Age-of-Information Minimization via Opportunistic Sampling by an Energy Harvesting Source

Akanksha Jaiswal, Arpan Chattopadhyay, Member, IEEE, and Amokh Varma

Abstract—Herein, minimization of time-averaged age-of-information (AoI) in an energy harvesting (EH) source setting is considered. The EH source opportunistically samples one or multiple processes over discrete time instants and sends the status updates to a sink node over a wireless fading channel. Each time, the EH node decides whether to probe the link quality and then decides whether to sample a process and communicate based on the channel probe outcome. The trade-off is between the freshness of information available at the sink node and the available energy at the source node. We use infinite horizon Markov decision process (MDP) to formulate the AoI minimization problem for two scenarios where energy arrival and channel fading processes are: (i) independent and identically distributed (i.i.d.), (ii) Markovian. In i.i.d. setting, after channel probing, the optimal source sampling policy is shown to be a threshold policy. Also, for unknown channel state and EH characteristics, a variant of the Q-learning algorithm is proposed for the two-stage action model, that seeks to learn the optimal policy. For Markovian system, the problem is again formulated as an MDP, and a learning algorithm is provided for unknown dynamics. Finally, numerical results demonstrate the policy structures and performance trade-offs.

Index Terms—Age-of-information, remote sensing, energy harvesting, Markov decision process (MDP), reinforcement learning.

I. INTRODUCTION

In recent years, the need for combining physical systems with the cyber world has attracted significant research interest. These cyber-physical systems (CPS) are supported by ultra-low power, low latency Internet of Things (IoT) networks, and encompass a large number of applications such as vehicle tracking, environment monitoring, intelligent transportation, industrial process monitoring, smart home systems, etc. Such systems often require the deployment of sensor nodes to monitor a physical process and send real-time status updates to a remote estimator over a wireless network. However, for such critical CPS applications, minimizing mean packet delay without accounting for delay jitter can often be detrimental to the system’s performance. Also, mean delay minimization does not guarantee delivery of the observation packets to the sink node in the same order in which they were generated, thereby often resulting in unnecessarily dedicating network resources towards delivering outdated observation packets despite the availability of a freshly generated observation packet in the system. Hence, it is necessary to take into account the freshness of information of the data packets, apart from the mean packet delay.

Recently, a metric named Age of Information (AoI) has been proposed [2] as a measure of the freshness of the information update. In this setting, a sensor monitoring system generates a time-stamped status update and sends it to the sink over a network. At the time $t$, if the latest monitoring information available to the sink node comes from a packet whose time-stamped generation instant was $t'$, then the AoI at the sink node is computed as $(t - t')$. Thus, AoI has emerged as an alternative metric to mean delays [3].

However, the timely delivery of status updates is often limited by energy and bandwidth constraints in the network. Recent efforts towards designing EH source nodes (e.g., source nodes equipped with solar panels) have opened a new research paradigm for IoT network operations. The energy generation process in such nodes is very uncertain and typically modeled as a stochastic process. The harvested energy is stored in an energy buffer as energy packets and used for sensing and communication as and when needed. This EH capability significantly improves network lifetime and eliminates the need for frequent manual battery replacement, but poses a new challenge towards network operations due to uncertainty in the available energy at the source nodes at any given time.

Motivated by the above challenges, we consider the problem of minimizing the time-averaged expected AoI in a remote sensing setting, where a single EH source probes the channel state, samples one or multiple processes, and sends the observation packets to the sink node over a fading channel. The energy generation process is modeled as a discrete-time i.i.d. process, and a finite energy buffer is considered. Two variants of the problem are considered: (i) single process with channel state information at the transmitter (CSIT), and (ii) multiple processes with CSIT. Channel state probing and process sampling for time-averaged expected AoI minimization is formulated as an MDP, and the threshold nature of the
optimal policy is established analytically for each case. Next, we propose reinforcement learning (RL) algorithms to find the optimal policy when the channel statistics and energy harvesting characteristics are unknown. We also extend these works to the setting where the system has a Markovian channel and Markovian energy harvesting process. In addition, we consider AoI minimization without channel probing, and compare its performance numerically with our algorithms that involve channel probing. Interestingly, channel probing yields smaller AoI if the energy required in probing is significantly smaller than sampling energy. Numerical results validate the theoretical results and intuitions.

A. Related Work

Initial efforts towards optimizing AoI mostly involved the analysis of various queueing models; e.g., [2] for analyzing a single source single server queueing system with FCFS service discipline, [4] for LCFS service discipline for M/M/1 queue, [5] and [6] for multi-source single sink system with M/M/1 queueing at each source, [7] for AoI performance analysis for multi-source single-sink infinite-buffer queueing system where unserved packets are substituted by available newer ones, etc. Additionally, the paper [8] has developed a scheduling policy for two users- a deadline-constrained user, and an AoI-oriented user that send their information to a single receiver through an error-prone channel (modeled as both i.i.d. and time-correlated channel). The authors formulated the average AoI minimization problem with timely throughput constraints as a constrained Markov decision process (CMDP). The authors in [9] have considered the AoI-minimization problem for the peak power-constrained base station (BS) which sends updates to multiple users through time-varying, Markovian channels, and proposed a truncated multi-user scheduling policy. The optimal threshold-based sampling policy for AoI minimization for a source sending samples to a remote estimator over an unreliable channel with feedback has been proposed in [10].

On the other hand, a number of papers have considered AoI minimization problem under EH setting: [11] for the derivation of average AoI for a single source having finite battery capacity, [12] for the derivation of the minimal age policy for EH two-hop network, [13] for average AoI expression for single source EH server, [14] for AoI minimization for a wirelessly powered user, [15] for sampling, transmission scheduling and transmit power selection for a single source single sink system over infinite time horizon where the delay is dependent on the packet transmission energy. The authors in [16] have considered two source nodes (power grid node and EH sensor node) sending different data packets to a common destination by using multiple access channel, and derived the delay and AoI of two source nodes respectively. The authors of [17] have analyzed the non-linear AoI for EH system, whereas [18], [19] focused on the characterization of moment generating function of AoI in EH status update systems under various queueing disciplines. In [20], average AoI and peak AoI (PAoI) are analyzed for a source and a relay network where energy is harvested from a power station.

There have also been several other works on developing optimal scheduling policy for minimizing AoI for EH sensor networks [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32]. For example, [21] has investigated optimal online policy for single sensor single sink system with a noiseless channel for infinite, finite and unit size battery capacity; for finite battery size, it has provided energy aware status update policy. The paper [22] has considered a multi-sensor single sink system with infinite battery size, and proposed a randomized myopic scheduling scheme. For an EH source with finite battery and Poisson energy arrival process, the authors in [23] have provided age-energy trade-off and optimal threshold policy for AoI minimization, but unlike our work, they have not considered probing/estimation of a fading channel while making a sampling decision. In [24], the optimal online status update policies to minimize the long-run average AoI for EH source with updating erasures have been proposed. It has been shown that the best-effort uniform updating policy is optimal when there is no feedback and the best-effort uniform updating with re-transmission (BUR) policy is optimal when feedback is available to the source. The authors of [25] have examined the problem of minimizing AoI under a constraint on the count of status updates. The authors of [26] have addressed the AoI minimization problem for cognitive radio communication systems with EH capability; they formulated optimal sensing and updating for perfect and imperfect sensing as a partially observable Markov decision process (POMDP). Information source diversity, i.e., multiple sources tracking the same process but with dissimilar energy cost, and sending status updates to the EH monitoring node with finite battery size, has been considered in [27] with an MDP formulation, but no structure was provided for the optimal policy. In [28], reinforcement learning has been used to minimize AoI for a single EH sensor with HARQ protocol where a decision is made whether to send new status updates or retransmit failed status updates and it has shown the existence of a threshold-based optimal policy. In [30], the authors have formulated optimal transmission schemes for an EH sensor that sends update packets to the BS by considering different transmission modes (controlled by power and error probability) and battery recovery setting as an MDP. The authors of [31] have developed a threshold policy for minimizing AoI for a single-sensor single-sink system with an erasure channel and no channel feedback. For a system with Poisson energy arrival, unit battery size, and error-free channel, it has shown that a threshold policy achieves an average age lower than that of a zero-wait policy; based on this, a lower bound on average age for general battery size and erasure channel has been derived.

The paper [33] has proposed a threshold-based sampling policy for AoI minimization in a multi-source system and also uses deep reinforcement learning to learn the optimal policy. Here, it is assumed that the source nodes harvest energy from the destination, but it cannot simultaneously harvest energy and transmit data by using downlink and uplink fading channels, respectively. The authors in [34] have proposed optimal threshold policy and RL algorithms for a system with cache-enabled edge node between users and EH sensors, which
either commands the sensor to send an update or retrieves the age measurement from the cache on user request. However, these papers [28], [33], [34] have not considered channel probing as done in our paper, which makes our system model completely different from them and it involves challenges to find out optimal probing and sampling decisions taken by the source node (capable of harvesting energy and transmit data simultaneously) at each time instant. Also, we have formulated the AoI minimization problem and proposed RL algorithms for both the i.i.d. system as well as the Markovian system which involves a two-stage action model (probing and sampling), which is not available in the literature.

Usually, a source node probes the channel by sending a pilot signal which consumes resources. Hence, the source node needs to identify the optimal channel probing and data transmission strategies, in order to optimize energy consumption. Several papers have considered channel probing policies for wireless communication systems [35], [36], [37], [38]. The authors in [39] have considered a system model where an energy-limited robot sends short status packets to a control center (CC) over a wireless fading channel which is estimated by sending a pilot signal during each coherence time. They have formulated the average AoI minimization problem as a CMDP with an average power consumption constraint and proposed a threshold-based status update scheme. However, these papers have not used channel probing policies for AoI minimization under EH remote sensing system, which motivates us to include channel probing before transmitting information updates, and the major challenges brought by the consideration of probing lies in the formulation of two-stage MDP model which involves two actions, i.e., channel probing and sampling decision-making at the same time. The value iteration used here for finding the optimal solution involves intermediate iterates which is non-standard. Also, this setup has not been considered before. Further, for unknown dynamics, Q-learning formulation involving a two-stage action model is not straightforward and different from that available in the literature. Later, in Fig. 7, we demonstrate that the channel probing is useful in minimizing AoI under practical system parameters settings.

B. Contributions

1) We formulate the problem of minimizing the time-averaged expected AoI in an EH remote sensing system with a single source monitoring one or multiple processes, as an MDP with a two-stage action model. The two stages of actions involve channel probing and process sampling at each time instant. The considered MDP with a two-stage action model is different from standard MDP in the literature since it involves using certain intermediate value function. Under the assumptions of i.i.d. time-varying channel with CSIT, channel state probing capability at the source, and finite battery size, for the single/multiple process(es), we derive the optimal sampling policy structures which turn out to be a simple threshold policy on the probed channel quality which is a function of the age of the process(es). The source node, depending on the current age of a process (the largest age of the processes), decides whether to probe the channel or not. Next, based on the channel probe outcome, the source node decides whether to sample the process (with the largest age) and send an observation packet, or to remain idle. Thus, the MDP involves taking action in two stages at each time instant where a sampling decision is taken only if the channel is probed.

2) We prove the convergence of an analogue of value iteration for this two-stage MDP. Our analysis directly shows that the difference between the value function iterate and the optimal value function decreases to 0 exponentially fast.

3) We prove certain interesting properties of various cost functions which are useful for deriving optimal threshold structures. For example, we prove that the primary cost value functions increase with age, and the intermediate cost-to-go value functions decrease with packet success probability in the probed channel.

4) We also formulate the AoI minimization problem for a Markovian system model with a source monitoring a single process as an MDP; the channel dynamics and energy harvesting characteristics are modeled as Markov chains in this case. We provide optimality equations for the MDP. Certain conjectures regarding the threshold structure of the optimal policy for this MDP are supported numerically. Note that, these results can be extended to multiple processes scenarios, but we have omitted this case due to lack of space.

5) For a source sensing a single process with unknown channel statistics and energy harvesting characteristics, we propose reinforcement learning algorithms that yield the optimal policies for the formulated MDPs. However, unlike standard Q-learning algorithm involving a single action at each time instant, we propose a variant of the Q-learning algorithm, that involves taking the action in two stages; first, the source decides whether to probe the channel quality and if the channel quality turns out to be good, then it further decides whether to sample and communicate based on its available energy and age in sampling. Similar techniques can also be used for multiple processes. The Q-learning algorithms are based on asynchronous stochastic approximation techniques [40].

C. Organization

The rest of the paper is organized as follows. The system model has been explained in Section II. AoI minimization for both single and multiple process case with i.i.d. system model is addressed in Section III, whereas AoI minimization for Markovian system model is considered in Section IV. Further, reinforcement learning for the i.i.d. and the Markovian system model is proposed in Section V. Lastly, numerical results are provided in Section VI, followed by the conclusions in Section VII. All proofs are provided in the appendices.
Probability of success of a transmitted packet. We denote by $\text{sink}$, where a particular fading state is characterized by the energy harvesting and packet transmission happen simultaneously.

## A. Channel model

We consider a fading channel between the source and the sink, where a particular fading state is characterized by the probability of success of a transmitted packet. We denote by $p(t) \in \{p_1, p_2, \ldots, p_m\}$ the probability of packet transmission success from the source to the sink at time $t$, and by $C(t) \in \{C_1, C_2, \ldots, C_m\}$ the corresponding channel state, where $m$ is the finite number of channel states. The packet success probability corresponding to channel state $C_j$ is given by $p(C_j) = p_j$. Let us also denote by $r(t) \in \{0, 1\}$ the indicator that the packet transmission from the source to the sink at time $t$ is successful. Obviously, $\mathbb{P}(r(t) = 1 | C(t) = C_j) = p_j$ for all $j \in \{1, 2, \ldots, m\}$. It is assumed that the channel state $C(t)$ is learnt perfectly via a channel probe. For many practical systems, the data packets that need to be transmitted are very long and the pilot signal used for channel estimation is short in duration and requires very less time as compared to the time required for data transmission. Therefore, the channel can be estimated instantaneously without a penalty in time.

## I.D. channel model

In this case, we assume that $\{p(t)\}_{t \geq 0}$ is i.i.d. across $t$, with $\mathbb{P}(p(t) = p_j) = q_j$ for all $j \in \{1, 2, \ldots, m\}$ [8].

**Markovian channel model:** Here, we model the fading channel as finite state Markov chain [8], [44], [45], [46], where the one-step state transition probability from state $C_i$ to state $C_j$ is denoted by $q_{i,j}$ and the $t$-step transition probability is given by $q_{i,j}(t) = \mathbb{P}(C(t) = C_j | C(0) = C_i)$ for all $i, j \in \{1, 2, \ldots, m\}$.

## B. Energy Harvesting Process

Let $A(t)$ denote the number of energy packet arrivals to the energy buffer at time $t$, and $E(t)$ denote the energy available to the source at time $t$, for all $t \geq 0$. We consider the following two models for energy harvesting:

**I.D. model:** The energy packet generation process $\{A(t)\}_{t \geq 0}$ in the energy buffer is assumed to be an i.i.d. process with known mean $\lambda > 0$; see [16], [26], [27], [34] for reference.

**Markov model:** To address the temporal correlation in energy generated by many renewable energy sources such as wind and solar, the papers [47], [48] have considered a two-state Markov chain to model the harvested energy. Motivated by this, we model the EH state of the source as a two-state Markov chain with states $\{H_1, H_2\}$, where $H_1$ is called the harvesting state and $H_2$ is called the non-harvesting state. When the source node is in the harvesting state, the energy packet generation process is considered to be a (conditionally) i.i.d. process (see [48]) with known mean $\lambda > 0$. When the source node is in a non-harvesting state, then the energy packet generation rate is zero. The sate transition probability from state $H_1$ to state $H_2$ is $h_{1,2}$ and from state $H_2$ to state $H_1$ is $h_{2,1}$ and it is governed by a Markov chain. Thus, the energy generation process is a Markov-modulated i.i.d. process and it exhibits temporal correlation.

We assume that, in case the energy buffer at the source node is full, the newly generated energy packets will not be accommodated unless $E_p$ unit of energy is spent in probing.

---

**Table 1: Key Notations Used**

| Notation Used | Description |
|---------------|-------------|
| $t$           | Discrete time index |
| $B$           | Energy buffer size |
| $E_p$         | Energy required for probing the channel |
| $E_s$         | Energy required for sensing and communicating a data packet |
| $N$           | Number of processes |
| $C_j$         | Channel state $j$, $v \in \{1, 2, \ldots, m\}$, where $m$ denotes a finite number |
| $p_j$         | Packet success probability of channel state $C_j$ |
| $q_j$         | Probability of being in a channel state $C_j$ |
| $A$           | Energy arrival rate |
| $A_s$         | Number of energy packet arrivals in a slot |
| $E$           | Energy available at the energy buffer |
| $A_t$         | Act of the $th$ process |
| $H_1$         | Harvesting state of the Markovian channel model |
| $H_2$         | Non-harvesting state of the Markovian channel model |
| $h_{1,2}$     | Transition probability from state $H_1$ to state $H_2$ |
| $r$           | Indicator of successful packet transmission |
| $s$           | Indicator of probing channel state |
| $a$           | Indicator of sampling a process |

---

**Fig. 1.** Pictorial representation of a remote sensing system where an EH source samples one of $N$ number of processes at a time and sends the observation packet to a sink node.
C. Decision Process and Policy

At time \( t \), let \( b(t) \in \{0, 1\} \) denote the indicator of deciding to probe the channel, and \( a(t) \in \{0, 1, \ldots, N\} \) denote the identity of the process being sampled, with \( a(t) = 0 \) meaning that no process is sampled, and \( b(t) = 0 \) meaning that the channel is not probed. Also, \( b(t) = 0 \) implies \( a(t) = 0 \). The set of possible actions or decisions is denoted by \( \mathcal{A} = \{\{0, 0\} \cup \{1 \times \{0, 1, 2, \ldots, N\}\}\} \), where a generic action at time \( t \) is denoted by \((b(t), a(t))\).

Let us denote by \( \tau_k(t) = \sup\{0 \leq \tau < t : a(\tau) = k, r(\tau) = 1\} \) the last time instant before time \( t \), when process \( k \) was sampled and the observation packet was successfully delivered to the sink. The age of information (AoI) for the \( k \)-th process at time \( t \) is given by \( T_k(t) = (t - \tau_k(t)) \). However, if \( a(t) = k \) and \( r(t) = 1 \), then \( T_k(t) = 0 \) since the current observation of the \( k \)-th process is available to the sink node. A generic Markov scheduling policy is a collection of mappings \( \{\mu_t\}_{t \geq 0} \), where \( \mu_t \) maps \((E(t), C(t), \{T_k(t)\}_{1 \leq k \leq N})\) to the action space \( \mathcal{A} \). If \( \mu_t = \mu \) for all \( t \geq 0 \), the policy is called stationary, else non-stationary.

We seek to find a stationary scheduling policy \( \mu \) that minimizes the expected sum AoI which is averaged over time:

\[
\min_{\mu} \frac{1}{T} \sum_{t=0}^{T} \sum_{k=1}^{N} \mathbb{E}_\mu(T_k(t)) \tag{1}
\]

III. I.I.D. SYSTEM

In this section, we derive the optimal channel probing, source activation, and data transmission policy for a single EH source sampling a single process \((N = 1)\) or multiple processes, when the channel and energy harvesting processes follow the i.i.d. model as described in Section II.

A. Single Process \((N = 1)\)

Here, we formulate (1) as a long-run average cost MDP with state space \( S = \{0, 1, \ldots, B\} \times \mathbb{Z}_+ \) and an intermediate state space \( V = \{0, 1, \ldots, B\} \times \mathbb{Z}_+ \times \{C_1, C_2, \ldots, C_m\} \). A generic primary state \( s = (E, T) \) means that the energy buffer has \( E \) energy packets, and the last successfully received packet was generated at the source node \( T \) slots ago. A generic intermediate state \( v = (E, T, C) \) additionally means that the current channel state obtained via probing is \( C \). The action space is \( \mathcal{A} = \{\{0, 0\} \cup \{1 \times \{0, 1\}\}\} \) with \( a(t), b(t) \in \{0, 1\} \). At each time, if the source node decides not to probe the channel state, then it will not perform sampling, and thus \( b(t) = 0, a(t) = 0 \), and the expected single-stage Aol cost is \( c(s(t), b(t), a(t)) = T \). However, if the source node decides to probe the channel state, the expected single-stage Aol cost is \( c(s(t), b(t) = 1, a(t) = 0) = T \), and \( c(s(t), b(t) = 1, a(t) = 1) = T(1 - p(C)) \), where the expectation is taken over packet success probability \( p(C) \). We first formulate the average-cost MDP problem as an \( \alpha \)-discounted cost MDP problem with \( \alpha \in (0, 1) \), and derive the optimal policy, from which the solution of the average cost minimization problem can be obtained by taking \( \alpha \to 1 \) [49, Sec. IV].

1) Optimality Equation: Let \( J^*(E, T) \) be the optimal value function for primary state \((E, T)\) in the discounted cost problem, and let \( W^*(E, T, C) \) be the cost-to-go from an intermediate state \((E, T, C)\). The Bellman equations [50, Proposition 7.2.1] for the \( \alpha \)-discounted cost MDP problem are given by (2) which are solved by standard value iteration technique [50, Proposition 7.3.1].

\[
J^*(E \geq E_p + E_s, T) = \min\left\{ T + \alpha E_A J^*(\min\{E + A, B\}, T + 1), V^*(E, T) \right\}
\]

\[
V^*(E, T) = \sum_{j=1}^{m} q_j W^*(E, T, C_j)
\]

\[
W^*(E, T, C) = \min\left\{ T + \alpha E_A J^*(\min\{E - E_p + A, B\}, T + 1), \right. \\
(1 - p(C)) + \alpha p(C) E_A J^*(\min\{E - E_p - E_s + A, B\}, T + 1) \\
+ \alpha(1 - p(C)) E_A J^*(\min\{E - E_p - E_s + A, B\}, T + 1) \left. \right\}
\]

\[
J^*(E \geq E_p + E_s, T) = T + \alpha E_A J^*(\min\{E + A, B\}, T + 1)
\]

\[
(2)
\]

The first expression in the minimization in the R.H.S. of the first equation in (2) is the cost of not probing channel state \((b(t) = 0)\), which includes single-stage Aol cost \( T \) and an \( \alpha \) discounted future cost for a random next state \((\min\{E + A, B\}, T + 1)\), averaged over the distribution of the number of energy packet generation \( A \). The quantity \( V^*(E, T) \) is the expected cost of probing the channel state, which explains the second equation in (2). At an intermediate state \((E, T, C)\), if \( a(t) = 0 \), a single stage Aol cost \( T \) is incurred and the next state becomes \( (\min\{E - E_p + A, B\}, T + 1) \); if \( a(t) = 1 \), the expected Aol cost \( T(1 - p(C)) \) (expectation taken over the packet success probability \( p(C) \)), and the next random state becomes \( (\min\{E - E_p - E_s + A, B\}, T + 1) \) and \( (\min\{E - E_p - E_s + A, B\}, T + 1) \) if \( r(t) = 1 \) and \( r(t) = 0 \), respectively. The last equation in (2) follows similarly since \( b(t) = 0, a(t) = 0 \) is the only possible action when \( E < E_p + E_s \).

Substituting the value of \( V^*(E, T) \) in the first equation of (2), we obtain the following Bellman equations:

\[
J^*(E \geq E_p + E_s, T) = \min\left\{ T + \alpha E_A J^*(\min\{E + A, B\}, T + 1), \right. \\
\mathbb{E}_C \left( \min\{T + \alpha E_A J^*(\min\{E - E_p + A, B\}, T + 1), \right. \\
T(1 - p(C)) + \alpha p(C) E_A J^*(\min\{E - E_p - E_s + A, B\}, T + 1) \\
+ \alpha(1 - p(C)) E_A J^*(\min\{E - E_p - E_s + A, B\}, T + 1) \left. \right\} \right. \\
\left. \right\}
\]

\[
J^*(E \geq E_p + E_s, T) = T + \alpha E_A J^*(\min\{E + A, B\}, T + 1)
\]

\[
(3)
\]

2) Policy Structure: We first provide the convergence proof of value iteration since we have a two-stage decision process as opposed to traditional MDP where a single action is taken.

Proposition 1: The value function \( J^{(t)}(s) \) converges to \( J^*(s) \) as \( t \to \infty \).

Proof: See Appendix A.

Next, we provide some properties of the value function, which will help in proving the structure of the optimal policy.

Lemma 1: For \( N = 1 \), \( J^*(E, T) \) is increasing in \( T \) and \( W^*(E, T, C) \) is decreasing in \( p(C) \).

Proof: See Appendix B.
Conjecture 1: For $N = 1$, the optimal probing policy for the $\alpha$-discounted AoI cost minimization problem is a threshold policy on $T$. For any $E \geq E_p + E_s$, the optimal action is to probe the channel state if and only if $T \geq T_{th}(E)$ for a threshold function $T_{th}(E)$ of $E$.

Remark 1: The optimal probing decision at $(E, T)$ is obtained by using the condition that the expected cost of probing is less than or equal to the expected cost of not probing. This condition, when further simplified, reduces to checking whether $f(E, T) \geq 0$ for a suitable function $f$. From this, one can prove Conjecture 1 if $f(E, T)$ can be shown to be increasing in $T$, which is difficult in this case. A similar difficulty is also faced in other subsequent system models.

Theorem 1: For $N = 1$, at any time, if the source decides to probe the channel, then the optimal sampling policy is a threshold policy on $p(C)$. For any $E \geq E_p + E_s$ and probed channel state, the optimal action is to sample the source node if and only if $p(C) \geq p_{th}(E, T)$ for a threshold function $p_{th}(E, T)$ of $E$ and $T$.

Proof: See Appendix C.

The policy structure supports the following two intuitions. Firstly, given $E$ and $T$, the source decides to probe the channel state if AoI is greater than some threshold value. Secondly, given $E$ and $T$ and probed channel state, if the channel quality is better than a threshold, then the optimal action is to sample the process and communicate the observation to the sink node. We will later numerically observe in Section VI some intuitive properties of $T_{th}(E)$ as a function of $E$ and $\lambda$, and $p_{th}(E, T)$ as a function of $E$, $T$ and $\lambda$.

3) No Probing Case: If we consider the case where the source samples the process directly without channel probing, then the Bellman equations for this case are given by:

\[
J^*(E \geq E_s, T) = \min \left\{ T + \alpha E_A J^*(\min \{E + A, B\}, T + 1),
\right.
\]

\[
E_C \left( T(1-p(C)) + \alpha p(C) E_A J^*(\min \{E - E_s + A, B\}, 1) + \alpha (1-p(C)) E_A J^*(\min \{E - E_s + A, B\}, T + 1) \right)
\]

\[
J^*(E < E_s, T) = T + \alpha E_A J^*(\min \{E + A, B\}, T + 1)
\]

(4)

Here also, we numerically observe an optimal threshold policy on $T$; see Section VI.

B. Multiple Processes ($N > 1$)

Here, we formulate the $\alpha$-discounted cost version of (1) as an MDP with a generic primary state $s = (E, T_1, T_2, \ldots, T_N)$ which means that the energy buffer has $E$ energy packets, and the latest (successfully) received packet from the $k$-th process was generated at the source $T_k$ slots ago. Also, a generic intermediate state is $v = (E, T_1, T_2, \ldots, T_N, C)$ which additionally means that the current channel state $C$ is learnt by probing, has packet success probability $p(C)$. The action space $A = \{\{0, 0\} \cup \{1 \times \{0, 1, 2, \ldots, N\}\}\}$ with $b(t) \in \{0, 1\}$ and $a(t) \in \{0, 1, \ldots, N\}$. At each time, if the source node decides not to probe the channel state then it will not sample any process, thus $b(t) = 0$, $a(t) = 0$, and the expected single-stage AoI cost is $c(s(t), b(t), a(t)) = \sum_{i=1}^{N} T_i$. However, if the source node decides to probe the channel state, the expected single-stage AoI cost is $c(s(t), b(t) = 1, a(t) = 0) = \sum_{i=1}^{N} T_i$ and $c(s(t), b(t) = 1, a(t) = k) = \sum_{i \neq k} T_i + T_k(1-p(C))$ where the expectation is taken over packet success probability $p(C)$.

1) Optimality Equation: In this case, the Bellman equations are given by:

\[
J^*(E \geq E_p + E_s, T_1, T_2, \ldots, T_N)
\]

\[
= \min \left\{ \sum_{i=1}^{N} T_i + \alpha E_A J^*(\min \{E + A, B\},
\right.
\]

\[
T_1 + 1, T_2 + 1, \ldots, T_N + 1, V^*(E, T_1, T_2, \ldots, T_N)
\]

\[
V^*(E, T_1, T_2, \ldots, T_N)
\]

\[
= \sum_{j=1}^{m} q_j W^*(E, T_1, T_2, \ldots, T_N, C_j)
\]

\[
W^*(E, T_1, T_2, \ldots, T_N, C)
\]

\[
= \min \left\{ \sum_{i=1}^{N} T_i + \alpha E_A J^*(\min \{E - E_p + A, B\}, T_1 + 1, T_2 + 1, \ldots, T_N + 1),
\right.
\]

\[
\min_{1 \leq i \leq N} \left\{ \sum_{i \neq k} T_i + T_k(1-p(C)) + \alpha p(C) E_A J^*(\min \{E - E_p + E_s + A, B\},
\right.
\]

\[
T_1 + 1, T_2 + 1, \ldots, T_k + 1, T_{k+1} + 1, \ldots, T_N + 1) + \alpha (1-p(C)) E_A J^*(\min \{E - E_p + E_s + A, B\}, T_1 + 1, T_2 + 1, \ldots, T_N + 1)
\]

\[
J^*(E < E_p + E_s, T_1, T_2, \ldots, T_N)
\]

\[
= \sum_{i=1}^{N} T_i + \alpha E_A J^*(\min \{E + A, B\}, T_1 + 1, T_2 + 1, \ldots, T_N + 1)
\]

(5)

The first expression in the minimization in the R.H.S. of the first equation in (2) is the cost of not probing channel state ($b(t) = 0$), which includes single-stage AoI cost $\sum_{i=1}^{N} T_i$ and an $\alpha$ discounted future cost with a random next state ($\min \{E + A, B\}$, $T_1 + 1, T_2 + 1, \ldots, T_N + 1$), averaged over the distribution of the number of energy packet generation $A$. The quantity $V^*(E, T_1, T_2, \ldots, T_N)$ is the optimal expected cost of probing the channel state, which explains the second equation in (2). At an intermediate state $(E, T_1, T_2, \ldots, T_N, C)$, if $a(t)=0$, a single stage AoI cost $\sum_{i=1}^{N} T_i$ is incurred and the next state becomes ($\min \{E - E_p + A, B\}$, $T_1 + 1, T_2 + 1, \ldots, T_N + 1$); if $a(t)=k$, the expected AoI cost is $\sum_{i \neq k} T_i + T_k(1-p(C))$ (expectation taken over the packet success probability $p(C)$), and the next random state becomes ($\min \{E - E_p - E_s - E_s + A, B\}$, $T_1 + 1, T_2 + 1, \ldots, T_k + 1, T_{k+1} + 1, \ldots, T_N + 1$) and ($\min \{E - E_p - E_s + A, B\}$, $T_1 + 1, T_2 + 1, \ldots, T_N + 1$) if $r(t) = 1$ and $r(t) = 0$, respectively. The last equation in (2) follows similarly since $b(t) = 0$, $a(t) = 0$ is the only possible action when $E < E_p + E_s$.

Substituting the value of $V^*(E, T_1, T_2, \ldots, T_N)$ in the first expression of (6), shown at the bottom of the next page, we obtain the Bellman equations (5).

2) Policy Structure:

Lemma 2: For $N > 1$, the value function $J^*(E, T_1, T_2, \ldots, T_N)$ is increasing in each of $T_1, T_2, \ldots, T_N$ and $W^*(E, T_1, T_2, \ldots, T_N, C)$ is decreasing in $p(C)$.
Proof: See Appendix D. 
Let us define \( T = [T_1, T_2, \ldots, T_N] \) and \( T_{-k} = [T_1, T_2, \ldots, T_{k-1}, T_{k+1}, \ldots, T_N] \).

Conjecture 2: For \( N > 1 \), the optimal probing policy for the \( \alpha \)-discounted AoI cost minimization problem is a threshold policy on \( \max_{1 \leq k \leq N} T_k \). For any \( E \geq E_p + E_s \), the optimal policy is to probe the channel state if and only if \( \max_{1 \leq k \leq N} T_k \geq T_{\text{th}}(E, T_{-k}) \) for a threshold function \( T_{\text{th}}(E, T_{-k}) \) of \( E \) and \( T_{-k} \).

We next show that sampling the process with the largest AoI is optimal.

Theorem 2: For \( N > 1 \), after probing the channel state, the optimal source activation policy for the \( \alpha \)-discounted cost problem is a threshold policy on \( p(C) \). For any \( E \geq E_p + E_s \) and probing channel state, the optimal policy is to sample the process \( \arg \max_{1 \leq k \leq N} T_k \) if and only if \( p(C) \geq p_{\text{th}}(E, T) \) for a threshold function \( p_{\text{th}}(E, T) \) of \( (E, T) \).

Proof: See Appendix E.

We will later numerically demonstrate some intuitive properties of \( T_{\text{th}}(E, T_{-k}) \) as a function of \( E, T_{-k} \) and \( \lambda \) and \( p_{\text{th}}(E, T_1, T_2, \ldots, T_N) \) as a function of \( E, (T_1, T_2, \ldots, T_N) \) and \( \lambda \) in Section VI.

IV. MARKOVIAN SYSTEM MODEL: SINGLE PROCESS

Here, we derive the optimal channel probing, source activation, and data transmission policy for an EH source sampling a single process with a Markovian channel and Markovian energy arrival process. We formulate the \( \alpha \)-discounted cost version of (1) as an MDP with state space \( S = \{0, 1, \ldots, B\} \times Z_+ \times Z_+ \times \{C_1, C_2, \ldots, C_m\} \times \{H_1, H_2\} \) where a generic primary state \( s = (E, T, \tau, C_{\text{prev}}, H_1) \) means that the energy buffer has \( E \) energy packets, the last successfully received packet was generated at the source node \( T \) slots ago, the channel state was last probed \( \tau \) slots ago where \( \tau \leq T \), the previous channel state obtained via probing is \( C_{\text{prev}} \in \{C_1, C_2, \ldots, C_m\} \), and the EH source is in harvesting state \( H_1 \). A generic primary state \( s = (E, T, \tau, C_{\text{prev}}, H_2) \) similarly means that the EH source is in non-harvesting state \( H_2 \). Also, \( V \) denotes intermediate state space where a generic intermediate state \( v = (E, T, C, H_1) \) or \( v = (E, T, C, H_2) \) is used for current channel state \( C \) (obtained after probing). The action space \( \mathcal{A} = \{0, 1\} \cup \{1 \times 0, 1, 1, 1\} \) with \( a(t), b(t) \in \{0, 1\} \). At each time, if \((b(t) = 0, a(t) = 0)\), then the expected single-stage AoI cost is \( c(s(t), b(t), a(t)) = T \).

However, if \((b(t) = 1)\), the expected single-stage AoI cost is \( c(s(t), b(t) = 1, a(t) = 0) = T \), and \( c(s(t), b(t) = 1, a(t) = 1) = T(1 - p(C)) \).

A. Optimality Equation

In this case, for primary state \((E, T, \tau, C_{\text{prev}}, H_1)\), the Bellman equations are given by:

\[
J^*(E \geq E_p + E_s, T, \tau, C_{\text{prev}}, H_1) = \min \left\{ T + \alpha E_A[H_1, H_{\text{next}}|H_1] J^*(\min\{E + A, B\}, T + 1, \tau + 1, C_{\text{prev}}, H_{\text{next}}), V^*(E, T, \tau, C_{\text{prev}}, H_1) \right\}
\]

\[
V^*(E, T, \tau, C_{\text{prev}}, H_1) = \mathbb{E}_{C \sim q^{(\tau)}(\cdot|C_{\text{prev}})} W^*(E, T, C, H_1)
\]

\[
W^*(E, T, C, H_1) = \min \left\{ T + \alpha E_A[H_1, H_{\text{next}}|H_1] J^*(\min\{E - E_p - E_s + A, B\}, T + 1, 1, C, H_{\text{next}}) + \alpha p(C) E_A[H_1, H_{\text{next}}|H_1] J^*(\min\{E - E_p - E_s + A, B\}, T + 1, 1, C, H_{\text{next}}) \right\}
\]

\[
J^*(E < E_p + E_s, T, \tau, C_{\text{prev}}, H_2) = T + \alpha E_A[H_1, H_{\text{next}}|H_1] J^*(\min\{E + A, B\}, T + 1, 1, C_{\text{prev}}, H_{\text{next}})
\]

(7)

The first expression in the minimization in the R.H.S. of the first equation in (7) is the cost when \( b(t) = 0 \). The quantity \( V^*(E, T, \tau, C_{\text{prev}}, H_1) \) is the expected cost when \( b(t) = 1 \), which explains the second equation in (7). Here \( q^{(\tau)} \) denote the \( \tau \)-step transition probability of the channel dynamics, starting from \( C_{\text{prev}} \).

Similarly, in (7), if we consider \( H_2 \) for primary state \((E, T, \tau, C_{\text{prev}}, H_2)\) then the same Bellman equations hold with \( A = 0 \).

B. Policy Structure

We conjecture the following properties regarding the value function and the optimal policy. These conjectures are supported by numerical results which are given in Section VI-C.

\[
J^*(E \geq E_p + E_s, T_1, T_2, \ldots, T_N) = \min_{i=1}^{N} T_i + \alpha E_A J^*(\min\{E + A, B\}, T_1 + 1, T_2 + 1, \ldots, T_N + 1),
\]

\[
E_C \left( \min_{i=1}^{N} T_i + \alpha E_A J^*(\min\{E - E_p - E_s + A, B\}, T_1 + 1, T_2 + 1, \ldots, T_N + 1) \right)
\]

\[
\min_{i=1}^{N} \left\{ T_i + \sum_{b \neq a} (1 - \alpha p(C)) E_A J^*(\min\{E - E_p - E_s + A, B\}, T_i + 1, T_{i+1} + 1, \ldots, T_N + 1) + \alpha (1 - p(C)) E_A J^*(\min\{E - E_p - E_s + A, B\}, T_i + 1, T_{i+1} + 1, \ldots, T_N + 1) \right\}
\]

\[
J^*(E < E_p + E_s, T_1, T_2, \ldots, T_N) = \sum_{i=1}^{N} T_i + \alpha E_A J^*(\min\{E + A, B\}, T_1 + 1, T_2 + 1, \ldots, T_N + 1)
\]

(6)
Conjecture 3: For the Markovian system model and $N = 1$, the value function $J^*(E, T, \tau, C_{prev}, H)$ is increasing in $T$ and $W^*(E, T, C, H)$ is decreasing in $p(C)$.

Conjecture 4: For the Markovian system model and $N = 1$, the optimal probing policy for the $\alpha$-discounted AoI cost minimization problem is a threshold policy on $T$. For any $E \geq E_p + E_s$, the optimal action is to probe the channel state if and only if $T \geq T_{th}(E, \tau, C_{prev}, H)$ for a threshold function $T_{th}(E, \tau, C_{prev}, H)$ of $E, \tau, C_{prev}, H$.

Conjecture 5: For the Markovian system model and $N = 1$, at any time, if the source decides to probe the channel, then the optimal sampling policy is a threshold policy on $p(C)$. For any $E \geq E_p + E_s$ and probing channel state, the optimal action is to sample the source node if and only if $p(C) \geq p_{th}(E, T, H)$ for a threshold function $p_{th}(E, T, H)$ of $E, T$ and $H$.

V. REINFORCEMENT LEARNING FOR AOI OPTIMIZATION: SINGLE PROCESS

If the channel statistics and the energy harvesting characteristics are not known at the source, then the source has to learn them with time. We adapt the Q-learning technique [51, Sec. 11.4.2] to find the optimal policy for our two-stage action model.

A. I.I.D. System

Here, we assume that the energy generation rate $\lambda$ and the channel state probabilities $\{q_j\}_{1 \leq j \leq m}$ are not known to the source. The packet success probabilities $\{p_j\}_{1 \leq j \leq m}$ are either known or unknown; the packet success indicator $r(t)$ is used to estimate $p_j$ if it is unknown.

1) Optimality Equation in Terms of Q-Function: The optimal Q-function is denoted by $Q^*(E, T, b)$ for a generic primary state-action pair $(E, T, b)$, where the primary state is $(E, T)$ and the action is $b \in \{0, 1\}$. The quantity $Q^*(E, T, b)$ denotes the expected cost obtained if the current primary state is $(E, T)$, the current action chosen is $b$, and an optimal policy is employed from the next time instant. Also, $Q^*(E, T, C, a)$ corresponds to the intermediate state $(E, T, C)$ and a corresponding action $a \in \{0, 1\}$. It is important to note that, we have to maintain two different types of Q-functions to handle primary states and intermediate states; this is not done in standard Q-learning. The optimal Q-value functions are given by:

$$Q^*(E \geq E_p + E_s, T, b = 0) = T + \alpha\mathbb{E}_A \min_{b \in \{0, 1\}} Q^*(\min\{E + A, B\}, T + 1, b)$$

(8)

Here, the terms in the R.H.S of the first equality operator in (8) is the immediate cost of not probing channel state action ($b = 0$) when primary state is $(E \geq E_p + E_s, T)$ and an $\alpha$ discounted future cost $\mathbb{E}_A J^*(\min\{E + A, B\}, T + 1)$ which is later substituted by $\mathbb{E}_A \min_{b \in \{0, 1\}} Q^*(\min\{E + A, B\}, T + 1, b)$ (averaged over the distribution of $A$). On the other hand,

$$Q^*(E \geq E_p + E_s, T, b = 1) = \sum_{j=1}^{m} q_j W^*(E, T, C_j)$$

(9)

Now,

$$Q^*(E, T, C, a = 0) = T + \alpha\mathbb{E}_A \min_{b \in \{0, 1\}} Q^*(\min\{E - E_p + A, B\}, T + 1, b) = J^*(\min\{E - E_p + A, B\}, T + 1)$$

(10)

and

$$Q^*(E, T, C, a = 1) = T(1 - p(C)) + \alpha p(C)\mathbb{E}_A \min_{b \in \{0, 1\}} Q^*(\min\{E - E_p - E_s + A, B\}, 1, b) + \alpha(1 - p(C))\mathbb{E}_A \min_{b \in \{0, 1\}} Q^*(\min\{E - E_p - E_s + A, B\}, T + 1, b)$$

(11)

Similarly, the optimal Q-function for primary state $(E < E_p + E_s, T)$ is given by (12) in which only not probing ($b = 0$) is a feasible action:

$$Q^*(E < E_p + E_s, T, b = 0) = T + \alpha\mathbb{E}_A J^*(\min\{E + A, B\}, T + 1)$$

(12)

2) Q-Learning Algorithm: The optimal Q-function is given by the zeros of (8)-(12). Here, we employ asynchronous stochastic approximation [40] to iteratively converge to the optimal Q-function, starting from any arbitrary Q-function. Online learning is facilitated by sequentially observing $C(t)$ (via channel probing), $r(t)$ (via transmission ACK/NACK), and the number $A(t)$ of energy packets harvested at time $t$.

In asynchronous stochastic approximation, we need step size sequence $d(t)$ ($t \geq 0$) which satisfies the following assumptions [52]:

Assumption 1: (i) $0 < d(t) \leq \bar{d}$ where $\bar{d} > 0$, (ii) $d(t + 1) \leq d(t)$ for all $t \geq 0$, (iii) $\sum_{t=0}^{\infty} d(t) = \infty$ and $\sum_{t=0}^{\infty} d(t)^2 < \infty$.

Let us denote by $\nu_t(s, b)$ the number of occurrences of the primary state-action pair $(s, b)$ up to iteration $t$ where $s \in S$ and $b \in \{0, 1\}$. A similar notation is assumed for any generic intermediate state $v$ and the corresponding action $a$. The following assumption is required for the convergence of asynchronous stochastic approximation [52]:

Assumption 2: $\lim inf_{t \to \infty} \frac{\nu_t(s, b)}{t} > 0$ almost surely $\forall (s, b)$, and $\lim inf_{t \to \infty} \frac{\nu_t(v, a)}{t} > 0$ almost surely $\forall (v, a)$.

The proposed algorithm maintains a look-up table $Q_t(\cdot, \cdot)$ for various state-action pairs, and iteratively updates its entries depending on the current state, the current action taken and the observed next state. However, Assumption 2 requires that all state-action pairs should be visited infinitely and comparatively often. This is ensured by taking a random action (uniformly chosen) with probability $\mathbb{E}$ at each decision instant, and taking the action $\arg \min_b Q(s, b)$ or $\arg \min_a Q(v, a)$ with probability $1 - \mathbb{E}$. 

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
One example of a Q-value update is the following, aimed at convergence to the solution of (8):

\[ Q_{t+1}(E \geq E_p + E_s, T, b = 0) = Q_t(E, T, b = 0) + d(\nu_t(E, T, b = 0)) \mathbf{1}_{s(t) = (E, T), b(t) = 0} \]

\[ + T + \alpha \min_{b \in \{0,1\}} Q_t(\min\{E + A(t), B\}, T + 1, b) - Q_t(E, T, b = 0) \]  

(13)

Here, \( \mathbf{1}_{\cdot} \) is the indicator function. It is important to note that, this Q-update can be performed only when the random next state (\( \min\{E + A(t), B\} \), \( T + 1 \)) is observed.

Similarly, (9)-(12) lead to the following Q-updates for various primary states and intermediate states:

\[ Q_{t+1}(E \geq E_p + E_s, T, b = 1) = Q_t(E, T, b = 1) + d(\nu_t(E, T, b = 1)) \mathbf{1}_{s(t) = (E, T), b(t) = 1} \]

\[ + \min_{a,0 \leq \alpha \leq 1} Q_t(E, T, C, a) - Q_t(E, T, b = 1) \]  

(14)

\[ Q_{t+1}(E, T, C, a = 0) = Q_t(E, T, C, a = 0) + d(\nu_t(E, T, C, a = 0)) \mathbf{1}_{s(t) = (E, T, C), a(t) = 0} \]

\[ + T + \alpha \min_{b \in \{0,1\}} Q_t(\min\{E - E_p + A(t), B\}, T + 1, b) - Q_t(E, T, C, a = 0) \]  

(15)

\[ Q_{t+1}(E, T, C, a = 1) = Q_t(E, T, C, a = 1) + d(\nu_t(E, T, C, a = 1)) \mathbf{1}_{s(t) = (E, T, C), a(t) = 1} \]

\[ + \min_{b \in \{0,1\}} Q_t(E, T, C, a = 1) - Q_t(E, T, C, a = 1) \]  

(16)

\[ Q_{t+1}(E < E_p + E_s, T, b = 0) = Q_t(E, T, b = 0) + d(\nu_t(E, T, b = 0)) \mathbf{1}_{s(t) = (E, T), b(t) = 0} \]

\[ + T + \alpha \min_{b \in \{0,1\}} Q_t(\min\{E + A(t), B\}, T + 1, b) - Q_t(E, T, b = 0) \]  

(17)

In (16), we consider the case where packet success probability \( p(C) \) for channel state \( C \) is known to the source. However, if these probabilities are unknown, then one can replace \( p(C) \) in (16) by \( r(t) \) to obtain a valid Q-update.

Q-learning type algorithm for \( N > 1 \) can also be formulated similarly.

### B. Markovian System

Here, we adapt Q-learning for the Markovian model with unknown channel statistics and unknown energy harvesting characteristics to obtain the optimal solutions for (7).

#### 1) Optimality Equation in Terms of Q-Function: The optimal Q-value functions are given by:

\[ Q^*(E \geq E_p + E_s, T, \tau, C_{prev}, H_1, b = 0) = T + \alpha \mathbb{E}_{A[H_1,H_{next}[H_1]} J^*(\min\{E + A, B\}, T + 1, \tau + 1, C_{prev}, H_{next}) \]

\[ T + \alpha \mathbb{E}_{A[H_1,H_{next}[H_1]} \min_{b \in \{0,1\}} Q^*(\min\{E + A, B\}, T + 1, \tau + 1, C_{prev}, H_{next}, b) \]  

(18)

On the other hand,

\[ Q^*(E \geq E_p + E_s, T, \tau, C_{prev}, H_1, b = 1) = \mathbb{E}_{C \sim q(\cdot)} W^*(E, T, C, H_1) \]

\[ = \mathbb{E}_{C \sim q(\cdot)} \min_{a \in \{0,1\}} Q^*(E, T, C, H_1, a) \]  

(19)

When \( b(t) = 1 \), for a current measured channel state \( C \), the optimal \( Q^* \)-functions for action \( a = 0 \) and \( a = 1 \) are given by:

\[ Q^*(E, T, C, H_1, a = 0) = T + \alpha \mathbb{E}_{A[H_1,H_{next}[H_1]} \min_{b \in \{0,1\}} Q^*(\min\{E - E_p + A, B\}, \]

\[ T + 1, 1, C, H_{next}, b) \]  

(20)

\[ Q^*(E, T, C, H_1, a = 1) = T(1 - p(C)) + \alpha p(C) \mathbb{E}_{A[H_1,H_{next}[H_1]} \min_{b \in \{0,1\}} Q^*(\min\{E - E_p - E_s + A, B\}, \]

\[ T + 1, 1, C, H_{next}, b) \]  

(21)

Also, the optimal Q-function for primary state \( (E < E_p + E_s, T, \tau, C_{prev}, H_1) \) is given by:

\[ Q^*(E < E_p + E_s, T, \tau, C_{prev}, H_1, b = 0) = T + \alpha \mathbb{E}_{A[H_1,H_{next}[H_1]} \min_{b \in \{0,1\}} Q^*(\min\{E + A, B\}, \]

\[ T + 1, \tau + 1, C_{prev}, H_{next}, b) \]  

(22)

Similarly, the Q-functions are given for primary state \( (E, T, \tau, C_{prev}, H_2) \) by putting \( A = 0 \).

2) Q-Learning Algorithm: Here also, we employ asynchronous stochastic approximation to find the solution of (18)-(22) as in Section V-A. The Q-updates for various primary states and intermediate states are given by the following:

\[ Q_{t+1}(E \geq E_p + E_s, T, \tau, C_{prev}, H_1, b = 0) = Q_t(E, T, \tau, C_{prev}, H_1, b = 0) + d(\nu_t(E, T, \tau, C_{prev}, H_1, b = 0)) \]

\[ \mathbf{1}_{s(t) = (E, T, \tau, C_{prev}, H_1), b(t) = 0} \]  

\[ + T + \alpha \min_{b \in \{0,1\}} Q_t(\min\{E - E_p + A(t), B\}, \]

\[ T + 1, \tau + 1, C_{prev}, H_{next}, b) - Q_t(E, T, \tau, C_{prev}, H_1, \]

\[ b = 0) \]  

(23)

\[ Q_{t+1}(E \geq E_p + E_s, T, \tau, C_{prev}, H_1, b = 1) = Q_t(E, T, \tau, C_{prev}, H_1, b = 1) + d(\nu_t(E, T, \tau, C_{prev}, H_1, b = 1)) \]

\[ \mathbf{1}_{s(t) = (E, T, \tau, C_{prev}, H_1), b(t) = 1} \]  

\[ + T + \alpha \min_{b \in \{0,1\}} Q_t(\min\{E - E_p - A(t), B\), \]

\[ T + 1, \tau + 1, C_{prev}, H_{next}, b) - Q_t(E, T, \tau, C_{prev}, H_1, \]

\[ b = 1) \]  

(24)

\[ Q_{t+1}(E < E_p + E_s, T, \tau, C_{prev}, H_1, b = 0) = Q_t(E, T, \tau, C_{prev}, H_1, b = 0) + d(\nu_t(E, T, \tau, C_{prev}, H_1, b = 0)) \]

\[ \mathbf{1}_{s(t) = (E, T, \tau, C_{prev}, H_1), b(t) = 0} \]  

\[ + T + \alpha \min_{b \in \{0,1\}} Q_t(\min\{E - E_p + A(t), B\), \]

\[ T + 1, \tau + 1, C_{prev}, H_{next}, b) - Q_t(E, T, \tau, C_{prev}, H_1, \]

\[ b = 0) \]  

(25)

\[ Q_{t+1}(E, T, C, H_1, a = 0) = Q_t(E, T, C, H_1, a = 0) + d(\nu_t(E, T, C, H_1, a = 0)) \mathbf{1}_{s(t) = (E, T, C, H_1, a = 0) \}

\[ + T + \alpha \min_{b \in \{0,1\}} Q_t(\min\{E - E_p + A(t), B\), \]

\[ T + 1, 1, C, H_{next}, b) - Q_t(E, T, C, H_1, a = 0) \]  

(26)

\[ Q_{t+1}(E, T, C, H_1, a = 1) = Q_t(E, T, C, H_1, a = 1) + d(\nu_t(E, T, C, H_1, a = 1)) \mathbf{1}_{s(t) = (E, T, C, H_1, a = 1) \}

\[ + T + \alpha \min_{b \in \{0,1\}} Q_t(\min\{E - E_p - A(t), B\), \]

\[ T + 1, 1, C, H_{next}, b) - Q_t(E, T, C, H_1, a = 1) \]  

(27)
min\{E - E_p - E_s + A(t), B\}, T + 1, 1, C, H_{arrival}, b\) - \(Q_i(E, τ, C, H, a = 1)\)

\[Q_i(E, τ, C, H, b = 0) = Q_i(E, τ, C, H, b = 0) + \sum_{k=1}^{\infty} \alpha_k Q_i(E, τ, C, H, b = 0)\]

\[1(s(t) = E, τ, C, H, b = 0) + \sum_{k=1}^{\infty} \alpha_k Q_i(E, τ, C, H, b = 0)\]

Similarly, the Q-value function iteration updates are given for primary state \(E, τ, C, H\) by taking \(A = 0\). Also, when \(p(C)\) is not known, it can be replaced by \(r(t)\) in (26).

VI. NUMERICAL RESULTS

Note: For simulation purpose, we assume that the maximum age of a process(es) is upper bounded by \(T_{max} = 100\).

A. Single Process, I.I.D. System, Known Dynamics

We consider five channel states \((m = 5)\) with channel state occurrence probabilities \(q = [0.2, 0.2, 0.2, 0.2, 0.2]\) and the corresponding packet success probabilities \(p = [0.9, 0.7, 0.5, 0.3, 0.1]\). Energy arrival process is i.i.d. Bernoulli(\(λ\)) with energy buffer size \(B = 12\), \(E_p = 1\) unit, \(E_s = 1\) unit [34]. Numerical exploration reveals that there exists a threshold policy on \(T\) in decision-making for channel state probing; see Fig. 2(a) which substantiates our conjecture in Section III-A2. It is observed that this \(T_{th}(E)\) decreases with \(E\) since higher available energy in the energy buffer allows the EH node to probe the channel state more aggressively. Similar reasoning explains the observation that \(T_{th}(E)\) decreases with \(λ\). For probed channel state, Fig. 2(b) shows the variation of \(p_{th}(E, T)\) with \(E, T, λ\). It is observed that \(p_{th}(E, T)\) decreases with \(E\) since the EH node tries to sample the process more aggressively if more energy is available in the buffer. Similarly, a higher value of \(T\) results in aggressive sampling, and hence \(p_{th}(E, T)\) decreases with \(T\). By similar arguments as before, we can explain the observation that this \(p_{th}(E, T)\) decreases with \(λ\).

Further, rigorous numerical analysis also reveals that the optimal decision rules for probing and sampling are threshold rules on \(E\), which we omit in this paper due to lack of space.

B. Multiple Processes \((N > 1)\), I.I.D. Model, Known Dynamics

We choose \(N = 3, B = 12, E_p = 1, E_s = 1\), and the same channel model and parameters as in Section VI-A. Fig. 3(a) provides support to our conjecture for optimal probing policy in Section III-B2 by demonstrating the variation on the threshold on \(T_1\) for given \(T_2, T_3\), for channel state probing. It is observed that \(T_{th}(E, T_2, T_3)\) decreases with \(E\) and \(λ\). Extensive numerical work also demonstrated that for the same \(λ\), this threshold decreases with each of \(T_2, T_3\), since an increase in the age of processes results in more frequent channel probing. For probed channel state, Fig. 3(b) shows that \(p_{th}(E, T_1, T_2, T_3)\) decreases with \(E\) and \(λ\). Further, the numerical analysis also demonstrated that for the same \(λ\), this threshold decreases with each of \(T_1, T_2, T_3\), since the source node becomes more aggressive in sampling.

C. Markovian Model, \(N = 1\), Known Dynamics

For numerical work, we have considered a two-state \((m = 2)\) Gilbert-Eliot channel model [53], [54] since it is practical as well as results in low computational complexity, and has been used in many recent papers involving AoI and EH source [8], [46], [55], [56]. While the Gilbert-Eliot channel model is a classic model, it is also highly relevant to millimeter-wave (mm-wave) communication which is a key driver for 5G and...
beyond wireless communications. Due to small wavelength, narrow beamwidth, and large propagation loss, mm-wave signals are highly susceptible to blocking and the channel can rapidly switch between on and off states due to the movement of obstacles and reflectors [55], [57], [58], [59]. Thus, the Gilbert-Eliot channel model can be used for mm-wave communication channels with the “On” and “Off” states representing Line of Sight (LOS) and Non-Line of Sight conditions, respectively. However, the procedure described in this paper can be used to numerically analyze any finite $m$-state channel model, and similar results can be obtained there too.

Here, we consider a two-state Markovian fading channel ($m = 2$), where $C_1$ is the good channel state and $C_2$ is the bad channel state. The channel state transition probabilities are $q_{1,2} = 0.1$ and $q_{2,1} = 0.1$ [8], [46], and the corresponding packet transmission success probabilities are $p_1 = 0.9$ and $p_2 = 0.4$. Also, for the Markovian energy harvesting process, we consider the transition probabilities $h_{1,2} = 0.3$ and $h_{2,1} = 0.3$ [28], [48]. When the source node is in a harvesting state, the energy packet generation process is i.i.d. Bernoulli($\lambda$) across time, and otherwise zero. We consider $\lambda = 9$, $E_p = 1$ unit, $E_s = 1$ unit, and $\lambda = 0.4$.

Our numerical exploration demonstrated the threshold nature of the optimal policy which supports conjectures predicted in Section IV-B. Fig. 4 shows that, for a fixed $\tau$, $T_{th}(E, \tau, C_{prev}, H)$ decreases with $E$, $T_{th}(E, \tau, C_{prev}, H_1) \leq T_{th}(E, \tau, C_{prev}, H_2)$ and $T_{th}(E, \tau, C_1, H) \leq T_{th}(E, \tau, C_2, H)$ (because probability of packet success in channel state $C_1$ is higher than that in $C_2$). Also, from Fig. 5 it is observed that for probed channel state, $p_{th}(E, T, H)$ decreases with $E$ and $T$, and $p_{th}(E, T, H_1) \leq p_{th}(E, T, H_2)$.

Also, numerical exploration revealed that, after probing is done, the optimal sampling policy can also be represented as a threshold policy on $T$ instead of $p(C)$. In this case, sampling is done when $T \geq T_{th}(E, C, H)$. The variation of $T_{th}(E, C, H)$ with $E, C, H$ is shown in Fig. 6, which supports many intuitive justifications provided earlier in this section. Similar result as in Fig. 6 can also be obtained for the i.i.d. system model which we omit due to space constraint.

D. Comparison Between Probing and Not Probing

1) I.I.D. System: We consider $N = 1$ and the same model as in Section VI-A, and compare the performance of our probing-based threshold policy against the optimal policy that involves sampling mandatorily without probing (as in Section III-A3) based on $T$ and $E$. The results are summarised in Fig. 7 (a); we note that AoI increases with $E_p$ and $E_s$ since each action becomes more costly for the source. One interesting observation is that when the sensing energy is close to probing energy ($E_s = 1$, $E_p = 1$), we see slightly higher time-averaged AoI values for the probing case compared to the case with $E_s = 1$ and no probing. This happens because the advantage offered by probing is being overshadowed by the extra amount of energy expended for it. However, when $E_s = 5$, $E_p = 1$, we notice that probing yields a lower time-averaged AoI. Obviously, the importance of probing is dependent on the trade-off between the advantage offered by probing and the extra energy required for it. Typically, $E_s$ is much smaller than $E_s$ since probing requires less energy as compared to sampling and transmitting a data packet. We have also observed numerically that, when there is sufficient variation in the channel quality, probing improves the AoI performance. On the other hand, when channel variation is not predominant, spending energy in probing tends to hit the AoI performance and can sometimes render it worse than no probing.

2) Markovian System: We consider the same system parameters as in Section VI-C. Here also, Fig. 7 (b) exhibits a
similar trade-off between the improvement offered by probing and the energy spent in it, as seen for the i.i.d. model.

E. Q-Learning

We consider the same settings as Sections VI-A and VI-C, except that we set $B = 5$ and an upper bound on the age as $T_{\text{max}} = 7$, in order to reduce the number of possible states. We then run an online Q-learning scheme across two different sample paths. For the sake of comparison, we also plot the time-averaged AoI of a uniform random strategy as well as that of the optimal policy derived through value iteration. Convergence of the time-averaged AoI under Q-learning to the optimal AoI for the i.i.d. channel system and the Markovian model can be seen in Fig. 8 (a) and Fig. 8 (b), respectively. We also observe significant improvement from the initial policy which always picks an action uniformly at random; this demonstrates the necessity of the MDP-based formulation for probing and sampling.

VII. CONCLUSION

In this paper, we have derived optimal policy structures for minimizing the time-averaged expected AoI under an energy-harvesting source. We considered single and multiple processes, i.i.d. and Markovian time-varying channels, time-varying characteristics for energy harvesting, and channel probing capability at the source. The optimal source sampling policy often turned out to be a threshold policy. We have also proposed RL algorithms for unknown channel statistics and energy harvesting characteristics. Numerical results demonstrated the power of channel probing, and also the performances and trade-offs for various algorithms. However, there are a number of issues that remain untouched, such as multi-source star topology and multi-hop network settings with multiple source and sink nodes. We plan to address these issues in our future research.

APPENDIX A

PROOF OF PROPOSITION I

We prove that $\max_{s \in S} |J^t(s) - J^*| \uparrow 0$ as $t \uparrow \infty$. Let us define the error $\epsilon_t = \max_{s \in S} |J^t(s) - J^*(s)|$ and $J^{(0)}(s)$ as initial estimate for $J^*(s)$.

For any state $s$, we seek to establish a relation between the error at time $t + 1$ to the error at time $t$.

\[
\begin{align*}
J^{(t+1)}(E \geq E_p + E_s, T) &= \min \left\{ T + \alpha E_A J^t(\min\{E - E_p + A, B\}, T + 1), \right. \nonumber \\
& \quad T(1 - p(C)) + \alpha p(C) E_A J^t(\min\{E - E_p - E_s + A, B\}, 1) \\
& \left. \quad + \alpha(1 - p(C)) E_A J^t(\min\{E - E_p - E_s + A, B\}, T + 1)\right\} \\
& = \min \left\{ T + \alpha E_A J^t(\min\{E - E_p + A, B\}, T + 1), \right. \nonumber \\
& \quad T(1 - p(C)) + \alpha p(C) E_A J^t(\min\{E - E_p - E_s + A, B\}, 1) \\
& \left. \quad + \alpha(1 - p(C)) E_A J^t(\min\{E - E_p - E_s + A, B\}, T + 1)\right\} \\
& \leq \min \left\{ T + \alpha E_A J^t(\min\{E - E_p + A, B\}, T + 1) + \epsilon_t, \right. \nonumber \\
& \quad T(1 - p(C)) + \alpha p(C) E_A J^t(\min\{E - E_p - E_s + A, B\}, 1) + \epsilon_t \\
& \left. \quad + \alpha(1 - p(C)) E_A J^t(\min\{E - E_p - E_s + A, B\}, T + 1) + \epsilon_t\right\}
\end{align*}
\]

We assume there exists an optimal value function $J^*(E, T)$ for the discounted cost problem. By using the fact that $J^t(E, T) \leq J^*(E, T) + \epsilon_t$ in (28), we obtain:

\[
\begin{align*}
J^{(t+1)}(E \geq E_p + E_s, T) & \leq \min \left\{ T + \alpha E_A J^t(\min\{E - E_p + A, B\}, T + 1) + \epsilon_t, \right. \nonumber \\
& \quad T(1 - p(C)) + \alpha p(C) E_A J^t(\min\{E - E_p - E_s + A, B\}, 1) + \epsilon_t \\
& \left. \quad + \alpha(1 - p(C)) E_A J^t(\min\{E - E_p - E_s + A, B\}, T + 1) + \epsilon_t\right\}
\end{align*}
\]

(29)
Similarly, using $J^{(t)}(E, T)$ by $J^{(t)}(E, T) \geq J^{(*)}(E, T) - \epsilon_t$ in (28), we obtain:

$$J^{(t+1)}(E \geq E_p + E_s, T) \geq \min \left\{ T + \alpha E_A J^{(1)}(\min \left\{ E + A, B \right\}, T + 1) - \epsilon_t, \right.$$ \[
E_C \left( \min \left\{ T + \alpha E_A J^{(1)}(\min \left\{ E - E_p + A, B \right\}, T + 1) - \epsilon_t, \right. \right.
\left. \left. T(1 - p(C)) + \alpha p(C) E_A J^{(*)}(\min \left\{ E - E_p - E_s + A, B \right\}, 1) + \alpha(1 - p(C)) E_A J^{(*)}(\min \left\{ E - E_p - E_s + A, B \right\}, T + 1) - \epsilon_t \right) \right\} \right) \]

Combining the results obtained from equations (29) and (30), we get:

$$J^{(*)}(E \geq E_p + E_s, T) - \alpha \epsilon_t \leq J^{(t+1)}(E \geq E_p + E_s, T) \leq J^{(*)}(E \geq E_p + E_s, T) + \alpha \epsilon_t$$
$$\Rightarrow \max_{s \in S} J^{(t+1)}(s) - J^{(*)}(s) \leq \alpha \epsilon_t$$
$$\Rightarrow \lim_{t \to \infty} \epsilon_t = 0$$

Hence, $J^{(t)}(s)$ converges to $J^{(*)}(s)$.

**APPENDIX B**
**PROOF OF LEMMA 1**

We prove this result by value iteration:

$$J^{(t+1)}(E \geq E_p + E_s, T)$$
$$= \min \left\{ T + \alpha E_A J^{(1)}(\min \left\{ E + A, B \right\}, T + 1), E_C \left( \min \left\{ T + \alpha E_A J^{(1)}(\min \left\{ E - E_p + A, B \right\}, T + 1), T(1 - p(C)) + \alpha p(C) E_A J^{(*)}(\min \left\{ E - E_p - E_s + A, B \right\}, 1) + \alpha(1 - p(C)) E_A J^{(*)}(\min \left\{ E - E_p - E_s + A, B \right\}, T + 1) \right) \right\} \right)$$

(32)

Let us start with $J^{(0)}(s) = 0$ for all $s \in S$. Clearly, $J^{(1)}(E \geq E_p + E_s, T) = \min \{ T, E_C(\min \{ T, T(1 - p(C)) \}) \} = \min \{ T, E_C(\min \{ T, T(1 - p(C)) \}) \}$ and $J^{(1)}(E < E_p + E_s, T) = T$. Hence, for any given $E$, the value function $J^{(1)}(E, T)$ is an increasing function of $T$. As an induction hypothesis, we assume that $J^{(1)}(E, T)$ is also an increasing function of $T$. Now, $J^{(t+1)}(E \geq E_p + E_s, T)$ is also increasing in $T$. The first term inside the minimization operation in (33) is increasing in $T$, by the induction hypothesis and from the fact that expectation is a linear operator. On the other hand, the second term has linear expectation over the channel state and another minimization operator. Also, the first and second terms inside the second minimization operation in (33) are increasing in $T$ by the induction hypothesis and the linearity of expectation operation. Thus, $J^{(t+1)}(E \geq E_p + E_s, T)$ is also increasing in $T$. By similar arguments, we can claim that $J^{(t+1)}(E < E_p + E_s, T)$ is increasing in $T$. Now, since $J^{(t)}(\cdot) \uparrow J^{(*)}(\cdot)$ as $t \uparrow \infty$ by proof of Proposition 1, $J^{(*)}(E, T)$ is also increasing in $T$. Hence, the first part of the lemma is proved.

For current probed channel state, the value function $W^{(t+1)}(E, T, C)$ is given by:

$$W^{(t+1)}(E, T, C)$$
$$= T + \alpha E_A J^{(1)}(\min \{ E - E_p + A, B \}, T + 1),$$
$$= T(1 - p(C)) + \alpha p(C) E_A J^{(*)}(\min \{ E - E_p - E_s + A, B \}, 1) + \alpha(1 - p(C)) E_A J^{(*)}(\min \{ E - E_p - E_s + A, B \}, T + 1)$$

(34)

The first term (cost of not sampling a source) inside the minimization operation in (34) is independent of $p(C)$, whereas the second term (cost of sampling a source) inside the minimization operation in (34) is given by:

$$T(1 - p(C)) + \alpha p(C) E_A J^{(*)}(\min \{ E - E_p - E_s + A, B \}, 1)$$
$$+ \alpha(1 - p(C)) E_A J^{(*)}(\min \{ E - E_p - E_s + A, B \}, T + 1) - p(C) \alpha E_A J^{(*)}(\min \{ E - E_p - E_s + A, B \}, 1) - E_A J^{(*)}(\min \{ E - E_p - E_s + A, B \}, T)$$

(35)

By the induction hypothesis and the linearity of expectation operation, $E_A J^{(*)}(\min \{ E - E_p - E_s + A, B \}, T) - E_A J^{(*)}(\min \{ E - E_p - E_s + A, B \}, T + 1)$ is non-negative. Thus, the second term inside the minimization operation in (35) is decreasing in $p(C)$. Now, since $W^{(*)}(\cdot) \uparrow W^{(*)}(\cdot)$ as $t \uparrow \infty$, $W^{(*)}(E, T, C)$ is also decreasing in $p(C)$.

**APPENDIX C**
**PROOF OF THEOREM 1**

From (3), it is obvious that for probed channel state the optimal decision for $E \geq E_p + E_s$ is to sample the source if and only if the cost of sampling is lower than the cost of not sampling the source, i.e., $T + \alpha E_A J^{(*)}(\min \{ E - E_p + A, B \}, T + 1) \geq \min \{ T, E_C(\min \{ T, T(1 - p(C)) \}) \} + \alpha(1 - p(C)) E_A J^{(*)}(\min \{ E - E_p - E_s + A, B \}, T + 1) - p(C) \alpha E_A J^{(*)}(\min \{ E - E_p - E_s + A, B \}, 1)$. Now, by Lemma 1, $E_A J^{(*)}(\min \{ E - E_p - E_s + A, B \}, T + 1) - E_A J^{(*)}(\min \{ E - E_p - E_s + A, B \}, 1)$ is non-negative. Thus the R.H.S. decreases with $p(C)$, whereas the L.H.S. is independent of $p(C)$. Hence, for probed channel state the optimal action is to sample if and only if $p(C) \geq p_{th}(E, T)$ for some suitable threshold function $p_{th}(E, T)$.

**APPENDIX D**
**PROOF OF LEMMA 2**

The proof is similar to the proof of Lemma 1 and it follows from the convergence of value iteration as given below:
\[ J^{(t+1)}(E \geq E_p + E_s, T_1, T_2, \ldots, T_N) = \min \left\{ \sum_{i=1}^{N} T_i + \alpha E_A J^{(i)}(\{E + A, B\}), T_1 + 1, T_2 + 1, \ldots, T_N + 1, \right\}_{1 \leq k \leq N} \]

\[ = \min \left\{ \sum_{i=1}^{N} T_i + \alpha E_A J^{(i)}(\{E - E_p + A, B\}), T_1 + 1, T_2 + 1, \ldots, T_N + 1, \right\}_{1 \leq k \leq N} \]

\[(T_1 + 1, T_2 + 1, \ldots, T_N + 1)\]

\[ J^{(t+1)}(E < E_p + E_s, T_1, T_2, \ldots, T_N) = \min \left\{ \sum_{i=1}^{N} T_i + \alpha E_A J^{(i)}(\{E - E_p + A, B\}), T_1 + 1, T_2 + 1, \ldots, T_N + 1, \right\}_{1 \leq k \leq N} \]

\[(T_1 + 1, T_2 + 1, \ldots, T_N + 1)\]

(36)

Let us start with \( J^{(0)}(s) = 0 \) for all \( s \in S \). Clearly, \( J^{(1)}(E \geq E_p + E_s, T_1, T_2, \ldots, T_N) = \min \left\{ \sum_{i=1}^{N} T_i, EC(\{E + A, B\}), T_1 + 1, T_2 + 1, \ldots, T_N + 1, \right\}_{1 \leq k \leq N} \)

\[ = \min \left\{ \sum_{i=1}^{N} T_i + \alpha E_A J^{(i)}(\{E + A, B\}), T_1 + 1, T_2 + 1, \ldots, T_N + 1, \right\}_{1 \leq k \leq N} \]

\[ (T_1 + 1, T_2 + 1, \ldots, T_N + 1) \]

(37)

We seek to show that \( J^{(t+1)}(E \geq E_p + E_s, T_1, T_2, \ldots, T_N) \) is also increasing in each of \( T_1, T_2, \ldots, T_N \). The first term inner to the minimization operation in (37) is increasing in each of \( T_1, T_2, \ldots, T_N \), utilizing the induction hypothesis and linear property of expectation operation. On the other hand, the second term has expectation over the channel state and another minimization operator. Also, first and second terms of the second minimization operator are increasing in each of \( T_1, T_2, \ldots, T_N \) by using the induction hypothesis and the linearity of expectation operation. Thus, \( J^{(t+1)}(E \geq E_p + E_s, T_1, T_2, \ldots, T_N) \) is also increasing in each of \( T_1, T_2, \ldots, T_N \). Similarly, we can assert that \( J^{(t+1)}(E < E_p + E_s, T_1, T_2, \ldots, T_N) \) is increasing in each of \( T_1, T_2, \ldots, T_N \). Now, since \( J^{(t)}(\cdot) \uparrow J^{*}(\cdot) \) as \( t \uparrow \infty \), \( J^{*}(E, T_1, T_2, \ldots, T_N) \) is also increasing in each of \( T_1, T_2, \ldots, T_N \). Hence, the first part of the lemma is proved. Proof of second part of Lemma 2 is similar to the proof of second part of Lemma 1.
M. Hatami, M. Leinonen, and M. Codreanu, “AoI minimization in status
Y. Khorsandmanesh, M. J. Emadi, and I. Krikidis, “Average peak age of
M. A. Abd-Elmagid and H. S. Dhillon, “Closed-form characterization
X. Zheng, S. Zhou, Z. Jiang, and Z. Niu, “Closed-form analysis
A. Arafa and S. Ulukus, “Age minimization in energy harvesting
2310 IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, VOL. 10, NO. 6, DECEMBER 2024
C. Tunc and S. Panwar, “Optimal transmission policies for energy
A. Arafa, J. Yang, and S. Ulukus, “Age-minimal online policies for
E. Gindullina, L. Badia, and D. Gündüz, “Age-of-information with
E. T. Ceran, D. Gunduz, and A. Gyorgy, “Learning to minimize age of
J. Yang, X. Wu, and J. Wu, “Optimal scheduling of collaborative sensing
M. A. Abd-Elmagid, H. S. Dhillon, and N. Pappas, “A reinforcement
S. Feng and J. Yang, “Age of information minimization for an energy
B. T. Bacinoglu, Y. Sun, E. Uysal-Biyikoglu, and V. Mutlu, “Achieving the age-energy tradeoff with a finite-battery energy har
S. Feng and J. Yang, “Age of information minimization for an energy
B. T. Bacinoglu, E. T. Ceran, and E. Uysal-Biyikoglu, “Age of information under energy replenishment constraints,” in Proc. Inf.
S. Leng and A. Yener, “Age of information minimization for an energy
B. T. Bacinoglu, Y. Sun, E. Uysal-Biyikoglu, and V. Mutlu, “Achieving the age-energy tradeoff with a finite-battery energy har
E. Gindullina, L. Badia, and D. Gündüz, “Age-of-information with information source diversity in an energy harvesting system,”
E. T. Ceran, D. Gunduz, and A. Gyorgy, “Learning to minimize age of information over an unreliable channel with energy harvesting,”
A. Arafa, J. Yang, and S. Ulukus, “Age-minimal online policies for energy harvesting sensors with random battery recharges,” in Proc. IEEE Int. Conf. Commun. (ICC), 2018, pp. 1–6.
C. Tunc and S. Panwar, “Optimal transmission policies for energy harvesting age of information systems with battery recovery,” in Proc. IEEE Int. Conf. Commun. (ICC), 2019, pp. 2012–2016.
B. T. Bacinoglu and E. Uysal-Biyikoglu, “Scheduling status updates to minimize age of information with an energy harvesting sensor,” in Proc. IEEE Int. Symp. Inf. Theory (ISIT), 2017, pp. 1122–1126.
A. Arafa, J. Yang, S. Ulukus, and H. V. Poor, “Timely status updating over erasure channels using an energy harvesting sensor: Single and multiple sources,” IEEE Trans. Green Commun. Netw., vol. 6, no. 1, pp. 6–19, Mar. 2022.
M. A. Abd-Elmagid, H. S. Dhillon, and N. Pappas, “A reinforcement learning framework for optimizing age of information in RF-powered communication systems,” IEEE Trans. Commun., vol. 68, no. 8, pp. 4747–4760, Aug. 2020.
M. Hatami, M. Leinonen, and M. Codreanu, “AoI minimization in status update control with energy harvesting sensors,” IEEE Trans. Commun., vol. 68, no. 7, pp. 3835–3851, Jul. 2020.
N. B. Chang and M. Liu, “Optimal channel probing and transmission scheduling for opportunistic spectrum access,” in Proc. 13th Annu. ACM Int. Conf. Mobile Comput. Netw., 2007, pp. 27–38.
A. Hentati, J.-F. Frigon, and W. Aijib, “Energy harvesting wireless sensor networks with channel estimation: Delay and packet loss performance analysis,” IEEE Trans. Veh. Technol., vol. 69, no. 2, pp. 1956–1969, Feb. 2020.
P. Chaporkar and A. Proutiere, “Optimal joint probing and transmission strategy for maximizing throughput in wireless systems,” IEEE J. Sel. Areas Commun., vol. 20, no. 12, pp. 1546–1551, Dec. 2002.
M. Johnston, E. Modiano, and I. Keslassy, “Channel probing in communication systems: Myopic policies are not always optimal,” in Proc. IEEE Int. Symp. Inf. Theory, 2013, pp. 1934–1938.
B. Yu, Y. Cai, X. Diao, and Y. Chen, “AoI minimization scheme for short packet communications in the energy-constrained IoT,” IEEE Internet Things J., vol. 10, no. 11, pp. 20138–20200, Nov. 2023.
V. S. Borkar, “Asynchronous stochastic approximations,” SIAM J. Control Optim., vol. 36, no. 3, pp. 840–851, 1998.
Z. Wang, D. Tisonet, S. Videv, and H. Haas, “On the design of a solar-panel receiver for optical wireless communications with simultaneous energy harvesting,” IEEE J. Sel. Areas Commun., vol. 33, no. 8, pp. 1612–1623, Aug. 2015.
S. Das, A. Sparks, E. Poves, S. Videv, J. Fakidis, and H. Haas, “Effect of sunlight on photovoltaics as optical wireless communication receivers,” J. Lightw. Technol., vol. 39, no. 19, pp. 6182–6190, Oct. 1, 2021.
S. Ulukus et al., “Energy harvesting wireless communications: A review of recent advances,” IEEE J. Sel. Areas Commun., vol. 33, no. 3, pp. 360–381, Mar. 2015.
H. S. Wang and N. Moayeri, “Finite-state Markov channel—a useful model for radio communication channels,” IEEE Trans. Veh. Technol., vol. 44, no. 1, pp. 163–171, Feb. 1995.
Q. Zhang and S. A. Kassam, “Finite-state Markov model for Rayleigh fading channels,” IEEE Trans. Commun., vol. 47, no. 11, pp. 1688–1692, Nov. 1999.
G. Yao, A. M. Bedewy, and N. B. Shroff, “Age-optimal low-power status update over time-correlated fading channel,” in Proc. IEEE Int. Symp. Inf. Theory (ISIT), 2021, pp. 2972–2977.
N. Michelusi, K. Stamatiou, and M. Zorzi, “Transmission policies for energy harvesting sensors with time-correlated energy supply,” IEEE Trans. Commun., vol. 61, no. 12, pp. 2988–3001, Dec. 2013.
B. Sombahu and S. Moharir, “Age-of-information aware scheduling under Markovian energy arrivals,” in Proc. Int. Conf. Signal Process. Commun. (SPCOM), 2020, pp. 1–5.
D. P. Bertsekas, Dynamic Programming and Optimal Control, vol. 2, 3rd ed. Belmont, MA, USA: Athena Sci., 2011.
D. P. Bertsekas, Dynamic Programming and Optimal Control, vol. 1, 3rd ed. Belmont, MA, USA: Athena Sci., 2005.
S. Bhatnagar, H. Prasad, and L. Prashanth, Stochastic Recursive Algorithms for Optimization: Simultaneous Perturbation Methods. London, U.K.: Springer, 2013, pp. 187–220.
S. Bhatnagar, “The Borkar–Meyn theorem for asynchronous stochastic approximations,” Syst. Control Lett., vol. 60, no. 7, pp. 472–478, 2011.
E. N. Gilbert, “Capacity of a burst-noise channel,” Bell Syst. Techn. J., vol. 39, no. 5, pp. 1253–1265, Sep. 1960.
M. Mushkin and I. Bar-David, “Capacity and coding for the Gilbert-Elliott channels,” IEEE Trans. Inf. Theory, vol. 35, no. 6, pp. 1277–1290, Nov. 1989.
J. Pan, A. M. Bedewy, Y. Sun, and N. B. Shroff, “Age-optimal scheduling over hybrid channels,” IEEE Trans. Mobile Comput., vol. 22, no. 12, pp. 7027–7043, Dec. 2023.
M. S. H. Abad, O. Ercetin, and D. Gündüz, “Channel sensing and communication over a time-correlated channel with an energy harvesting transmitter,” IEEE Trans. Green Commun. Netw., vol. 2, no. 1, pp. 114–126, Mar. 2018.
G. Yao, M. Hashemi, and N. B. Shroff, “Integrating sub-6 GHz and millimeter wave to combat blockage: Delay-optimal scheduling,” in Proc. Int. Symp. Model. Optim. Mobile, Ad Hoc, Wireless Netw. (WoMoPT), 2019, pp. 1–8.
A. Bildea, O. Alphand, F. Rousseau, and A. Duda, “Link quality estimation with the Gilbert-Elliott model for wireless sensor networks,” in Proc. IEEE 26th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC), 2015, pp. 2049–2054.
Z. He and S. Mao, “Adaptive multiple description coding and transmission of uncompressed video over 60GHz networks,” ACM SIGMOBILE Mobile Comput. Commun. Rev., vol. 18, no. 1, pp. 14–24, 2014.
E. Dias et al., “Millimeter-wave testbed and modeling in next generation URLLC communications,” in Proc. IEEE 24th Int. Symp. World Wireless, Mobile Multimedia Networks (WoWMoM), 2023, pp. 480–482.