Aging Wireless Bandits: Regret Analysis and Order-Optimal Learning Algorithm

Aging Bandits: Regret Analysis and Order-Optimal Learning Algorithm for Wireless Networks with Stochastic Arrivals

Eray Unsal Atay, Igor Kadota, and Eytan Modiano

***Abstract*—We consider a single-hop wireless network with sources transmitting time-sensitive information to the destination over multiple unreliable channels. Packets from each source are generated according to a stochastic process with known statistics and the state of each wireless channel (ON/OFF) varies according to a stochastic process with unknown statistics. The reliability of the wireless channels is to be learned through observation. At every time slot, the learning algorithm selects a single pair (source, channel) and the selected source attempts to transmit its packet via the selected channel. The probability of a successful transmission to the destination depends on the reliability of the selected channel. The goal of the learning algorithm is to minimize the Age-of-Information (AoI) in the network over T time slots. To analyze the performance of the learning algorithm, we introduce the notion of AoI regret, which is the difference between the expected cumulative AoI of the learning algorithm under consideration and the expected cumulative AoI of a genie algorithm that knows the reliability of the channels a priori. The AoI regret captures the penalty incurred by having to learn the statistics of the channels over the T time slots. The results are two-fold: first, we consider learning algorithms that employ well-known solutions to the stochastic multi-armed bandit problem (such as ϵ-Greedy, Upper Confidence Bound, and Thompson Sampling) and show that their AoI regret scales as $\Theta(\log T)$; second, we develop a novel learning algorithm and show that it has $O(1)$ regret. To the best of our knowledge, this is the first learning algorithm with bounded AoI regret.**

I. Introduction

Age-of-Information (AoI) is a performance metric that captures the freshness of the information from the perspective of the destination. AoI measures the time that elapsed since the generation of the packet that was most recently delivered to the destination. This performance metric has been receiving attention in the literature [1], [2], [3] for its application in communication systems that carry time-sensitive data. In this paper, we consider a network with M sources transmitting time-sensitive information to the destination over N unreliable wireless channels, as illustrated in Fig. 1. Packets from each source are generated according to an i.i.d. stochastic process with known statistics and the state of each wireless channel (ON/OFF) varies according to an i.i.d. stochastic process with *unknown statistics*. At every time slot, the learning algorithm schedules a single pair (source, channel) and the selected source attempts to transmit its packet via the selected wireless channel. When a packet with fresh information is successfully transmitted to the destination, the AoI associated with the selected source is reduced. The goal of the scheduler is to keep the information associated with every source in the network as fresh as possible, i.e. to minimize the AoI in the network. To decide which pair to select in a time slot, the scheduler takes into account: i) the packet generation processes at the M sources; ii) the current values of AoI at the destination; and iii) the estimated reliability of the N wireless channels.

In this sequential decision problem, the outcomes of previous transmission attempts are used to estimate the reliability of the wireless channels. This statistical learning problem is closely related to the stochastic multi-armed bandit (MAB) problem in which the wireless channels are the bandits that give i.i.d. rewards and the scheduler is the player that attempts to learn the statistics of the bandits in order to maximize the reward accumulated over time. The main challenge in the stochastic MAB problem is to strike a balance between exploiting the bandit that gave the highest rewards in the past and exploring other bandits that may give high rewards in the future. To evaluate the performance of different learning algorithms, we define regret. Regret is the difference between the expected cumulative reward of a *genie algorithm* (that knows the statistics of the bandits a priori) and the expected cumulative reward of the *learning algorithm* under consideration. The regret captures the penalty incurred by having to learn the statistics of the bandits over time. Some well-known order-optimal learning algorithms in terms of regret are: ϵ-Greedy, Upper Confidence Bound (UCB), and Thompson Sampling (TS). The regret of these policies was shown to increase no more than logarithmically in time [4], [5], [6], $O(\log T)$, and this bound was shown to be tight [7].

We refer to our problem as the *Aging Bandit problem*. An important distinction between the stochastic MAB problem and the Aging Bandit problem is the reward structure. In the stochastic MAB problem, the player selects a bandit in each time slot and receives a reward that is i.i.d. over time and depends only on the probability distribution associated with the selected bandit. In the Aging Bandit problem, the scheduler selects a pair (source, channel) and the reward is the AoI reduction that results from a packet transmission to the destination. This reward depends on the state of the selected channel (which is i.i.d. over time), since a failed transmission gives zero reward, and it also depends on the history of previous packet deliveries and packet generations. In particular, if the selected source has recently delivered a fresh information update to the destination, then the reduction in AoI may be small. In contrast, if the selected source has not updated the destination for a long period, then the AoI

reduction may be large. The reward structure of Aging Bandits is closely related to the AoI evolution (formally defined in Sec. II) which is history-dependent. This intricate reward structure has significant impact on the analysis of regret and on the development of learning algorithms when compared to the analysis of the traditional stochastic MAB.

The literature on MAB problems is vast, dating more than eight decades [8]. For surveys on different types of MAB problems, we refer the readers to [9], [10], [11], [12]. Most relevant to this work are [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24]. The authors in [13], [14], [15] considered the problem of minimizing the expected queue-length in a system with a single queue and multiple servers with unknown service rates. In [13], the authors introduced the concept of queue-length regret, developed a learning algorithm inspired by Thompson Sampling, and analyzed its regret. In [14], [15], the authors used information particular to the queue evolution to develop a learning algorithm with $O(1)$ queue-length regret.

The authors in [16], [17], [18], [19], [20], [21], [22], [23], [24] considered the problem of minimizing the average AoI in a single-hop wireless network with unreliable channels. In [16], [17], [18], [19], [20], [21], the authors posed the AoI minimization problem in a network with multiple sources and *known channel statistics* as a restless MAB problem, developed the associated Whittle's Index scheduling policy, and evaluated its performance in terms of the average AoI. In [22], the authors considered the AoI minimization problem in a network with a single source-destination pair and unknown channel statistics, introduced the concept of AoI regret, and showed that the AoI regret of UCB and TS scale as $O(\log T)$. In [23], the authors obtained similar results as in [22] for the more challenging case of correlated wireless channels. In [24], the authors considered the AoI minimization problem in a network with multiple sources that generate and transmit fresh packets at every time slot through (possibly) different channels with unknown statistics. The authors in [24] showed that the AoI regret of a UCB-based distributed learning algorithm scales as $O(\log^2 T)$. An important modelling assumption common to [22], [23], [24] is that sources generate and transmit fresh packets at every time slot. The more realistic assumptions of random packet generation and scheduled transmissions have significant impact on the AoI evolution, on the analysis of AoI regret, and on the development of learning algorithms. For example, in Sec. IV, we leverage the random packet generation to develop a learning algorithm with $O(1)$ AoI regret.

In this paper, we study learning algorithms that attempt to minimize AoI in a network with multiple sources generating packets according to stochastic processes and transmitting these packets to the destination over wireless channels with initially unknown statistics. At every time slot, the learning algorithm schedules a single pair (source, channel) and the selected source attempts to transmit a packet through the selected channel. Note that the source policy, which selects a source at each time slot, and the channel policy, which selects the channel to be used in each time slot, can be naturally decoupled, as the optimal channel is independent of the source selected. In this paper, we focus on the exploration-exploitation dilemma faced by the channel policy. In particular, we consider learning algorithms employing the *optimal source policy* and *different channel policies*. Our main contributions include:

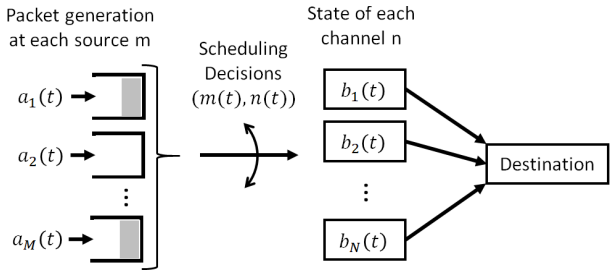

Fig. 1. Illustration of the wireless network with M sources, N channels, and a destination.

- we analyze the performance of channel policies based on traditional MAB algorithms including ϵ-Greedy, UCB, and TS, and show that their AoI regret scales as $\Theta(\log T)$. These results generalize the analysis in [22] to networks with multiple sources generating packets randomly. The analysis of the AoI regret is more challenging in this network setting since the AoI evolution depends on both the source policy and the stochastic packet generation process. These challenges are discussed in Sec. III;
- we develop a novel learning algorithm and establish that it has $O(1)$ AoI regret. The key insight is that when packets are generated randomly, the learning algorithm can utilize times when the network has no packets to transmit, in order to learn the statistics of the channel. To the best of our knowledge, this is the first learning algorithm with bounded AoI regret.

The remainder of this paper is outlined as follows. In Sec. II, the network model and performance metrics are formally presented. In Sec. III, we analyze the AoI regret of traditional learning algorithms. In Sec. IV, we develop an order-optimal learning algorithm and analyze its AoI regret. In Sec. V, we compare the AoI regret of different learning algorithms using simulations. The paper is concluded in Sec. VI. Some of the technical proofs have been omitted due to the space constraint, and will be made available in a technical report.

II. System Model

Consider a single-hop wireless network with M sources, N channels and a single destination, as illustrated in Fig. 1. Each source generates packets containing time-sensitive information and these packets are to be transmitted to the destination through one of the wireless channels. Let the time be slotted, with slot index $t \in \{1, 2, \cdots, T\}$, where T is the time horizon of this discrete-time system. The slot duration allows for a single packet transmission. We normalize the slot duration to unity.

At the beginning of every slot t, each source generates a packet with probability $\lambda \in (0, 1)$. Let $a_m(t) \in \{0, 1\}$ be the indicator function that is equal to 1 when source $m \in$

$\{1, 2, \cdots, M\}$ generates a packet in slot t, and $a_m(t) = 0$ otherwise. This Bernoulli process with parameter λ is i.i.d. over time and independent across different sources, with $\mathrm{P}\left(a_m(t) = 1\right) = \lambda, \forall m, t$. A packet that is generated in slot t can be transmitted during the same slot t. We denote the vector of packet generations in slot t by $\vec{a}(t) = [a_1(t) \ \cdots \ a_M(t)]^{\mathsf{T}}$.

Each source has a transmission queue to store its packets. Sources keep *only* the most recently generated packet, i.e. the freshest packet, in their queue. When source m generates a new packet at the beginning of slot t, older packets (if any) are discarded from its queue. Notice that delivering the most recently generated packet provides the freshest information to the destination. This queueing discipline is known to optimize the AoI in a variety of contexts [25], [26], [27]. After a packet delivery from source m, the queue remains empty until the next packet generation from the same source. However, while the queue is empty, a *dummy packet* can be transmitted for the purpose of probing the channels.

The networked system is empty during slot t if there are no data packets available for transmission, i.e. if the M queues are empty. Let $E(t) \in \{0, 1\}$ be the indicator function that is equal to 1 if the system is empty during slot t, and $E(t) = 0$ otherwise. Notice that if there is a packet generation at the beginning of slot t, then the system is nonempty during slot t and $E(t) = 0$. Recall that when the system is empty, sources can still transmit dummy packets.

In a slot, the learning algorithm selects a single pair (m, n), where $m \in \{1, 2, \cdots, M\}$ is the index of the source and $n \in \{1, 2, \cdots, N\}$ is the index of the wireless channel. Then, during this slot, source m transmits a packet to the destination through channel n. If channel n is ON, then the packet is successfully transmitted to the destination, and if channel n is OFF, then the transmission fails. The learning algorithm does not know the channel states while making scheduling decisions, and the outcome of a transmission attempt during slot t is known at the beginning of slot $t + 1$. Let $b_n(t) \in \{0, 1\}$ be the indicator function that represents the state of channel n during slot t. The channel is ON, $b_n(t) = 1$, with probability $\mu_n \in (0, 1]$, and the channel is OFF, $b_n(t) = 0$, with probability $1 - \mu_n$. The channel state process is i.i.d. over time and independent across different channels.

The *reliability of channel* n is represented by the probability of this channel being ON, μ_n. Let $\vec{\mu} = [\mu_1 \ \cdots \ \mu_N]^{\mathsf{T}}$ be the vector of channel reliabilities. Let μ^* be the maximum channel reliability and let n^* be the index of the corresponding channel, i.e. $\mu^* = \max_n \mu_n = \mu_{n^*}$. For simplicity, we assume that the optimal channel n^* is unique. Naturally, if the channel reliabilities were known by the learning algorithm in advance, then the algorithm would select channel n^* in every slot t. However, since the channel reliabilities $\vec{\mu}$ are initially unknown, the learning algorithm has to estimate μ_n using observations from previous transmission attempts, while at the same time attempting to minimize the AoI in the network. Next, we formulate the AoI minimization problem.

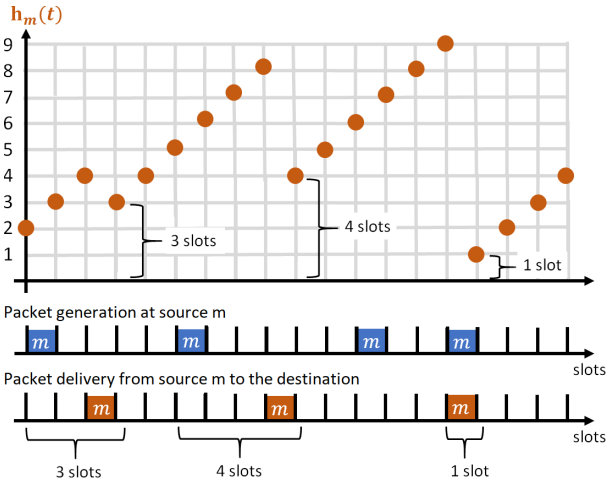

Fig. 2. The blue and orange rectangles at the bottom represent packets generated at source m and successful packet transmissions from source m, respectively. The orange curve shows the AoI evolution $h_m(t)$ associated with source m.

A. Age of Information

The AoI captures how old the information is from the perspective of the destination. Let $h_m(t)$ be a positive integer that represents the AoI associated with source m at the beginning of slot t. By definition, we have $h_m(t) := t - \tau_m(t)$, where $\tau_m(t)$ is the generation time of the latest packet successfully transmitted from source m to the destination[1]. If the destination does not receive a fresh packet from source m during slot t, then in the next slot we have $h_m(t + 1) = h_m(t) + 1$, since the information at the destination is one slot older. In contrast, if the destination receives a fresh packet from source m during slot t, then in the next slot the value of $\tau_m(t + 1)$ is updated to the generation time of the received packet and the AoI is reduced by $\tau_m(t + 1) - \tau_m(t)$. This difference is the "freshness gain" associated with the received packet. The evolution of $h_m(t)$ over time is illustrated in Fig. 2. We define the vector of AoI in slot t as $\vec{h}(t) = [h_1(t) \ \cdots \ h_M(t)]^{\mathsf{T}}$.

For capturing the information freshness of the entire network, we consider the *expected total AoI* $\bar{h}(T)$, which is defined as the expected sum of the AoI over all sources and over time, namely

$$\bar{h}(T) = \mathrm{E}\left[\sum_{m=1}^{M}\sum_{t=1}^{T} h_m(t)\right], \tag{1}$$

where the expectation is with respect to the randomness in the channel states $b_n(t)$, packet generation process $\vec{a}(t)$, and scheduling decisions (m, n). The learning algorithm schedules pairs (m, n) over time so as to minimize the expected total AoI $\bar{h}(T)$. Recall that in this sequential decision problem, the channel reliabilities μ_n are initially unknown by the learning algorithm and should be estimated over time. Next, we discuss the class of learning algorithms considered in this paper.

[1]We define $\tau_m(t) = 0$ prior to the first packet delivery from source m.

B. Learning Algorithm

In this section, we present three important concepts associated with the learning algorithm: the channel policy, the source policy, and the AoI regret. Prior to discussing these concepts, we introduce some notation. In each slot t, the learning algorithm selects a single source and a single channel. Let $m(t)$ be the index of the source selected during slot t and let $n(t)$ be the index of the channel selected during slot t. Then, the pair selected in each slot can be denoted as $(m(t), n(t))$. Notice that the learning algorithm can be divided into two components: the source policy, which selects $m(t)$, and the channel policy, which selects $n(t)$. Let $b(t) = b_{n(t)}(t)$ be the state of the channel selected during slot t, and recall that $\vec{a}(t)$ is the vector of packet generations and $\vec{h}(t)$ is the vector of AoI in slot t. Using this notation, we define the *channel policy* and the *source policy*.

The *channel policy* may (or may not) take into account the status of the transmission queues at the sources (in particular $E(t)$) in making scheduling decisions $n(t)$. Hence, we define two types of channel policies: queue-independent channel policies and queue-dependent channel policies. Let Π_B be the class of admissible *queue-independent channel policies* π_b. In slot t, an arbitrary policy $\pi_b \in \Pi_B$ selects $n(t)$ using information about the outcome of previous transmission attempts. In particular, the queue-independent channel history in slot t is given by $H_B(t) = \{n(1), b(1), \cdots, n(t-1), b(t-1)\}$. Let $\bar{\Pi}_B$ be the class of admissible *queue-dependent channel policies* $\bar{\pi}_b$. In slot t, an arbitrary policy $\bar{\pi}_b \in \bar{\Pi}_B$ selects $n(t)$ using information about the outcome of previous transmission attempts and about the current status of the transmission queues. In particular, the queue-dependent channel history in slot t is given by $\bar{H}_B(t) = H_B(t) \cup \{E(t)\}$. In Sec. IV, we show that this small amount of information, namely $E(t)$, can have a significant impact on the performance of the channel policy. It is easy to see that both the optimal queue-independent channel policy π_b^* and the optimal queue-dependent channel policy $\bar{\pi}_b^*$ select the channel with highest reliability μ^* at every slot t. However, since the reliabilities $\vec{\mu}$ are not known a priori, the channel policies have to estimate $\vec{\mu}$ over time. In Sec. III, we consider queue-independent channel policies and in Sec. IV, we consider queue-dependent channel policies.

The *source policies* considered in this paper are work-conserving, i.e. policies that never transmit dummy packets when there are undelivered data packets in the system. Let Π_A be the class of admissible work-conserving source policies π_a. In slot t, an arbitrary source policy $\pi_a \in \Pi_A$ selects $m(t)$ using information about the current AoI and the generation times of the packets waiting to be transmitted at the sources' queues. In particular, the source history in slot t is given by $H_A(t) = \{\vec{a}(1), \vec{h}(1), \cdots, \vec{a}(t), \vec{h}(t)\}$. The optimal source policy $\pi_a^* \in \Pi_A$ is the transmission scheduling policy that minimizes the expected total AoI in (1). A few works in the literature [18], [19], [28], [29] have addressed the problem of finding the transmission scheduling policy that minimizes AoI in wireless networks with stochastic packet generation and unreliable channels with *known statistics*. Despite those efforts, a full characterization of the optimal source policy is still an open problem.

In this paper, we consider learning algorithms π that are a composition of a source policy and a channel policy $\pi = (\pi_a, \pi_b)$. Our goal is to study the exploration-exploitation dilemma faced by the channel policy. To that end, we analyze the AoI regret of learning algorithms employing the optimal source policy and different channel policies. To analyze the AoI regret of learning algorithms without the full characterization of the optimal source policy π_a^*, we derive lower and upper bounds on the regret. These bounds are discussed in Proposition 2, Proposition 3, and Theorem 7, where we assumed that the optimal source policy π_a^* is the same irrespective of the queue-independent channel policy π_b under consideration, namely

$$\pi_a^* = \underset{\pi_a \in \Pi_A}{\arg\min}\, \mathrm{E}\left[\sum_{m=1}^{M}\sum_{t=1}^{T} h_m^{(\pi_a,\pi_b)}(t)\right], \ \forall \pi_b \in \Pi_B\ , \quad (2)$$

where $h_m^{(\pi_a,\pi_b)}(t)$ denotes the AoI associated with source m in slot t when the learning algorithm $\pi = (\pi_a, \pi_b)$ is employed. An analogous assumption is utilized for the case of queue-dependent channel policies $\bar{\pi}_b \in \bar{\Pi}_B$.

The *AoI regret* of a learning algorithm π with queue-independent channel policy π_b is defined as the difference between the expected total AoI $\bar{h}^{\pi}(T)$ when $\pi = (\pi_a, \pi_b)$ is employed and the expected total AoI $\bar{h}^{*}(T)$ when the optimal algorithm $\pi^* = (\pi_a^*, \pi_b^*)$ is employed, namely

$$R^{\pi}(T) = \mathrm{E}\left[\sum_{m=1}^{M}\sum_{t=1}^{T} h_m^{\pi}(t) - \sum_{m=1}^{M}\sum_{t=1}^{T} h_m^{*}(t)\right], \quad (3)$$

where the expectation is with respect to the randomness in the channel states $b(t)$, packet generation process $\vec{a}(t)$, and scheduling decisions $(m(t), n(t))$. The definition of AoI regret for a learning algorithm $\bar{\pi}$ with queue-dependent channel policy $\bar{\pi}_b$ is analogous to (3). Next, we analyze the AoI regret of learning algorithms with *queue-independent channel policies*.

III. Regret Analysis

The problem of learning channel reliabilities over time is closely related to the stochastic MAB problem. A natural class of channel policies to consider are traditional MAB algorithms such as ϵ-Greedy, UCB, and TS. In this section, we derive bounds on the AoI regret of learning algorithms that employ *queue-independent channel policies*. Notice that the class of queue-independent channel policies Π_B includes traditional MAB algorithms. We describe a learning algorithm employing TS as its channel policy in Algorithm 1.

Scheduling decisions of a learning algorithm π might differ from those of π^* both in the source and in the channel, which makes the analysis of the AoI regret $\sum_{t=1}^{T}\sum_{m=1}^{M} \mathrm{E}[h_m^{\pi}(t) - h_m^{*}(t)]$ challenging. To alleviate this challenge, we use stochastic coupling to create *equivalent coupled channel state*

Algorithm 1: Learning Algorithm employing TS as its channel policy

Initialization: time $t = 1$, estimates $\hat{\mu}_n = 0$, counters $T_n = 0$, parameters $\alpha_n = \beta_n = 1$, $\forall n \in \{1, \cdots, N\}$;
while $1 \leq t \leq T$ **do**
  Optimal source policy selects $m \in \{1, 2, \cdots, M\}$;
  $\theta_n \sim \text{Beta}(\alpha_n, \beta_n)$;
  $n = \arg\max_{n' \in \{1, \cdots, N\}} \theta_{n'}$;
  Source m transmits packet through channel n and observes channel state b;
  if $b = 1$ **then**
    $\alpha_n = \alpha_n + 1$;
  else
    $\beta_n = \beta_n + 1$;
  end
  Compute new estimate $\hat{\mu}_n = \dfrac{\hat{\mu}_n T_n + b}{T_n + 1}$;
  $T_n = T_n + 1$;
  $t = t + 1$;
end

processes that are simpler to analyze. Similar coupling arguments were employed in [13], [22].

Remark 1 (Coupled Channel States)**.** *Let $\{U(t)\}_{t=1}^{T}$ be a sequence of i.i.d. random variables uniformly distributed in the interval $[0, 1]$. In each slot t, the channel states $b_n(t)$ are determined as follows*

$$b_n(t) = 1 \iff 0 \leq U(t) \leq \mu_n \ , \ \forall n \ . \tag{4}$$

By construction, the coupled channel states are no longer independent. In particular, if a channel is ON during slot t, then all channels with higher reliability μ_n are also ON during that slot. Notice that, in each slot t, each coupled channel n has the same probability distribution as the associated original channel n, namely $\text{P}\left(b_n(t) = 1\right) = \mu_n, \forall n, t$. Hence, given the scheduling decision $(m(t), n(t))$ of π during any slot t, the probability of a successful transmission attempt from source $m(t)$ through channel $n(t)$ is the same for both the coupled and original channel states. It follows that the probability distribution of $h_m^{\pi}(t)$ also remains the same for all slots t and for all sources m and, thus, the AoI regret $R^{\pi}(T)$ in (3) also remains the same for both the coupled and original channel state processes. *For simplicity of analysis, henceforth in this paper, we assume that the channel state processes are coupled as described in Remark 1.*

In Proposition 2, Proposition 3, and Corollary 4, we derive bounds on the AoI regret of a learning algorithm π with respect to its *expected number of suboptimal channel choices*, namely

$$\text{E}[K^{\pi}(T)] = \text{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{n^{\pi}(t) \neq n^*\right\}\right] \ , \tag{5}$$

where $\mathbb{1}\left\{n^{\pi}(t) \neq n^*\right\} = 1$ if $n^{\pi}(t) \neq n^*$, and $\mathbb{1}\left\{n^{\pi}(t) \neq n^*\right\} = 0$ otherwise. We consider two classes of admissible learning algorithms

$$\Pi = \{\pi = (\pi_a, \pi_b) : \pi_a \in \Pi_A, \pi_b \in \Pi_B\} \ ; \tag{6}$$

$$\Pi^* = \{\pi = (\pi_a, \pi_b) : \pi_a = \pi_a^*, \pi_b \in \Pi_B\} \ . \tag{7}$$

Both classes employ queue-independent channel policies. The difference is that Π employs any admissible source policy $\pi_a \in \Pi_A$, while Π^* employs the optimal source policy π_a^*. Naturally, we have $\Pi^* \subset \Pi$.

Proposition 2 (Lower Bound)**.** *For any given network configuration $(\lambda, \vec{\mu})$, the AoI regret of any learning algorithm $\pi \in \Pi$ scales at least on the order of its expected number of suboptimal channel choices, namely*[2]

$$R^{\pi}(T) = \Omega\left(\text{E}\left[K^{\pi}(T)\right]\right) \ . \tag{8}$$

Proof outline. In addition to the suboptimal channel choices, source choices $m^{\pi}(t)$ of algorithm $\pi \in \Pi$ can also differ from the source choices $m^*(t)$ of π^*. To overcome this challenge, we construct an auxiliary algorithm $\hat{\pi}^*$ with optimal channel policy and a source policy that selects the same source[3] $m^{\pi}(t)$ as π in every slot t. Then, we focus on the *auxiliary AoI regret* $\sum_{t=1}^{T}\sum_{m=1}^{M} \text{E}[h_m^{\pi}(t) - h_m^{\hat{\pi}^*}(t)]$ associated with the auxiliary algorithm $\hat{\pi}^*$, which we show to be not greater than the original AoI regret $\sum_{t=1}^{T}\sum_{m=1}^{M} \text{E}[h_m^{\pi}(t) - h_m^*(t)]$. We then observe that each suboptimal channel choice of π results in a penalty to the auxiliary AoI regret, and we show that this penalty is lower bounded by a constant. Using this constant, we obtain the desired lower bound on the original AoI regret in (8). The details are omitted due to the space constraint.

Proposition 3 (Upper Bound)**.** *For any given network configuration $(\lambda, \vec{\mu})$, the AoI regret of any learning algorithm $\pi \in \Pi^*$ scales at most on the order of its expected number of suboptimal channel choices, namely*[4]

$$R^{\pi}(T) = O\left(\text{E}\left[K^{\pi}(T)\right]\right) \ . \tag{9}$$

Proof outline. Despite the fact that both learning algorithms $\pi \in \Pi^*$ and π^* employ the same optimal source policy π_a^*, they might select different sources $m^{\pi}(t) \neq m^*(t)$ over time, due to their different channel policies. To address this challenge, we use an approach similar to the proof of Proposition 2. We construct an auxiliary algorithm $\hat{\pi} \in \Pi^*$ with a source policy that selects the same source $m^*(t)$ as π^* in every slot t, and with a channel policy that selects the same channel $n^{\pi}(t)$ as π in every slot t. Then, we show that the auxiliary AoI regret $\sum_{t=1}^{T}\sum_{m=1}^{M} \text{E}[h_m^{\hat{\pi}}(t) - h_m^*(t)]$ associated with the auxiliary algorithm $\hat{\pi}$ is not lower than the original AoI regret $\sum_{t=1}^{T}\sum_{m=1}^{M} \text{E}[h_m^{\pi}(t) - h_m^*(t)]$. To derive an upper bound on the auxiliary AoI regret, we analyze the penalty that results from each suboptimal channel choice of $\hat{\pi}$. During a slot t where $\hat{\pi}$ makes a suboptimal channel choice, if channel $n^{\hat{\pi}}(t)$ is OFF and channel n^* is ON, then a *discrepancy* is

[2] $f(t) = \Omega(g(t)) \iff \exists C > 0 \ \exists t_0 \ \forall t > t_0 : f(t) \geq C \cdot g(n)$

[3] Notice that if the selected source $m^{\pi}(t)$ has no packet in its transmission queue, then the auxiliary algorithm attempts to transmit a dummy packet.

[4] $f(t) = O(g(t)) \iff \exists C > 0 \ \exists t_0 \ \forall t > t_0 : f(t) \leq C \cdot g(n)$

added to the difference between the AoI of $\hat{\pi}$ and the AoI of π^*, i.e. $h_m^{\hat{\pi}}(t+1) - h_m^*(t+1) > h_m^{\hat{\pi}}(t) - h_m^*(t)$. This discrepancy lasts until the next successful transmission of a packet from source m by the auxiliary algorithm $\hat{\pi}$, after which the values of $h_m^{\hat{\pi}}(\cdot)$ and $h_m^*(\cdot)$ become equal[5]. We refer to the duration of the discrepancy as its *length*. The penalty that results from a suboptimal channel choice is the product of the discrepancy and its length. We characterize the auxiliary AoI regret by expressing it as the sum of the penalties arising from suboptimal channel choices. Then, using discrete phase-type distributions, we upper bound the discrepancies and the lengths by constants (in the expected sense) to obtain the result in (9). The details are omitted due to the space constraint.

Corollary 4. *For any given network configuration $(\lambda, \vec{\mu})$, the AoI regret of any learning algorithm $\pi \in \Pi^*$ scales with its expected number of suboptimal channel choices, namely*[6]

$$R^{\pi}(T) = \Theta\left(\mathrm{E}\left[K^{\pi}(T)\right]\right) . \tag{10}$$

Corollary 4 follows directly from Propositions 2 and 3. Notice that the bounds in Proposition 3 and Corollary 4 are not valid for the broader class of learning algorithms Π which includes suboptimal source policies. This is because suboptimal source choices may add to the AoI regret, possibly making it grow faster than $\mathrm{E}\left[K^{\pi}(T)\right]$.

Prior to analyzing the AoI regret of learning algorithms that employ ϵ-Greedy, UCB, and TS as their channel policy, we define *α-consistent learning algorithms* [12], [13] and discuss a few of their properties. Let $\mathrm{E}[T_n^{\pi}(T)]$ be the expected number of times channel n is selected by $\pi \in \Pi$ in the first T slots, namely

$$\mathrm{E}[T_n^{\pi}(T)] = \mathrm{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{n^{\pi}(t) = n\right\}\right] . \tag{11}$$

Definition 5 (α-consistency)**.** *For a given $\alpha \in (0,1)$, a learning algorithm $\pi \in \Pi$ is classified as* α-consistent *if, for any network configuration $(\lambda, \vec{\mu})$, we have $E\left[T_n^{\pi}(T)\right] = O(T^{\alpha})$ for* all suboptimal channels $n \neq n^*$.

Intuitively, a learning algorithm $\pi \in \Pi$ is α-consistent if its channel policy has good performance in *every network configuration*. Consider a learning algorithm with a trivial channel policy that selects $n(t) = 1$ in every slot t. In network configurations with $n^* = 1$, this channel policy never selects suboptimal channels, i.e. $\mathrm{E}\left[T_n^{\pi}(T)\right] = O(T^{\alpha}), \forall n \neq n^*$. However, in network settings with $n^* \neq 1$, this channel policy is such that $\mathrm{E}\left[T_1^{\pi}(T)\right] = T$, which violates the definition of α-consistency. In the remainder of this section, we focus on channel policies that have good performance in every network configuration. In particular, we analyze the AoI regret of α-consistent learning algorithms with queue-independent channel policies.

[5]Recall from Remark 1 that channel states are coupled. Hence, if channel $n^{\hat{\pi}}(t)$ is ON, then channel n^* is also ON.

[6]$f(t) = \Theta(g(t)) \iff f(t) = O(g(t)) \wedge f(t) = \Omega(g(t)) \iff \exists C_1, C_2 > 0\ \exists t_0\ \forall t > t_0 : C_1 \cdot g(n) \leq f(t) \leq C_2 \cdot g(n)$

Remark 6 (AoI regret of α-consistent algorithms)**.** *In [13, Corollary 20], the authors show that any learning algorithm $\pi \in \Pi$ that is α-consistent has an expected number of suboptimal channel choices that scales as $E\left[K^{\pi}(T)\right] = \Omega(\log T)$, for any network configuration $(\lambda, \vec{\mu})$. Hence, it follows from the lower bound in Proposition 2 that the associated AoI regret scales as*

$$R^{\pi}(T) = \Omega(\log T) , \tag{12}$$

for any network configuration $(\lambda, \vec{\mu})$.

Notice that the lower bound in Remark 6 applies to α-consistent learning algorithms with queue-independent channel policies that do not know the statistics of the channels in advance.

Learning algorithms that employ ϵ-Greedy, UCB, and TS as their channel policy are known to have suboptimal channel choices scaling as $\mathrm{E}[K^{\pi}(T)] = O(\log T)$ for any network configuration $(\lambda, \vec{\mu})$ [4], [30], which implies that they are α-consistent. Hence, it follows from the upper bound in Proposition 3 and from (12) that the AoI regret of these learning algorithms scale as

$$R^{\pi}(T) = \Theta(\log T) . \tag{13}$$

In [22], the authors derived lower and upper bounds on the AoI regret of learning algorithms employing queue-independent channel policies, including UCB and TS, in networks with a single source generating and transmitting fresh packets in every slot t. Propositions 2 and 3 generalize the results in [22] to networks with multiple sources generating packets according to stochastic processes. The analysis of the AoI regret is more challenging in this network setting for the following reasons: i) the optimal source policy π_a^* is unknown and there is no closed-form expression for the expected total AoI (1) of the optimal algorithm $\pi^* = (\pi_a^*, \pi_b^*)$; and ii) the learning algorithm under consideration $\pi = (\pi_a, \pi_b)$ can make suboptimal choices both in terms of sources $m(t)$ and channels $n(t)$, and these two types of suboptimal choices affect the AoI regret $R^{\pi}(T)$ differently. Next, we develop a learning algorithm that leverages information about the status of the transmission queues in making scheduling decisions $n(t)$, and show that this new learning algorithm has $O(1)$ AoI regret.

IV. Order-Optimal Learning Algorithm

In this section, we develop a learning algorithm $\bar{\eta} \in \bar{\Pi}$ with a *queue-dependent channel policy* that selects $n(t)$ using information about the outcome of previous transmission attempts, namely $H_B(t) = \{n(1), b(1), \cdots, n(t-1), b(t-1)\}$, and about the current status of the transmission queues, $E(t)$. Then, we derive an upper bound on its AoI regret. In particular, we show that the AoI regret of $\bar{\eta}$ is such that $R^{\bar{\eta}}(T) = O(1)$. Notice that the only difference between the learning algorithms $\pi \in \Pi$ in Sec. III and the order-optimal learning algorithm $\bar{\eta}$ is the knowledge of $E(t)$. This seemingly modest addition led to the reduction of the AoI regret from $R^{\pi}(T) = \Omega(\log T)$ to $R^{\bar{\eta}}(T) = O(1)$. To the best of our knowledge, this is the first learning algorithm with bounded AoI regret.

Algorithm 2: Order-Optimal Learning Algorithm

Initialization: time $t = 1$, estimates $\hat{\mu}_n = 0$, counters $T_n = 0$, $\forall n \in \{1, \cdots, N\}$;
while $1 \leq t \leq T$ **do**
  Optimal source policy selects $m \in \{1, 2, \cdots, M\}$;
  if *system is empty* **then**
    $n = \text{Unif}\{1, \cdots, N\}$;
    Source m transmits dummy packet through channel n and observes channel state b;
    $\hat{\mu}_n = \dfrac{\hat{\mu}_n T_n + b}{T_n + 1}$;
    $T_n = T_n + 1$;
  else
    $n = \arg\max_{n' \in \{1, \cdots, N\}} \hat{\mu}_{n'}$;
    Source m transmits data packet through channel n and observes channel state b;
  end
  $t = t + 1$;
end

The key insight is that when packets are generated randomly, the learning algorithm $\bar{\eta}$ can utilize times when the network has no data packets to transmit, i.e. when $E(t) = 1$, to transmit dummy packets and learn the statistics of the channels without incurring an opportunity cost. The order-optimal learning algorithm $\bar{\eta} = (\eta_a, \bar{\eta}_b)$ has optimal source policy $\eta_a = \pi_a^*$ and a channel policy $\bar{\eta}_b \in \bar{\Pi}_B$ that operates as follows: when the system is empty, $E(t) = 1$, the policy chooses a channel uniformly at random and uses the outcome of the transmission attempt to update its estimates of the channel reliabilities and, when the system is nonempty, $E(t) = 0$, the policy chooses the channel with the current highest estimated reliability. Notice that the channel policy only updates its estimates of the channel reliabilities when the system is empty. A similar channel policy was used in [14], [15] to develop a learning algorithm with bounded queue-length regret. The order-optimal learning algorithm $\bar{\eta}$ is described in Algorithm 2. The upper bound on the AoI regret is established in the theorem that follows.

Theorem 7. *For any given network configuration $(\lambda, \vec{\mu})$, the AoI regret of the order-optimal learning algorithm $\bar{\eta}$ is bounded, namely*

$$R^{\bar{\eta}}(T) = O(1) \, . \tag{14}$$

Proof. Recall that the two components of the order-optimal learning algorithm are $\bar{\eta} = (\eta_a, \bar{\eta}_b)$. Similarly to the proof of Proposition 3, we start by constructing an auxiliary algorithm $\hat{\eta} = (\hat{\eta}_a, \bar{\eta}_b)$ which has a source policy $\hat{\eta}_a$ that selects the same source $m^*(t)$ as π^* in every slot t. Since $\eta_a = \pi_a^*$ is optimal, it follows that $\hat{\eta}_a$ is suboptimal, which implies that

$$\sum_{m=1}^{M} \sum_{t=1}^{T} \mathrm{E}[h_m^{\bar{\eta}}(t) - h_m^*(t)] \leq \sum_{m=1}^{M} \sum_{t=1}^{T} \mathrm{E}[h_m^{\hat{\eta}}(t) - h_m^*(t)] \, . \tag{15}$$

We denote the RHS of (15) as the *auxiliary AoI regret*. Prior to deriving the upper bound on the auxiliary AoI regret, we introduce some definitions that are particular to the channel policy $\bar{\eta}_b$.

Consider the time slots when the system becomes empty, i.e. time slots t such that $E(t-1) = 0$ and $E(t) = 1$. We denote the time interval between two such slots as a *period* and we divide time $t \in \{1, 2, \cdots, T\}$ into successive periods, with period index $p \in \{1, 2, \cdots, P\}$. By definition, the system is empty, $E(t) = 1$, in the beginning of each period p and it remains empty until the first packet generation. Once the first packet is generated, the system becomes nonempty, $E(t) = 0$, and it remains nonempty until the end of the period. Hence, each period p has two phases: an *empty phase* and a *nonempty phase*, with each phase having at least one slot. Let s_p and f_p be the first and the last slots of period p, respectively, with $s_1 = 1$ and $s_{p+1} = f_p + 1, \forall p$. Then, the cumulative AoI of source m during period p can be written as

$$y_m^{\hat{\eta}}(p) = \sum_{t=s_p}^{f_p} h_m^{\hat{\eta}}(t) \, . \tag{16}$$

Recall from Algorithm 2 that estimates of the channel reliabilities are only updated during empty phases. Within a nonempty phase, the estimates do not change and, thus, the selected channel also does not change. Let $\bar{n}(p)$ be the channel selected by policy $\bar{\eta}_b$ during the entire *nonempty phase* of period p. If $\bar{n}(p) = n^*$, we refer to period p as an *optimal period*. Otherwise, we refer to period p as a *suboptimal period*. Next, we derive an upper bound on the auxiliary AoI regret in terms of the expected AoI contributions of the suboptimal periods.

Lemma 8. *The auxiliary AoI regret is upper bounded by*

$$\sum_{m=1}^{M} \sum_{t=1}^{T} \mathrm{E}[h_m^{\hat{\eta}}(t) - h_m^*(t)] \leq \sum_{m=1}^{M} \sum_{p=1}^{T} \mathrm{E}\left[y_m^{\hat{\eta}}(p) \,\middle|\, \bar{n}(p) \neq n^*\right] \mathrm{P}\left(\bar{n}(p) \neq n^*\right) \, . \tag{17}$$

To establish Lemma 8, we first show that if period p is an optimal period, then $h_m^{\hat{\eta}}(t) = h_m^*(t), \forall m, \forall t \in \{s_p, \cdots, f_p\}$, which implies that optimal periods do not contribute to the auxiliary AoI regret. Then, we obtain the upper bound in (17) by manipulating the expression of the auxiliary AoI regret. The complete proof of Lemma 8 can be found in Appendix A.

In Lemmas 9 and 10, we derive upper bounds on the first and second terms on the RHS of (17), respectively.

Lemma 9. *There exists a constant C_y such that*

$$\mathrm{E}\left[y_m^{\hat{\eta}}(p) \,\middle|\, \bar{n}(p) \neq n^*\right] \leq C_y \, . \tag{18}$$

To establish Lemma 9, we first show that the cumulative AoI $y_m^{\hat{\eta}}(p)$ of source m in period p can be upper bounded by

$$y_m^{\hat{\eta}}(p) = \sum_{t=s_p}^{f_p} h_m^{\hat{\eta}}(t) \leq \sum_{i=0}^{f_p - s_p} \left(h_m^{\hat{\eta}}(s_p) + i \right)$$
$$= h_m^{\hat{\eta}}(s_p)[f_p - s_p + 1] + \frac{1}{2}[(f_p - s_p)^2 + f_p - s_p] \, . \tag{19}$$

Then, we derive an upper bound on the conditional expectation of (19). In particular, we show that $h_m^{\hat{\eta}}(s_p)$ can be upper bounded by a geometric random variable. Then, we show that the random variable $f_p - s_p$, which represents the length of period p, follows a discrete phase-type distribution. The upper bound on the conditional expectation of (19) follows from the fact that the geometric random variable has finite second moment and the phase-type random variable has finite first and second moments. The details are omitted due to the space constraint.

Lemma 10. *There exists a constant C_p such that*

$$\sum_{p=1}^{T} \mathrm{P}\left(\bar{n}(p) \neq n^*\right) \leq C_p \, . \tag{20}$$

To establish Lemma 10, we use Hoeffding's inequality to upper bound $\mathrm{P}\left(\bar{n}(p) = n\right)$ by an exponential function of $-p$, for every suboptimal channel n. The result in (20) follows directly from this upper bound. The complete proof of Lemma 10 can be found in Appendix B.

From the upper bound on the AoI regret in (15) and the results in Lemmas 8, 9 and 10, we have

$$R^{\bar{\eta}}(T) \leq \sum_{m=1}^{M} \sum_{t=1}^{T} \mathrm{E}[h_m^{\hat{\eta}}(t) - h_m^*(t)]$$
$$\leq \sum_{m=1}^{M} \sum_{p=1}^{T} \mathrm{E}\left[y_m^{\hat{\eta}}(p) \;\middle|\; \bar{n}(p) \neq n^* \right] \mathrm{P}\left(\bar{n}(p) \neq n^*\right)$$
$$\leq C_y M C_p \tag{21}$$

which establishes the bound in (14). □

In the particular case of a network with sources generating fresh packets at every slot t, i.e. $\lambda = 1$, the algorithm $\bar{\eta}$ cannot utilize slots in which the system is empty to learn the channel reliabilities without incurring a cost in terms of AoI regret, which results in a $R^{\bar{\eta}}(T)$ that grows over time. The upper bound in Theorem 7 is only valid for the network models described in Sec. II, in which $\lambda \in (0, 1)$. Next, we evaluate the AoI regret of the different learning algorithms discussed in this paper using MATLAB simulations and we propose a heuristic algorithm that leverages the fast learning rates of TS and the bounded regret of the order-optimal algorithm.

V. Simulations

In this section, we evaluate the performance of learning algorithms in terms of the AoI regret in (3). We compare learning algorithms employing the Age-Based Max-Weight source policy [29, Sec. 5] and different channel policies, namely: i) ϵ-Greedy; ii) UCB; iii) TS; iv) Optimal; and v) Hybrid. The Age-Based Max-Weight source policy selects, in each slot t, the source m associated with the packet that gives the largest AoI reduction, $\tau_m(t+1) - \tau_m(t)$, if the transmission in slot t is successful. Intuitively, this policy is selecting the source with highest potential reward in terms of AoI. In [29], the authors evaluate the performance of the Age-Based Max-Weight source policy both analytically and using simulations, and show that it achieves near optimal AoI. The first three channel policies, namely ϵ-Greedy, UCB, and TS, were discussed in Sec. III. The Optimal policy is the order-optimal channel policy $\bar{\eta}_b$ developed in Sec. IV. The Hybrid policy employs TS for a fixed period in the beginning of the simulation and then employs the Optimal policy in the remaining slots.

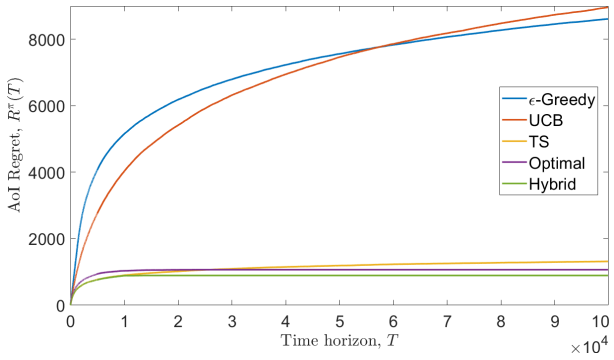

Fig. 3. Simulation of a network with $\lambda = 0.1$.

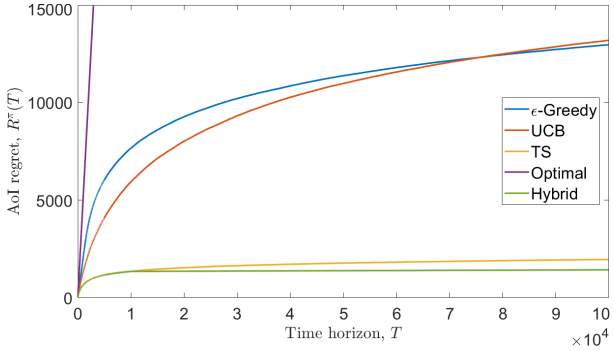

Fig. 4. Simulation of a network with $\lambda = 0.75$.

We simulate a network with a time horizon of $T = 10^5$ slots, $M = 3$ sources, each generating packets according to a Bernoulli process with rate λ, and $N = 5$ channels with reliabilities $\vec{\mu} = [0.4 \;\; 0.45 \;\; 0.5 \;\; 0.55 \;\; 0.6]^{\mathsf{T}}$. Figures 3 and 4 show simulation results of the evolution of the AoI regret over time for $\lambda = 0.1$ and $\lambda = 0.75$, respectively. Figure 5 shows simulation results of the evolution of the reliability estimates associated with the channels with $\mu_4 = 0.55$ and $\mu_5 = 0.6$ over time for $\lambda = 0.75$. Each data point in Figs. 3, 4, and 5 is an average over the results of 10^3 simulations.

The results in Figs. 3 and 4 suggest that, as expected, the AoI regret associated with Optimal and Hybrid is bounded, while the AoI regrets associated with ϵ-Greedy, UCB and TS grow over time. By comparing the AoI regret of Optimal

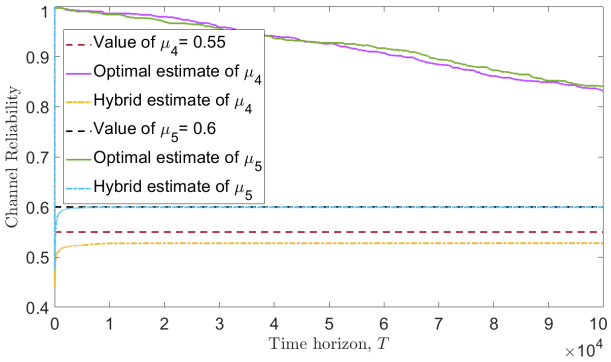

Fig. 5. Simulation of a network with $\lambda = 0.75$.

and TS in Figs. 3 and 4, it is clear that the AoI regret of the Optimal channel policy varies significantly with λ. In particular, for $T = 10^5$, when λ increases from 0.1 to 0.75, the AoI regret of TS increases by a factor of 1.5 (from $1,318$ to $1,963$), while the AoI regret of Optimal increases by a factor of 491.0 (from $1,068$ to $481,700$). A main reason for this performance degradation is that when λ increases, empty systems with $E(t) = 1$ occur less often and, as a result, the Optimal channel policy takes longer to learn the reliability of the channels, as can be seen in Fig. 5. To improve the performance of the Optimal policy for networks with large λ, we propose a heuristic policy called Hybrid channel policy, which employs TS in the first 10^4 slots to quickly learn the reliability of the channels, and then shifts to the Optimal policy which has bounded AoI regret in the long term. Figure 5 illustrates the difference in the learning rates between Optimal and Hybrid. Notice in Fig. 5 that there are extended periods of time in which the Optimal channel policy assigns a larger estimated reliability to a suboptimal channel, which leads to the large AoI regret shown in Fig. 4. However, as established in Theorem 7, for a long enough time-horizon T, the Optimal policy will eventually converge to the true reliabilities, at which point the AoI regret will stop increasing.

VI. Conclusion

This paper considers a single-hop wireless network with M sources transmitting time-sensitive information to the destination over N unreliable channels. Packets from each source are generated according to a Bernoulli process with known rate λ and the state of channel n (ON/OFF) varies according to a Bernoulli process with unknown rate μ_n. The reliabilities $\vec{\mu}$ of the wireless channels is to be learned through observation. At every slot t, the learning algorithm selects a single pair $(m(t), n(t))$ and the selected source $m(t)$ attempts to transmit its packet via the selected channel $n(t)$. The goal of the learning algorithm is to minimize the expected total AoI $\bar{h}(T)$. To analyze the performance of the learning algorithm, we derive bounds on the AoI regret $R^{\pi}(T)$ associated with different learning algorithms. Our main contributions include: i) analyzing the performance of learning algorithms that employ channel policies based on traditional MAB algorithms (ϵ-Greedy, UCB, and TS) and showing that their AoI regret scales as $\Theta(\log T)$; and ii) developing a novel learning algorithm and establishing that it has $O(1)$ AoI regret. To the best of our knowledge, this is the first learning algorithm with bounded AoI regret. Interesting extensions of this work include consideration of sources with unknown packet generation rates and channels with time-varying statistics.

References

[1] A. Kosta, N. Pappas, and V. Angelakis, "Age of information: A new concept, metric, and tool," *Foundations and Trends in Networking*, vol. 12, no. 3, pp. 162–259, 2017.

[2] Y. Sun, I. Kadota, R. Talak, and E. Modiano, *Age of Information: A New Metric for Information Freshness*. Morgan & Claypool, 2019.

[3] R. D. Yates, Y. Sun, D. R. B. III, S. K. Kaul, E. Modiano, and S. Ulukus, "Age of information: An introduction and survey," 2020.

[4] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2–3, p. 235–256, May 2002.

[5] A. Garivier and O. Cappé, "The KL-UCB algorithm for bounded stochastic bandits and beyond," in *Proceedings of the 24th Annual Conference on Learning Theory*, vol. 19, 2011.

[6] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *Proceedings of the 25th Annual Conference on Learning Theory*, vol. 23, 2012.

[7] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[8] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, pp. 285–294, 1933.

[9] A. Slivkins, "Introduction to multi-armed bandits," *Foundations and Trends in Machine Learning*, vol. 12, pp. 1–286, 2019.

[10] J. Gittins, K. Glazebrook, and R. Weber, *Multi-armed Bandit Allocation Indices*, 2nd ed. Wiley, Mar. 2011.

[11] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. USA: Cambridge University Press, 2006.

[12] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends in Machine Learning*, vol. 5, pp. 1–122, 2012.

[13] S. Krishnasamy, R. Sen, R. Johari, and S. Shakkottai, "Learning unknown service rates in queues: A multiarmed bandit approach," *Operations Research*, 2020.

[14] T. Stahlbuhk, B. Shrader, and E. Modiano, "Learning algorithms for minimizing queue length regret," in *Proceedings of IEEE ISIT*, 2018, pp. 1001–1005.

[15] T. B. Stahlbuhk, "Control of wireless networks under uncertain state information," Ph.D. dissertation, Massachusetts Institute of Technology, 2018.

[16] I. Kadota, A. Sinha, E. Uysal-Biyikoglu, R. Singh, and E. Modiano, "Scheduling policies for minimizing age of information in broadcast wireless networks," *IEEE/ACM Transactions on Networking*, 2018.

[17] I. Kadota, A. Sinha, and E. Modiano, "Optimizing age of information in wireless networks with throughput constraints," in *Proceedings of IEEE INFOCOM*, 2018.

[18] Y.-P. Hsu, "Age of information: Whittle index for scheduling stochastic arrivals," in *Proceedings of IEEE ISIT*, 2018.

[19] Y.-P. Hsu, E. Modiano, and L. Duan, "Scheduling algorithms for minimizing age of information in wireless broadcast networks with random arrivals," *IEEE Transactions on Mobile Computing*, vol. 19, no. 12, pp. 2903–2915, 2020.

[20] J. Sun, Z. Jiang, B. Krishnamachari, S. Zhou, and Z. Niu, "Closed-form whittle's index-enabled random access for timely status update," *IEEE Transactions on Communications*, vol. 68, no. 3, pp. 1538–1551, 2020.

[21] V. Tripathi and E. Modiano, "A whittle index approach to minimizing functions of age of information," in *Proceedings of IEEE Allerton*, 2019, p. 1160–1167.

[22] S. Fatale, K. Bhandari, U. Narula, S. Moharir, and M. K. Hanawal, "Regret of age-of-information bandits," 2020.

[23] I. Juneja, S. Fatale, and S. Moharir, "Correlated age-of-information bandits," 2020.

[24] A. Prasad, V. Jain, and S. Moharir, "Decentralized age-of-information bandits," 2020.

[25] M. Costa, M. Codreanu, and A. Ephremides, "On the age of information in status update systems with packet management," *IEEE Transactions on Information Theory*, vol. 62, no. 4, pp. 1897–1910, 2016.
[26] S. Kaul, R. D. Yates, and M. Gruteser, "Status updates through queues," in *Proceedings of IEEE CISS*, 2012.
[27] A. M. Bedewy, Y. Sun, and N. B. Shroff, "The age of information in multihop networks," *IEEE/ACM Transactions on Networking*, vol. 27, no. 3, pp. 1248–1257, 2019.
[28] I. Kadota and E. Modiano, "Minimizing the age of information in wireless networks with stochastic arrivals," in *Proceedings of ACM MobiHoc*, 2019.
[29] ——, "Minimizing the age of information in wireless networks with stochastic arrivals," *IEEE Transactions on Mobile Computing*, 2019.
[30] S. Agrawal and N. Goyal, "Further optimal regret bounds for thompson sampling," in *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, vol. 31, 2013, pp. 99–107.

APPENDIX A
PROOF OF LEMMA 8

To establish Lemma 8, we first show that *optimal periods* do not contribute to the auxiliary AoI regret defined in (15). Then, we obtain an upper bound on the contribution of *suboptimal periods* by manipulating the expression of the auxiliary AoI regret.

Lemma 11. *If period p is an optimal period, then for any slot t within period p and for any source m, we have $h^{\hat{\eta}}_m(t) = h^*_m(t)$.*

Proof. Consider the time slots preceding period p in a network employing the auxiliary learning algorithm $\hat{\eta}$. In slot s_p-1, the algorithm $\hat{\eta}$ delivers the last packet in the system and in slot s_p the system becomes empty, with $E(s_p) = 1$. Let t'_m be the slot in which the latest packet generated from source m was delivered by $\hat{\eta}$. It follows that $t'_m \leq s_p - 1$ and source m did not generate new packets during the time interval $[t'_m + 1, s_p]$.

Recall that (by construction) algorithms $\hat{\eta}$ and π^* select the same source $m^*(t)$ at every slot t and (due to the coupling argument in Remark 1) when a transmission by $\hat{\eta}$ is successful, the transmission by π^* is also successful. Hence, it follows that algorithm π^* also delivered the latest packet generated from source m during (or before) slot t'_m, which implies that $\tau^{\hat{\eta}}_m(s_p) = \tau^{\pi^*}_m(s_p)$ and, as a result, we have $h^{\hat{\eta}}_m(s_p) = s_p - \tau^{\hat{\eta}}_m(s_p) = s_p - \tau^{\pi^*}_m(s_p) = h^*_m(s_p)$.

Since $h^{\hat{\eta}}_m(s_p) = h^*_m(s_p)$ for every source m and for the first slot of every period p, it follows that if period p is an optimal period (in which $\hat{\eta}$ and π^* select the same source and the same channel in every slot t) then $h^{\hat{\eta}}_m(t) = h^*_m(t)$ in every slot t within period p and for every source m. □

From Lemma 11, we have that if period p is an optimal period, then

$$\sum_{m=1}^{M}\sum_{t=s_p}^{f_p} h^{\hat{\eta}}_m(t) = \sum_{m=1}^{M}\sum_{t=s_p}^{f_p} h^*_m(t)\ . \tag{22}$$

Using this result in the expression of the auxiliary AoI regret, we obtain

$$\begin{aligned}
&\mathrm{E}\left[\sum_{m=1}^{M}\sum_{t=1}^{T}\left[h^{\hat{\eta}}_m(t) - h^*_m(t)\right]\right] = \\
&= \mathrm{E}\left[\sum_{m=1}^{M}\sum_{t=1}^{T}\underbrace{\left[h^{\hat{\eta}}_m(t) - h^*_m(t)\right]}_{\leq h^{\hat{\eta}}_m(t)}\mathbb{1}\left\{n(t)\neq n^*\cap E(t)=0\right\} + \right.\\
&\qquad \left. + \underbrace{\left[h^{\hat{\eta}}_m(t) - h^*_m(t)\right]\mathbb{1}\left\{n(t)=n^*\cup E(t)=1\right\}}_{=0}\right]\\
&\overset{(a)}{\leq} \mathrm{E}\left[\sum_{m=1}^{M}\sum_{p=1}^{T} y^{\hat{\eta}}_m(p)\ \mathbb{1}\left\{\bar{n}(p)\neq n^*\right\}\right]\\
&= \sum_{m=1}^{M}\sum_{p=1}^{T}\mathrm{E}\left[y^{\hat{\eta}}_m(p)\ \mathbb{1}\left\{\bar{n}(p)\neq n^*\right\}\right]\\
&\overset{(b)}{=} \sum_{m=1}^{M}\sum_{p=1}^{T}\mathrm{E}\left[y^{\hat{\eta}}_m(p)\ \middle|\ \bar{n}(p)\neq n^*\right]\mathrm{P}\left(\bar{n}(p)\neq n^*\right)
\end{aligned}\tag{23}$$

where (a) follows from the fact that each period p has duration of at least 1 time slot, and (b) follows from the law of total expectation.

APPENDIX B
PROOF OF LEMMA 10

To establish Lemma 10, we first use Hoeffding's inequality to upper bound $\mathrm{P}\left(\bar{n}(p) = n\right)$ by an exponential function of $-p$, for every suboptimal channel n. The result then follows directly from this upper bound.

Lemma 12. *For every suboptimal channel index $n \neq n^*$, the probability of channel policy $\bar{\eta}_b$ selecting channel n in the nonempty phase of period p is bounded by*

$$\mathrm{P}\left(\bar{n}(p) = n\right) \leq 2\exp\left(-\frac{1}{2N^2}p\right) + 2\exp\left(-\frac{\Delta_n^2}{4N}p\right)\ , \tag{24}$$

where $\Delta_n = \mu^* - \mu_n$.

Proof. The channel policy $\bar{\eta}_b$ uses the empty phases of each period p to explore the channels and update its estimate of the channel reliabilities. We know that at the beginning of each period p there is an empty phase with at least one exploration slot. Let $N_n(p)$ be the number of exploration slots in which channel n is selected within the first p periods and let p' be the total number of exploration slots within the first p periods. It follows that

$$\sum_{n=1}^{N} N_n(p) = p' \geq p \quad \text{and} \quad \mathrm{E}\left[N_n(p)\right] = \frac{1}{N}p'\ . \tag{25}$$

Let $\hat{\mu}_n(p')$ be the estimate of the reliability of channel n after a total of p' exploration slots. Then, for any suboptimal channel $n \neq n^*$, we have

$$\begin{aligned}
&\mathrm{P}\left(\bar{n}(p)=n\right) \leq \mathrm{P}\left(\hat{\mu}_n(p') \geq \hat{\mu}^*(p')\right) \\
&\overset{(a)}{\leq} \mathrm{P}\left(\hat{\mu}_n(p') \geq \frac{\mu_n+\mu^*}{2}\right) + \mathrm{P}\left(\frac{\mu_n+\mu^*}{2} \geq \hat{\mu}^*(p')\right) \\
&= \mathrm{P}\left(\hat{\mu}_n(p')-\mu_n \geq \frac{\Delta_n}{2}\right) + \mathrm{P}\left(\hat{\mu}^*(p')-\mu^* \leq -\frac{\Delta_n}{2}\right) .
\end{aligned} \tag{26}$$

where (a) follows from the union bound. Denote $\hat{\mu}_n(p')-\mu_n = I_n$. Then,

$$\begin{aligned}
&\mathrm{P}\left(\hat{\mu}_n(p')-\mu_n \geq \frac{\Delta_n}{2}\right) \\
&= \mathrm{P}\left(I_n \geq \frac{\Delta_n}{2},\, N_n(p) \leq \frac{p'}{2N}\right) + \mathrm{P}\left(I_n > \frac{\Delta_n}{2},\, N_n(p) > \frac{p'}{2N}\right) \\
&\leq \mathrm{P}\left(N_n(p) \leq \frac{p'}{2N}\right) + \mathrm{P}\left(I_n > \frac{\Delta_n}{2} \,\middle|\, N_n(p) > \frac{p'}{2N}\right) .
\end{aligned} \tag{27}$$

Next, we upper bound each of the last two terms by Hoeffding's inequality.

Notice that:

- $N_n(p)$ is the sum of p' i.i.d. Bernoulli random variables with mean $\frac{1}{N}$; and
- $\hat{\mu}_n(p')$ is the average of $N_n(p)$ i.i.d. Bernoulli random variables with mean μ_n.

Thus, by Hoeffding's inequality

$$\begin{aligned}
\mathrm{P}\left(N_n(p) \leq \frac{1}{2N}p'\right) &= \mathrm{P}\left(\frac{N_n(p)}{p'} - \frac{1}{N} \leq -\frac{1}{2N}\right) \\
&\leq \exp\left(-\frac{1}{2N^2}p'\right)
\end{aligned} \tag{28}$$

and

$$\mathrm{P}\left(I_n > \frac{\Delta_n}{2} \,\middle|\, N_n(p) > \frac{1}{2N}p'\right) \leq \exp\left(-\frac{\Delta_n^2}{4N}p'\right) . \tag{29}$$

Inequalities (27), (28) and (29) imply

$$\mathrm{P}\left(\hat{\mu}_n(p')-\mu_n \geq \frac{\Delta_n}{2}\right) \leq \exp\left(-\frac{1}{2N^2}p'\right) + \exp\left(-\frac{\Delta_n^2}{4N}p'\right) . \tag{30}$$

Analogously, we have

$$\mathrm{P}\left(\hat{\mu}^*(p')-\mu^* \leq \frac{\Delta_n}{2}\right) \leq \exp\left(-\frac{1}{2N^2}p'\right) + \exp\left(-\frac{\Delta_n^2}{4N}p'\right) . \tag{31}$$

Now, inequalities (26), (30) and (31) imply that

$$\begin{aligned}
\mathrm{P}\left(\bar{n}(p')=n\right) &\leq 2\exp\left(-\frac{1}{2N^2}p'\right) + 2\exp\left(-\frac{\Delta_n^2}{4N}p'\right) \\
&\leq 2\exp\left(-\frac{1}{2N^2}p\right) + 2\exp\left(-\frac{\Delta_n^2}{4N}p\right)
\end{aligned} \tag{32}$$

□

Using Lemma 12, we obtain

$$\begin{aligned}
\sum_{p=1}^{T} \mathrm{P}\left(\bar{n}(p) \neq n^*\right) &= \sum_{n\neq n^*}\sum_{p=1}^{T} \mathrm{P}\left(\bar{n}(p)=n\right) \\
&\leq \sum_{n\neq n^*}\sum_{p=1}^{T}\left[2\exp\left(-\frac{1}{2N^2}p\right) + 2\exp\left(-\frac{\Delta_n^2}{4N}p\right)\right] \\
&\leq \sum_{n\neq n^*}\sum_{p=1}^{\infty}\left[2\exp\left(-\frac{1}{2N^2}p\right) + 2\exp\left(-\frac{\Delta_n^2}{4N}p\right)\right] \\
&= \sum_{n\neq n^*}\left[\frac{2}{\exp\left(\frac{1}{2N^2}\right)-1} + \frac{2}{\exp\left(\frac{\Delta_n^2}{4N}\right)-1}\right] \\
&\leq \frac{2(N-1)}{\exp\left(\frac{1}{2N^2}\right)-1} + \frac{2(N-1)}{\exp\left(\frac{\Delta_{min}^2}{4N}\right)-1}
\end{aligned} \tag{33}$$

which is a constant, where $\Delta_{min} = \mu^* - \max_{n\neq n^*}\mu_n$.