Decentralized Age-of-Information Bandits

Archiki Prasad
Department of Electrical Engineering
Indian Institute of Technology Bombay
archiki@iitb.ac.in

Vishal Jain
Department of Electrical Engineering
Indian Institute of Technology Bombay
vishalj409@gmail.com

Sharayu Moharir
Department of Electrical Engineering
Indian Institute of Technology Bombay
sharayum@ee.iitb.ac.in

***Abstract*—Age-of-Information (AoI) is a performance metric for scheduling systems that measures the freshness of the data available at the intended destination. AoI is formally defined as the time elapsed since the destination received the recent most update from the source. We consider the problem of scheduling to minimize the cumulative AoI in a multi-source multi-channel setting. Our focus is on the setting where channel statistics are unknown and we model the problem as a distributed multi-armed bandit problem. For an appropriately defined AoI regret metric, we provide analytical performance guarantees of an existing UCB-based policy for the distributed multi-armed bandit problem. In addition, we propose a novel policy based on Thomson Sampling and a hybrid policy that tries to balance the trade-off between the aforementioned policies. Further, we develop AoI-aware variants of these policies in which each source takes its current AoI into account while making decisions. We compare the performance of various policies via simulations.**

I. Introduction

We consider the problem of scheduling in a multi-source multi-channel system, focusing on the metric of Age-of-Information (AoI), introduced in [1]. AoI is formally defined as the time elapsed since the destination received the recent most update from the source. It follows that AoI is a measure of the freshness of the data available at the intended destination which makes it a suitable metric for time-sensitive systems like smart homes, smart cars, and other IoT based systems. Since its introduction, AoI has been used in areas like caching, scheduling, and channel state information estimation. A comprehensive survey of AoI-based works is available in [2].

The work in [3] shows the performance of AoI bandits for a single source and multiple channels, where the source acts as the "bandit" which pulls one of the arms in every time-slot, i.e., selects one of the channels for communication. The aim is to find the best arm (channel) while minimizing the AoI (instead of the usual reward maximization) for that source. In this work, we address more practical scenarios in which systems have multiple sources looking to simultaneously communicate through a common, limited pool of channels. Here, we need to ensure that the total sum of the AoIs of all the sources is minimized. Moreover, we consider a decentralized setting where the sources cannot share information with each other, meaning that chances of collisions (attempting to communicate through the same channel in a time-slot) can be very high. Thus, the task of designing policies that minimize total AoI while ensuring fairness among all sources and avoiding collisions is a challenging problem. Prior works on decentralized multi-player MAB problems primarily discuss reward-based policies, however, minimizing AoI is more challenging, as the impact of sub-optimal decisions gets accumulated over time.

The decentralized system comprises of multiple sources and multiple channels, where at every time-slot, each of the M decentralized users searches for idle channels to send a periodic update. The probability of an attempted update succeeding is independent across communication channels and independent and identically (i.i.d) distributed across time-slots for each channel. AoI increases by one on each failed update and resets to one on each successful update. These distributed players can only learn from their local observations and collide (with a reward penalty) when choosing the same arm. The desired objective is to develop a sequential policy running at each user to select one of the channels, without any information exchange, in order to minimize the cumulative AoI over a finite time-interval of T consecutive time-slots.

A. Our Contribution

Optimality of Round-Robin policy: We describe an oracle Round-Robin policy and characterize its optimality.

Upper-Bound on AoI regret of DLF [4]: We characterize a generic expression for the upper bound of the total cumulative AoI for any policy. Further, we show that the AoI regret of the DLF policy scales as $\mathrm{O}(M^2 N \log^2 T)$.

New AoI-agnostic policies: We propose a Thompson Sampling [5] based policy. We also present a new hybrid policy trading-off between Thompson Sampling and DLF.

New AoI-aware policies: We propose AoI-aware variants for all the AoI-agnostic policies. When AoI values are below a certain threshold, the variants mimic the original policy. Otherwise, the variants exploit local past observations. Through simulations, we show that these variants exhibit lower AoI regrets as compared to their agnostic counterparts.

B. Related Work

In this section, we discuss the prior work most relevant to our setting. AoI-based scheduling has been explored by [6]–[10], where the channel statistics were assumed to be known and an infinite time-horizon steady-state performance-based approach is adopted. The work in [3] explores the setting where the channel statistics are unknown, for a single source and multiple channels. We consider a decentralized system with multiple sources, which is a much more complex setting.

Time Division Fair Sharing algorithm proposed in [11] addresses the fair-access problem in a distributed multi-player setting. This policy was outperformed by the policy proposed in [4]. Recent work in Multi-Player MABs (MPMABs) include [12]–[14]. The policies in [12] consider an alternate channel allotment in the event of a collision. Although these ideas can be adopted in all our policies, in this work, we did not wish to make any assumptions about the feasibility of these alternate allotments. The settings in [13], [14] are significantly different and cannot be readily adapted to our setting. Further, the policies may not be directly consistent with the AoI metric.

II. Setting

A. Our System

We consider a system with M sources and N channels ($N \geq M$). The sources track/measure different time-varying quantities and relay their measurements to a monitoring station via available communication channels. Time is divided into slots. In each time-slot, each of the M decentralized users (sources) selects an arm (channel) only based on its own observation histories under a decentralized policy, and then attempts to transmit via that channel.[1] Each attempted communication via channel i is successful with probability μ_i and unsuccessful otherwise, independent of all other channels and across time-slots. The values of the μ_is are unknown to the sources. Without loss of generality, we assume the μ_is to be in descending order, i.e., $\mu_1 > \mu_2 > ... > \mu_N$.

A decentralized setting means that when a particular arm i is selected by user j, the reward (whether the transmission was successful or not) is only observed by user j, and if no other user is playing the same arm, a reward is obtained with probability μ_j. Else, if multiple users are playing the same arm (which is possible in a decentralized setting), then we assume that, due to collision, exactly one of the conflicting users can use the channel to transmit, while the other users get zero reward. This "winner" is chosen at random since we assume all sources to have equal priority. This is consistent with network protocols like CSMA with perfect sensing.

B. Metric: Age-of-information Regret

The age-of-information is a metric that measures the freshness of information available at the monitoring station. It is formally defined as follows:

Definition 1 (Age-of-Information (AoI)). *Let $a(t)$ denote the AoI at the monitoring station in time-slot t and $u(t)$ denote the index of the time-slot in which the monitoring station received the latest update from the source before the beginning of time-slot t. Then,*

$$a(t) = t - u(t).$$

By definition,

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

Let $a_m^{\mathcal{P}}(t)$ be the AoI of source m in time-slot t under a given policy $\mathcal{P}$, and let $a_m^*(t)$ be the corresponding AoI under the oracle policy, discussed in detail in section III-A. We define the AoI regret at time T as the cumulative difference in expected AoI for the two policies in time-slots 1 to T summed over all the sources. Hence, the lesser the AoI a policy accumulates, the better it is. All policies operate under Assumption 1, similar to the initial conditions in [3].

Definition 2 (Age-of-Information Regret (AoI Regret)). *AoI regret under policy $\mathcal{P}$ is denoted by $\mathcal{R}_{\mathcal{P}}(T)$ and*

$$\mathcal{R}_{\mathcal{P}}(T) = \sum_{m=1}^{M}\sum_{t=1}^{T}\mathbb{E}[a_m^{\mathcal{P}}(t) - a_m^*(t)]. \tag{1}$$

Assumption 1 (Initial Conditions). *The system starts operating at time-slot $t = -\infty$, but for any policy, decision making begins at $t = 1$. It does not use information from observations in time-slots $t \leq 0$ to make decisions in time-slots $t \geq 1$.*

III. Analytical Results

In this section, we state and discuss our main theoretical results. We characterize our oracle/genie policy and prove its optimality. We also provide an upper-bound for the regret of a UCB [15] based policy for multi-user and multi-armed setting, called Distributed Learning with Fairness (DLF) [4].

A. Oracle Policy

We will consider two candidates for the oracle policy, namely I.I.D and Round-Robin policy described in Definitions 3 and 4 respectively. Then, Theorem 1 describes the optimal I.I.D policy.

Definition 3 (I.I.D Policy). *Any policy $\mathcal{P}$ where scheduling decisions are independently and identically distributed across time is called an I.I.D policy.*

Theorem 1 (Optimal I.I.D Policy). *Consider the set of all I.I.D policies, the policy that schedules the M sources on a random permutation of the set of arms $\mathcal{S}(M) = \{1, \cdots, M\}$ at every instant (uniformly at random) is optimal.*

Definition 4 (Round-Robin Policy). *Let $\mathcal{S}(k) = \{1, \cdots, k\}$ be the set of k channels with the highest success probabilities. For a problem instance $(M, N, \boldsymbol{\mu})$, consider the index set $\mathcal{I}$ which is a random permutation of the arms in the set $\mathcal{S}(M)$. Then, in time-slot t, the round-robin policy schedules a source m on the channel $\mathcal{I}_{((m+t) \bmod M)+1}$.*

Theorems 2 and 3 characterize the optimality of the Round-Robin policy under certain conditions. This leads us to Conjecture 1, where we generalize the oracle to be optimal when there are more than two simultaneous sources.

Theorem 2 (Optimality of Round-Robin Policy for $M = 2$). *For any problem instance $(M, N, \boldsymbol{\mu})$ such that $M = 2, N \geq M$, a policy that schedules source $m \in [M]$ on the channel $((m+t) \bmod M) + 1$ is optimal.*

Theorem 3 (I.I.D. vs. Round-Robin). *The round-robin policy has a smaller expected total AoI value than the I.I.D. policy.*

[1] We use the terms source and user, and arm and channel interchangeably

Conjecture 1. *Generalizing Theorem 2 (for $M > 2$), over all possible permutation of arms in the set $\mathcal{S}(M)$, the oracle policy is optimal for any problem instance $(M, N, \boldsymbol{\mu})$.*

Hereafter the regret of all the policies is calculated with respect to the round-robin policy, that is, the round-robin policy is our oracle. The set $\mathcal{S}(M)$ is used as our pre-decided index set $\mathcal{I}$. Standard single-source bandit policies try to pick the best-estimated arm. Since we have multiple sources with equal priority, we try to pick the "k^{th} best" arm, where k changes in a round-robin fashion for each source, thus ensuring fairness. However, this can still lead to collisions, as the estimates of the channel means are independently maintained by each source, and no information is shared. Thus, we check at every instant if the channel is acquired by the source (based on our collision model), and update the mean estimates only when the source has access to the channel. The regret for any policy $\mathcal{P}$ arises if a source m, for an index k, chooses an arm other than the desired one (which is used by the oracle) or due to a collision, the source is unable to acquire the channel. Thus, there is a trade-off between exploration and the number of collisions.

B. Distributed Learning with Fairness (DLF)

DLF is an equal-priority multi-source and multi-channel policy based on UCB-1. This policy tries to mimic our oracle policy while avoiding collisions, by estimating the channel $\boldsymbol{\mu}$ through a two-step process- estimating the set of best-M channels and then selecting the channel with k^{th} highest value of μ from the set, where the index k is determined by the user m and current slot t. This is formally described in Algorithm 1.

Algorithm 1: Distributed Learning Algorithm with Fairness for M Sources and N Channels at Source m (DLF)

Initialize: Set $\hat{\mu}_n^m = 0$ to be the estimated success probability of Channel n, $T_n^m(1) = 0$ $n \in [N]$.
while $1 \le t \le N$ **do**
  Schedule an update to a link $n(t)$ such that
  $n(t) = ((m + t) \bmod M) + 1$,
  Receive reward $X_{n(t)}^m(t) \sim \text{Ber}(\mu_{n(t)}^m)$
  $\hat{\mu}_n^m = X_{n(t)}^m(t)$,
  $T_n^m(t) = 1$
while $t \ge N + 1$ **do**
  Set index $k = ((m + t) \bmod M) + 1$
  Let the set $\mathcal{O}_k$ contain the k arms with the k largest values in (2)

$$\hat{\mu}_n^m + \sqrt{\frac{2 \log t}{T_n^m(t-1)}} \tag{2}$$

  Schedule an update to a link $n(t)$ such that

$$n(t) = \arg\min_{n \in \mathcal{O}_k} \hat{\mu}_n^m - \sqrt{\frac{2 \log t}{T_n^m(t-1)}}$$

  if *Channel Acquired* **then**
    Receive reward $X_{n(t)}^m(t) \sim \text{Ber}(\mu_{n(t)}^m)$
    $\hat{\mu}_{n(t)}^m = \frac{\hat{\mu}_{n(t)}^m \cdot T_{n(t)}^m(t-1) + X_{n(t)^m}(t)}{T_{k(t)}(t-1)+1}$
    $T_{n(t)}^m(t) = T_{n(t)}^m(t-1) + 1$
  else
    $\hat{\mu}_{n(t)}^m = \hat{\mu}_{n(t)}^m$
    $T_{n(t)}^m(t) = T_{n(t)}^m(t-1)$
  $t = t + 1$

Theorem 4 (Performance of DLF). *Consider any problem instance $(M, N, \boldsymbol{\mu})$, such that $\mu_{min} = \min_{i=1:N} \mu_i > 0$, $\Delta = \min_{i,j \in [M]; i > j} \mu_i - \mu_j$, and a constant c. Then, for any sufficiently large $T > N$, under Assumption 1,*

$$\mathcal{R}_{DLF}(T) \le \frac{M^2}{\mu_{min}} + \frac{M^2 c \log T}{\mu_{min}} \left[1 + (N-1)\left(\frac{8 \log T}{\Delta^2} + 1 + \frac{2\pi^2}{3} \right) \right].$$

From Theorem 4, we conclude that $\mathcal{R}_{\text{DLF}}(T)$ scales as $\text{O}(M^2 N \log^2 T)$. The proof first characterizes the source wise regret as a function of the expected number of times a non-desired channel is scheduled or another source acquires the desired channel under the policy. The result then follows using known upper-bounds for this quantity, from [4] for DLF.

IV. Other Policies

In this section, we extend Thompson Sampling [5] to our setting. We also propose a new hybrid policy, as well as AoI-Aware variants for each of these policies, and compare their performances via simulations.

A. AoI-Agnostic Policies

Definition 5 (AoI-Agnostic Policies). *A policy is AoI-agnostic if, given past scheduling decisions and the number of successful updates sent via each of the N channels in the past, it does not explicitly use the age of information of any source in a time-slot to make scheduling decisions.*

While AoI is a new metric, to devise AoI-agnostic policies one can use the myriad of policies used for Bernoulli rewards, most commonly the Upper Confidence Bound (UCB) [15] policy and Thompson Sampling [5]. These policies can be readily applied to a one-source setting, i.e. $M = 1$ as done in [3]. DLF [4], as described in Algorithm 1, is based on UCB and tries to mimic our oracle policy while avoiding collisions. Similarly, we extend Thompson Sampling for our setting, and we term this new policy as *Distributed Learning-based Thompson Sampling (DL-TS)*, the algorithm for which can be found in [16]. Further, we combine these two policies to propose a hybrid policy - *Distributed Learning-based Hybrid policy (DLH)*, detailed in Algorithm 2.

Empirically, we observe that DLF exhibits a lesser number of collisions while being more exploratory, and DL-TS is more exploitative but leads to a higher number of collisions. The Distributed Learning-based Hybrid policy (DLH), shown

Algorithm 2: Distributed Learning-based Hybrid policy for M Sources and N Channels at Source m (DLH)

Initialize: Set $\hat{\mu}_n^m = 0$ to be the estimated success probability of Channel n, $T_n^m(1) = 0$ $n \in [N]$.
while $t \geq 1$ **do**
  Let $E(t) \sim$ Ber $\left(\min\left\{1, mn\frac{\log t}{t}\right\}\right)$.
  if $E(t) = 1$ **then**
    DLF: Schedule an update on a channel chosen by the DLF policy given in Algorithm 1.
  else
    Thompson Sampling:
    $\alpha_n^m(t) = \hat{\mu}_n^m(t) T_n^m(t-1) + 1$,
    $\beta_n^m(t) = (1 - \hat{\mu}_n^m(t)) T_n^m(t-1) + 1$,
    For each $n \in [N]$, pick a sample $\hat{\theta}_n^m(t)$ of distribution,
$$\hat{\theta}_n^m(t) \sim \text{Beta}(\alpha_n^m(t), \alpha_n^m(t)) \tag{3}$$
    Set index $k = ((m + t) \bmod M) + 1$
    Schedule an update to a link $n(t)$ such that it is the arm with the k^{th} largest value in (3).
  if *Channel Acquired* **then**
    Receive reward $X_{n(t)}^m(t) \sim \text{Ber}(\mu_{n(t)}^m)$
    $\hat{\mu}_{n(t)}^m = \frac{\hat{\mu}_{n(t)}^m \cdot T_{n(t)}^m(t-1) + X_{n(t)^m}(t)}{T_{k(t)}(t-1)+1}$
    $T_{n(t)}^m(t) = T_{n(t)}^m(t-1) + 1$
  else
    $\hat{\mu}_{n(t)}^m = \hat{\mu}_{n(t)}^m$
    $T_{n(t)}^m(t) = T_{n(t)}^m(t-1)$
  $t = t + 1$

in Algorithm 2, switches between DLF and DL-TS using a Bernoulli random variable to balance this trade-off. As the chance of a collision is higher in the initial time-slots, here the algorithm uses DLF with a higher probability than DL-TS. However, the probability of using DLF decreases with an increase in the number of time-slots elapsed.

B. AoI-Aware Policies

In this section, we propose AoI-aware variants of the three policies discussed in the previous section. In the classical MAB with Bernoulli rewards, the contribution of a time-slot to the overall regret is upper bounded by one, but in AoI bandits it can be more than one. Also, unlike the MAB, for AoI bandits, the difference between AoIs under a candidate policy and the oracle policy in a time-slot can be unbounded. This motivates the need to take the current AoI value into account when making scheduling decisions.

Intuitively, it makes sense to explore when AoI is low and exploit when AoI is high since the cost of making a mistake is much higher when AoI is high. We use this intuition to design AoI-aware policies. The key idea behind these policies is that they mimic the original policies when AoI is below a threshold

Algorithm 3: AoI-Aware Distributed Learning-based Hybrid Policy for M Sources and N Channels at Source m (DLH-AA)

Initialize: Set $\hat{\mu}_n^m = 0$ to be the estimated success probability of Channel n, $T_n^m(1) = 0$ $n \in [N]$.
while $t \geq 1$ **do**
  $\alpha_n^m(t) = \hat{\mu}_n^m(t) T_n^m(t-1) + 1$,
  $\beta_n^m(t) = (1 - \hat{\mu}_n^m(t)) T_n^m(t-1) + 1$,
  Set index $k = ((m + t) \bmod M) + 1$
  Let *limit(t)* $= k^{th} \min\limits_{n \in [N]} \frac{\alpha_n^m(t) + \beta_n^m(t)}{\alpha_n^m(t)}$
  if $a(t-1) >$ *limit(t)* **then**
    Exploit: Select channel with the k^{th} highest estimated success probability
  else
    Explore:
    Let $E(t) \sim$ Ber $\left(\min\left\{1, mn\frac{\log t}{t}\right\}\right)$.
    if $E(t) = 1$ **then**
      DLF: Schedule an update on a channel chosen by the DLF policy given in Algorithm 1.
    else
      Thompson Sampling: Schedule an update on a channel chosen by the DL-TS policy given in [16].
  if *Channel Acquired* **then**
    Receive reward $X_{n(t)}^m(t) \sim \text{Ber}(\mu_{n(t)}^m)$
    $\hat{\mu}_{n(t)}^m = \frac{\hat{\mu}_{n(t)}^m \cdot T_{n(t)}^m(t-1) + X_{n(t)^m}(t)}{T_{k(t)}(t-1)+1}$
    $T_{n(t)}^m(t) = T_{n(t)}^m(t-1) + 1$
  else
    $\hat{\mu}_{n(t)}^m = \hat{\mu}_{n(t)}^m$
    $T_{n(t)}^m(t) = T_{n(t)}^m(t-1)$
  $t = t + 1$

and exploit when AoI is equal to or above a threshold, for an appropriately chosen threshold.

In each policy, at a source m, we maintain an estimate of success probability of the arms, denoted by $\hat{\boldsymbol{\mu}}^m$. For an index k at time t, we mimic the original policy if the AoI is not more than $\frac{1}{\hat{\mu}_k^m}$ (which is the expected AoI value if the k^{th} arm was used throughout). Otherwise, we exploit the "k^{th} best" arm (based on past observations) as per the source. The AoI-Aware modification to DLH is given in Algorithm 3. For brevity, the AoI-aware variants of DLF and DL-TS are detailed in [16].

V. Simulations

In this section, we will present the simulation results for all the six policies discussed in section IV. All the simulations hereafter are conducted for $T = 2 \times 10^4$ time-slots with each data point averaged across 200 iterations. For the purpose of simulating the policies, we choose $\boldsymbol{\mu}$ such that consecutive elements are equidistant, and the difference is denoted by Δ.

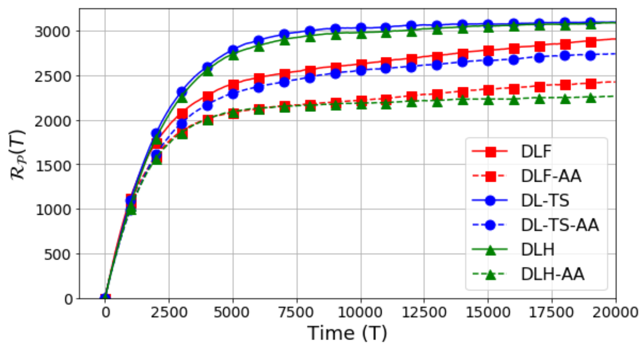


Figure 1: AoI regret for $M = 3, N = 5, \boldsymbol{\mu} = [0.6; 0.8]$

Table I: Distribution of Channels scheduled across sources for various policies when $M = 3, N = 5$ and $\boldsymbol{\mu} = [0.6; 0.8]$

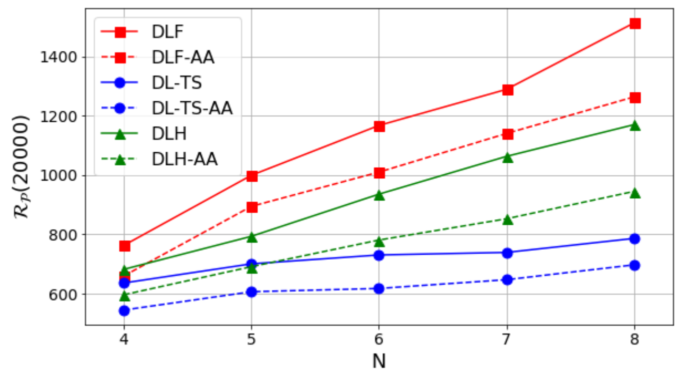
		$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
DLF-AA	$m=1$	6621	6543	4598	1511	727
	$m=2$	6627	6524	4631	1487	731
	$m=3$	6624	6535	4634	1479	728
DLTS-AA	$m=1$	6640	6524	6023	655	158
	$m=2$	6585	6557	6039	644	175
	$m=3$	6581	6573	6096	580	170

Figure 1 shows the regret, $\mathcal{R}_\mathcal{P}(T)$, for all the six policies mentioned previously for a particular problem instance $(M, N, \boldsymbol{\mu})$. We observe that *AoI-Aware* policies have significantly lower regret values (and fewer collisions) than their *AoI-Agnostic* counterparts.

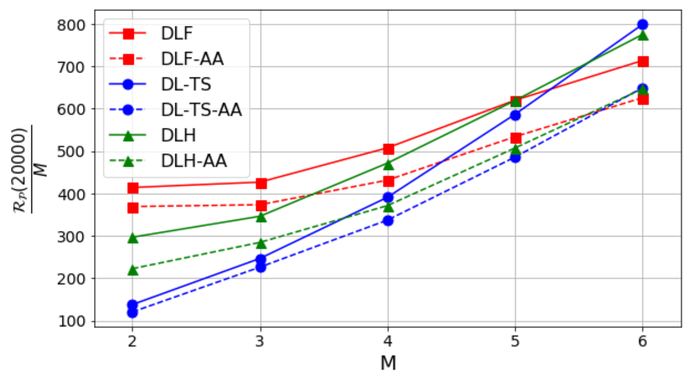
In Table I, we compare the distribution of source-wise channels scheduled under DLF-AA and DLTS-AA for the problem instance $(M = 3, N = 5, \boldsymbol{\mu} = [0.6; 0.8])$. We observe that the number of pulls of sub-optimal arms, ie., arms $\notin \mathcal{S}(3)$ is higher in DLF-AA as compared to DLTS-AA. At the same time, the former policy averages about 1071 net collisions, whereas the latter averages about 1478 net collisions.

The same trend occurs when comparing DLF with DL-TS. This indicates that there is a trade-off between the two causes of regret: sub-optimal arm pulls and the number of collisions. This served as the primary motivation behind the hybrid policies- DLH and DLH-AA, which lead to an intermediate number of both sub-optimal arm pulls and number of collisions. As one can see from Figure 1, DLH-AA performs the best in terms of cumulative regret, given that Δ is small ($\Delta = 0.1$). For higher Δ, we observed that the DL-TS policies performed better since $\boldsymbol{\mu}$ being farther apart will lead to lesser collisions. DLH always resulted in intermediate values of regret, in between regret values for DLF and DL-TS. However, DLH is 'closer' to DL-TS as compared to DLF, which is justified by the construction of the policy. Readers can find more regret plots with different values of $(M, N, \boldsymbol{\mu})$ in [16].

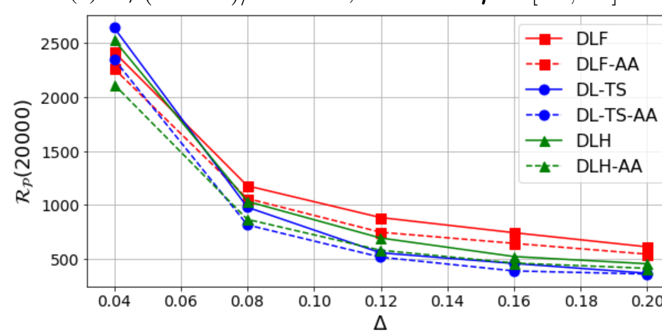
In Table I, there is a sizable difference in the pulls of the M^{th} arm as compared to the others in the set $\mathcal{S}(M)$. We believe that this is because across polices a majority of the sub-optimal pulls arise from scheduling the M^{th} channel. This trend can be seen in Figure 3(b) of [4]. Note that although the order of magnitude of the difference in the arm pulls is similar to ours, the absolute number is higher as their simulations are conducted for $T = 10^6$. On increasing T, the number of arm pulls is close to the numbers presented in [4].



(a) $\mathcal{R}_\mathcal{P}(2 \times 10^4)$ vs. N; $M = 3$, $\mu_1 = 0.9$ and $\Delta = 0.1$.



(b) $\mathcal{R}_\mathcal{P}(2 \times 10^4)/M$ vs. M; $N = 7$ and $\boldsymbol{\mu} = [0.3; 0.9]$



(c) $\mathcal{R}_\mathcal{P}(2 \times 10^4)$ vs. Δ; $M = 3, N = 5$ and $\mu_1 = 0.9$

Figure 2: Variation in regret, $\mathcal{R}_\mathcal{P}(T)$, at $T = 2 \times 10^4$, with different values of N, M and Δ (or $\boldsymbol{\mu}$), in that order.

Figure 2 shows the variation of regret values of policies at $T = 2 \times 10^4$ with different parameters. In Figure 2a, we observe that the regret scales approximately linearly with N, with the DLF policies having the highest values of both slope and regret, and DL-TS the lowest. This is consistent with the degree of exploration in the policies. For variation with M, in Figure 2b, we plot $\mathcal{R}_\mathcal{P}(T)/M$ instead of $\mathcal{R}_\mathcal{P}(T)$. The almost linear nature of these plots suggests that regret scales with

M^2. We notice that the trend in the magnitude of the slopes is the reverse of that in Figure2a, because as N increases, DL-TS becomes more susceptible to collisions, thus accumulating higher regret. Figure 2c indicates an inverse relationship of regret $\mathcal{R}_{\mathcal{P}}(T)$ with Δ in all the six policies. We note that when $\Delta = 0.04$, DLH and DLF based policies do better than the DL-TS based ones, with DLH-AA being the best policy. As Δ increases, DL-TS based policies have the lowest regret and the DLF based policies have the highest regret.

VI. Proof Outlines

In this section, we present the proof outlines for all the theorems. For the detailed proofs, we refer readers to [16].

A. Theorem 1

- We first establish that any candidate for an optimal policy must satisfy two properties. It should not lead to any collisions and it should not allot any channel outside the "best-M" channels or $\mathcal{S}(M)$. (Lemmas 1 and 2 in [16])
- We then obtain an expression for the expected total age-of-information of an I.I.D policy. (Lemma 3 in [16])
- Finally, through KKT conditions, we show that uniform I.I.D policy is the best I.I.D policy proving the theorem.

B. Theorems 2 and 3

- We devise two scheduling policies: Schedule A is the round-robin policy and Schedule B mimics Schedule A except for an M-length (in time-slots) deviation in which at least one channel is repeated.
- We show that the expected total AoI at any time T under Schedule B is greater than that under Schedule A.
- For Theorem 2, we enumerate the cases and use the arithmetic and geometric means inequality, whereas, for Theorem 3 we use the generalized Muirhead's inequality.

C. Theorem 4

- We first upper bound the expected cumulative source-wise AoI for any schedule by the expected cumulative source-wise AoI of an alternative schedule in which all uses of sub-optimal channels in the original schedule are replaced by all sources using (and thereby colliding on) the worst channel. Only one of these sources, at random, acquires the channel. (Lemma 4 in [16])
- We further upper bound the expected cumulative source-wise AoI with another schedule where all uses of the worst channel are clustered together starting from $T = 1$, followed by the oracle allotments. (Lemma 4 in [16])
- We can express the resultant upper bound on the source-wise expected cumulative AoI in terms of the number of deviations from the oracle allotments and the number of times another source acquires the same channel as the source (resulting in a collision). (Lemma 4 in [16])
- We substitute the known bounds for these quantities under the DLF policy and use this to obtain an upper bound on the expected cumulative total AoI. (Lemma 5 in [16])

VII. Conclusion

We model a multi-source multi-channel setting using Age-of-Information bandits with decentralized users. We first characterize the oracle policy, namely, the round-robin policy, and demonstrate that the upper-bound of AoI-regret of the existing Distributed Learning with Fairness (DLF) policy scales as $\mathrm{O}(M^2 N \log^2 T)$. However, proving the optimality of the oracle policy for the general setting is still an open problem. We then propose two other AoI-agnostic policies (DL-TS and DLH) with different trade-offs between exploration and the number of collisions. These policies only utilize the past arm-pulls in deciding the next arm. We also present AoI-aware policies that incorporate the current value of AoI into the arm selections. Via simulations, we show that the AoI-aware policies lead to fewer collisions and thus, outperform their AoI-agnostic counterparts.

References

[1] 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.

[2] 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.

[3] K. Bhandari, S. Fatale, U. Narula, S. Moharir, and M. K. Hanawal, "Age-of-information bandits," in *2020 18th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT)*, 2020, pp. 1–8.

[4] Y. Gai and B. Krishnamachari, "Distributed Stochastic Online Learning Policies for Opportunistic Spectrum Access," *IEEE Transactions on Signal Processing*, vol. 62, no. 23, pp. 6184–6193, 2014.

[5] 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.

[6] 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*, 2018, pp. 696–698.

[7] 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.

[8] V. Tripathi and E. Modiano, "A whittle index approach to minimizing functions of age of information," in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2019, pp. 1160–1167.

[9] P. R. Jhunjhunwala and S. Moharir, "Age-of-information aware scheduling," in *2018 International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 2018, pp. 222–226.

[10] 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.

[11] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pp. 5667–5681, 2010.

[12] L. Besson and E. Kaufmann, "Multi-player bandits revisited," in *Algorithmic Learning Theory*, 2018, pp. 56–92.

[13] M. K. Hanawal and S. J. Darak, "Multi-player bandits: A trekking approach," *arXiv preprint arXiv:1809.06040*, 2018.

[14] E. Boursier and V. Perchet, "SIC-MMAB: synchronisation involves communication in multiplayer multi-armed bandits," in *Advances in Neural Information Processing Systems*, 2019, pp. 12 071–12 080.

[15] 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.

[16] A. Prasad, V. Jain, and S. Moharir, "Decentralized age-of-information bandits," *arXiv preprint arXiv:2009.12961*, 2020. [Online]. Available: https://arxiv.org/abs/2009.12961