Efficient Learning-based Scheduling for Information Freshness in Wireless Networks

Bin Li
Department of Electrical, Computer, and Biomedical Engineering
University of Rhode Island, Kingston, RI 02881, USA
Email: binli@uri.edu

***Abstract*—Motivated by the recent trend of integrating artificial intelligence into the Internet-of-Things (IoT), we consider the problem of scheduling packets from multiple sensing sources to a central controller over a wireless network. Here, packets from different sensing sources have different values or degrees of importance to the central controller for intelligent decision making. In such a setup, it is critical to provide timely and valuable information for the central controller. In this paper, we develop a parameterized maximum-weight type scheduling policy that combines both the AoI metrics and Upper Confidence Bound (UCB) estimates in its weight measure with parameter η. Here, UCB estimates balance the tradeoff between exploration and exploitation in learning and are critical for yielding a small cumulative regret. We show that our proposed algorithm yields the running average total age at most by $O(N^2\eta)$. We also prove that our proposed algorithm achieves the cumulative regret over time horizon T at most by $O(NT/\eta+\sqrt{NT\log T})$. This reveals a tradeoff between the cumulative regret and the running average total age: when increasing η, the cumulative regret becomes smaller, but is at the cost of increasing running average total age. Simulation results are provided to evaluate the efficiency of our proposed algorithm.**

I. Introduction

With the recent advances in artificial intelligence (AI), there is a trend for incorporating AI into the Internet-of-Things (IoT) consisting of multiple wireless sensing sources to provide wise decisions. In such an IoT system, it is critical to make sure that the received sensing information is valuable and timely. As such, in this paper, we consider the problem of scheduling packets from multiple sensing sources to a central controller over a wireless network as shown in Fig. 1, where packets from different sensing sources have different values or degrees of importance to the central controller for decision making.

In particular, we assume that each sensing source constantly generates packets with random values independently and identically distributed (i.i.d.) with an unknown distribution. The value of a packet is revealed only after the central controller successfully receives it. On the one hand, we would like to deliver packets from the most important sensing sources to the central controller for making a better decision subject to the wireless interference constraints. However, the controller does not have any prior knowledge of the degree of importance of these sensing sources and requires to gradually learn these statistics while scheduling the best sensing sources (a.k.a. exploration-exploitation tradeoff in online learning). On the other hand, we should also ensure that the received packets have a low Age-of-Information (AoI) that measures the duration between the packet generation time and its received time. This is because the stale information is less useful to the central controller and might even mislead the controller to make harmful decisions. To that end, we aim to develop a scheduling algorithm to achieve this dual objective, which is complicated by the strong coupling between the learning and AoI performance.

This research has been supported in part by NSF grants: CNS-1717108, CNS-1815563, and CNS-1942383.

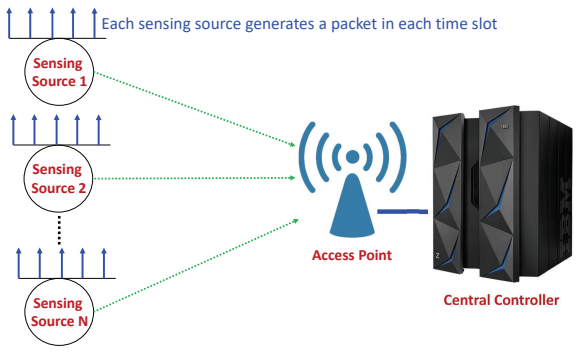


Fig. 1: An intelligent Internet-of-Things (IoT) system.

Without the AoI constraint, the considered problem can be formulated as a combinatorial multi-armed bandit (MAB) problem (e.g., [1], [2], [3], [4]), where each arm corresponds to a sensing source, and the goal is to minimize the cumulative regret over a finite time horizon (i.e., the difference between the optimal cumulative reward and the cumulative reward under an algorithm). The combinatorial MAB algorithms allow to play multiple arms simultaneously in each time slot instead of one arm as in classical MAB algorithms such as Upper-Confidence-Bound (UCB [5]), Kullback-Leibler UCB (KL-UCB [6]), and Thompson sampling [7]. The efficient combinatorial MAB algorithms should quickly identify the set of best arms and keep pulling them. This, however, leads to the large AoI for other relatively poor arms. In particular, the AoI keeps increasing over time and this implies that the received information from relatively poor arms is outdated. This motivates us to incorporate the AoI metric into the learning algorithm design for combinatorial MAB problems.

While there are some recent works on AoI-efficient wireless

scheduling (e.g, [8], [9], [10] and see [11] for an overview), their goal was to minimize AoI while guaranteeing the desired throughput. They did not consider the MAB setting with unknown system statistics, which is typical in intelligent IoT systems. As such, in this paper, we integrate the main ideas of UCB algorithm (e.g., [5]) and AoI-efficient scheduling (e.g., [8]), and propose a Learning-based Age-Efficient Scheduling (LAES) Algorithm that utilizes both the UCB estimates and AoI metrics. While there are some recent works in combinatorial bandits with fairness constraints (e.g., [12]), they focused on the long-term fairness constraint, i.e., each arm should at least be played for a fixed fraction of times on average. The main approach is to maintain a virtual queue for each arm that keeps track of its debt and prioritizes arms with high virtual queue lengths, typically referred to as virtual queue techniques (e.g., see [13] for an overview). However, the AoI captures the short-term dynamic of the system and thus its evolution is fundamentally different from the virtual queue length. In particular, it has an unbounded decrement whenever a packet is successfully delivered, which has a significant impact on the performance analysis of the proposed algorithm. Our contributions in this work are summarized as follows:

- We develop a parameterized maximum-weight type scheduling policy that combines both the AoI metric and UCB estimate in its weight measure (cf. Section IV). In particular, we use the parameter η to balance the AoI metric and UCB estimate. The larger the η, the more emphasis on the UCB estimate and thus leads to the smaller regret, but it is at the cost of the larger AoI.
- We derive an upper bound on the running average total age under our proposed algorithm with any $\eta > 0$ (cf. Proposition 1), which linearly increases with the parameter η. Such an upper bound is tight in some cases in the sense that the average total average under our proposed algorithm linearly scales with the parameter η.
- We show that the cumulative regret over a finite time horizon T can be bounded from above by $O(NT/\eta + \sqrt{NT\log T})$[1] under our proposed algorithm (cf. Proposition 2). Here, the second term has the same order as that of the UCB algorithm and is attributed to the cost for exploration/exploitation in online learning, while the first term NT/η is the cost paid for improving AoI performance. This, together with the derived upper bound on the running average total age, reveals a tradeoff: when increasing η, the regret upper bound decreases, but the upper bound on running average total age increases.
- We support our analytical results with extensive simulations (cf. Section V), which demonstrates the superior performance of our proposed algorithm over both UCB algorithm and age-based algorithm (i.e., our proposed algorithm with $\eta = 0$). Simulation results also confirm a tradeoff between the cumulative regret and the running average total age. The desired tradeoff can be achieved by tuning the value of parameter η.

[1] $f(x) = O(x)$ if there exists a positive real number M such that $f(x) \leq Mx, \forall x \geq 0$.

The remainder of this paper is organized as follows: Section II reviews related work. Section III introduces system model and problem statement. Section IV introduces our proposed algorithm, and analyzes both its AoI and regret performance. Section V presents simulation results and Section VI concludes this paper.

II. Related Work and Context

In this section, we overview two main areas that are closely related to our work: multi-armed bandit and age of information, and further provide a brief discussion of our design methodology in the context of prior work.

(a) Multi-Armed Bandit: The MAB problem models an agent that attempts to learn system statistics while optimizing its decision based on existing learning experiences, and has wide applications in recommender systems, healthcare, finance, and computer networks. As such, it has been received extensive research efforts (e.g., [14], [5], [6], [7]). The seminal work of Lai and Robbins [14] established a fundamental logarithmic lower bound on the cumulative regret (i.e., the difference between the optimal cumulative reward and the cumulative reward under an algorithm) over a finite time horizon under a class of uniformly good policies and developed a UCB algorithm that asymptotically achieves this fundamental lower bound. Such a logarithmic regret bound has been shown to be achieved by the sample-mean-based UCB algorithm and ϵ-greedy policy (see [5]), Kullback-Leibler UCB (KL-UCB [6]), and Thompson sampling [7].

Subsequent works extended the classical MAB problem to various settings that account for different applications. The one closest to ours is combinatorial MAB (e.g., [1], [2], [3], [4]), where a subset of arms can be played simultaneously at each time. More recent works considered the combinatorial MAB with fairness constraint (e.g., [12], [15]), where each arm should at least be played for a certain fraction of time on average. The authors introduced the virtual-queue-length to address fairness constraint and incorporated it into the algorithm design. However, all these MAB works did not address the AoI performance and thus yet unbounded AoI over time, as demonstrated in our simulations (cf. Section V).

(b) Age of Information: AoI measures the duration between the time when the information was generated and its received time. It directly captures the information freshness and thus has received great attention in recent years. Unlike the traditional queueing delay that is negligible in the case with a low sampling rate (i.e., low arrival rate), the AoI is dominated by the inter-arrival time and thus is rather large in the low sampling rate regime. This key difference has spurred AoI research in several aspects in recent years, e.g., AoI analysis and optimization (e.g., [16], [17]), AoI in vehicular networks (e.g., [18], [19]), online sampling and remote estimation (e.g., [20], [21]), AoI and energy harvesting (e.g., [22], [23], [24]), just to name a few.

The one that is closest to our research is the AoI-efficient scheduling in wireless networks (e.g., [8], [9], [10] and see [11] for an overview) that aims to develop wireless scheduling algorithms with the goal of minimizing AoI. For example, the authors in [8] developed an age-based scheduler for real-time traffic that achieves not only desired timely throughput but also guaranteed AoI performance. Our research differs from this line of research in that we explicitly incorporate AoI metrics into the MAB algorithm design, which is desirable in the emerging intelligent IoT applications. This key difference poses significant challenges in guaranteeing information freshness in the MAB setting that is unseen in existing AoI research. While a recent work [25] considered the AoI performance in the MAB setting, it focused on the single-user setting and did not consider the case with multiple users and wireless interference constraints.

(c) Our Design Philosophy: In this paper, we extend a UCB-type algorithm to our setting that demands desired AoI performance while minimizing cumulative regret over time. One extreme is to serve arms with the largest UCB estimates in order to minimize the cumulative regret, but it can result in increasing AoI over time. The other extreme is to serve arms with the largest ages, yet this could lead to a large regret. This is because it does not learn any system statistics nor exploit the best arms so far. Therefore, it is clear that one should tradeoff the benefits of these two approaches. The natural idea is to integrate both UCB estimates and AoI metrics into the scheduling decisions. However, the AoI metric in our work is fundamentally different from the virtual queue length, since it has an unbounded decrement whenever a packet is successfully delivered. Such an abrupt dynamic poses a significant challenge in characterizing AoI performance. The main contribution of this paper is to develop a parameterized learning-based age-efficient algorithm and to show that such an algorithm achieves a tradeoff between the cumulative regret and average total age, which can be tuned by our algorithmic parameter.

III. SYSTEM MODEL

We consider a wireless network with N links, where each link represents a transmitter-receiver pair that are within the transmission range of each other. We assume that the system operates in slotted time with normalized slots $t \in \{0, 1, 2, \ldots\}$. In each time slot t, the transmitter of link n ($n = 1, 2, \ldots, N$) generates a packet with a random value $X_n(t) \in [0, 1]$, which is independently and identically distributed (i.i.d.) with an unknown mean μ_n. Here, $X_n(t)$ represents the *reward* when a packet is successfully delivered over link n in time slot t. Due to the wireless interference constraints, only a subset of links can transmit in each time slot. We use $S_n(t) = 1$ if link n is scheduled for transmission in time slot t, and $S_n(t) = 0$ otherwise. We call $\mathbf{S}(t) \triangleq (S_n(t))_{n=1}^N$ the *feasible schedule* denoting the set of links that can be active simultaneously in time slot t. Let $\mathcal{S}$ be the collection of all feasible schedules. We assume that each link n experiences i.i.d. ON-OFF channel fading over time with $C_n(t) = 1$ denoting that the channel of link n is ON in time slot t. Let $p_n \triangleq \Pr\{C_n(t) = 1\}$ be the probability that link n has an available channel in time slot t. We assume that each link has a non-zero probability that its channel is ON, i.e., $p_{\min} \triangleq \min_n p_n > 0$. Hence, the received reward $R(t)$ in each time slot t can be expressed as $R(t) \triangleq \sum_{n=1}^N X_n(t)C_n(t)S_n(t)$. We consider the case where the channel state is known via channel probing at the beginning of each time slot[2].

Our goal is to maximize the cumulative reward $\sum_{t=0}^{T-1} R(t)$ until the T^{th} time slot while guaranteeing the desired information freshness. If the statistics of rewards (i.e., $\{\mu_n, n = 1, 2, \ldots, N\}$) are known in advance, then the first objective can be achieved by solving the following optimization problem:

$$\mathbf{S}^*(t) \triangleq (S_n^*(t))_{n=1}^N \in \arg\max_{\mathbf{S}\in\mathcal{S}} \sum_{n=1}^N \mu_n C_n(t) S_n. \quad (1)$$

That is, it serves a set of non-interfering and available links with the maximum sum of mean rewards in each time slot. Unfortunately, the statistics of rewards are unknown. This requires the algorithm not only to learn these statistics (also known as (a.k.a.) exploration) but also to select the best schedule so far (a.k.a. exploitation). Our first goal is equivalent to minimizing the *cumulative regret* over consecutive T time slots, which is the gap between the accumulated reward and the optimal reward, i.e.,

$$\mathrm{Reg}(T) \triangleq \sum_{t=0}^{T-1} \sum_{n=1}^{N} \left(\mathbb{E}\left[\mu_n C_n(t) S_n^*(t)\right] - \mathbb{E}\left[\mu_n C_n(t) S_n(t)\right]\right).$$

To address our second goal for the desired information freshness, we introduce $Z_n(t)$ to denote the *age* of information received from the n^{th} link in time slot t, which increases by one if a packet is not received by the receiver of link n in time slot t and reset to one otherwise, i.e.,

$$Z_n(t+1) = \begin{cases} Z_n(t) + 1 & \text{if } S_n(t)C_n(t) = 0; \\ 1 & \text{if } S_n(t)C_n(t) = 1. \end{cases} \quad (2)$$

Fig. 2 shows one sample path of age of link n. We can obverse from Fig. 2 that $Z_n(t)$ resets to one whenever there is a successful packet delivery. We note that the dynamics of the age is similar to that of Time-Since-Last-Service (TSLS) counter in [26], [27], [28], [29].

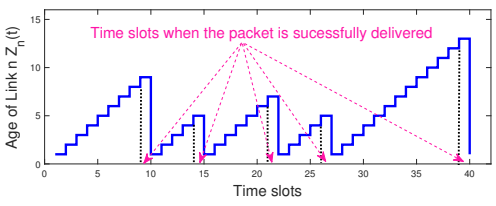


Fig. 2: The evolution of age of link n.

[2]Our algorithm design and its analysis can be easily adapted to the case with unknown channel state.

Our second goal is to keep information as fresh as possible, i.e., minimizing $\sum_{n=1}^{N}\mathbb{E}[Z_n(t)]$. We achieve this dual objective by developing a parametric class of wireless schedulers that efficiently utilize a combination of UCB estimates for minimizing the cumulative regret and ages in its decision.

IV. Algorithm Design and Performance Analysis

In this section, we develop a learning-based wireless scheduler by integrating the key idea of the well-known UCB algorithms (see [5]) and age metrics. In particular, the UCB is utilized to deal with the fundamental exploitation-exploration tradeoff in online learning and aims to achieve minimum cumulative regret. On the other hand, age metrics are employed to guarantee desired information freshness.

In order to obtain the weight for exploitation and exploration, we introduce the following notations. Let $H_n(t)$ be the number of times link n has successfully received a packet until time slot t, i.e., $H_n(t)\triangleq\sum_{\tau=0}^{t-1}C_n(\tau)S_n(\tau)$. We set $H_n(0)=0$ due to the fact that the system starts at $t=0$. We use $\overline{\mu}_n(t)$ to denote the sample mean of the received rewards of link n until time slot t, i.e., $\overline{\mu}_n(t)\triangleq\left(\sum_{\tau=0}^{t-1}X_n(\tau)C_n(\tau)S_n(\tau)\right)/H_n(t)$. If $H_n(t)=0$ (i.e., link n has not successfully received a packet yet until time slot t), we set $\overline{\mu}_n(t)=1$. Let $w_n(t)$ denote the *UCB estimate* of link n in time slot t and is defined as follows:

$$w_n(t)\triangleq\min\left\{\overline{\mu}_n(t)+\sqrt{\frac{3\log t}{2H_n(t)}},1\right\}, \tag{3}$$

where $\sqrt{3\log t/(2H_n(t))}$ is the exploration term that measures the uncertainty of the received reward of link n until time slot t. Indeed, the smaller the $H_n(t)$, the less exploitation of link n and thus less accuracy of its sample mean estimation, in which case link n should get a higher priority to be scheduled. Here, we use the truncated version of the UCB estimate, since the actual reward of each link is at most 1. Again, when $H_n(t)=0$, we set $w_n(t)=1$. That is, if link n has not been scheduled yet until time slot t, it has the highest priority to get served.

In order to achieve a low cumulative regret, we prefer to serve links with large UCB estimates in each time slot. Indeed, we would like to serve links with high sample mean rewards and links with large uncertainties of received rewards due to fewer explorations. In order to address information freshness guarantees, we also need to incorporate age metrics into the scheduling design. In particular, the links with large ages should get high priorities to be scheduled. This naturally motivates the following algorithm.

Note that η is a parameter that balances the age metrics and the UCB estimates. In particular, if $\eta=0$, then the LAES coincides with the age-based policy (see [11, Ch. 4.5.4]). In the presence of fully-connected networks with non-fading channels (i.e., at most one link can be scheduled in each time slot and $C_n(t)=1,\forall n,t\geq 0$), the age-based policy is equivalent to the well-known Round-Robin policy that serves links in turn. In such a case, the age-based policy, in fact, minimizes the average total age (see [26, Proposition 2]). The larger the η, the higher priority the UCB estimates and hence yields a smaller cumulative regret.

Algorithm 1 Learning-based Age-Efficient Scheduling (LAES) Algorithm

In each time slot t, given the channel state $(C_n(t))_{n=1}^N$, select a schedule $\widehat{\mathbf{S}}[t]\triangleq(\widehat{S}_n(t))_{n=1}^N$ satisfying

$$\widehat{\mathbf{S}}[t]\in\arg\max_{\mathbf{S}\in\mathcal{S}}\sum_{n=1}^{N}\left(Z_n(t)+\eta w_n(t)\right)C_n(t)S_n, \tag{4}$$

where $\eta\geq 0$ is some control parameter.

Next, we characterize the age performance of the proposed LAES Algorithm.

Proposition 1: [Information Freshness Guarantee] If $Z_n(0)=0,\forall n=1,2,\ldots,N$, then, under the LAES algorithm with any $\eta\geq 0$, the running average total age can be bounded from above as follow:

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_{n=1}^{N}\mathbb{E}\left[Z_n(t)\right]\leq\frac{(\eta+1)N^2}{p_{\min}},$$

holding for any $T\geq 1$, where $p_{\min}\triangleq\min_n p_n>0$.

Proof: We consider the Lyapunov function $V(t)$

$$V(t)\triangleq\sum_{n=1}^{N}Z_n(t). \tag{5}$$

and study its drift. Using telescoping techniques as in the classical Lyapunov drift analysis (e.g., [13]), we obtain an upper bound on the running average total age. Please see Appendix A for the detailed proof. ∎

Remarks 1: From Proposition 1, we can see that the running average total age is bounded under the LAES Algorithm with any $\eta\geq 0$, which is desirable since the central controller always demands a certain degree of information freshness. In addition, the derived upper bound on the running average total age linearly increases with the parameter η. This matches our intuition on the LAES Algorithm that a large η implies a smaller weight on the age metric and thus deteriorates the AoI performance.

Remarks 2: The derived upper bound on the running average total age linearly scales with the parameter η, which might be tight in some cases. Indeed, consider two interfering non-fading links, where $C_n(t)=1,\forall n=1,2,\forall t$, and at most one link can be scheduled in each time slot. Suppose $\mu_1>\mu_2$, and assume that both links are scheduled sufficiently many times. In such a case, both $w_1(t)$ and $w_2(t)$ are close to μ_1 and μ_2, respectively. As such, under the LAES Algorithm, link 2 is scheduled roughly one every $\lceil\eta(\mu_1-\mu_2)\rceil$ time slots and link 1 is scheduled in all other time slots. Hence, the average age of link 2 in each time slot is roughly equal to $(1+2+3+\ldots+\lceil\eta(\mu_1-\mu_2)\rceil)/\lceil\eta(\mu_1-\mu_2)\rceil=(1+\lceil\eta(\mu_1-\mu_2)\rceil)/2$. On the other hand, the average age of link 2 in each time slot is equal to $(\lceil\eta(\mu_1-\mu_2)\rceil-1+2)/\lceil\eta(\mu_1-\mu_2)\rceil=1+1/\lceil\eta(\mu_1-\mu_2)\rceil$. Hence, the average total age in each time slot is $O(\eta)$.

Remarks 3: In the case that all links have a non-zero probability of the channel being OFF (i.e., $p_n < 1, \forall n = 1, 2, \ldots, N$), the upper bound on the mean total age is independent of the parameter η. Indeed, if the event $\mathcal{F}_n(\tau) \triangleq \{C_n(\tau) = 1, C_{n'}(\tau) = 0, \forall n' \neq n\}$ happens for some $\tau \in [t-m+1, t)$, then under the LAES Algorithm, link n should be scheduled at least once during the past m time slots, and thus $Z_n(t) < m$. This implies that

$$\begin{aligned}&\Pr\{Z_n(t) \geq m\}\\ &\leq \Pr\{\mathcal{F}_n(\tau) \text{ does not happen for all } \tau \in [t-m+1, t)\}\\ &\overset{(a)}{=} \nu_n^m \overset{(b)}{\leq} \nu^m,\end{aligned} \tag{6}$$

where step (a) is true for $\nu_n \triangleq 1 - p_n \Pi_{n' \neq n}(1-p_{n'}) \in (0,1)$ under the assumption that $p_n < 1, \forall n$, and follows from the fact that channel rates are independently distributed across links and i.i.d. over time for each link, and (b) holds for $\nu \triangleq \max_n \nu_n$. Hence, we have

$$\mathbb{E}[Z_n(t)] = \sum_{m=1}^{\infty} \Pr\{Z_n(t) \geq m\} \leq \sum_{m=1}^{\infty} \nu^m = \frac{\nu}{1-\nu}. \tag{7}$$

As such, the average total age in each time slot is upper bounded by $N\nu/(1-\nu)$, which is independent of parameter η. However, such an upper bound is extremely large, as demonstrated in Section V, and thus it does not say too much on the dependence of the average total age on the parameter η when the average age is moderate or small.

Lastly, we provide an upper bound on the cumulative regret under the LAES Algorithm with $\eta > 0$.

Proposition 2: [Upper Bound on Regret] If $Z_n(0) = 0, \forall n = 1, 2, \ldots, N$, then, under the LAES Algorithm with $\eta > 0$, the cumulative regret $\text{Reg}(T)$ until time slot $T > 0$ can be bounded from above as follows:

$$R(T) \leq \frac{NT}{\eta} + 2\sqrt{6N|\mathbf{S}|_{\max} T \log T} + N\left(1 + \frac{5\pi^2}{12}\right),$$

where $|\mathbf{S}|_{\max}$ denotes the maximum number of links that can be scheduled simultaneously in each time slot.

Proof: We first perform drift-plus-penalty analysis on

$$\mathbb{E}[V(t+1) - V(t)] + \eta \Delta R(t), \tag{8}$$

where $\Delta R(t) \triangleq \sum_{n=1}^{N} \mathbb{E}\left[\mu_n C_n(t) S_n^*(t) - \mu_n C_n(t) \widehat{S}_n(t)\right]$ and the cumulative regret $\text{Reg}(T) \triangleq \sum_{t=0}^{T-1} \Delta R(t)$. Then, we carefully incorporate the regret analysis for classical UCB algorithm (e.g., [5]) into our analysis. The analysis is similar to the line of regret analysis in [30] and [12], and is available in Appendix B. ∎

Remarks 4: The derived upper bound on the cumulative regret consists of two terms: (i) $2\sqrt{6N|\mathbf{S}|_{\max} T \log T} + N(1 + 5\pi^2/12)$ has the same order $O(\sqrt{NT \log T})$ as the instance-independent upper bound for the classical UCB algorithm (see [31, Ch. 2.4.3]) and thus this term is attributed to the cost involved in the exploration/exploitation process in online learning; (ii) NT/η decreases as parameter η increases. This also matches our intuition on the LAES Algorithm: the larger the η, the larger the weight put on the UCB estimate and thus yields a smaller cumulative regret.

This together with Proposition 1 reveals a tradeoff between the running average total age and the regret performance in the general network setup under the LAES Algorithm: when increasing η, the regret upper bound decreases, but the upper bound on running average total age increases. That is, the improvement of the cumulative regret is at the cost of increasing running average total age. Moreover, it can be easily derived that the product of the upper bound on the running average total age and the upper bound on the cumulative regret is on the order of $O(N^3 T)$ for any $\eta \leq O(\sqrt{NT/\log T})$. In Table I, we provide three different η values to illustrate the tradeoff between the running average total age and cumulative regret.

Parameter η	Regret	Age
$O(\sqrt{NT/\log T})$	$O(\sqrt{NT \log T})$	$O(\sqrt{N^5 T/\log T})$
$O(\sqrt[3]{NT})$	$O(\sqrt[3]{N^2 T^2})$	$O(\sqrt[3]{N^7 T})$
$O(1)$	$O(NT)$	$O(N^2)$

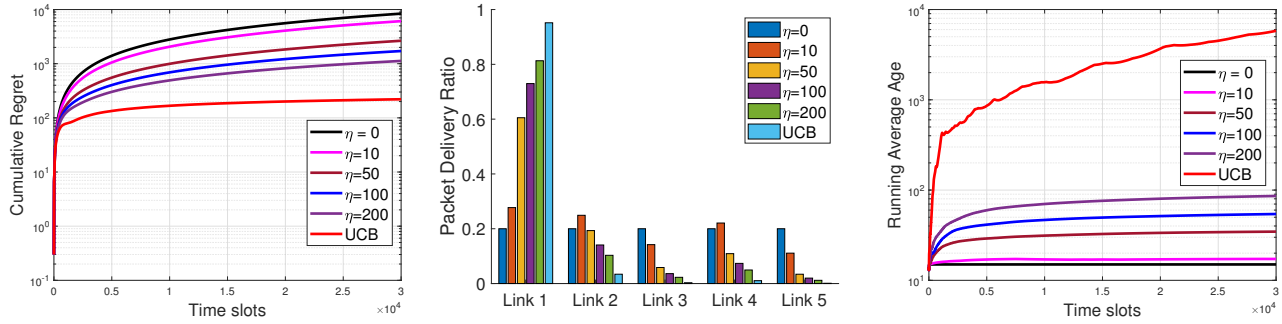
TABLE I: Cumulative Regret vs. Running Average Age.

From Table I, we can see that in order for the cumulative regret to be on the same order as that for UCB algorithm (i.e., $\text{Reg}(T) = O(\sqrt{NT \log T})$), the running average total age should be on the order of $O(\sqrt{N^5 T/\log T})$ under the LAES Algorithm. Nevertheless, we can use a relatively large η (e.g., $\eta = 200$) and achieve both low regret and low average age, as demonstrated in Section V.

V. Simulations

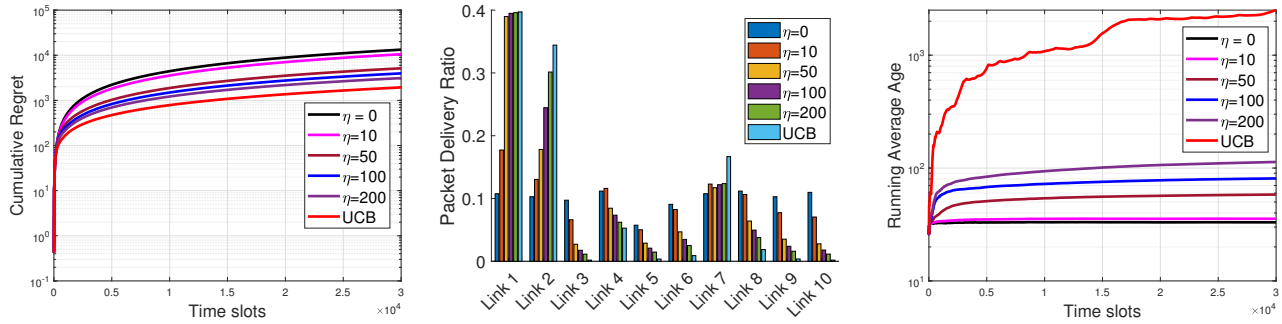
In this section, we perform simulations to evaluate the performance of our proposed LAES Algorithm. We consider the following two network setups: (i) a fully-connected non-fading network with $N = 5$ links (at most one link can be scheduled in each time slot and $C_n(t) = 1, \forall n, t \geq 0$), and (ii) a $10-$link ON-OFF fading network where at most two links can be scheduled in each time slot. For the first setup, the mean reward vector is $\boldsymbol{\mu} = (0.9, 0.8, 0.5, 0.7, 0.2)$. For the second network setup, we set the mean reward vector $\boldsymbol{\mu} = (0.9, 0.8, 0.4, 0.7, 0.5, 0.6, 0.75, 0.65, 0.5, 0.4)$ and heterogeneous ON-OFF channel fading parameters $\mathbf{p} = (0.8, 0.7, 0.6, 0.9, 0.2, 0.5, 0.8, 0.9, 0.7, 0.85)$. We compare the LAES Algorithm with $\eta \in \{0, 10, 50, 100, 200\}$ with the UCB algorithm that makes the decision only based on the UCB estimates. Note that the LAES with $\eta = 0$ coincides with the age-based scheduler. We run 500 experiments, each of which has simulated 3×10^4 time slots.

Fig. 3 shows the performance of UCB algorithm and LAES Algorithm with different η values in the fully-connected non-fading network. We can observe from Fig. 3a that UCB algorithm outperforms the LAES Algorithm with all η values in terms of cumulative regret performance. The larger the η, the smaller the cumulative regret. This is because the larger η puts more weight on the UCB estimates and thus the LAES



(a) Cumulative Regret (b) Successful Delivery Ratio (c) Total Average Age

Fig. 3: Performance of the LAES Algorithm in a fully-connected non-fading network.



(a) Cumulative Regret (b) Successful Delivery Ratio (c) Total Average Age

Fig. 4: Performance of the LAES Algorithm in a $10-$link ON-OFF fading network.

Algorithm with an extremely large η value should have similar regret as the UCB algorithm. Indeed, we can see from Fig. 3b that the packet successful delivery ratio of the best link (i.e., link 1) increases as η increases.

However, the cumulative regret performance improvement is at the cost of increasing running average total age, as shown in Fig. 3c that shows the running average of total age over time. Indeed, we can observe from Fig. 3c that under the LAES Algorithm, the age becomes larger as η increases. Nevertheless, it is worth pointing out that the age keeps increasing over time under the UCB algorithm while it is always bounded under the LAES Algorithm with all fixed η values. We can observe similar phenomena in a relatively complicated network with 10 links, as shown in Fig. 4, despite the derived upper bound on the average total average (cf. (7)) in each time slot is independent of the parameter η. This is because such an upper bound is equal to 4.6×10^7 (much larger than 100).

VI. Conclusion

In this paper, we considered the problem of scheduling packets from multiple sensing sources to a central controller over a wireless network with the goal of minimizing cumulative regret over time while guaranteeing desired AoI performance. We developed a parameterized maximum-weight type scheduling policy that combines both the AoI metrics and UCB estimates in its weight measure with parameter η. We derived an upper bound on the running average total age, which linearly increases with the parameter η. We also derived an upper bound on the cumulative regret under our proposed algorithm. These derived upper bounds reveal a tradeoff: the improvement of the cumulative regret is at the cost of increasing running average total age. Simulation results were provided to confirm such a tradeoff and to demonstrate the superior performance of our proposed algorithm over the UCB algorithm and the age-based algorithm.

Appendix A
Proof of Proposition 1

Select the Lyapunov function

$$V(t) \triangleq \sum_{n=1}^{N} Z_n(t). \tag{9}$$

Then, under the LAES Algorithm, we have

$$\begin{aligned} V(t+1) &= \sum_{n=1}^{N} Z_n(t+1) \\ &\overset{(a)}{=} \sum_{n=1}^{N} \Big((Z_n(t)+1)(1 - C_n(t)\widehat{S}_n(t)) + C_n(t)\widehat{S}_n(t) \Big) \end{aligned}$$

$$= \sum_{n=1}^{N} Z_n(t) - \sum_{n=1}^{N} Z_n(t) C_n(t) \widehat{S}_n(t) + N$$
$$\overset{(b)}{=} V(t) - \sum_{n=1}^{N} Z_n(t) C_n(t) \widehat{S}_n(t) + N, \tag{10}$$

where step (a) uses the dynamics of the age (cf. (2)) and (b) follows from the definition of the Lyapunov function $V(t)$. Let $\mathbf{Z}(t) \triangleq (Z_n(t))_{n=1}^N$. Then, we have

$$\mathbb{E}\left[V(t+1) - V(t) | \mathbf{Z}(t)\right]$$
$$= - \mathbb{E}\left[\sum_{n=1}^{N} Z_n(t) C_n(t) \widehat{S}_n(t) \middle| \mathbf{Z}(t)\right] + N. \tag{11}$$

Given the age vector $\mathbf{Z}(t)$ and channel state $\mathbf{C}(t) \triangleq (C_n(t))_{n=1}^N$, according to the definition of the LAES Algorithm, we have

$$\sum_{n=1}^{N} \left(Z_n(t) + \eta w_n(t)\right) C_n(t) \widehat{S}_n(t)$$
$$\geq \left(Z_{n^*(t)}(t) + \eta w_{n^*(t)}\right) C_{n^*(t)}(t)$$
$$\geq Z_{n^*(t)}(t) C_{n^*(t)}(t), \tag{12}$$

where the first step is true for $n^*(t) \in \arg\max_n Z_n(t)$. This implies that

$$\mathbb{E}\left[\sum_{n=1}^{N} Z_n(t) C_n(t) \widehat{S}_n(t) \middle| \mathbf{Z}(t)\right]$$
$$\overset{(a)}{\geq} \mathbb{E}\left[Z_{n^*(t)}(t) C_{n^*(t)}(t) - \eta \sum_{n=1}^{N} w_n(t) C_n(t) \widehat{S}_n(t) \middle| \mathbf{Z}(t)\right]$$
$$\overset{(b)}{\geq} p_{\min} Z_{n^*(t)}(t) - \eta N, \tag{13}$$

where step (a) uses (12), and (b) uses the fact that $w_n(t) \leq 1$ $C_n(t) \leq 1$ and $\widehat{S}_n(t) \leq 1, \forall n, t \geq 0$ and $p_{\min} \triangleq \min_n p_n > 0$.

By substituting (13) into (11), we have

$$\mathbb{E}\left[V(t+1) - V(t) | \mathbf{Z}(t)\right]$$
$$\overset{(a)}{\leq} - p_{\min} Z_{\max}(t) + (\eta + 1) N$$
$$\overset{(b)}{\leq} - \frac{p_{\min}}{N} \sum_{n=1}^{N} Z_n(t) + (\eta + 1) N \tag{14}$$

where step (a) is true for $Z_{\max}(t) \triangleq \max_n Z_n(t) = Z_{n^*(t)}(t)$ and (b) follows from the fact that $Z_{\max}(t) \geq \frac{1}{N} \sum_{n=1}^{N} Z_n(t)$.

Taking the expectation on both sides of (14), we have

$$\mathbb{E}\left[V(t+1) - V(t)\right] \leq - \frac{p_{\min}}{N} \sum_{n=1}^{N} \mathbb{E}\left[Z_n(t)\right] + (\eta + 1) N.$$

Summing the above inequality over time $t = 0, 1, \ldots, T-1$, we have

$$\mathbb{E}\left[V(T) - V(0)\right] \leq - \frac{p_{\min}}{N} \sum_{t=0}^{T-1} \sum_{n=1}^{N} \mathbb{E}\left[Z_n(t)\right] + (\eta + 1) N T,$$

which implies

$$\frac{1}{T} \sum_{t=0}^{T-1} \sum_{n=1}^{N} \mathbb{E}\left[Z_n(t)\right] \leq \frac{(\eta + 1) N^2}{p_{\min}}. \tag{15}$$

Here, we use the fact that $V(0) = 0$ and $V(T) \geq 0$.

Appendix B
Proof of Proposition 2

We rewrite the regret of the LAES Algorithm as

$$\mathrm{Reg}(T) \triangleq \sum_{t=0}^{T-1} \sum_{n=1}^{N} \left(\mathbb{E}\left[\mu_n C_n(t) S_n^*(t)\right] - \mathbb{E}\left[\mu_n C_n(t) \widehat{S}_n(t)\right]\right)$$
$$= \sum_{t=0}^{T-1} \Delta R(t), \tag{16}$$

where $\Delta R(t) \triangleq \sum_{n=1}^{N} \mathbb{E}\left[\mu_n C_n(t) S_n^*(t) - \mu_n C_n(t) \widehat{S}_n(t)\right]$.

We add the term $\eta \Delta R(t)$ on both sides of the drift of Lyapunov function $V(t)$ (cf. (11)) and obtain

$$\mathbb{E}\left[V(t+1) - V(t)\right] + \eta \Delta R(t)$$
$$= - \sum_{n=1}^{N} \mathbb{E}\left[Z_n(t) C_n(t) \widehat{S}_n(t)\right] + N + \eta \sum_{n=1}^{N} \mathbb{E}\left[\mu_n C_n(t) S_n^*\right]$$
$$- \eta \sum_{n=1}^{N} \mathbb{E}\left[\mu_n C_n(t) \widehat{S}_n(t)\right]$$
$$= N + \sum_{n=1}^{N} \mathbb{E}\left[\left(Z_n(t) + \eta \mu_n\right) C_n(t) \left(S_n^* - \widehat{S}_n(t)\right)\right]$$
$$- \sum_{n=1}^{N} \mathbb{E}\left[Z_n(t) C_n(t) S_n^*\right]$$
$$\leq N + \sum_{n=1}^{N} \mathbb{E}\left[\left(Z_n(t) + \eta \mu_n\right) C_n(t) \left(S_n^* - \widehat{S}_n(t)\right)\right], \tag{17}$$

where the last step is true since $\sum_{n=1}^{N} \mathbb{E}\left[Z_n(t) C_n(t) S_n^*\right] \geq 0$.

Summing (17) over $t = 0, 1, 2, \ldots, T-1$, we have

$$\sum_{t=0}^{T-1} \mathbb{E}\left[V(t+1) - V(t)\right] + \eta \sum_{t=0}^{T-1} \Delta R(t)$$
$$\leq N T + \sum_{t=0}^{T-1} \sum_{n=1}^{N} \mathbb{E}\left[\left(Z_n(t) + \eta \mu_n\right) C_n(t) \left(S_n^* - \widehat{S}_n(t)\right)\right],$$

which implies

$$\mathrm{Reg}(T) \triangleq \sum_{t=0}^{T-1} \Delta R(t) \leq \frac{N T}{\eta}$$
$$+ \frac{1}{\eta} \sum_{t=0}^{T-1} \sum_{n=1}^{N} \mathbb{E}\left[\left(Z_n(t) + \eta \mu_n\right) C_n(t) \left(S_n^* - \widehat{S}_n(t)\right)\right] \tag{18}$$

Here, we use the fact that $V(0) = 0$, $V(T) \geq 0$, and the definition of $\mathrm{Reg}(T)$.

Next, we focus on the term

$$\sum_{n=1}^{N}\left(Z_n(t)+\eta\mu_n\right)C_n(t)\left(S_n^*-\widehat{S}_n(t)\right).$$

Then, we have

$$\begin{aligned}
&\sum_{n=1}^{N}\left(Z_n(t)+\eta\mu_n\right)C_n(t)\left(S_n^*-\widehat{S}_n(t)\right)\\
\overset{(a)}{\leq}&\sum_{n=1}^{N}\left(Z_n(t)+\eta\mu_n\right)C_n(t)\widetilde{S}_n(t)\\
&\qquad-\sum_{n=1}^{N}\left(Z_n(t)+\eta\mu_n\right)C_n(t)\widehat{S}_n(t)\\
\overset{(b)}{\leq}&\sum_{n=1}^{N}\left(Z_n(t)+\eta\mu_n\right)C_n(t)\widetilde{S}_n(t)\\
&\qquad-\sum_{n=1}^{N}\left(Z_n(t)+\eta\mu_n\right)C_n(t)\widehat{S}_n(t)\\
&+\sum_{n=1}^{N}\left(Z_n(t)+\eta w_n(t)\right)C_n(t)\widehat{S}_n(t)\\
&\qquad-\sum_{n=1}^{N}\left(Z_n(t)+\eta w_n(t)\right)C_n(t)\widetilde{S}_n(t)\\
=&\eta\sum_{n=1}^{N}(w_n(t)-\mu_n)C_n(t)\widehat{S}_n(t)\\
&\qquad+\eta\sum_{n=1}^{N}(\mu_n-w_n(t))C_n(t)\widetilde{S}_n(t),
\end{aligned}\tag{19}$$

where step (a) is true for

$$\widetilde{\mathbf{S}}(t)\triangleq(\widetilde{S}_n(t))_{n=1}^N\in\arg\max_{\mathbf{S}\in\mathcal{S}}\sum_{n=1}^{N}\left(Z_n(t)+\eta\mu_n\right)C_n(t)S_n,$$

and (b) uses the definition of $\widehat{\mathbf{S}}(t)$.

By substituting (19) into (18), we have

$$\begin{aligned}
\mathrm{Reg}(T)\leq\frac{NT}{\eta}+&\underbrace{\sum_{t=0}^{T-1}\sum_{n=1}^{N}\mathbb{E}\left[\left(w_n(t)-\mu_n\right)C_n(t)\widehat{S}_n(t)\right]}_{\triangleq G_1(T)}\\
&+\underbrace{\sum_{t=0}^{T-1}\sum_{n=1}^{N}\mathbb{E}\left[\left(\mu_n-w_n(t)\right)C_n(t)\widetilde{S}_n(t)\right]}_{\triangleq G_2(T)}.
\end{aligned}\tag{20}$$

Next, we focus on $G_1(T)$ and $G_2(T)$, respectively. Let $t_{n,\tau}$ denote the time slot at which link n successfully received a packet, i.e., $C_n(t_{n,\tau})\widehat{S}_n(t_{n,\tau})=1$ and $C_n(t_{n,\tau})\widehat{S}_n(t_{n,\tau})=0$ if $t\neq t_{n,\tau},\tau=1,2,\ldots,H_n(T)$. Therefore, we have $H_n(t_{n,\tau})=\tau-1$.

Let $G_{n,1}(T)\triangleq\sum_{t=0}^{T-1}\mathbb{E}\left[\left(w_n(t)-\mu_n\right)C_n(t)\widehat{S}_n(t)\right]$ and thus $G_1(T)=\sum_{n=1}^{N}G_{n,1}(T)$.

Hence, we have

$$\begin{aligned}
G_{n,1}(T)\overset{(a)}{\leq}&\sum_{t=0}^{T-1}\mathbb{E}\left[(w_n(t)-\mu_n)C_n(t)\widehat{S}_n(t)\mathbb{1}_{\mathcal{F}_n(t)}\right]\\
\overset{(b)}{\leq}&\mathbb{E}\left[\sum_{\tau=1}^{H_n(T)}(w_n(t_{n,\tau})-\mu_n)\mathbb{1}_{\mathcal{F}_n(t_{n,\tau})}\right]\\
\overset{(c)}{\leq}&1+\mathbb{E}\left[\sum_{\tau=2}^{H_n(T)}(w_n(t_{n,\tau})-\mu_n)\mathbb{1}_{\mathcal{F}_n(t_{n,\tau})}\right]\\
\overset{(d)}{\leq}&1+\mathbb{E}\left[\sum_{\tau=2}^{H_n(T)}(w_n(t_{n,\tau})-\mu_n)\mathbb{1}_{\mathcal{F}_n(t_{n,\tau})\cap\mathcal{G}_n(t_{n,\tau})}\right]\\
&+\mathbb{E}\left[\sum_{\tau=2}^{H_n(T)}\mathbb{1}_{\overline{\mathcal{G}}_n(t_{n,\tau})}\right],
\end{aligned}\tag{21}$$

where step (a) is true for $\mathcal{F}_n(t)\triangleq\{w_n(t)\geq\mu_n\}$ and $\mathbb{1}_{\{\cdot\}}$ being an indicator function; (b) uses the definition of $t_{n,\tau}$, and the fact that $C_n(t)\leq 1$ and $\widehat{S}_n(t)\leq 1,\forall t\geq 0$; (c) follows from the fact that $w_n(t)\leq 1,\forall t\geq 0$; (d) is true for

$$\mathcal{G}_n(t)\triangleq\left\{\overline{\mu}_n(t)-\mu_n\leq\sqrt{\frac{3\log t}{2H_n(t)}}\right\},$$

and $\overline{\mathcal{G}}_n(t)$ being the complement of the event $\mathcal{G}_n(t)$.

Next, we consider the second term on the right hand side (RHS) of (21).

$$\begin{aligned}
&\mathbb{E}\left[\sum_{\tau=2}^{H_n(T)}(w_n(t_{n,\tau})-\mu_n)\mathbb{1}_{\mathcal{F}_n(t_{n,\tau})\cap\mathcal{G}_n(t_{n,\tau})}\right]\\
\overset{(a)}{\leq}&\mathbb{E}\left[\sum_{\tau=2}^{H_n(T)}2\sqrt{\frac{3\log t_{n,\tau}}{2H_n(t_{n,\tau})}}\right]\\
\overset{(b)}{\leq}&\sqrt{6\log T}\mathbb{E}\left[\sum_{\tau=2}^{H_n(T)}\frac{1}{\sqrt{\tau-1}}\right]\\
\leq&\sqrt{6\log T}\left(1+\int_1^{H_n(T)}\frac{1}{\sqrt{x}}dx\right)\\
\leq&2\sqrt{6\log T}\mathbb{E}\left[\sqrt{H_n(T)}\right],
\end{aligned}\tag{22}$$

where step (a) uses the definition of $w_n(t)$ and $\mathcal{G}_n(t)$, and (b) follows from the fact that $t_{n,\tau}\leq T$ and the definition of $t_{n,\tau}$. With regard to the third term on the RHS of (21), we have

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{\overline{\mathcal{G}}_n(t_{n,\tau})}\right]=\Pr\{\overline{\mathcal{G}}_n(t_{n,\tau})\}\\
\overset{(a)}{\leq}&\Pr\left\{\bigcup_{m=\tau-1}^{T-1}\left\{\overline{\mu}_n(m)-\mu_n>\sqrt{\frac{3\log m}{2(\tau-1)}}\right\}\right\}\\
\leq&\Pr\left\{\bigcup_{m=\tau-1}^{T-1}\left\{\overline{\mu}_n(m)-\mu_n>\sqrt{\frac{3\log m}{2m}}\right\}\right\}
\end{aligned}$$

$$\overset{(b)}{\leq} \sum_{m=\tau-1}^{T-1} \Pr\left\{\overline{\mu}_n(m) - \mu_n > \sqrt{\frac{3\log m}{2m}}\right\}$$
$$\overset{(c)}{\leq} \sum_{m=\tau-1}^{T-1} \frac{1}{m^3} \leq \frac{1}{(\tau-1)^3} + \int_{\tau-1}^{\infty} \frac{1}{x^3} dx \overset{(d)}{\leq} \frac{3}{2(\tau-1)^2},$$

where step (a) follows from the fact that

$$\overline{\mathcal{G}}_n(t_{n,\tau}) \subset \bigcup_{m=\tau-1}^{T-1} \left\{\overline{\mu}_n(m) - \mu_n > \sqrt{\frac{3\log m}{2(\tau-1)}}\right\};$$

(b) uses the union bound; (c) follows from the Chernoff-Hoeffding Bound (see, e.g., [5, Fact 1]), i.e., for $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with common range $[0,1]$ and mean μ, then for any $a \geq 0$, we have

$$\Pr\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq \mu + a\right\} \leq e^{-2na^2}, \tag{23}$$

(d) is true for $\tau \geq 2$.

Hence, the third term on the RHS of (21) can be bounded as follows.

$$\mathbb{E}\left[\sum_{\tau=2}^{H_n(T)} \mathbb{1}_{\overline{\mathcal{G}}_n(t_{n,\tau})}\right] \leq \mathbb{E}\left[\sum_{\tau=2}^{H_n(T)} \frac{3}{2(\tau-1)^2}\right] \leq \sum_{\tau=1}^{\infty} \frac{3}{2\tau^2} = \frac{\pi^2}{4}, \tag{24}$$

where the last step use the fact that $\sum_{n=1}^{\infty} 1/n^2 = \pi^2/6$. By substituting (22) and (24) into (21) and using the definition of $G_1(T)$, we have

$$G_1(T) \leq N\left(1 + \frac{\pi^2}{4}\right) + 2\sqrt{6\log T}\sum_{n=1}^{N} \mathbb{E}\left[\sqrt{H_n(T)}\right]$$
$$\overset{(a)}{\leq} N\left(1 + \frac{\pi^2}{4}\right) + 2N\sqrt{6\log T}\,\mathbb{E}\left[\sqrt{\frac{1}{N}\sum_{n=1}^{N} H_n(T)}\right]$$
$$\overset{(b)}{\leq} N\left(1 + \frac{\pi^2}{4}\right) + 2\sqrt{6N|\mathbf{S}|_{\max} T\log T}, \tag{25}$$

where step (a) uses the Jensen's inequality, and (b) is true since $\sum_{n=1}^{N} H_n(T) \leq T|\mathbf{S}|_{\max}$ and $|\mathbf{S}|_{\max}$ is the maximum number of links that can be scheduled in each time slot.

Next, we consider the term $G_2(T)$. First, we note that

$$G_2(T) \leq \sum_{t=0}^{T-1}\sum_{n=1}^{N} \mathbb{E}\left[(\mu_n - w_n(t))\widetilde{S}_n(t)\mathbb{1}_{\overline{\mathcal{F}}_n(t)}\right], \tag{26}$$

where we recall that $\mathcal{F}_n(t) \triangleq \{w_n(t) \geq \mu_n\}$. Note that for $t \leq t_{n,1}$, we have $w_n(t) = 1$ and thus $\mathcal{F}_n(t)$ happens. Therefore, we have

$$G_2(T) \leq \sum_{n=1}^{N} \mathbb{E}\left[\sum_{t=t_{n,1}+1}^{T-1} (\mu_n - w_n(t))\,\widetilde{S}_n(t)\mathbb{1}_{\overline{\mathcal{F}}_n(t)}\right]$$
$$\overset{(a)}{\leq} \sum_{n=1}^{N} \mathbb{E}\left[\sum_{t=t_{n,1}+1}^{T-1} \Pr\left\{\overline{\mu}_n(t) - \mu_n \leq -\sqrt{\frac{3\log t}{2H_n(t-1)}}\right\}\right]$$
$$\leq \sum_{n=1}^{N}\sum_{\tau=1}^{T-1}\sum_{m=1}^{\tau} \Pr\left\{\frac{1}{m}\sum_{i=1}^{m} X(i) - \mu_n \leq -\sqrt{\frac{3\log\tau}{2m}}\right\}$$
$$\overset{(b)}{\leq} \sum_{n=1}^{N}\sum_{\tau=1}^{T-1}\sum_{m=1}^{\tau} \frac{1}{\tau^3} = \sum_{n=1}^{N}\sum_{\tau=1}^{T-1} \frac{1}{\tau^2} \overset{(c)}{\leq} \frac{N\pi^2}{6}, \tag{27}$$

where step (a) follows from the fact that $\mu_n \leq 1$ and $\widetilde{S}_n(t) \leq 1$ as well as the definition of $\overline{\mathcal{F}}_n(t)$; (b) again uses the Chernoff-Hoeffding Bound (cf. (23)); (c) is true since $\sum_{\tau=1}^{T-1} 1/\tau^2 \leq \sum_{\tau=1}^{\infty} 1/\tau^2 = \pi^2/6$.

Hence, by substituting (25) and (27) into (20), we have the desired result.

REFERENCES

[1] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards," *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 968–976, 1987.

[2] Y. Gai, B. Krishnamachari, and R. Jain, "Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations," *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, pp. 1466–1478, 2012.

[3] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework and applications," in *International Conference on Machine Learning*, 2013, pp. 151–159.

[4] R. Combes, M. S. T. M. Shahi, A. Proutiere *et al.*, "Combinatorial bandits revisited," in *Advances in Neural Information Processing Systems*, 2015, pp. 2116–2124.

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

[6] 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*, 2011, pp. 359–376.

[7] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *Conference on learning theory*, 2012, pp. 39–1.

[8] N. Lu, B. Ji, and B. Li, "Age-based scheduling: Improving data freshness for wireless real-time traffic," in *Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 2018, pp. 191–200.

[9] I. Kadota, A. Sinha, and E. Modiano, "Optimizing age of information in wireless networks with throughput constraints," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 1844–1852.

[10] I. Kadota and E. Modiano, "Minimizing the age of information in wireless networks with stochastic arrivals," *IEEE Transactions on Mobile Computing*, 2019.

[11] Y. Sun, I. Kadota, R. Talak, and E. Modiano, "Age of information: A new metric for information freshness," *Synthesis Lectures on Communication Networks*, vol. 12, no. 2, pp. 1–224, 2019.

[12] F. Li, J. Liu, and B. Ji, "Combinatorial sleeping bandits with fairness constraints," *IEEE Transactions on Network Science and Engineering*, 2019.

[13] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures on Communication Networks*, vol. 3, no. 1, pp. 1–211, 2010.

[14] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[15] V. Patil, G. Ghalme, V. Nair, and Y. Narahari, "Achieving fairness in the stochastic multi-armed bandit problem." in *AAAI*, 2020, pp. 5379–5386.
[16] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *2012 Proceedings IEEE INFOCOM*. IEEE, 2012, pp. 2731–2735.
[17] E. Altman, R. El-Azouzi, D. S. Menasche, and Y. Xu, "Forever young: Aging control for hybrid networks," in *Proceedings of the Twentieth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 2019, pp. 91–100.
[18] S. Kaul, M. Gruteser, V. Rai, and J. Kenney, "Minimizing age of information in vehicular networks," in *2011 8th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*. IEEE, 2011, pp. 350–358.
[19] B. Choudhury, V. K. Shah, A. Dayal, and J. H. Reed, "Experimental analysis of safety application reliability in v2v networks," *arXiv preprint arXiv:2005.13031*, 2020.
[20] K. Nar and T. Başar, "Sampling multidimensional wiener processes," in *53rd IEEE Conference on Decision and Control*. IEEE, 2014, pp. 3426–3431.
[21] T. Z. Ornee and Y. Sun, "Sampling for remote estimation through queues: Age of information and beyond," *arXiv preprint arXiv:1902.03552*, 2019.
[22] R. D. Yates, "Lazy is timely: Status updates by an energy harvesting source," in *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2015, pp. 3008–3012.
[23] Y. Sun, E. Uysal-Biyikoglu, R. D. Yates, C. E. Koksal, and N. B. Shroff, "Update or wait: How to keep your data fresh," *IEEE Transactions on Information Theory*, vol. 63, no. 11, pp. 7492–7508, 2017.
[24] Y. Dong, P. Fan, and K. B. Letaief, "Energy harvesting powered sensing in iot: Timeliness versus distortion," *IEEE Internet of Things Journal*, 2020.
[25] S. Fatale, K. Bhandari, U. Narula, S. Moharir, and M. K. Hanawal, "Regret of age-of-information bandits," *arXiv*, pp. arXiv–2001, 2020.
[26] R. Li, A. Eryilmaz, and B. Li, "Throughput-optimal wireless scheduling with regulated inter-service times," in *2013 Proceedings IEEE INFOCOM*. IEEE, 2013, pp. 2616–2624.
[27] B. Li, R. Li, and A. Eryilmaz, "Throughput-optimal scheduling design with regular service guarantees in wireless networks," *IEEE/ACM Transactions on Networking*, vol. 23, no. 5, pp. 1542–1552, 2014.
[28] B. Li, A. Eryilmaz, and R. Srikant, "Emulating round-robin in wireless networks," in *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 2017, pp. 1–10.
[29] ——, "Emulating round-robin for serving dynamic flows over wireless fading channels," in *Proceedings of the Twenty-First International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, 2020, pp. 231–240.
[30] W.-K. Hsu, J. Xu, X. Lin, and M. R. Bell, "Integrate learning and control in queueing systems with uncertain payoffs," *Purdue University, available at https://engineering. purdue. edu/% 7elinx/papers. html, Tech. Rep*, 2018.
[31] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *arXiv preprint arXiv:1204.5721*, 2012.