Correlated Age-of-Information Bandits

Ishank Juneja, Santosh Fatale, and Sharayu Moharir
Department of Electrical Engineering, Indian Institute of Technology Bombay

***Abstract*—We consider a system composed of a sensor node tracking a time varying quantity. In every discretized time slot, the node attempts to send an update to a central monitoring station through one of K communication channels. We consider the setting where channel realizations are correlated across channels. This is motivated by mmWave based 5G systems where line-of-sight which is critical for successful communication is common across all frequency channels while the effect of other factors like humidity is frequency dependent.**

The metric of interest is the Age-of-Information (AoI) which is a measure of the freshness of the data available at the monitoring station. In the setting where channel statistics are unknown but stationary across time and correlated across channels, the algorithmic challenge is to determine which channel to use in each time-slot for communication. We model the problem as a Multi-Armed bandit (MAB) with channels as arms. We characterize the fundamental limits on the performance of any policy. In addition, via analysis and simulations, we characterize the performance of variants of the UCB and Thompson Sampling policies that exploit correlation.

I. Introduction

Future communication technologies including 5G are likely to use the millimeter band (30GHz to 300GHz) for communication. The available bandwidth is partitioned into frequency channels for communication. Factors such as frequency dependent atmospheric attenuation affect propagation in the millimeter band. In addition, the availability of a line-of-sight path between the source and receiver is critical for successful communication in this band [1]. Since the existence of a line-of-sight path is frequency agnostic, channel realizations across different frequency channels at a given time are correlated.

Age of Information (AoI), introduced in [2], is a freshness of data metric that measures the time elapsed since the most recent successful update sent from a source was received at the intended destination. For time-critical applications like self-driving cars, smart homes and other up and coming IoT applications, it is imperative that the data used by the control unit to make decisions is as recent as possible. In these cases, AoI is a suitable performance metric.

The system we study builds on the setting studied in [3] and consists of a single source node tracking a time-varying quantity. The source attempts communicating an update to a central monitoring station. At every discretized time step t, an update is sent through one among K available channels. Each channel has a certain probability of success which is assumed to remain static across the period of operation. The channel statistics, that is the probability of success or failure of communication for a certain channel, are not known to the scheduler. However, it is known that the successes and the failures across the K channels are correlated with one another through an underlying stochastic state X. Depending on the state in which X finds itself, certain channels are successful whereas others are not (Figure 1). Thereby the model accounts for correlation between the performances of channels.

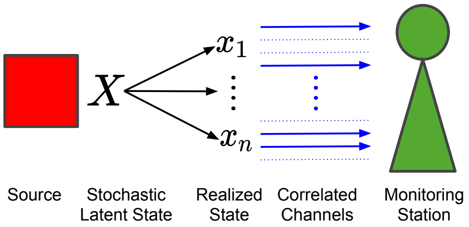

Fig. 1: The source node is attempting to communicate with the monitoring station. Depending on the state in which X finds itself, only certain channels work.

The algorithmic challenge is to determine which channel to use for communication in each time step in order to minimize cumulative AoI over a finite horizon T. The difference between expected cumulative AoI under the chosen scheduling policy, and under an oracle's optimal strategy of choosing the best channel k^* at all time steps, is called *AoI Regret*. We model the correlation between the performance of channels using the *Correlated Multi-Armed Bandit* framework introduced by [4]. For the AoI metric, scheduling decisions taken at any time step have a downstream effect across all future time slots. Hence, a new analysis for the AoI regret metric is needed to tackle problem instances drawn from the Correlated Bandit framework.

A. Our Contributions

Lower bound on AoI regret: We show a lower bound of $\Omega(\log T)$ on AoI regret for instances that have at least one *strictly competitive* arm (formally defined in Section II).

Performance of variants of UCB and Thompson Sampling: We show that the AoI regret for Correlated-UCB (CUCB) and Correlated-Thompson Sampling (CTS) (proposed in [4]) is $\mathrm{O}(C \log T) + \mathrm{O}(1)$. Here C is the number of sub-optimal *competitive arms* (formally defined in Section II).

Empirical validation: Through simulations, we compare the performance of UCB, Thompson Sampling, CUCB, CTS, and their *AoI-Aware* variants proposed in [3].

B. Related Work

Multi-Armed Bandits (MABs) are a sequential decision making framework where at every time step t, a choice has to be made between K possible *bandit arms* with unknown statistics. UCB [5] and Thompson Sampling [6] are two widely

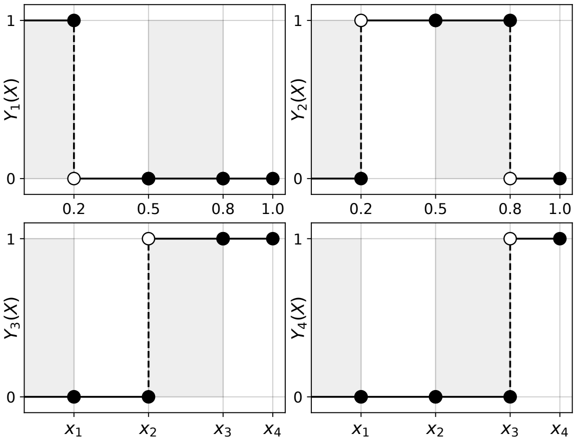

Fig. 2: Bandit instance I_1

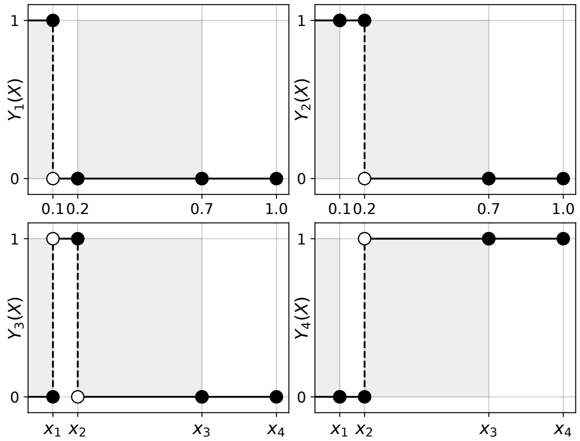

Fig. 3: Bandit instance I_2

studied algorithms for the MAB problem. In this work, we study variants of these policies more suited for our setting.

Recently, variants of the traditional MAB framework that are capable of incorporating additional structure into the decision making problem have been introduced. In addition to observing rewards, the *contexual bandit* framework [7], learns a mapping between a context vector θ and the best arm k^*. Another example is the structured bandit framework [8], in which the mean rewards for all arms as a function of the context θ are known but θ itself is hidden. The Correlated Multi-Armed Bandit framework of [4] is a variant of the MAB problem that presents the scheduler with arms whose rewards are not independent of each another. That is, sampling arm k can reveal information about the rewards we can expect from another arm ℓ. In addition to observing the rewards obtained from sampling an arm k, scheduling algorithms cognizant of correlation can track additional side information to identify some arms as sub-optimal without sampling them often, thereby reducing the accumulation of regret.

AoI or Age-of-Information is measure of the freshness of data available at the central monitoring station. AoI has been the focus of a variety of work, and [9] can be referred to for a comprehensive survey. Previously, a large focus of the work on AoI has been on problems where channel statistics are known [10]–[13]. In our work, we take channel statistics to be unknown and apply Multi-Armed Bandits to the scheduling problem, which is the approach taken by [3] for minimizing AoI regret. Building upon the work of [3], we drop the assumption of independence between channels and instead take that they follow the Correlated Multi-Armed Bandit model.

II. Setting

A. Correlated Bandit Model

For a system with K communication channels we construct a Correlated Bandit instance with the same number of arms. The random variable X captures the underlying state that determines the rewards for the correlated arms. The reward obtained on playing arm k is denoted by a Bernoulli random variable $Y_k(X)$, where Y_k is a known deterministic function. We define $\mu_k = \mathbb{E}_X[Y_k(X)]$. The optimal arm is denoted by k^*, and has mean μ^*. The difference between the largest mean and the mean of a sub-optimal arm is called the sub-optimality gap and is given by $\Delta_k = \mu^* - \mu_k$.

Since the distribution of X is unknown, the means $\mu_1, \mu_2, \ldots, \mu_K$ are also not known. MAB algorithms such as UCB empirically estimate the means $\hat{\mu}_k$ for use in decision making. Under the Correlated Bandit model, rewards for all the arms are functions of the same random variable X. Hence, we can infer the possible rewards that an arm ℓ could have returned if we had chosen arm ℓ instead of arm k. To exploit this, [4] introduces the notion of *pseudo rewards.*

Definition 1 (Expected pseudo reward and pseudo gap)**.** *Pseudo reward for arm ℓ with respect to arm k is given by,*

$$s_{\ell,k}(r) = \sup_{x:Y_k(x)=r} Y_\ell(x). \tag{1}$$

Expected pseudo reward in turn is defined as,

$$\phi_{\ell,k} = \mathbb{E}_X[\, s_{\ell,k}(Y_k(X))\,]. \tag{2}$$

The pseudo gap is defined as $\tilde{\Delta}_{\ell,k^*} = \mu^* - \phi_{\ell,k^*}$.

We say arm k is *competitive* if $\tilde{\Delta}_{\ell,k^*} \leq 0$ and *strictly competitive* if the inequality is strict. If $\tilde{\Delta}_{\ell,k^*} > 0$, we call arm k non-competitive. We use C to denote the number of competitive arms excluding arm k^*.

Example 1. *Two examples of Correlated Bandit instances* I_1 *and* I_2 *are shown in Figures 2 and 3 respectively. Both instances have $K = 4$ arms and have X as a discrete random variable taking values in the abstract alphabet $\{x_1, x_2, x_3, x_4\}$. In the figures, the length of the interval corresponding to x_i is equal to $\mathbb{P}\{X = x_i\}$. Consider* I_1*, in it, the mean reward for the bandit arms is given by, $\mu_1 = 1 \times 0.2$, $\mu_2 = 1 \times 0.3 + 1 \times 0.3$, $\mu_3 = 1 \times 0.3 + 1 \times 0.2$ and $\mu_4 = 1 \times 0.3$. Hence arm 2 is optimal with $\mu^* = 0.6$. The expected pseudo rewards can be computed using Definition 1 to obtain $\phi_{1,2} = 0.4$, $\phi_{3,2} = 1.0$ and $\phi_{4,2} = 0.4$. Hence, for instance* I_1 *arm 2 is optimal and arm 3 is the only competitive* *sub-optimal arm. Therefore $C = 1$ for instance I_1. Performing a similar analysis on I_2, we find that arm 4 is optimal and no other arm is competitive. In other words $C = 0$ for I_2.*

B. The AoI Regret Metric

Here, the notions of Age-of-Information (AoI) and AoI regret, that were earlier described informally, are made precise.

Definition 2 (Age-of-Information (AoI))**.** *At the start of time slot t, let $a(t)$ denote the AoI at the central monitoring station and let $u(t)$ denote the time index at which the recent most successful update was received by the monitoring station. Then, $a(t) = t - u(t)$. Alternatively,*

$$a(t) = \begin{cases} 1 & \textit{if the update in } t-1 \textit{ succeeds} \\ a(t-1)+1 & \textit{otherwise.} \end{cases}$$

Under a given scheduling policy ρ, let $a_\rho(t)$ denote the AoI in time slot t. Further, let $a^*(t)$ be the AoI under an oracle's policy that uses the optimal arm k^* in all time slots. The AoI regret at time T is the cumulative difference in expected AoI for the two policies from time-slots 1 to T.

Definition 3 (Age-of-Information Regret (AoI Regret))**.** *AoI regret for a policy ρ, over T slots is given by,*

$$R_\rho(T) = \sum_{t=1}^{T} \mathbb{E}[a_\rho(t) - a^*(t)] = \sum_{t=1}^{T} \mathbb{E}[a_\rho(t)] - \frac{T}{\mu^*}, \quad (3)$$

where (3) follows from the expectation of a geometric random variable with parameter μ^*.

III. Main Results and Discussion

In this section we present our main contributions and their implications. First, we provide a lower bound on AoI regret for a certain class of policies, then we examine an upper bound on AoI regret for two policies, namely CUCB and CTS (proposed in [4]). Lastly, we derive the conditions under which the upper and lower bounds on AoI regret are order-wise equal.

A. Lower Bound for Correlated Bandit Instances

First, we define the class of α-consistent policies and then in Theorem 1 provide a lower bound on the AoI regret achievable by any policy ρ belonging to this class.

Definition 4 (α-consistent policies [5])**.** *Let k_s denote the index of the channel scheduled in time-slot s. The index k^* denotes the index of the optimal channel. A scheduling policy is called α-consistent, for a constant $\alpha \in (0, 1)$, if there exists an instance dependent constant M such that,*

$$\mathbb{E}\Big[\sum_{s=1}^{t} \mathbb{1}\{k_s = k\}\Big] \le Mt^\alpha, \; \forall\, k \ne k^*. \quad (4)$$

Theorem 1 (Lower bound on AoI regret)**.** *If a bandit instance I has at least one competitive arm k with $\tilde{\Delta}_{k,k^*} < 0$, then for any α-consistent policy ρ, we have,*

$$R_\rho(T) \ge \max_{k\in\mathcal{C}'} \frac{\Delta_k}{\mathrm{D}(P_k, P'_k)} \frac{(1-\alpha)\log T - \log(4M)}{\mu^*}.$$

Otherwise, if $\tilde{\Delta}_{k,k^*} \ge 0 \;\forall\, k \in [K]$, $R_\rho(T) \ge 0$.

Here $\mathrm{D}(P_k, P'_k)$ is the KL divergence between the reward distribution of arm k and a suitably chosen perturbed reward distribution. The set $\mathcal{C}'$ is the set of strictly competitive arms and M is an instance dependent constant as in Definition 4.

B. Upper Bound for Correlated Bandit Instances

We now characterize upper bounds on AoI regret for the policies CUCB and CTS proposed in [4]. The key idea behind these policies is to exploit the correlation in the rewards of various arms to obtain a set of competitive arms in each step. The algorithms then play an arm from this set. For the sake of completeness, we provide details of these policies in the appendix (Algorithms 1 and 2 respectively).

Theorem 2 (Upper bound on AoI regret under CUCB)**.** *For a Correlated Bandit instance, let the number of competitive sub-optimal arms be equal to C. Further, let $\mathcal{C}$ denote the set of competitive arms inclusive of the optimal arm k^*. Let $\mu_{\min} = \min_k \mu_k$,*

$$t_o = \inf\Big\{\tau \ge 2 : \Delta_{\min}, \tilde{\Delta}_{k,k^*} \ge 4\sqrt{\frac{2K\log\tau}{\tau}}\Big\},$$

$$U^{(nc)}_{k,\mathrm{CUCB}} = Kt_0 + K^3 \sum_{t=Kt_0}^{T} 2\Big(\frac{t}{K}\Big)^{-2} + \sum_{t=1}^{T} 3t^{-3},$$

$$U^{(c)}_{k,\mathrm{CUCB}} = 8\frac{\log(T)}{\Delta_k^2} + \Big(1 + \frac{\pi^2}{3}\Big) + \sum_{t=1}^{T} 2Kt\exp\Big(-\frac{t\Delta^2_{\min}}{2K}\Big).$$

Then, for $T > t_0$,

$$\begin{aligned}\mathbb{E}[R_{\mathrm{CUCB}}(T)] &\le \frac{1-\mu^*}{\mu^*\mu_{\min}} + \Big(\frac{1}{\mu_{\min}} - \frac{1}{\mu^*}\Big)\times \\ &\quad \Big(\sum_{k'\in[K]\setminus\mathcal{C}} \Delta_{k'} U^{(nc)}_{k,\mathrm{CUCB}} + \sum_{k\in\mathcal{C}\setminus\{k^*\}} \Delta_k U^{(c)}_{k,\mathrm{CUCB}}\Big) \\ &= O(1) + O(C\log T),\end{aligned}$$

and for $T \le t_0$, $\mathbb{E}[R_{\mathrm{CUCB}}(T)] \le \big(\frac{1}{\mu_{\min}} - \frac{1}{\mu^*}\big)T$.

Theorem 3 (Upper bound on AoI regret under CTS)**.** *For a Correlated Bandit instance, let the number of competitive sub-optimal arms be equal to C and let $\mathcal{C}$ denote the set of competitive arms inclusive of the optimal arm k^*. Further, let $\mu_{\min} = \min_k \mu_k$,*

$$t_b = \inf\Big\{\tau \ge \exp(11\beta) : \Delta_{\min}, \tilde{\Delta}_{k,k^*} \ge 6\sqrt{\frac{2K\beta\log\tau}{\tau}}\Big\},$$

$$\begin{aligned}U^{(nc)}_{k,\mathrm{CTS}} &= Kt_b + \sum_{t=1}^{T} 3t^{-3} \\ &\quad + K^2 \sum_{t=Kt_b}^{T} \Big((2K+3)\Big(\frac{t}{K}\Big)^{-2} + \Big(\frac{t}{K}\Big)^{1-2\beta}\Big),\end{aligned}$$

$$\begin{aligned}U^{(c)}_{k,\mathrm{CTS}} &= 18\frac{\log(T\Delta_k^2)}{\Delta_k^2} + \exp(11\beta) + \frac{9}{\Delta_k^2} \\ &\quad + \sum_{t=1}^{T} 2Kt\exp\Big(-\frac{t\Delta^2_{\min}}{2K}\Big).\end{aligned}$$

Then, for any choice of $\beta > 1$ and for $T > t_b$,

$$\mathbb{E}[R_{\text{CTS}}(T)] \leq \frac{1-\mu^*}{\mu^*\mu_{\min}} + \Big(\frac{1}{\mu_{\min}} - \frac{1}{\mu^*}\Big)\times$$
$$\Bigg(\sum_{k' \in [K]\setminus\mathcal{C}} \Delta_{k'} U^{(nc)}_{k,\text{CTS}} + \sum_{k \in \mathcal{C}\setminus\{k^*\}} \Delta_k U^{(c)}_{k,\text{CTS}}\Bigg)$$
$$= O(1) + O(C \log T),$$

and for $T \leq t_b$, $\mathbb{E}[R_{\text{CTS}}(T)] \leq \big(\frac{1}{\mu_{\min}} - \frac{1}{\mu^*}\big)T$.

C. Discussion of Implications

From Theorems 2 and 3 it is clear that both CUCB and CTS are α-consistent and therefore AoI regret under these policies will satisfy Theorem 1. Since, the bounds on AoI regret depend on the number of sub-optimal competitive arms C, we consider different possibilities for C to understand the regret bounds. If $C > 0$ for a Correlated Bandit instance, and if at least one arm is strictly competitive, then the lower bound and upper bound on AoI regret are both $\mathrm{O}(\log T)$. However, when there are no competitive arms, that is when $C = 0$, then there is no meaningful lower bound on the expected AoI regret. The $C = 0$ case agrees with the fact that the set $\mathcal{C}\setminus\{k^*\}$ being empty results in a constant $\mathrm{O}(1)$ upper bound on AoI regret. Hence for both these cases of Correlated Bandit instances, the bounds are order-optimal.

The Correlated Bandit model used in this work assumes the knowledge of deterministic reward functions. In practice, it may be challenging to determine these functions, and while conducting such an exercise we may even be able to learn the distribution of the underlying state X itself, thereby doing away with the need for any Bandit algorithm. However, the strength of this model is that once these functions are determined, they can be utilized in other communication systems with a similar configuration but a different and unknown distribution of X. Occlusions of different nature can repeatedly disrupt multiple channels at the same time engendering correlation in their performances. Communication systems similar to the one considered in this work might be placed in very different environments while following a standardized configuration. Due to differences in the environment, occlusion events analogous to the abstract entries in $\{x_1, x_2, \ldots, x_n\}$ used in this work, would occur with different relative frequencies at every installation. This variety in environments would make it difficult to scale solutions customized for every location. The distribution agnostic model and algorithms considered in this work would be highly beneficial in such scenarios.

IV. Simulations

In this section, we compare the performance of UCB, TS, CUCB, and CTS via simulations. In addition, we also compare the performance of *AoI-aware* variants of these policies [3].

Whenever current AoI $a(t)$ is low, the AoI-aware variant of a policy makes decisions identical to its parent algorithm. However, when $a(t)$ exceeds a certain threshold specified in [3], the AoI-aware variant selects the arm k with the highest empirical mean $\hat{\mu}_k$.

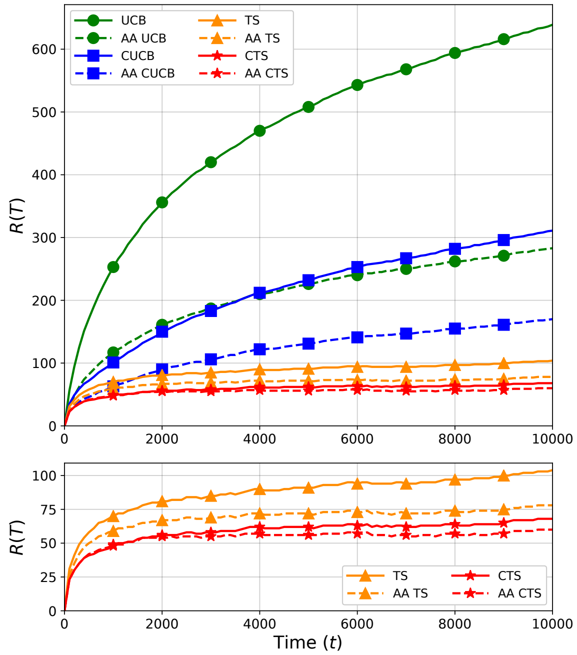

Fig. 4: AoI regret results for bandit instance I-1

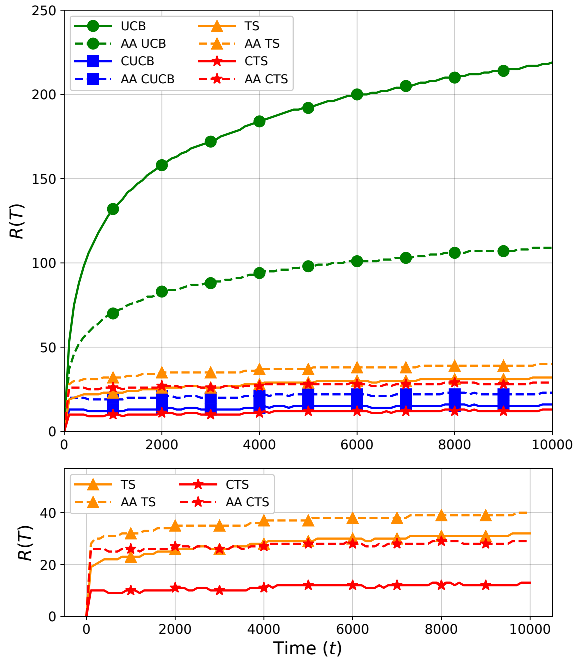

Fig. 5: AoI regret results for bandit instance I-2

We plot AoI regret for the bandit instances I_1 and I_2, described in Section II in Figures 4 and 5 respectively. It can be observed that for both I_1 and I_2, CTS and its AoI-aware variant perform the best on AoI regret. Further, CUCB and CTS have significantly lower AoI regret as compared to UCB and TS respectively. This is because of the fact that CUCB and CTS exploit correlation between arms to avoid sampling some non-competitive sub-optimal arms numerous times. Even though TS does not utilize correlation, its performance is still better than CUCB for I_1. This is because, even though both UCB and TS have $\log T$ AoI regret scaling, the pre-factor for $\log T$ in UCB, and consequently in CUCB, is much larger than the prefactor in TS.

As seen in Section II instance I_2 had no competitive sub-optimal arms. Hence, as predicted by the bounds in Theorems 2 and 3, both CUCB and CTS have constant expected AoI regret regardless of the horizon. Interestingly, through Figure 5, we also observe that the AoI-aware variant of a policy need not perform better than its parent policy as is the case for CUCB, CTS and TS. The regret bounds stated in Theorem 3 were for Thompson Sampling with Gaussian priors, however, the simulation results for TS and CTS in this section were with Beta priors since the latter performs better on AoI regret.

V. Proofs

A. Proof of Theorem 1

Lemma 1 (Lower bound on AoI regret for any policy). *Let $n_k(t)$ be the number of times arm k was scheduled in the time slots 1 to $t-1$. Then, AoI regret $R_\rho(T)$ under any policy ρ is lower bounded as,*

$$R_\rho(T) \geq \frac{1}{\mu^*} \sum_{i \neq k^*} \Delta_i \mathbb{E}[n_i(T+1)],$$

where $\Delta_i = \mu^* - \mu_i$.

Proof. Let $S_\rho(t)$ and $S^*(t)$ be indicator random variables denoting a successful update in time slot t by policy ρ and the oracle's policy respectively. Then, from Lemma 1 of [3] we have,

$$R_\rho(T) \geq \frac{1}{\mu^*} \sum_{t=1}^{T} \mathbb{E}[S^*(t) - S_\rho(t)]. \tag{5}$$

Further, from [3], we also have,

$$\mathbb{E}[S^*(t) - S_\rho(t)] \geq \sum_{i \neq k^*} (\mu^* - \mu_k) \mathbb{P}(\mathbb{1}\{k_t = i\} = 1) \tag{6}$$

$$= \sum_{i \neq k^*} \Delta_i \mathbb{P}(\mathbb{1}\{k_t = i\} = 1). \tag{7}$$

Therefore from inequality (5) we have,

$$R_\rho(T) \geq \frac{1}{\mu^*} \sum_{t=1}^{T} \sum_{i \neq k^*} \Delta_i \mathbb{P}(\mathbb{1}\{k_t = i\} = 1) \tag{8}$$

$$= \frac{1}{\mu^*} \sum_{i \neq k^*} \Delta_i \mathbb{E}[n_i(T+1)]. \tag{9}$$

■

Lemma 2 (Bretagnolle-Huber Inequality [14]). *Consider two probability measures P and Q, both absolutely continuous with respect to a given measure. Then for any event $\mathcal{A}$,*

$$P(\mathcal{A}) + Q(\mathcal{A}^c) \geq \frac{1}{2} \exp\big(-\mathrm{KL}(P||Q)\big).$$

Lemma 3 (Divergence Decomposition Lemma [15]). *Let $\nu = (P_1, \ldots, P_K)$ be the reward distributions associated with one K-armed bandit, and let $\nu' = (P'_1, \ldots, P'_K)$ be the reward distributions associated with another K-armed bandit. Let $\mathbb{P}^t_\nu = \mathbb{P}^t_{\nu\rho}$ and $\mathbb{P}^t_{\nu'} = \mathbb{P}^t_{\nu'\rho}$ be the joint distributions corresponding to the schedule of bandit arms chosen under policy ρ and the rewards received. Then,*

$$\mathrm{KL}(\mathbb{P}^t_\nu || \mathbb{P}^t_{\nu'}) = \mathrm{D}(\mathbb{P}^t_\nu, \mathbb{P}^t_{\nu'}) = \sum_{i=1}^{K} \mathbb{E}_\nu[n_i(t+1)] \mathrm{D}(P_i, P'_i),$$

where $\mathbb{E}_\nu[n_i(t+1)]$ is the expected number of pulls of arm i in t rounds of play for the bandit instance described by ν.

Proof of Theorem 1. Consider a Correlated Bandit instance I having just two arms, the optimal arm with index 1 and a lone sub-optimal arm with index 2. If the sub-optimal arm is strictly competitive, that is if $\tilde{\Delta}_{2,1} < 0$, then from Theorem 3 in [4] we can construct a perturbed bandit instance I' such that $\mathbb{E}_{X'}[\tilde{Y}_2(X)] > \mu_1$. Let $\mathbb{P}^t_{\mathrm{I}}$ and $\mathbb{P}^t_{\mathrm{I}'}$ be the distributions corresponding to rewards and scheduled arms in the first t time steps for instances I and I' respectively. By construction, arm 2 will be the optimal arm for the bandit instance I'. Now, using Definition 4, for α-consistent policies ρ, there exists a constant M such that,

$$\mathbb{E}^t_{\mathrm{I}}\Big[\sum_{\tau=1}^{t} \mathbb{1}\{k_\tau = 2\}\Big] \leq M t^\alpha \tag{10}$$

$$\mathbb{E}^t_{\mathrm{I}'}\Big[\sum_{\tau=1}^{t} \mathbb{1}\{k_\tau = 1\}\Big] \leq M t^\alpha. \tag{11}$$

Defining the event $\mathcal{A} = \{n_2(t+1) > t/2\}$ and using inequalities (10) and (11), the following Markov inequalities hold,

$$\mathbb{P}^t_{\mathrm{I}}(\mathcal{A}) \leq \frac{2M}{t^{1-\alpha}} \tag{12}$$

$$\mathbb{P}^t_{\mathrm{I}'}(\mathcal{A}^c) \leq \frac{2M}{t^{1-\alpha}}. \tag{13}$$

Now, using Lemma 2 we can write,

$$\mathrm{D}(\mathbb{P}^t_{\mathrm{I}}, \mathbb{P}^t_{\mathrm{I}'}) \geq (1-\alpha)\log t - \log(4M). \tag{14}$$

Next, using Lemma 3 we can expand $\mathrm{D}(\mathbb{P}^t_{\mathrm{I}}, \mathbb{P}^t_{\mathrm{I}'})$ as,

$$\mathrm{D}(\mathbb{P}^t_{\mathrm{I}}, \mathbb{P}^t_{\mathrm{I}'}) = \mathbb{E}_{\mathrm{I}}[n_2(t+1)] \mathrm{D}(P_2, P'_2). \tag{15}$$

Combining inequality (15) and Lemma 1, we get the following lower bound on AoI regret for instance I,

$$R_\rho(T) \geq \frac{\Delta_2\Big((1-\alpha)\log T - \log(4M)\Big)}{\mu^* \mathrm{D}(P_2, P'_2)}. \tag{16}$$

If a Correlated Bandit instance has more than one strictly competitive sub-optimal arm, then the expected number of sub-optimal pulls, and by extension, the lower bound on AoI regret, can only be higher due to the greater exploration required. Hence, in general, if the collection of strictly competitive arms is given by $\mathcal{C}'$, then,

$$R_\rho(T) \geq \max_{k \in \mathcal{C}'} \frac{\Delta_k}{\mathrm{D}(P_k, P'_k)} \frac{(1-\alpha)\log T - \log(4M)}{\mu^*}. \tag{17}$$

If among the sub-optimal arms, there is no strictly competitive arm, then the lower bound on AoI Regret is simply 0. ■

B. Proof of Theorems 2 and 3

As in ordinary MAB instances, the reward $Y_k(X)$ of each arm k in Correlated Bandit instances is distributed according to the Bernoulli distribution. Hence, under the technical Assumption 1 of [3] the following Lemma would apply to Correlated Bandit instances.

Lemma 4 (Lemma 4 in [3])**.** *Given the expected number of pulls of sub-optimal arms, the AoI regret over a run of T rounds is upper bounded by,*

$$\sum_{t=1}^{T}\mathbb{E}[a(t)]-\frac{T}{\mu^*}\le\frac{1-\mu^*}{\mu^*\mu_{\min}}+\Big(\frac{1}{\mu_{\min}}-\frac{1}{\mu^*}\Big)\mathbb{E}\Big[\sum_{t=1}^{T}\mathbb{1}_{k_t\neq k^*}\Big].$$

Lemma 5 (Expected number of pulls of a non-competitive arm, Theorem 1 in [4])**.** *The expected number of pulls of a non-competitive sub-optimal arm under CUCB is upper bounded by,*

$$\begin{aligned}\mathbb{E}[n_k^{(nc)}(T)]&\le Kt_0+K^3\sum_{t=Kt_0}^{T}2\Big(\frac{t}{K}\Big)^{-2}+\sum_{t=1}^{T}3t^{-3}\\&=U_{k,\mathrm{CUCB}}^{(nc)}=O(1),\end{aligned}$$

and under CTS with Gaussian priors by,

$$\begin{aligned}\mathbb{E}[n_k^{(nc)}(T)]&\le Kt_b+\sum_{t=1}^{T}3t^{-3}\\&\quad+K^2\sum_{t=Kt_b}^{T}\left((2K+3)\Big(\frac{t}{K}\Big)^{-2}+\Big(\frac{t}{K}\Big)^{1-2\beta}\right)\\&=U_{k,\mathrm{CTS}}^{(nc)}=O(1),\end{aligned}$$

where $t_0, t_b > 0$ are constants defined as,

$$t_o=\inf\Big\{\tau\ge 2:\Delta_{\min},\tilde{\Delta}_{k,k^*}\ge 4\sqrt{\frac{2K\log\tau}{\tau}}\Big\},$$

$$t_b=\inf\Big\{\tau\ge\exp{(11\beta)}:\Delta_{\min},\tilde{\Delta}_{k,k^*}\ge 6\sqrt{\frac{2K\beta\log\tau}{\tau}}\Big\},$$

where $\beta > 1$ is a parameter in CTS (Gaussian priors).

Lemma 6 (Expected number of pulls of a competitive arm, Theorem 2 in [4])**.** *The expected number of pulls of a competitive sub-optimal arm under CUCB is upper bounded by,*

$$\begin{aligned}\mathbb{E}[n_k^{(c)}(T)]&\le 8\frac{\log(T)}{\Delta_k^2}+\Big(1+\frac{\pi^2}{3}\Big)\\&\quad+\sum_{t=1}^{T}2Kt\exp\Big(-\frac{t\Delta_{\min}^2}{2K}\Big)\\&=U_{k,\mathrm{CUCB}}^{(c)}=O(\log T),\end{aligned}$$

and under CTS with Gaussian priors by,

$$\begin{aligned}\mathbb{E}[n_k^{(c)}(T)]&\le 18\frac{\log(T\Delta_k^2)}{\Delta_k^2}+\exp{(11\beta)}+\frac{9}{\Delta_k^2}\\&\quad+\sum_{t=1}^{T}2Kt\exp\Big(-\frac{t\Delta_{\min}^2}{2K}\Big)\\&=U_{k,\mathrm{CTS}}^{(c)}=O(\log T),\end{aligned}$$

where $\beta > 1$ is a parameter in CTS (Gaussian priors).

Proof of Theorems 2 and 3. The results follow by substituting the appropriate expression for the total number of sub-optimal pulls from Lemmas 5 and 6 into Lemma 4. ■

VI. Conclusions and Future Work

With next generation communication technologies relying on higher carrier frequencies, the problem of line-of-sight occlusion becomes more significant. In systems of the kind considered in this work, occlusions can effect multiple channels simultaneously resulting in their performances being correlated. This correlation can be exploited to identify certain channels as sub-optimal within a few time steps. We have shown theoretical bounds on AoI regret to corroborate this fact. Moreover, through simulations we observed that policies capable of exploiting correlation perform significantly better than those that disregard the presence of correlation.

The Correlated Bandit framework used in this work does not put constraints on the nature of rewards received. It would likely be beneficial to exploit the fact that in our system, rewards are always either 0 or 1. Additionally, the theoretical upper bound on AoI regret for CTS was for *Gaussian* priors. An analogous result remains to be shown for CTS with *Beta* priors. These topics will be the focus of future work.

References

[1] K. C. Huang and Z. Wang, *Millimeter Wave Characteristics.* John Wiley & Sons, Ltd, 2011, ch. 1, pp. 1–31.

[2] S. Kaul, R. Yates, and M. Gruteser, “Real-time status: How often should one update?” in *2012 Proceedings IEEE INFOCOM*, March 2012, pp. 2731–2735.

[3] S. Fatale, K. Bhandari, U. Narula, S. Moharir, and M. K. Hanawal, “Regret of age-of-information bandits,” *arXiv preprint arXiv:2001.09317*, 2020.

[4] S. Gupta, S. Chaudhari, G. Joshi, and O. Yağan, “Multi-armed bandits with correlated arms,” *arXiv preprint arXiv:1911.03959*, 2019.

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

[6] W. R. Thompson, “On the likelihood that one unknown probability exceeds another in view of the evidence of two samples,” *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.

[7] L. Zhou, “A survey on contextual multi-armed bandits,” 2015.

[8] V. Dani, T. Hayes, and S. Kakade, “Stochastic linear optimization under bandit feedback.” 01 2008, pp. 355–366.

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

[10] V. Tripathi and S. Moharir, “Age of Information in Multi-Source Systems,” in *GLOBECOM 2017-2017 IEEE Global Communications Conference.* IEEE, 2017, pp. 1–6.

[11] P. R. Jhunjhunwala and S. Moharir, “Age-of-Information aware scheduling,” *SPCOM*, 2018.

[12] I. Kadota, A. Sinha, and E. Modiano, “Optimizing age of information in wireless networks with throughput constraints,” in *Proc. INFOCOM*, 2018.

[13] B. Sombabu and S. Moharir, “Age-of-Information Aware Scheduling for Heterogeneous Sources,” in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom ’18. New York, NY, USA: ACM, 2018, pp. 696–698.

[14] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, ser. Springer series in statistics. Springer, 2009.

[15] T. Lattimore and C. Szepesvári, *Bandit Algorithms.* Cambridge University Press, 2020.

Appendix

Algorithm 1: Correlated UCB (CUCB)

```
1  Input: Pseudo-rewards $s_{\ell,k}(r)$
2  Initialize: Set $\hat{\mu}_k$, $\hat{\phi}_{\ell,k}$ and $n_k$ as $0 \ \forall \ k \in [K]$.
3  while $1 \leq t \leq K$ do
4      Schedule update on Channel $k_t = t$
5      Receive reward $r_t$ drawn from Ber($\mu_{k_t}$)
6      $\hat{\mu}_{k_t} = r_t$
7      $n_{k_t}(t) = 1$
8      $t = t + 1$
9  while $t \geq K + 1$ do
10     Find $\mathcal{S}_t = \{k : n_k(t-1) \geq \frac{t-1}{K}\}$, the set of arms
         pulled a significant number of times till $t - 1$.
         Define $k^{\text{emp}}(t) = \arg\max_{k \in \mathcal{S}_t} \hat{\mu}_{k_t}$
11     Initialize the empirically competitive set $\mathcal{A}_t$ as $\{\}$
12     for $k \in [K]$ do
13         if $\min_{\ell \in S_t} \hat{\phi}_{k,\ell}(t) \geq \hat{\mu}_{k^{\text{emp}}}(t)$ then
14             Add empirically competitive arms $k$ to the
                 set: $\mathcal{A}_t = \mathcal{A}_t \cup \{k\}$
15     Schedule update on Channel $k_t$ such that,
         $k_t = \arg\max_{k \in \mathcal{A}_t \cup \{k^{\text{emp}}(t)\}} \hat{\mu}_{k_t} + \sqrt{\frac{2 \log t}{n_k(t-1)}}$
16     Receive reward $r_t$ drawn from Ber($\mu_{k_t}$)
17     $\hat{\mu}_{k_t} = (\hat{\mu}_{k_t} \cdot n_{k_t}(t-1) + r_t)/(n_{k_t}(t-1) + 1)$
18     $n_{k_t}(t) = n_{k_t}(t-1) + 1$
19     $\hat{\phi}_{k,k_t} = \sum_{\tau : k_\tau = k_t} s_{k,k_\tau}(r_\tau)/n_{k_t}(t) \ \forall \, k \neq k_t$
20     $t = t + 1$
```

Algorithm 2: Correlated TS (CTS)

```
1  Input: Pseudo-rewards $s_{\ell,k}(r)$
2  Initialize: Set the number of successes $S_k(t)$, failures
     $F_k(t)$, and the quantities in CUCB as $0 \ \forall \ k \in [K]$.
3  while $t \geq 1$ do
4      Perform the steps 10 - 14 as in Algorithm 1
5      For each $k$ in $[K]$, draw a sample $\theta_k(t)$, where,
         $\theta_k(t) \sim$ Beta($S_k(t-1) + 1, F_k(t-1) + 1$)
6      Schedule update on Channel $k_t$ such that
         $k_t = \arg\max_{k \in \mathcal{A}_t \cup \{k^{\text{emp}}(t)\}} \theta_k(t)$
7      Receive reward $r_t$ drawn from Ber($\mu_{k_t}$)
8      $S_{k_t}(t) = S_{k_t}(t-1) + r_t$
9      $F_{k_t}(t) = F_{k_t}(t-1) + (1 - r_t)$
10     Perform the steps 17 - 20 as in Algorithm 1
```