a Graduate Research Assistant with the Telecommunications and Networks Laboratory, Institute of Computer Science, Foundation for Research and Technology-Hellas, and a Visiting Scholar with the Institute of Systems Research, University of Maryland at College Park, College Park, MD, USA. From 2012 to 2014, he was a Post-Doctoral Researcher with the Department of Telecommunications, Supélec, France. Since 2014, he has been with Linköping University, Norrköping, Sweden, as a Marie Curie Fellow (IAPP). He is currently an Associate Professor in mobile telecommunications with the Department of Science and Technology, Linköping University. His main research interests include the field of wireless communication networks with emphasis on the stability analysis, energy harvesting networks, network-level cooperation, age-of-information, network coding, and stochastic geometry. From 2013 to 2018, he was an Editor of the IEEE COMMUNICATIONS LETTERS. He is currently an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE/KICS JOURNAL OF COMMUNICATIONS AND NETWORKS, and the IEEE Open Journal of Communications Society. He is also a guest editor of the IEEE Internet of Things Journal for the Special Issue "Age of Information and Data Semantics for Sensing, Communication and Control Co-Design in IoT".