Online Learning for Optimizing AoI-Energy Tradeoff under Unknown Channel Statistics

Mohamed A. Abd-Elmagid
Department of ECE
The Ohio State University
Columbus, OH, USA
abd-elmagid.1@osu.edu

Ming Shi
Department of EE
University at Buffalo
Buffalo, NY, USA
mshi24@buffalo.edu

Eylem Ekici
Department of ECE
The Ohio State University
Columbus, OH, USA
ekici.2@osu.edu

Ness B. Shroff
Departments of ECE and CSE
The Ohio State University
Columbus, OH, USA
shroff.11@osu.edu

Abstract

We consider a real-time monitoring system where a source node (with energy limitations) aims to keep the information status at a destination node as fresh as possible by scheduling status update transmissions over a set of channels. The freshness of information at the destination node is measured in terms of the Age of Information (AoI) metric. In this setting, a natural tradeoff exists between the transmission cost (or equivalently, energy consumption) of the source and the achievable AoI performance at the destination. This tradeoff has been optimized in the existing literature under the assumption of having a complete knowledge of the channel statistics. In this work, we develop online learning-based algorithms with finite-time guarantees that optimize this tradeoff in the practical scenario where the channel statistics are unknown to the scheduler. In particular, when the channel statistics are known, the optimal scheduling policy is first proven to have a threshold-based structure with respect to the value of AoI (i.e., it is optimal to drop updates when the AoI value is below some threshold). This key insight was then utilized to develop the proposed learning algorithms that surprisingly achieve an order-optimal regret (i.e., $O(1)$) with respect to the time horizon length.

CCS Concepts

• **Networks → Network performance evaluation**; *Network performance analysis.*

Keywords

Age of information, communication networks, online learning.

ACM Reference Format:
Mohamed A. Abd-Elmagid, Ming Shi, Eylem Ekici, and Ness B. Shroff. 2025. Online Learning for Optimizing AoI-Energy Tradeoff under Unknown Channel Statistics. In *The Twenty-sixth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc '25), October 27–30, 2025, Houston, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3704413.3764460

1 Introduction

Timely delivery of real-time status updates is necessary for many critical and emerging applications including healthcare, factory automation, intelligent transportation systems, and smart homes, to name a few. A typical real-time status update system consists of an energy-constrained source node (e.g., a small sensor) that generates status updates about some physical process of interest, and then sends them through a communication system to a destination node. Clearly, excessive transmissions of status updates can maintain the freshness of information available at the destination at the price of quickly exhausting the limited energy available at the source. Therefore, there exists a natural tradeoff between maintaining the freshness of information available at the destination and the transmission cost (or energy cost) of the source. Scheduling the transmissions of status updates to optimize this tradeoff is challenging especially since in practice the statistics of the channels between the source and destination nodes are often unknown to the scheduler. In this paper, we address this open problem by developing novel online learning-based scheduling algorithms with provable guarantees.

We employ AoI as a metric to quantify the freshness of information at the destination about the process observed by the source. Specifically, AoI is defined as the time elapsed since the last successfully received status update at the destination was generated at the source [15]. There have been two main research directions in the AoI research area. The first direction aimed to analyze/characterize AoI in different queueing-theoretic models/disciplines, and the second direction was focused on the optimization of AoI in different communication systems that deal with time-sensitive information. Interested readers are advised to refer to [19] for a comprehensive book and [26] for a recent survey. Since this paper belongs to the second research direction, we next discuss the most closely-relevant prior optimization-based studies on AoI.

The scheduling problem to minimize AoI in single-hop wireless networks with unreliable channels was studied in [9, 10, 13, 14, 22, 23]. In particular, the problem was formulated as a restless

MobiHoc '25, Houston, TX, USA

ACM ISBN 979-8-4007-1353-8/2025/10
https://doi.org/10.1145/3704413.3764460

multi-armed bandit (MAB) problem for which Whittle Index-based scheduling policies were developed. A common assumption considered in [9, 10, 13, 14, 22, 23] was that the statistics of the channels and/or the channel state information are known to the scheduler. Also, none of these studies accounted for the energy limitations at the source node(s). Further, these prior works have mostly been focused on the study of the infinite horizon model, whereas we develop in this paper scheduling algorithms with provable finite-time guarantees.

For the case when the channel statistics are unknown to the scheduler, online learning-based scheduling algorithms with provable finite-time guarantees were developed in [2, 7, 12, 20]. The authors of [7] considered a system setting where the source is connected to the destination through a set of unreliable channels (i.e., each channel is associated with a different reliability or successful transmission probability). The study in [7] was extended in [12] to the case of having correlated unreliable channels, in [20] to the multi-source setting where each source generates a status update every time slot, and in [2] to the multi-source setting with random status update arrivals at different sources. The scheduling of status updates over different channels was formulated as a multi-armed bandit problem in [2, 7, 12, 20] where each channel corresponds to one arm. The reward obtained from selecting one arm in some time slot is a function of the reliability of that arm and the AoI value at the beginning of that time slot (without accounting for the transmission cost of sending status updates). The regret of UCB [3] and Q-UCB [16] algorithms (with respect to the optimal policy that knows the channel statistics a priori) were shown in [7] to scale as $O(\log T)$ and $O(\log^3 T)$, respectively, where T is the time horizon length. The authors of [20] developed a UCB-based distributed learning algorithm that scales as $O(\log^2 \mathrm{T})$, and the authors of [2] utilized the knowledge about the system being empty (i.e., there are no status updates to transmit) or not to develop a learning algorithm that achieves a bounded regret with respect to T (i.e., $O(1)$) when the arrival rates at different sources are relatively small.

A key distinction between [2, 7, 12, 20] and this paper is the structure of the optimal policy to which the proposed learning algorithms are compared (to obtain the regret). In particular, the optimal policy for the settings studied in [2, 7, 12, 20] is to simply send a status update over the channel with the highest successful transmission probability every time slot (whenever the system is not empty), and hence the scheduling problem could be formulated as a multi-armed bandit problem. Since this paper accounts for the transmission costs of sending status updates over different channels, the simple structure of the optimal policy in [2, 7, 12, 20] does not hold here anymore. In particular, it may be optimal in our setting to remain idle in some time slots (and drop the generated status updates). Thus, the decision of sending a status update should also depend on the AoI value, and hence the multi-armed bandit problem formulation in [2, 7, 12, 20] is not sufficient to study the scheduling problem considered in this paper. This key difference between the structures of the optimal policies has significant impact on the development of the learning algorithms in this paper and makes the regret analysis much more challenging. Before going into more details about our contributions, it is instructive to note that scheduling problems to jointly optimize AoI and transmission cost or other costs have been studied in a variety of settings [1, 5, 8, 17, 18, 21, 24, 25]. However, none of these works considered that the channel statistics are unknown to the scheduler, and most of them were focused on the study of the infinite horizon model.

Contributions. This paper presents novel online learning-based scheduling algorithms with provable finite-time guarantees to optimize AoI for energy-constrained communications under unknown channel statistics. In particular, we study a system setting in which an energy-constrained source node is connected to a destination node through a set of channels, where the channel statistics are assumed to be unknown to the scheduler. Towards developing AoI-aware online learning-based algorithms for this setting, we first analyze the structure of the optimal policy (that knows the channel statistics a priori) for the infinite time average-cost problem. Specifically, this optimal policy is proven to have a threshold-based structure with respect to the value of AoI (i.e., it is optimal to drop updates when the AoI is below some threshold). This key insight is then utilized to develop the proposed learning algorithms for the finite-time horizon model under consideration. Our proposed AoI-aware learning algorithms (with and without an exploration bonus) are proven to surprisingly have a bounded regret performance with respect to the time horizon length (i.e., $O(1)$). Extensive simulations are conducted to show the impact of different system design parameters on the empirical performance of the proposed learning algorithms. *To the best of our knowledge, this paper makes the first attempt towards developing AoI-aware learning algorithms with a provable order-optimal regret performance for optimizing the fundamental AoI-energy tradeoff.*

2 System Model and Problem Statement

2.1 Network Model

We consider a real-time monitoring system where a source node is connected to a destination node through C communication channels (C_i denotes the i-th channel). Without loss of generality, we consider a discrete time finite horizon composed of T slots of unit length. Hence, the terms power and energy are used interchangeably throughout the paper. At the beginning of each time slot, the source generates a fresh status update, and either transmits it to the destination using one of the channels or drops it. A power cost P is associated with each transmission attempt over any of the channels, and the transmission power cost of time slot t is denoted by $P(t)$. Note that $P(t)$ is equal to zero when the status update generated at the beginning of time slot t is dropped, and is equal to P otherwise. The freshness of information at the destination node is measured using the AoI metric. In particular, the AoI measures the time elapsed since the generation time of the latest successfully received status update at the destination node. Let $A(t)$ denote the AoI value at the beginning of time slot t. Without loss of generality, we assume that $A(t)$ is upper bounded by a finite value A_{m} which can be chosen to be arbitrarily large. When $A(t)$ reaches A_{m}, it means that the available information at the destination node is too stale to be of any use. A status update transmission over channel C_i is successful with probability μ_i, independent of all other channels and across time slots. The values of $\{\mu_i\}$ are assumed to be unknown to the scheduler. The total cost of time slot t is defined as

$$C(t) = \alpha A(t) + (1-\alpha)P(t), \tag{1}$$

where $\alpha \in [0,1]$. Our intention behind using a weighted cost function [1] is to provide a generic problem formulation that allows the scheduler to set the importance weights of AoI and power consumption in the optimization problem.

State and action spaces. At the beginning of time slot t, the state of the system $s(t)$ is represented by the AoI value $A(t)$, i.e., $s(t) = A(t) \in \mathcal{S} = \{1, 2, \cdots, A_{\mathrm{m}}\}$. Based on the state $s(t)$, the action taken in slot t is given by $a(t) \in \mathcal{A} = \{0, 1, \cdots, C\}$. In particular, when $a(t) = 0$, the status update generated by the source at the beginning of slot t is dropped, and $A(t+1) = \min\{A_{\mathrm{m}}, A(t)+1\}$. On the other hand, when $a(t) = i > 0$, the generated status update is transmitted over channel C_i. Further, $A(t+1)$ is given by

$$A(t+1) = \begin{cases} 1, & \text{with probability } \mu_i, \\ \min\{A_{\mathrm{m}}, A(t)+1\}, & \text{with probability } 1-\mu_i. \end{cases} \tag{2}$$

Based on the above definitions, the total cost of time slot t in (1) can be expressed as

$$C(s(t), a(t)) = \alpha s(t) + (1-\alpha)P\mathbb{1}\left(a(t) \neq 0\right), \tag{3}$$

where $\mathbb{1}(\cdot)$ is the indicator function.

2.2 Problem Statement

A policy $\pi = \{\pi_1, \pi_2, \cdots, \pi_T\}$ is a sequence of mappings from the state space to the action space over different time slots, i.e., $\pi_i : \mathcal{S} \longrightarrow \mathcal{A}, \forall i$. We also use $\pi^* = \{\pi_1^*, \pi_2^*, \cdots, \pi_T^*\}$ to refer to the optimal policy, which has a complete knowledge of the statistics of the channels (or the probabilities $\{\mu_i\}$) a priori. In absence of any knowledge about $\{\mu_i\}$, the accumulated cost over T time slots under a policy π starting from state s is given by

$$C(\pi, s, T) = \sum_{t=1}^{T} C(t). \tag{4}$$

The total regret of a policy π with respect to π^* after T time slots is defined as

$$R^{\pi}(T) = \mathbb{E}[C(\pi, s, T)] - \mathbb{E}[C(\pi^*, s, T)], \tag{5}$$

where the expectation is taken with respect to the statistics of the channels. Our goal is to develop a learning algorithm which determines π such that a tight upper bound on the total regret in (5) is obtained.

3 The Case When the Statistics of the Channels Are Known

The first step towards developing a learning algorithm that achieves a tight upper bound on the total regret in (5) is to understand the structure of the optimal policy π^* (that has a complete knowledge of the successful transmission probabilities over different channels $\{\mu_i\}$). Although π^* can be evaluated using the standard backward induction algorithm, it is not possible to obtain analytical insights about its structure in the finite time horizon problem under consideration. Because of that, we first characterize the structure of π^* for the infinite time average-cost problem in this section, and then utilize the obtained insights to develop the learning algorithms for the finite time horizon problem (when the probabilities $\{\mu_i\}$ are unknown) in the next section. Specifically, the expected long-term average cost under policy π can be expressed as

$$\rho(\pi, s) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}[C(\pi, s, T)]. \tag{6}$$

Due to the nature of the evolution of AoI (as described by (2)), and the independence of channel statistics over time, the problem can be modeled as an infinite horizon average-cost MDP with finite state and action spaces $\mathcal{S}$ and $\mathcal{A}$, respectively. Since there exists an optimal stationary deterministic policy (minimizing $\rho(\pi, s)$) for solving MDPs with finite state and action spaces [6], we aim at investigating the structure of that stationary deterministic policy in the sequel.

Lemma 1. *The stationary deterministic optimal policy $\pi^\star$ can be evaluated by solving the following Bellman's equations for average-cost MDPs [6]:*

$$\rho^* + V(s) = \min_{a\in\mathcal{A}} Q(s,a), s \in \mathcal{S}, \tag{7}$$

where $\rho^* = \min_{\pi} \rho(\pi, s)$, $V(s)$ is the value function, and $Q(s,a)$ is the Q-function (also referred to as the Q-factors, $\forall s \in \mathcal{S}$ and $a \in \mathcal{A}$), which is the expected cost resulting from taking action a in state s, i.e.,

$$Q(s,a) = \alpha s + (1-\alpha)P\mathbb{1}\left(a \neq 0\right) + \sum_{s'\in\mathcal{S}} \mathbb{P}(s' \mid s, a)V(s'), \tag{8}$$

where $\mathbb{P}(s' \mid s, a)$ is the transition probability of moving from state s to state s' as a result of taking action a, which can be evaluated from (2) as

$$\mathbb{P}(s' \mid s, a) = \begin{cases} 1, & s' = s^+ \text{ and } a = 0, \\ 1-\mu_i, & s' = s^+ \text{ and } a = i > 0, \\ \mu_i, & s' = 1 \text{ and } a = i > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

where $s^+ = \min\{A_{\mathrm{m}}, s+1\}$. In addition, the optimal action taken at state s is given by

$$\pi^*(s) = \arg\min_{a\in\mathcal{A}} Q(s,a). \tag{10}$$

Since the weak accessibility condition holds for our problem, a solution for the Bellman's equations in Lemma 1 is guaranteed to exist [6]. We will now analytically characterize the structure of the stationary deterministic optimal policy π^* using the Value Iteration Algorithm (VIA). According to the VIA, the value function $V(s)$ can be evaluated iteratively such that $V(s)$ at iteration n, $n = 1, 2, \cdots$, is computed as

$$\begin{aligned} V(s)^{(n)} &= \min_{a\in\mathcal{A}} Q(s,a)^{(n-1)}, \\ &= \min_{a\in\mathcal{A}} \alpha s + (1-\alpha)P\mathbb{1}\left(a \neq 0\right) + \sum_{s'\in\mathcal{S}} \mathbb{P}(s' \mid s, a)V(s')^{(n-1)}, \end{aligned} \tag{11}$$

where $s \in \mathcal{S}$. Hence, the optimal policy at iteration n is given by

$$\pi^{*(n)}(s) = \arg\min_{a\in\mathcal{S}} Q(s,a)^{(n-1)}. \tag{12}$$

[1] Note that the results obtained in this paper (for the structure of the optimal policy in Section 3 as well as the regret bounds in Section 4) using the linear age cost function in (1) can be readily extended to the case of having a non-decreasing age function $\mathcal{F}(A(t))$, i.e., $C(t) = \alpha\mathcal{F}(A(t)) + (1-\alpha)P(t)$. In Section 4, we also provide numerical results demonstrating the bounded regret performance of our proposed order-optimal learning algorithm for a non-linear age cost function.

As per the VIA, under any initialization of the value function $V(s)^{(0)}$, the sequence $\left\{V(s)^{(n)}\right\}$ converges to $V(s)$ which satisfies the Bellman's equation in (7), i.e.,

$$\lim_{n\to\infty} V(s)^{(n)} = V(s). \tag{13}$$

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

Lemma 2. *The value function $V(s)$, satisfying the Bellman's equation in (7) and corresponding to the optimal policy π^*, is non-decreasing with respect to s.*

Proof. Consider two states s_1 and s_2 such that $s_1 \le s_2$. Hence, the objective is to show that $V(s_1) \le V(s_2)$. According to (13), it is then sufficient to show that $V(s_1)^{(n)} \le V(s_2)^{(n)}, \forall n$, which we prove using mathematical induction. Particularly, the relation holds by construction for $n = 0$ since it corresponds to the initial values for the value function which can be chosen arbitrary. Now, we assume that $V(s_1)^{(n)} \le V(s_2)^{(n)}$ holds for some n, and then show that it holds for $V(s_1)^{(n+1)} \le V(s_2)^{(n+1)}$ as well. Particularly, according to (11) and (12), $V(s_1)^{(n+1)}$ and $V(s_2)^{(n+1)}$ can be respectively expressed as

$$\begin{aligned} V(s_1)^{(n+1)} &= \alpha s_1 + (1-\alpha)P\mathbb{1}\left(\pi^{*(n)}(s_1) \neq 0\right) \\ &\quad + \sum_{s'\in\mathcal{S}} \mathbb{P}(s' \mid s_1, \pi^{*(n)}(s_1))V(s')^{(n)}, \\ &\overset{(a)}{\le} \alpha s_1 + (1-\alpha)P\mathbb{1}\left(\pi^{*(n)}(s_2) \neq 0\right) \\ &\quad + \sum_{s'\in\mathcal{S}} \mathbb{P}(s' \mid s_1, \pi^{*(n)}(s_2))V(s')^{(n)}, \end{aligned} \tag{14}$$

$$\begin{aligned} V(s_2)^{(n+1)} &= \alpha s_2 + (1-\alpha)P\mathbb{1}\left(\pi^{*(n)}(s_2) \neq 0\right) \\ &\quad + \sum_{s'\in\mathcal{S}} \mathbb{P}(s' \mid s_2, \pi^{*(n)}(s_2))V(s')^{(n)}, \end{aligned} \tag{15}$$

where step (a) follows since it is not optimal to take action $\pi^{*(n)}(s_2)$ in state s_1. From (9), note that we have

$$\begin{aligned} \sum_{s'\in\mathcal{S}} \mathbb{P}(s' \mid s_i, a)V(s')^{(n)} &= \mathbb{1}\,(a=0)\, V(s_i^+)^{(n)} \\ + \mathbb{1}\,(a \neq 0) &\left[\mu_a V(1)^{(n)} + (1-\mu_a)\, V(s_i^+)^{(n)}\right]. \end{aligned} \tag{16}$$

Since $s_1 \le s_2$, we have $V(s_1^+) \le V(s_2^+)$, and hence we observe from (16) that

$$\sum_{s'\in\mathcal{S}} \left[\mathbb{P}(s' \mid s_1, \pi^{*(n)}(s_2)) - \mathbb{P}(s' \mid s_2, \pi^{*(n)}(s_2))\right] V(s')^{(n)} \le 0.$$

Thus, the right hand side of (14) is less than or equal to $V(s_2)^{(n+1)}$, which leads to having $V(s_1)^{(n+1)} \le V(s_2)^{(n+1)}$. This completes the proof. □

Let k^* denote the index of the channel with the highest successful transmission probability, i.e., $\mu_{k^*} > \mu_k, \forall k \in \mathcal{A} \setminus \{k^*\}$. Based on Lemma 2, the following Lemma characterizes the structure of the optimal policy π^*.

Lemma 3. *The optimal policy π^* has the following structural properties:*
(i) When $s_1 \ge s_2$, if $\pi^*(s_1) = 0$, then $\pi^*(s_2) = 0$.
(ii) When $s_1 \le s_2$, if $\pi^*(s_1) = k^*$, then $\pi^*(s_2) = k^*$.

Proof. We start the proof by showing that $\pi^*(s) \in \{0, k^*\}, \forall s \in \mathcal{S}$. In particular, for $k > 0$, we have

$$Q(s,0) = \alpha s + V(s^+), \tag{17}$$

$$Q(s,k) = \alpha s + V(s^+) + (1-\alpha)P - \mu_k\left[V(s^+) - V(1)\right]. \tag{18}$$

From Lemma 2, we have $V(s^+) - V(1) \ge 0$, and hence we observe from (18) that $k^* = \arg\min_{k\in\mathcal{A}\setminus\{0\}} Q(s,k), \forall s \in \mathcal{S}$. Hence, $\pi^*(s) \in \{0, k^*\}, \forall s \in \mathcal{S}$. Now, note that proving that $\pi^*(s_1) = a$ leads to $\pi^*(s_2) = a'$ is equivalent to showing that

$$Q(s_2,a) - Q(s_2,a') \le Q(s_1,a) - Q(s_1,a'), \forall a' \neq a, \tag{19}$$

where this holds since if a is optimal in state s_1, then we have $Q(s_1,a) - Q(s_1,a') \le 0, \forall a'$, which leads to $Q(s_2,a) \le Q(s_2,a'), \forall a'$, i.e., taking action a is optimal in state s_2. Hence, (i) is proven ((ii) is proven) if (19) holds when $a = 0$ and $a' = k^*$ ($a = k^*$ and $a' = 0$). Therefore, in the remaining, we focus on the proof of (i) while (ii) can be proven similarly. Since $s_1 \ge s_2$ and based on Lemma 2, we have $V(s_1^+) \ge V(s_2^+)$. Hence (19) holds for $a = 0$ and $a' = k^*$, which completes the proof of (i). □

Remark 1. *According to Lemma 3, the optimal policy π^* has a threshold-based structure, where it is optimal to transmit a status update only when the AoI/state is above some threshold value A_{th} (i.e., the updates are dropped when the AoI is less than or equal to A_{th}). In addition, the scheduler uses the channel with the highest successful transmission probability (i.e., C_{k^*}) for each transmission attempt. Further, from (17) and (18), if $P < P_{\min} = \frac{\mu_{k^*}}{1-\alpha}\left(V(2) - V(1)\right)$, then $\pi^*(s) = k^*, \forall s$, whereas if $P > P_{\max} = \frac{\mu_{k^*}}{1-\alpha}\left(V(A_{\text{m}}) - V(1)\right)$, then $\pi^*(s) = 0, \forall s$.*

Based on Lemma 3 and Remark 1, $1 \le A_{\text{th}} < A_{\text{m}}$ when $P \in [P_{\min}, P_{\max}]$. In that case, the optimal value of the long-term average cost ρ^* is obtained in closed-form in the following Lemma.

Lemma 4. *The optimal value of the long-term average cost associated with the optimal policy π^* (with $1 \le A_{\text{th}} < A_{\text{m}}$) is given by*

$$\rho^* = \frac{1}{A_{\text{th}} + \mu_{k^*}^{-1}} \sum_{i=1}^{3} \beta_i, \tag{20}$$

where

$$\beta_1 = \frac{\alpha A_{\text{th}}\left(A_{\text{th}}+1\right)}{2}, \tag{21}$$

$$\beta_2 = \frac{\alpha\left(A_{\text{th}}+1\right) + (1-\alpha)P + \alpha\beta\left(1-\mu_{k^*}\right)^{\beta}}{\mu_{k^*}}, \tag{22}$$

$$\beta_3 = \frac{\alpha\left(1-\mu_{k^*}\right)}{\mu_{k^*}^2}\left[\frac{\left(1-\mu_{k^*}\right)(1-\beta) - \beta}{\left(1-\mu_{k^*}\right)^{\beta-1}} + 1\right], \tag{23}$$

$$\beta = A_{\text{m}} - \left(A_{\text{th}}+1\right). \tag{24}$$

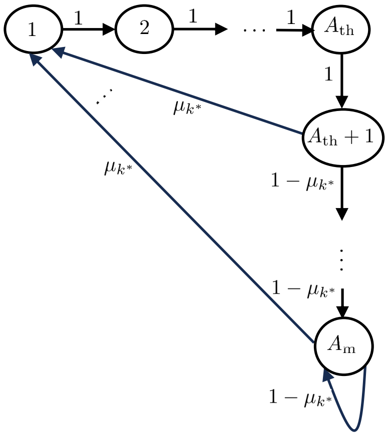


Figure 1: The discrete time Markov chain induced by the optimal policy π^*.

Proof. According to Lemma 3 and Remark 1, the discrete time Markov chain (representing the system state) induced by the optimal policy π^* is depicted in Fig. 1. Thus, ρ^* can be expressed as

$$\rho^* = \sum_{s\in\mathcal{S}} C_s^* \gamma_s, \tag{25}$$

where C_s^* is the cost of being in state s under the optimal policy π^*, and $\{\gamma_s\}_{s\in\mathcal{S}}$ is the stationary distribution of the discrete time Markov chain in Fig. 1 (induced by π^*). Note that C_s^* can be expressed as

$$C_s^* = \begin{cases} \alpha s, & s \le A_{\rm th}, \\ \alpha s + (1-\alpha) P, & \text{otherwise.} \end{cases} \tag{26}$$

Thus, what remains is to obtain the probabilities $\{\gamma_s\}$. The balance equations associated with the Markov chain in Fig. 1 are given by

$$\begin{aligned} &\gamma_s = \gamma_{s-1}, && 2 \le s \le A_{\rm th}+1, \\ &\gamma_s = (1-\mu_{k^*})\,\gamma_{s-1}, && A_{\rm th}+2 \le k < A_{\rm m}, \\ &\mu_{k^*}\gamma_{A_{\rm m}} = (1-\mu_{k^*})\,\gamma_{A_{\rm m}-1}. \end{aligned} \tag{27}$$

From (27), we get

$$\begin{aligned} &\gamma_i = \gamma_{A_{\rm th}+1}, \quad 1 \le i \le A_{\rm th}, \\ &\gamma_{A_{\rm th}+1+i} = (1-\mu_{k^*})^i\,\gamma_{A_{\rm th}+1}, \quad 0 \le i \le A_{\rm m} - A_{\rm th} - 2, \\ &\gamma_{A_{\rm m}} = \mu_{k^*}^{-1}\,(1-\mu_{k^*})^{A_{\rm m}-(A_{\rm th}+1)}\,\gamma_{A_{\rm th}+1}. \end{aligned} \tag{28}$$

By applying $\sum_{s\in\mathcal{S}} \gamma_s = 1$ to the set of equations in (28), we obtain

$$\gamma_{A_{\rm th}+1} = \frac{1}{A_{\rm th} + \mu_{k^*}^{-1}}. \tag{29}$$

The final expression of ρ^* in (20) is derived from substituting $\{C_s^*\}$ and $\{\gamma_s\}$ from (26)-(29) into (25), followed by some algebraic simplifications. This completes the proof. □

4 Order-Optimal Learning Algorithms

In this section, we use the insights obtained about the structure of the optimal policy in the previous section to develop order-optimal learning algorithms with provable finite-time guarantees for the case when the channel statistics are unknown to the scheduler. For ease of presentation of our proposed algorithms, we will use an equivalent normalized reward function $r(s(t), a(t)) \in [0, 1]$ to the cost function $C(s(t), a(t))$ defined in (3). In particular, we have

$$r(s(t), a(t)) = \frac{\alpha A_{\rm m} + (1-\alpha)P - C(s(t), a(t))}{\alpha(A_{\rm m}-1) + (1-\alpha)P}. \tag{30}$$

According to (30), the minimum value of $C(s(t), a(t))$, i.e., $C(1, 0)$, is mapped to $r(1, 0) = 1$, whereas the maximum value of $C(s(t), a(t))$, i.e., $C(A_{\rm m}, a(t) \neq 0)$, is mapped to $r(A_{\rm m}, a(t) \neq 0) = 0$. We also consider an episodic finite horizon MDP setting, where the finite time horizon T is divided into K episodes with equal length H (i.e., each episode has H time steps/slots and $T = KH$). Let $s^\pi_{k,h}$ and $s^*_{k,h}$ ($a^\pi_{k,h}$ and $a^*_{k,h}$) denote the state of the system (the action) at the h-th time step in episode k under the learning algorithm π and the optimal policy π^*, respectively. Note that since the channel statistics are assumed to be unknown, the transition probability matrix of the MDP under consideration is unknown to the scheduler (as can also be observed from (9)). In particular, as the scheduler interacts with the MDP over time, it observes the states, actions and rewards generated by the unknown system dynamics (or transition probability matrix). This leads to the fundamental exploration-exploitation tradeoff where the scheduler needs to balance between exploring poorly-understood state-action pairs (to gain information and improve future performance) and exploiting its current knowledge about the system dynamics (to optimize short-run rewards).

Before going into more details about our proposed AoI-aware order-optimal learning algorithms, it is worth noting that an AoI-agnostic learning algorithm that can handle the setting of episodic finite horizon MDPs with unknown system dynamics is the upper confidence bound value iteration (UCBVI) algorithm [4]. The key idea of the UCBVI algorithm is to directly add an exploration bonus term (to strike a balance between exploration and exploitation) to the Q-values, rather than building confidence sets for the transition probabilities and rewards (as in UCRL2 [11]). This leads to an improvement in the achievable regret bound by the UCBVI algorithm (compared to UCRL2). By directly applying the analysis of the UCBVI algorithm in [4] to our problem, with probability $1-\delta$ ($0 < \delta < 1$), the regret can be upper bounded as follows

$$R(T) \le O\left([\alpha(A_{\rm m}-1) + (1-\alpha)P] \sqrt{HA_{\rm m}CT} \right). \tag{31}$$

The regret bound of the UCBVI algorithm in (31) is near-optimal since it matches the established regret lower bound of [11] for this MDP problem up to logarithmic factors:

$$R(T) \ge \Omega\left([\alpha(A_{\rm m}-1) + (1-\alpha)P] \sqrt{HA_{\rm m}CT} \right). \tag{32}$$

In the sequel, we significantly improve the dependency of the regret on T by developing novel AoI-aware order-optimal learning algorithms that utilize the structure of the optimal policy. In particular, our proposed learning algorithms achieve provably $O(1)$ regret bounds (i.e., the regret is bounded with respect to the increase in T).

Algorithm 1 Order-optimal algorithm with exploration bonus.

```
for $k = 1, \cdots, K$ do
    Compute, for all $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,
    $N_k(s, a, s') = \sum_{\tau=1}^{k-1} \sum_{s_\tau, a_\tau, s'_\tau} \mathbb{1}(s_\tau = s, a_\tau = a, s'_\tau = s')$
    $N_k(s, a) = \sum_{s' \in \mathcal{S}} N_k(s, a, s')$
    $\hat{T}_k(a) = \frac{\sum_{s,s'=s^+} N_k(s,a,s')}{\sum_s N_k(s,a)}, \forall a \in \mathcal{A} \setminus \{0\}$
    for $s \in \mathcal{S}$
        if $a = 0$ then
            $\hat{\mathbb{P}}'_k(s^+|s, a) = 1$
        else
            $\hat{\mathbb{P}}'_k(s^+|s, a) = \hat{T}_k(a)$ and $\hat{\mathbb{P}}'_k(1|s, a) = 1 - \hat{T}_k(a)$
        end if
    end for
    Initialize $V_{k,H+1}(s) = 0$ for all $s \in \mathcal{S}$
    for $h = H, \cdots, 1$ do
        for $(s, a) \in \mathcal{S} \times \mathcal{A}$
            $Q_{k,h}(s, a) = \min\{Q_{k-1,h}(s, a), H, r(s, a)$
                $+\mathbb{E}_{\hat{\mathbb{P}}'_k}[V_{k,h+1}(s')|s, a]$
                $+7H \ln(5SAT/\delta)\sqrt{\frac{1}{N_k(s,a)}}\mathbb{1}(\theta_0 - k \geq 0)\}$,
            $V_{k,h}(s) = \max_{a \in \mathcal{A}} Q_{k,h}(s, a)$
        end for
    end for
    for $h = 1, \cdots, H$ do
        Set $s^\pi_{k,1} \in \mathcal{S}$ arbitrarily
        Take action $a^\pi_{k,h} = \underset{a \in \mathcal{A}}{\operatorname{argmax}} Q_{k,h}(s^\pi_{k,h}, a)$ and observe
        $s^\pi_{k,h+1}$
        if $a^\pi_{k,h} = 0$ then
        Send a pilot signal over a uniformly randomly-chosen
        channel $c$ and record $\left(s^\pi_{k,h}, 0, s^{\pi,+}_{k,h}\right)$ or $\left(s^\pi_{k,h}, 0, 1\right)$
        based on the outcome of the transmission
        end if
    end for
end for
```

4.1 An Order-Optimal Learning Algorithm with Exploration Bonus

As a consequence of the insights obtained in Section 3 for the infinite time average-cost problem, it is possible that the optimal policy (that knows the channel statistics beforehand) for the finite horizon model drops the generated status updates in certain time slots (e.g., when the AoI value is relatively small). This is in contrast to the optimal policy for the settings studied in [2, 7, 12, 20] (in which the transmission costs of sending status updates are ignored), where it was optimal to send a status update over the channel with the highest reliability every time slot. This key insight is utilized to develop our proposed order-optimal learning algorithms. In particular, when the action is to drop the generated status update in a certain time slot (i.e., the channels are idle), it would be useful to utilize that slot for exploring the status of one of the channels at a negligible power cost (by sending a pilot signal).

Our first order-optimal learning algorithm is described in Algorithm 1. Specifically, prior to each episode k, we obtain some estimates for the system dynamics (or the state transition probabilities denoted by $\{\hat{\mathbb{P}}'_k(s'|s, a)\}_{s,a,s'}$ based on the counts of state-action pairs experienced prior to episode k. These estimated transition probabilities are then fed into a backward induction algorithm with an exploration bonus term (directly added to the Q-values, similar to the UCBVI algorithm) to evaluate the policy to be executed within episode k. Here, the exploration bonus term is given by $7H \ln(5SAT/\delta)\sqrt{\frac{1}{N_k(s,a)}}$, which was obtained in [4] based on the Chernoff-Hoeffding's concentration inequality. Finally, the evaluated policy is executed within episode k while utilizing that whenever the action is to drop the generated status update, an opportunity for exploration is created by sending a pilot signal over a uniformly randomly-chosen channel. The achievable regret by Algorithm 1 is stated in the following Theorem.

Theorem 1. *With probability* $1 - \delta$, *the regret of Algorithm 1 is upper bounded as follows:*

$$R(T) \leq O\left([\alpha(A_m - 1) + (1 - \alpha)P]H\sqrt{\theta_0}\right), \tag{33}$$

where $\theta_0 = \Theta(C^2 \ln \frac{2C}{\delta})$.

Remark 2. *Theorem 1 shows that with high probability* $1 - \delta$, *the regret of Algorithm 1 with bonus terms is* $O(1)$, *especially that it does not increase with the total time horizon* T *(or equivalently, the number of episodes* K*). Particularly, compared with existing MDP solutions (such as the UCBVI algorithm), we reduce the regret from* $O(\sqrt{T})$ *to* $O(1)$, *and the intuition behind that is as follows. During the early phase* ($k \leq \lceil\theta_0\rceil$) *where we need to carefully handle the exploration-exploitation tradeoff, the bonus term is added to encourage learning. However, thanks to the idea of sending a pilot signal when the action is to drop the generated status update, the success probabilities over different channels* $\{\mu_i\}$ *can be estimated much faster than the case where we do not use pilot signals (e.g., refer to (36) where the use of pilot signals help us choose optimal actions with high probability in a faster way). Because of that, in the later phase* ($k > \lceil\theta_0\rceil$), *we can choose the optimal action greedily and with high probability.*

Due to space limitations, we will provide next a proof sketch of Theorem 1.

Proof Sketch of Theorem 1: We analyze the phases before and after episode $\lceil\theta_0\rceil$ separately. First, before episode $\lceil\theta_0\rceil$, Algorithm 1 is similar to the UCBVI algorithm [4], but with the additional possibility of sending a pilot signal in each time step where the action is to drop the generated status update. Thus, we can still apply mathematical induction to the V-value function, (similar to the proof of [4, Lemma 18]). We first define the event

$$\Omega = \{V_{k,h} \geq V^*_h, \forall k, h\}. \tag{34}$$

Under Ω, all computed $V_{k,h}$ values in Algorithm 1 are upper bounds on the optimal value function $V^*_h = \max_\pi \mathbb{E}[\sum_{j=h}^{H} r(s^\pi_{k,j}, a^\pi_{k,j})]$. Using backward induction on steps h and concentration inequalities, we can prove that Ω holds with high probability. Thus, the regret in the early phase before episode $\lceil\theta_0\rceil$ is $\sum_{k=1}^{\lceil\theta_0\rceil}\left[V^*_1(s_{k,1}) - V'_{k,1}(s_{k,1})\right]$, where $V'_{k,1}(s_{k,1}) = \mathbb{E}[\sum_{h=1}^{H} r(s^\pi_{k,h}, a^\pi_{k,h})]$ represents the expected cumulative reward of Algorithm 1 in episode k. Under the event Ω, we have

$$\begin{aligned}&\sum_{k=1}^{\lceil\theta_0\rceil}\left[V_1^*(s_{k,1}) - V'_{k,1}(s_{k,1})\right],\\ &\le \sum_{k=1}^{\lceil\theta_0\rceil}\left[V_{k,1}(s_{k,1}) - V'_{k,1}(s_{k,1})\right],\\ &\le O\left(\bar{R}\sum_{k=1}^{\lceil\theta_0\rceil}\sum_{(s,a)} H\ln(SAT/\delta)\sqrt{\frac{1}{N_k(s,a)}}\right),\\ &\le O\left([\alpha(A_m-1)+(1-\alpha)P]H\sqrt{\theta_0}\right), \qquad (35)\end{aligned}$$

where $\bar{R} = [\alpha(A_m-1)+(1-\alpha)P]$, the second inequality is because the difference between $V_{k,1}(s_{k,1})$ and $V'_{k,1}(s_{k,1})$ is upper bounded by the bonus term $7H\ln(5SAT/\delta)\sqrt{\frac{1}{N_k(s,a)}}$ used by Algorithm 1, and the last inequality is because of the pigeon hole principle. Second, after episode $\lceil\theta_0\rceil$, we can show that with high probability, for each episode k, when $a^*_{k,h}=0$, $a^\pi_{k,h}=0$, or when $a^*_{k,h}>0$, $a^\pi_{k,h}>0$. In other words, Algorithm 1 will not mistakenly send or drop a status update. Specifically, according to Hoeffding's inequality, we have that after episode $\lceil\theta_0\rceil$,

$$\begin{aligned}&\sum_{k>\lceil\theta_0\rceil,h}\mathbb{P}\left\{P(a^\pi_{k,h}) \neq P(a^*_{k,h})\right\},\\ &\le O\left(\exp\left(-\frac{1}{2C^2}\lceil\theta_0\rceil\right)\right) \le O(\delta), \qquad (36)\end{aligned}$$

where the last inequality is true because $\theta_0 = \Theta(C^2\ln\frac{2C}{\delta})$. Also, according to Hoeffding's inequality and [2, Lemma 12], the regret due to choosing suboptimal channels after episode $\lceil\theta_0\rceil$ can be upper bounded by $2C\cdot\left[\exp\left(-\frac{1}{2C^2}k\right)+\exp\left(-\frac{\Delta^2_{min}}{4C}k\right)\right]$. For completeness, [2, Lemma 12] states that for every suboptimal channel index $i\neq i^*$, the probability of choosing this suboptimal channel is bounded by

$$\mathbb{P}\{a^\pi_{k,h}=i\} \le 2C\cdot\left[\exp\left(-\frac{1}{2C^2}k\right)+\exp\left(-\frac{\Delta^2_{min}}{4C}k\right)\right]. \qquad (37)$$

According to (37), the probability that our algorithm does not choose the optimal channel decreases in k exponentially. This benefits from the fact that each episode has at least one pilot signal transmission (which occurs at the last time step of the episode where the action is to drop the generated status update, based on the implementation of the backward induction algorithm). The final regret in (33) is obtained by taking the sum over the episode index k. □

The empirical performance of Algorithm 1 is compared to that of the UCBVI algorithm in Figs. 2 and 3. Note that the tight regret bounds in [4, Theorems 1 and 2] were obtained under the condition that $H \le SA$, where S and A are the sizes of the state and action spaces, respectively. Thus, for fair comparison with the UCBVI algorithm, we consider values of H that satisfy this condition. Also, we consider a similar exploration bonus term to that of the UCBVI algorithm (i.e., we focus on the performance of the early phase of Algorithm 1 in Figs. 2 and 3). While the initial state $s_{k,1}$ may change arbitrarily from one episode to the next [4], the curves with the abbreviation "NR" refer to the specific case where the initial state (or the initial AoI value) of each episode is set to be the AoI value at the end of its preceding episode. On the other hand, the initial state $s_{k,1}$ is chosen uniformly at random in the curves without the abbreviation "NR".

A couple of key observations can be noticed from Fig. 2. First, Algorithm 1 significantly outperforms the UCBVI algorithm in terms of the achievable regret. This demonstrates/quantifies the significant impact of the exploration opportunities created through sending pilot signals (when the channels are idle, i.e., the action is to drop the generated status update) on the achievable regret performance. Second, the achievable regret by Algorithm 1 approaches a bounded regret value as the number of episodes K increases. However, this convergence of the regret to a bounded value appears to be relatively slow, which is mainly due to the existence of the exploration bonus term. This motivates us to develop a variant of Algorithm 1 with a similar provable theoretical guarantee (in terms of achieving a bounded regret with respect to K), but with a much better empirical performance (in terms of the fast convergence of the regret to a bounded value even when the sizes of state and action spaces are quite large). It can also be observed from Fig. 3 that as the importance weight of AoI α increases, the gap between the achievable regrets by Algorithm 1 and the UCBVI algorithm slightly decreases (since the exploration opportunities of sending pilot signals become less).

4.2 An Order-Optimal Learning Algorithm without Exploration Bonus

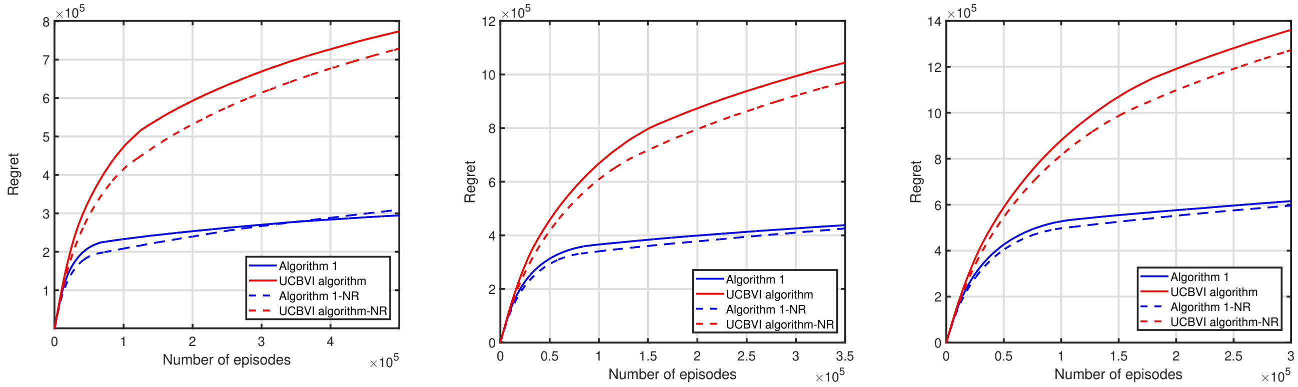
To overcome the limitation in the empirical performance of Algorithm 1 (related to the slow convergence of the regret to a bounded value), we develop another AoI-aware order-optimal learning algorithm (referred to as Algorithm 2) with a significantly better empirical performance. Different from Algorithm 1, our second order-optimal learning algorithm eliminates the exploration bonus term from the Q-values. To evaluate the policy to be executed within episode k, this algorithm just feeds the transition probabilities estimated prior to episode k into the backward induction algorithm. Specifically, the complete description of our second order-optimal learning algorithm is similar to Algorithm 1 with the only difference that the Q-values are evaluated as: $Q_{k,h}(s,a) = r(s,a) + \mathbb{E}_{\hat{\mathbb{P}}'_k}[V_{k,h+1}(s')|s,a]$. The achievable regret by Algorithm 2 is stated in the following Theorem.

Theorem 2. *The regret of Algorithm 2 is upper bounded as follows:*

$$\begin{aligned}&R(T)\\ &\le O\left(H^2A_{\mathrm{m}}PC\left[\frac{3}{\frac{1}{2C^2}-1}+\frac{2}{\frac{\delta^2_{min}}{4C}-1}+\frac{1}{\frac{\Delta^2_{min}}{4C}-1}\right]\right), \qquad (38)\end{aligned}$$

where $\delta_{min} = \min_i\left|\frac{1-\alpha}{\alpha A_{\mathrm{m}}}P-\mu_i\right|$ *and* $\Delta_{min} = \min_i|\mu_{k^*}-\mu_i|$.

Remark 3. *Theorem 2 shows that the regret of our order-optimal learning algorithm without exploration bonus is $O(1)$, especially that it does not increase with the total time horizon T (or equivalently, the number of episodes K). It is worth noting that [2] developed a learning algorithm without exploration bonus that achieves a bounded regret with respect to T (i.e., $O(1)$) when the power cost of sending status updates are ignored and the status updates arrive at the source nodes*



(a)

(b)

(c)

Figure 2: Comparison between Algorithm 1 and the UCBVI algorithm. We use $A_{\mathrm{m}} = 10, P = 15, \alpha = 0.4$ and $C = 4$. The successful transmission probabilities over different channel are equally spaced from 0.2 to 0.8 (i.e., the probabilities are $\{0.2, 0.4, 0.6, 0.8\}$). We consider: i) $H = 30$ in (a), ii) $H = 40$ in (b), and iii) $H = 50$ in (c).

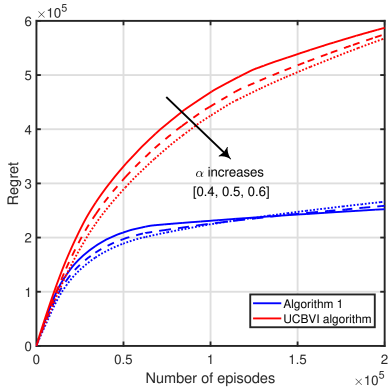


Figure 3: Impact of the importance weight of AoI α on the achievable regret by Algorithm 1. We use $H = 30$. Other parameters are same as Fig. 2.

according to a Bernoulli random process. Unlike the regret analysis in [2], accounting for the power costs of sending status updates in this paper introduces several technical challenges to the regret analysis. First, the regret due to the time steps in which the learning algorithm mistakenly sends the generated status update or drops it needs to be carefully quantified. Second, in [2], whenever the optimal channel is chosen by the learning algorithm, the regret is 0*. However, this does not hold in our setting where we consider the power cost as well. Third, different from [2] where the only reason that the optimal action is not chosen is channel estimation bias, an additional reason in our setting could be saving the power cost P.*

Due to space limitations, we will provide next a proof sketch of Theorem 2.

Proof Sketch of Theorem 2: Recall that the total cost at time step t contains two parts, the AoI $s(t)$ and the power cost $P(t)$. We analyze each of them separately. First, the power regret is

$$(1-\alpha)\sum_{k=1}^{K}\mathbb{E}\left[\sum_{h=1}^{H}\left[P(a^{\pi}_{k,h})-P(a^{*}_{k,h})\right]\right]. \tag{39}$$

According to the law of total expectation, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{h=1}^{H}\left[P(a^{\pi}_{k,h})-P(a^{*}_{k,h})\right]\right]\\
&=\sum_{h=1}^{H}\mathbb{E}\left[P(a^{\pi}_{k,h})-P(a^{*}_{k,h})|a^{*}_{k,h}=a^{\pi}_{k,h}\right]\cdot\mathbb{P}(a^{*}_{k,h}=a^{\pi}_{k,h})\\
&+\sum_{h=1}^{H}\mathbb{E}\left[P(a^{\pi}_{k,h})-P(a^{*}_{k,h})|a^{*}_{k,h}\neq a^{\pi}_{k,h}\right]\cdot\mathbb{P}(a^{*}_{k,h}\neq a^{\pi}_{k,h}).
\end{aligned} \tag{40}$$

Note that when $a^{*}_{k,h} = a^{\pi}_{k,h}$, the power regret is 0, i.e.,

$$\mathbb{E}\left[P(a^{\pi}_{k,h})-P(a^{*}_{k,h})|a^{*}_{k,h}=a^{\pi}_{k,h}\right]=0$$

.

Therefore, we focus on the case when $a^{*}_{k,h} \neq a^{\pi}_{k,h}$. Since sending a status update over any of the channels has the same power cost P, it is sufficient to focus on the time steps where either our learning algorithm or the optimal policy drops the generated status update. Thus, we have,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{h=1}^{H}\left[P(a^{\pi}_{k,h})-P(a^{*}_{k,h})\right]\right]\leq HP\\
&\cdot\mathbb{P}\left(\exists h, s.t., \{a^{*}_{k,h}=0, a^{\pi}_{k,h}>0\}\cup\{a^{*}_{k,h}>0, a^{\pi}_{k,h}=0\}\right).
\end{aligned} \tag{41}$$

When the estimation error $\max_i\{\hat{\mu}_i - \mu_i\}$ of the transition probability, i.e., the channel reliability, is larger than the gap between the weighted AoI and power costs, the algorithm will mistakenly send the status update (instead of dropping the update and sending a pilot signal). Thus, according to Hoeffding's inequality, we have

$$\mathbb{E}\left[\sum_{h=1}^{H}\left[P(a_{k,h}^{\pi}) - P(a_{k,h}^{*})\right]\right],$$
$$\leq 2HPC \cdot \left[\exp\left(-\frac{1}{2C^2}k\right) + \exp\left(-\frac{\delta_{min}^2}{4C}k\right)\right], \tag{42}$$

where $\delta_{min} = \min_i \left|\frac{1-\alpha}{\alpha A_{\mathrm{m}}}P - \mu_i\right|$. By summing over the episode index k, we get the term depending on δ_{min} in the final regret.

Second, the AoI regret is

$$\alpha\sum_{k=1}^{K}\mathbb{E}\left[\sum_{h=1}^{H}s_{k,h}^{\pi} - \sum_{h=1}^{H}s_{k,h}^{*}\right]. \tag{43}$$

Similarly, according to the law of total expectation, we have

$$\mathbb{E}\left[\sum_{h=1}^{H}s_{k,h}^{\pi} - \sum_{h=1}^{H}s_{k,h}^{*}\right]$$
$$= \sum_{h=1}^{H}\mathbb{E}\left[s_{k,h}^{\pi} - s_{k,h}^{*}|a_{k,h}^{*} = a_{k,h}^{\pi}\right]\cdot\mathbb{P}(a_{k,h}^{*} = a_{k,h}^{\pi})$$
$$+ \sum_{h=1}^{H}\mathbb{E}\left[s_{k,h}^{\pi} - s_{k,h}^{*}|a_{k,h}^{*} \neq a_{k,h}^{\pi}\right]\cdot\mathbb{P}(a_{k,h}^{*} \neq a_{k,h}^{\pi}). \tag{44}$$

There are two scenarios in which we may have $a_{k,h}^{*} \neq a_{k,h}^{\pi}$. First, the learning algorithm (or the policy π) may mistakenly drop the generated status update to save the power cost P. This occurs with a probability that can be upper bounded in a way similar to what we discussed above, i.e., $2HPC \cdot \left[\exp\left(-\frac{1}{2C^2}k\right) + \exp\left(-\frac{\delta_{min}^2}{4C}k\right)\right]$. Second, if the channel estimation error is too large, the policy π may mistakenly choose a suboptimal channel to send a status update over. According to Hoeffding's inequality and [2, Lemma 12], the probability of this second scenario can be upper bounded by $2C\cdot\left[\exp\left(-\frac{1}{2C^2}k\right) + \exp\left(-\frac{\Delta_{min}^2}{4C}k\right)\right]$. Moreover, since $s_{k,h} \leq A_{\mathrm{m}}$, we have that

$$\sum_{h=1}^{H}E\left[s_{k,h}^{\pi} - s_{k,h}^{*}|a_{k,h}^{*} \neq a_{k,h}^{\pi}\right] \leq HA_{\mathrm{m}}. \tag{45}$$

Since sending a status update over any of the channels has the same power cost, and both the optimal policy and π start from the same initial state/AoI in each episode, if the event $\{a_{k,h}^{*} = 0, a_{k,h}^{\pi} > 0\}\cup\{a_{k,h}^{*} > 0, a_{k,h}^{\pi} = 0\}$ and the event $\{a_{k,h}^{*} \neq a_{k,h}^{\pi}, a_{k,h}^{*} > 0, a_{k,h}^{\pi} > 0\}$ do not occur in an episode, we have $\mathbb{E}\left[s_{k,h}^{\pi} - s_{k,h}^{*}|a_{k,h}^{*} = a_{k,h}^{\pi}\right] = 0$. The final regret in (38) follows by combining (39)-(45) and taking the sum over the episode index k. □

The empirical performance of Algorithm 2 is shown in Fig. 4. It can be observed that the regret of Algorithm 2 (without exploration bonus) converges to a small bounded value very quickly even when the sizes of state and action spaces are relatively large. Further, it can be noticed from Fig. 4 that the bounded value of the regret slightly increases with the increase in the size of the action space (or equivalently, the number of channels C). In addition, the empirical performance of Algorithm 2 for a non-linear age function $\mathcal{F}(A(t)) = \exp(\psi A(t))$ is shown in Figs. 5 and 6. Similarly, the regret quickly converges to a small bounded value, and it can be noticed that this bounded value increases with either the rate ψ of the exponential age function (Fig. 5) or the size of action space C (Fig. 6).

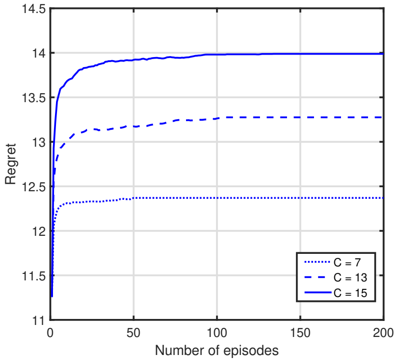


Figure 4: Empirical performance of Algorithm 2. We use $A_{\mathrm{m}} = 100, P = 2, \alpha = 0.5$ and $H = 50$. The successful transmission probabilities over different channel are equally spaced from 0.2 to 0.8.

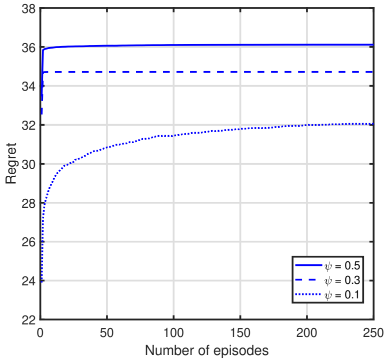


Figure 5: Empirical performance of Algorithm 2 for a non-linear age function $\mathcal{F}(A(t)) = \exp(\psi A(t))$. We use $A_{\mathrm{m}} = 20, C = 15, P = 2, \alpha = 0.5$ and $H = 50$. The successful transmission probabilities over different channel are equally spaced from 0.2 to 0.8.

5 Conclusion

This paper proposed novel AoI-aware online learning-based algorithms for optimizing the fundamental AoI-energy tradeoff under unknown channel statistics. In particular, we considered a system setting in which an energy-constrained source node is connected to a destination node through a set of channels, where the channel statistics were assumed to be unknown to the scheduler. For this setting, the optimal policy (that knows the channel statistics a priori) for the infinite-time average-cost problem was first proven to have a threshold-based structure with respect to the value of AoI. We then utilized this key insight to develop AoI-aware learning

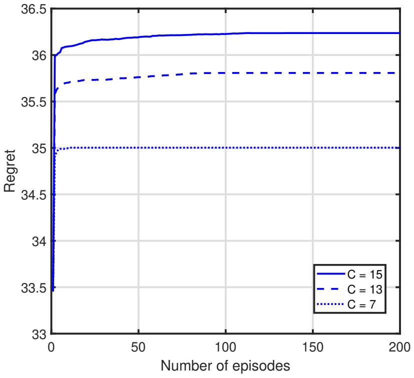


Figure 6: Empirical performance of Algorithm 2 for a non-linear age function $\mathcal{F}(A(t)) = \exp(\psi A(t))$. We use $A_{\mathrm{m}} = 20, \psi = 0.3, P = 2, \alpha = 0.5$ and $H = 50$. The successful transmission probabilities over different channel are equally spaced from 0.2 to 0.8.

algorithms with a provable order-optimal regret performance for the finite time horizon model under consideration. In particular, our proposed learning algorithms with (Algorithm 1) and without (Algorithm 2) an exploration bonus were proven to surprisingly have a bounded regret performance with respect to the time horizon length (i.e., $O(1)$).

Several system design insights were drawn from our simulation results. For instance, our results quantified the significant improvement of our proposed learning algorithms over the UCBVI algorithm in terms of the achievable regret performance. They also revealed that compared to Algorithm 1, Algorithm 2 has a much better empirical performance in terms of the fast convergence of the regret to a bounded value even when the sizes of state and action spaces are quite large. The results also showed that the bounded value of the achievable regret by Algorithm 2 slightly increases with the increase in the number of channels.

An interesting extension of this work is to investigate the possibility of developing AoI-aware order-optimal learning algorithms for the multi-source system setting in which each source is associated with an AoI process. The study of this multi-source setting adds another layer of complexity to the analysis related to scheduling the status update transmissions from different sources. It would also be interesting to extend the proposed algorithms in this paper to account for the possibility of having: i) stochastic status update arrivals at the source node(s) [2], and ii) time-varying unknown cost functions of AoI [24].

Acknowledgments

This work has been supported in part by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-23-2-0225, and by the U.S. National Science Foundation under the grants: NSF AI Institute (AI-EDGE) 2112471, CNS-2312836, CNS-2225561, and CNS-2239677. Additionally, this research was supported by the Office of Naval Research under grant N00014-24-1-2729. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] Mohamed A Abd-Elmagid, Harpreet S Dhillon, and Nikolaos Pappas. 2020. A Reinforcement Learning Framework for Optimizing Age of Information in RF-powered Communication Systems. *IEEE Trans. on Commun.* 68, 8 (Aug. 2020), 4747 – 4760.

[2] Eray Unsal Atay, Igor Kadota, and Eytan Modiano. 2021. Aging wireless bandits: Regret analysis and order-optimal learning algorithm. In *Proc., IEEE WiOpt*. IEEE.

[3] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47 (2002), 235–256.

[4] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. 2017. Minimax regret bounds for reinforcement learning. In *International conference on machine learning*. PMLR, 263–272.

[5] Ahmed M Bedewy, Yin Sun, Sastry Kompella, and Ness B Shroff. 2021. Optimal sampling and scheduling for timely status updates in multi-source networks. *IEEE Trans. on Info. Theory* 67, 6 (2021), 4019–4034.

[6] Dimitri P Bertsekas. 2011. Dynamic programming and optimal control 3rd edition, volume II. *Belmont, MA: Athena Scientific* (2011).

[7] Santosh Fatale, Kavya Bhandari, Urvidh Narula, Sharayu Moharir, and Manjesh K Hanawal. 2021. Regret of age-of-information bandits. *IEEE Trans. on Communications* 70, 1 (2021), 87–100.

[8] Emmanouil Fountoulakis, Nikolaos Pappas, Marian Codreanu, and Anthony Ephremides. 2020. Optimal sampling cost in wireless networks with age of information constraints. In *Proc., IEEE INFOCOM Workshops*. IEEE.

[9] Yu-Pin Hsu. 2018. Age of information: Whittle index for scheduling stochastic arrivals. In *Proc., IEEE ISIT*. IEEE.

[10] Yu-Pin Hsu, Eytan Modiano, and Lingjie Duan. 2019. Scheduling algorithms for minimizing age of information in wireless broadcast networks with random arrivals. *IEEE Trans. on Mobile Computing* 19, 12 (2019), 2903–2915.

[11] Thomas Jaksch, Ronald Ortner, and Peter Auer. 2010. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research* 11 (2010), 1563–1600.

[12] Ishank Juneja, Santosh Fatale, and Sharayu Moharir. 2021. Correlated age-of-information bandits. In *Proc., IEEE WCNC*. IEEE.

[13] Igor Kadota, Abhishek Sinha, and Eytan Modiano. 2018. Optimizing age of information in wireless networks with throughput constraints. In *Proc., IEEE INFOCOM*. IEEE.

[14] Igor Kadota, Abhishek Sinha, Elif Uysal-Biyikoglu, Rahul Singh, and Eytan Modiano. 2018. Scheduling policies for minimizing age of information in broadcast wireless networks. *IEEE/ACM Trans. on Networking* 26, 6 (2018), 2637–2650.

[15] Sanjit Kaul, Roy Yates, and Marco Gruteser. 2012. Real-time status: How often should one update?. In *Proc., IEEE INFOCOM*.

[16] Subhashini Krishnasamy, Rajat Sen, Ramesh Johari, and Sanjay Shakkottai. 2016. Regret of queueing bandits. *Proc., Adv. in Neural Inf. Processing Syst.* (2016).

[17] Zhongdong Liu, Bin Li, Zizhan Zheng, Y Thomas Hou, and Bo Ji. 2023. Toward Optimal Tradeoff Between Data Freshness and Update Cost in Information-Update Systems. *IEEE Internet of Things Journal* 10, 16 (2023), 13988–14002.

[18] Zhongdong Liu, Keyuan Zhang, Bin Li, Yin Sun, Y Thomas Hou, and Bo Ji. 2024. Learning-augmented Online Minimization of Age of Information and Transmission Costs. (2024). available online: arxiv.org/abs/2403.02573.

[19] Nikolaos Pappas, Mohamed A Abd-Elmagid, Bo Zhou, Walid Saad, and Harpreet S Dhillon. 2023. *Age of Information: Foundations and Applications*. Cambridge University Press.

[20] Archiki Prasad, Vishal Jain, and Sharayu Moharir. 2021. Decentralized age-of-information bandits. In *Proc., IEEE WCNC*. IEEE.

[21] Kumar Saurav and Rahul Vaze. 2021. Minimizing the sum of age of information and transmission cost under stochastic arrival model. In *Proc., IEEE INFOCOM*.

[22] Jingzhou Sun, Zhiyuan Jiang, Bhaskar Krishnamachari, Sheng Zhou, and Zhisheng Niu. 2019. Closed-form Whittle's index-enabled random access for timely status update. *IEEE Trans. on Communications* 68, 3 (2019), 1538–1551.

[23] Vishrant Tripathi and Eytan Modiano. 2019. A whittle index approach to minimizing functions of age of information. In *Proc., IEEE Allerton*. IEEE.

[24] Vishrant Tripathi and Eytan Modiano. 2021. An online learning approach to optimizing time-varying costs of aoi. In *Proc., MobiHoc*.

[25] Yi-Hsuan Tseng and Yu-Pin Hsu. 2019. Online energy-efficient scheduling for timely information downloads in mobile networks. In *Proc., IEEE ISIT*. IEEE.

[26] Roy D. Yates, Yin Sun, D. Richard Brown, Sanjit K. Kaul, Eytan Modiano, and Sennur Ulukus. May 2021. Age of Information: An Introduction and Survey. *EEE Journal on Selected Areas in Commun.* 39, 5 (May 2021), 1183–1210.