Online Wireless Scheduling for Throughput Maximization under Unknown Channel Statistics

***Abstract*—We consider a wireless scheduling problem in downlink wireless networks with unknown channel statistics. Scheduling performance relies heavily on accurate channel state information (CSI), which is often costly to acquire. In this study, CSI is obtained from ACK/NACK feedback, only after each scheduled transmission. Due to limited channel resources, all users cannot be scheduled for transmission simultaneously. Hence, the most recently observed CSI can be outdated. The traditional approach to solving scheduling problems using outdated CSI is to utilize belief states, which are calculated using the time correlation statistics of channels. However, channel statistics are often unknown; consequently, belief states can be uncountable and this approach becomes infeasible. In this paper, we introduce a new sufficient statistic-the latest observed CSI and its *Age of Channel State Information (AoCSI)*, which characterizes the CSI staleness, to make the scheduling decisions. Accordingly, we are able to significantly reduce the state space. We develop an online Maximum Gain First (Online-MGF) policy which achieves sub-linear regret on the number of episodes. Numerical results demonstrate that Online-MGF policy converges to MGF and Whittle index policies with known channel statistics within a very few episodes. In addition, Online-MGF outperforms Maximum AoCSI First (MAF) and random policies.**

I. Introduction

6G and Future-Generation (Future-G) wireless networks aims to support a massive number of users and a broad range of applications, such as interconnected Internet of Things (IoT) devices, vehicle to vehicle and vehicle to infrastructure communications, and Unmanned Aerial Vehicle (UAV) deployments. A fundamental challenge in these applications is the efficient allocation of limited wireless resources among a massive number of users. This difficulty is further increased when the wireless channel states fluctuate over time.

One strategy to address the difficulty that arises from fluctuating channel conditions is to exploit the time correlation of the channel state information (CSI) to predict the current CSI based on outdated CSI [1]–[7]. For example, a BS can anticipate the future capacity of a channel by understanding its time-varying characteristics. However, accurate time correlation statistics for a wireless channel are often unknown in practice. The lack of complete channel statistical knowledge makes the resource allocation problem highly complicated.

Motivated by these practical limitations, we investigate an online wireless scheduling problem for throughput maximization under unknown channel statistics and Imperfect CSI in this paper. Our system consists of a BS that transmits data to N users and operates in a discrete-time environment. Due to limited communication resources, the BS selects M out of N users for transmission at any time-slot t. To model the time correlations in fading channels, we assume that the wireless channel between the BS and each scheduled user evolves as a discrete-time, two-state Markov chain (ON/OFF channels). The BS obtains the CSI from an ACK/NACK feedback after each scheduled transmission. Therefore, the BS utilizes the past available observations to predict the CSI in the current time-slot before making the scheduling decisions. Moreover, the channel statistics are unknown to the BS in our model. To that end, we answer the following research question in this study:

> *How to design an efficient online scheduling algorithm for maximizing throughput while satisfying an instantaneous resource constraint with unknown channel statistics and imperfect CSI?*

The contributions of this study are summarized below:

- The optimal scheduling problem for maximizing throughput can be formulated as an RMAB [8]. Existing studies [1]–[4], [7], [9]–[11] utilize belief states as sufficient statistics which can lead to an uncountable number of possible states, specifically when the channel statistics are unknown. We are able to significantly reduce the state space of this problem by utilizing the sufficient statistic of the history [12]. To that end, we leverage a metric called *Age of Channel State Information (AoCSI)* [13], which is defined as the time difference between the current time and the last time CSI is updated. We show that the latest observed CSI and its AoCSI form a sufficient statistic of the information history to make the scheduling decisions for each user (see Theorem 1). Therefore, by leveraging the latest observed CSI and its AoCSI as states, we replace the Partially Observed Markov Decision Process (POMDP) or belief MDP framework used in [1]–[4], [7], [9]–[11] with a more tractable MDP which exhibits a significantly smaller and countable state space.
- By using (i) relaxation and Lagrangian decomposition and (ii) an optimistic Upper Confidence Bound (UCB) approach, we develop an online Maximum Gain First (Online-MGF) policy (see Algorithm 2). Unlike the Whittle index-based policy [8], our policy does not require satisfying any indexability condition. Utilizing standard relaxation and Lagrangian decomposition, we decouple the original problem into multiple independent MDPs. Because we are able to reduce the state space of each MDP by finding a sufficient statistic, we can simplify the search space of the online learning algorithm. Unlike the existing online learning approaches developed in [14]–[17], we do not need to estimate the entire transition

probabilities (see Remarks 1 and 2). This also reduces the computational complexity of our policy compared to the online learning policies studied in [14]–[17].
- The proposed Online-MGF policy achieves sublinear regret $O(\sqrt{K \log K})$ on the number of episodes K (see Theorem 2).
- Numerical results show that the Online-MGF policy converges faster to the MGF and Whittle index policies with known channel statistics and achieves good performance over Maximum AoCSI First (MAF) and Random policies (see Figures 3-6). In one setup, the Online-MGF policy outperforms the Whittle index policy with known channel statistics (see Figure 5).

II. Related Work

Opportunistic scheduling algorithms have been widely used and extensively studied in the literature under perfect and imperfect CSI [1], [3]–[7], [9], [18]. These studies addressed the scheduling problem using the POMDP (or belief MDP) formulation which considers belief states [2]–[7], [9], [18]. Such formulations can lead to an uncountable number of state spaces, specifically when the channel statistics are unknown. Age of Information (AoI) has emerged as an important metric to study sampling and scheduling problems, because it captures the "freshness (or rather staleness)" of information [10], [19]–[27]. Building upon this concept, *Age of Channel State Information (AoCSI)* was proposed in [13] to analyze the impact of CSI staleness on the overall system performance. Subsequently, several papers [28]–[31] studied different scheduling schemes and the impact of AoCSI and different functions of AoCSI. When the channel statistics are unknown, using POMDP (or belief MDP) formulations with belief states makes the solution computationally intractable and introduces significant challenges for online learning algorithms. Unlike [1], [3]–[7], [9], [18], we do not require utilizing POMDP (or belief MDP) formulation with belief states. We obtain a sufficient statistic of the history in this study, which contains the last observed CSI and its AoCSI. Therefore, instead of utilizing belief states, we can use a significantly smaller state space.

Moreover, restless bandits are a popular framework to solve resource allocation and sequential decision-making problems. Numerous scheduling problems have been formulated as RMABs and Whittle index policies were found for contexts such as multi-access channels, network utility maximization, age penalty minimization, remote estimation, etc [2], [8], [20]–[22]. Whittle index policy is proven to be asymptotically optimal [32] under an indexability condition, which is often very challenging to establish in practice. Consequently, non-indexable scheduling policies have been analyzed in recent years [11], [23], [27], [33]–[35]. Unlike [11], [23], [27], [33]–[35], we consider unknown channel statistics, and thus the methods established in existing studies cannot be used to solve our problem. A gain index-based asymptotically optimal policy was proposed in [11], [27], [34]–[36] with known system dynamics. When system dynamics are unknown, online learning is a promising approach to solve RMABs, where a decision-maker simultaneously estimates the unknown system dynamics and makes scheduling decisions. There exist several online learning policies that studied RMAB problems with unknown system dynamics [15]–[17], [37]–[40]. Without the "restless" setting, [16], [37], [38] studied online learning policies for Multi-armed Bandits (MABs). Under indexability, [40] proposed a Thompson sampling-based online Whittle index policy and [15], [17] proposed UCB-based online Whittle index policies. Unlike [15], [17], [40], we develop a UCB-based online Maximum Gain First (Online-MGF) policy that does not require satisfying any indexability condition, which often does not hold in practice. Moreover, compared to the existing studies, our proposed Online-MGF policy requires us to estimate only one variable instead of the entire transition probabilities as we find a sufficient statistic of the history.

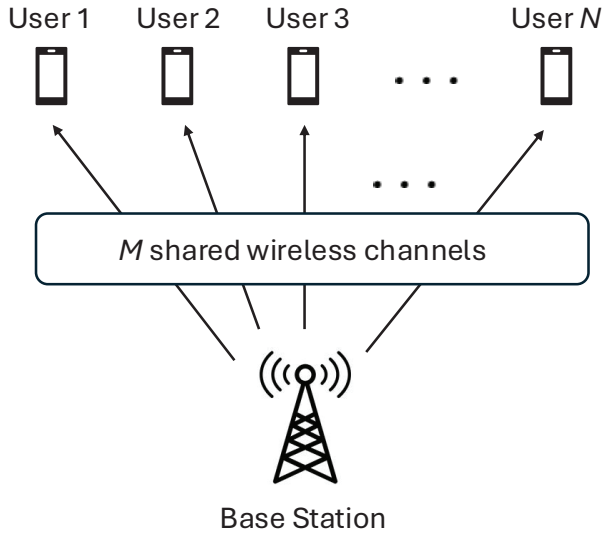

Fig. 1: System Model.

III. Model and Problem Formulation

A. System Model

Our system consists of a wireless network with one Base Station (BS) and N users with M shared wireless channels as depicted in Figure 1. The system is time-slotted and operates within a finite time horizon T. At any time-slot t, the BS can simultaneously transmit to at most M users without interference, where $M < N$.

The wireless channel $C_n(t)$ between the BS and user n is independent across users, i.e., $C_n(t), t = 1, 2, \ldots, T$, and $C_m(t), t = 1, 2, \ldots, T$ are independent for all $n \neq m$. The CSI $C_n(t)$ remains static within each time-slot. However, $C_n(t)$ evolves stochastically between successive time-slots. The CSI $C_n(t)$ is modeled as a discrete-time, two-state time-homogeneous Markov chain (ON/OFF channels), where $C_n(t) \in \{0, 1\}$. If $C_n(t) = 0$, the channel of user n is in the OFF state; otherwise, if $C_n(t) = 1$, the channel of user n is in the ON state. The state transition of the channels for user n

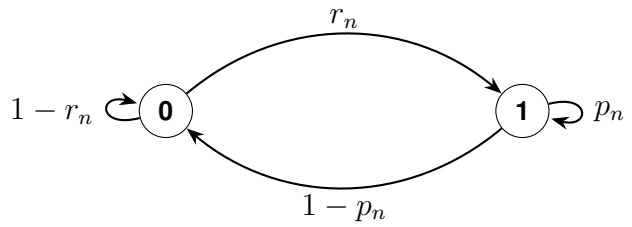

Fig. 2: Two-state Markov chain model for channel of user n.

is provided in Figure 2. Their corresponding transition matrix is given by

$$\mathbb{P}_n = \begin{bmatrix} P_n(1|1) & P_n(0|1) \\ P_n(1|0) & P_n(0|0) \end{bmatrix} = \begin{bmatrix} p_n & 1-p_n \\ r_n & 1-r_n \end{bmatrix}, \quad (1)$$

where

$$p_n = P_n(C_n(t)=1|C_n(t-1)=1), \quad (2)$$
$$r_n = P_n(C_n(t)=1|C_n(t-1)=0), \quad (3)$$

and $P_n(c'|c)$ is the state transition probability from CSI c to CSI c' for user n. The transition matrix $\mathbb{P}_n$ is unknown in our model.

At the beginning of each time-slot t, the BS does not know the actual CSI $C_n(t)$ while making the scheduling decisions. Let $a_n(t) \in \{0,1\}$ denotes the scheduling decision to schedule user n at time-slot t, defined as

$$a_n(t) = \begin{cases} 1, \text{ if user } n \text{ is scheduled in time-slot } t, \\ 0, \text{ otherwise.} \end{cases} \quad (4)$$

Whenever user n receives a data from the BS, it sends a zero-delay ACK/NACK feedback to the BS to inform a successful or failed transmission. Denote $\beta_n(t) \in \{0,1\}$ as the delivery indicator for user n at time-slot t. If user n successfully receives the data from the BS in time-slot t, then $\beta_n(t)=1$. However, the feedback is only received at the end of the time-slot after the data is transmitted. The delivery indicators $\beta_n(t)$ are independent across both users and time-slots. Consequently, when the BS makes the scheduling decision in time-slot t, the most recent information available of user n is the last observed CSI $C_n(t-\Delta_n(t))$, which is observed at time-slot $t-\Delta_n(t)$. The time difference between the current time t and the last CSI update time $t-\Delta_n(t)$ is called the *Age of Channel State Information (AoCSI)* [13], defined as

$$\Delta_n(t+1) = \begin{cases} 1, & \text{if } a_n(t)=1\,, \\ \Delta_n(t)+1, & \text{otherwise.} \end{cases} \quad (5)$$

B. Problem Formulation with Known Channel Statistics

Let $\pi = (a_n(1), a_n(2), \ldots, a_n(T))_{n=1}^N$ denote a scheduling policy, where $a_n(t)$ is defined in (4) for all $t=1,2,\ldots,T$. Our goal is to find the optimal policy π that maximizes the sum of throughput over the finite time horizon T, subject to the constraint on the number of users selected in each time-slot t. The sum-throughput optimization problem is formulated as follows

$$\max_{\pi\in\Pi} \sum_{t=1}^{T}\sum_{n=1}^{N} \mathbb{E}_\pi \left[C_n(t)a_n(t) \middle| \mathcal{H}_n(t) \right] \quad (6)$$

$$\text{s.t.} \sum_{n=1}^{N} a_n(t) \le M,\ t=1,2,\ldots,T, \quad (7)$$

where Π is the set of all causal scheduling policies, $C_n(t)a_n(t)$ is the amount of user-n data served in time-slot t, and $\mathcal{H}_n(t)$ is the information history available for the BS at time-slot t, defined as

$$\mathcal{H}_n(t) = (C_n(\tau-\Delta_n(\tau)), \Delta_n(\tau), a_n(\tau-1), \beta_n(\tau-1))_{\tau=1}^{t}, \quad (8)$$

which includes all previously observed channel states, their AoCSI values, scheduling decisions, and delivery indicators of user n up to time-slot t.

Problem (6)-(7) can be formulated as a Partially Observed Markov Decision Process (POMDP) [41]. This is because the scheduler has no knowledge of the current CSI at time-slot t while making the scheduling decisions. Existing studies utilize belief states to solve similar kind of POMDP problems [1]–[4], [7], [9]–[11]. However, with unknown channel statistics, evaluating belief states are significantly challenging for online learning algorithms as the belief state space can be uncountable. To address this, we provide a simplification of problem (6)-(7) by reducing its state space. The details are presented in Section IV.

IV. Problem Simplification

We are able to simplify problem (6)-(7) by utilizing the sufficient statistic of the history [12].

Theorem 1. *If $C_n(t) \leftrightarrow C_n(t-1) \leftrightarrow C_n(t-2) \leftrightarrow \ldots$ is a Markov chain, then the last observed CSI and its AoCSI $(C_n(t-\Delta_n(t)), \Delta_n(t))_{n=1}^N$ is a sufficient statistic of the information history $(\mathcal{H}_n(t))_{n=1}^N$ for making the scheduling decisions $(a_n(t))_{n=1}^N$ at time-slot t in problem* (6)-(7)*.*

Proof. See Appendix A. □

Theorem 1 implies that $(C_n(t-\Delta_n(t)), \Delta_n(t))_{n=1}^N$ is a sufficient statistic to find the optimal scheduling decisions at time-slot t. Using Theorem 1, we can reformulate problem (6)-(7) as

$$\max_{\pi\in\Pi} \sum_{t=1}^{T}\sum_{n=1}^{N} \mathbb{E}_\pi \Big[f_n(C_n(t-\Delta_n(t)), \Delta_n(t), a_n(t)) \Big] \quad (9)$$

$$\text{s.t.} \sum_{n=1}^{N} a_n(t) \le M,\ t=1,2,\ldots,T, \quad (10)$$

where

$$\begin{aligned} &f_n(C_n(t-\Delta_n(t)), \Delta_n(t), a_n(t)) \\ =&a_n(t)\mathbb{E}[C_n(t)|C_n(t-\Delta_n(t)), \Delta_n(t)]. \end{aligned} \quad (11)$$

If $a_n(t)=0$, then $f_n(C_n(t-\Delta_n(t)), \Delta_n(t), 0) = 0$; If $a_n(t)=1$, then $f_n(C_n(t-\Delta_n(t)), \Delta_n(t), 1)$ can be further

simplified: We denote the probability of the current CSI $C_n(t) = c'$ from the outdated CSI $C_n(t - \Delta_n(t)) = c$ and AoCSI $\Delta_n(t) = \delta$ by $P_n(c'|c, \delta)$. Given $C_n(t - \Delta_n(t)) = c$ and $\Delta_n(t) = \delta$, we have

$$f_n(c, \delta, 1) = P_n(1|c, \delta), \tag{12}$$

where $P_n(1|c, \delta)$ can be calculated as

$$P_n(1|0, \delta) = \frac{r_n(1 - (p_n - r_n)^\delta)}{1 - p_n + r_n}, \tag{13}$$

$$P_n(1|1, \delta) = \frac{r_n + (1 - p_n)(1 - (p_n - r_n)^\delta)}{1 - p_n + r_n}. \tag{14}$$

In problem (9)-(10), the state space is reduced from an uncountable number of belief states to a countable number of states. Moreover, the state space can be made finitely countable using a large truncated AoCSI value [1], [21]. We let τ denote the maximum AoCSI value in the truncated space.

Problem (9)-(10) is an RMAB where each user n is represented as an arm and each arm n is described as an MDP. Each MDP associated with arm n has two actions at every time-slot t: active ($a_n(t) = 1$) and passive ($a_n(t) = 0$). The problem (9)-(10) is "restless" because the AoCSI $\Delta_n(t)$ evolves even when the n-th arm is passive [21]. Since we are able to simplify the original problem (6)-(7), the RMAB (9)-(10) has a reduced state space compared to (6)-(7).

However, even with reduced state space, solving RMAB problems is significantly challenging, and it is intractable to find the optimal solution [42]. A Whittle index policy is known to be an efficient approach to solve RMAB problems [8]. This policy is known to be asymptotically optimal under a complicated condition called indexability [32]. Nevertheless, indexability is often very difficult to establish in practice. Recently, another policy called the gain index policy has been proven to be asymptotically optimal for RMABs [11], [27], [43]. A key advantage of the gain index-based policy is that it does not require proving indexability. In this paper, we provide a gain index-based policy for the problem (9)-(10). Furthermore, in Section VII, we show numerically that our developed gain index-based policy performs equally well compared to the Whittle index policy; Even in one setup (Figure 5 with $p_1 = 0.8$), the developed gain index-based policy performs better than the Whittle index policy.

In addition, the transition matrix $\mathbb{P}_n$ in (1) is unknown in our model. Therefore, we first study a gain index-based policy for problem (9)-(10) with known channel state transition matrix $\mathbb{P}_n$. Utilizing the insights from the solution with known channel statistics, we will develop an online algorithm called "Online Maximum Gain First (Online-MGF)" policy when the channel statistics are unknown. To that end, we first utilize Lagrangian dual decomposition and decompose problem (9)-(10) to N independent per-arm problems. The details are provided in Section IV-A.

A. Relaxation and Lagrangian Decomposition

Following the standard relaxation and Lagrangian decomposition technique for RMABs [8], we first relax the constraint (10) and obtain the following relaxed problem:

$$\max_{\pi \in \Pi} \sum_{t=1}^{T} \sum_{n=1}^{N} \mathbb{E}_\pi \Big[f_n(C_n(t - \Delta_n(t)), \Delta_n(t), a_n(t)) \Big] \tag{15}$$

$$\text{s.t.} \sum_{t=1}^{T} \sum_{n=1}^{N} \mathbb{E}_\pi [a_n(t)] \leq MT, \tag{16}$$

where the relaxed constraint (16) needs to be satisfied in total finite time T, instead of satisfying at every time-slot t. Next, we apply Lagrange multiplier $\lambda \geq 0$ (also known as dual cost) to the relaxed constraint (16) and obtain the following Lagrangian dual problem:

$$L(\pi^*, \mathbf{P}, \lambda) = \max_{\pi \in \Pi} \sum_{n=1}^{N} \mathbb{E}_\pi \Bigg[\sum_{t=1}^{T} f_n(C_n(t - \Delta_n(t)), \Delta_n(t), a_n(t)) - \lambda(a_n(t) - MT) \Bigg], \tag{17}$$

where $\mathbf{P} = (\mathbb{P}_n)_{n=1}^N$, and the term λMT is excluded in (17) as it is a constant, and does not affect the policy π, and therefore can be removed. For a given λ, problem (17) can be decomposed into N independent per-arm problems, where each per-arm problem associated with arm n is given by

$$L_n(\pi_n^*, \mathbb{P}_n, \lambda) = \max_{\pi_n \in \Pi_n} \mathbb{E}_{\pi_n} \Bigg[\sum_{t=1}^{T} f_n(C_n(t - \Delta_n(t)), \Delta_n(t)), a_n(t)) - \lambda a_n(t) \Bigg], \tag{18}$$

where $L_n(\pi_n^*, \mathbb{P}_n, \lambda)$ is the optimum value of (18), $\pi_n = (a_n(1), a_n(2), \ldots, a_n(T))$ is the sub-scheduling policy of arm n, and Π_n is the set of all causal sub-scheduling policies of arm n.

We first solve (18) for given λ. Following the solution of the decoupled problem (18), we will develop a solution for the original RMAB (9)-(10). Then we obtain the optimal Lagrange multiplier from $\lambda^* = \text{argmin}_\lambda \, L(\pi^*, \mathbf{P}, \lambda)$, where $L(\pi^*, \mathbf{P}, \lambda)$ is defined in (17).

Using dynamic programming [12], (18) can be solved. Given $C_n(t - \Delta_n(t)) = c$ and $\Delta_n(t) = \delta$ at time-slot t, the Bellman optimality equation for the MDP (18) is given by

$$V_{n,t}(c, \delta; \lambda) = \max_{a \in \mathcal{A}} Q_{n,t}(c, \delta, a; \lambda), \tag{19}$$

where $Q_{n,t}(c, \delta, a; \lambda)$ is the action-value function defined as

$$Q_{n,t}(c, \delta, a; \lambda) = \begin{cases} f_n(c, \delta, 1) + \sum_{c' \in \{0,1\}} P_n(c'|c, \delta) V_{n,t+1}(c', 1; \lambda) - \lambda, & \text{if } a = 1, \\ V_{n,t+1}(c, \delta + 1; \lambda), & \text{if } a = 0. \end{cases} \tag{20}$$

Using (12), the action-value function (20) can be further

simplified as

$$Q_{n,t}(c,\delta,a;\lambda) = \begin{cases} P_n(1|c,\delta)+\sum_{c'\in\{0,1\}} P_n(c'|c,\delta)V_{n,t+1}(c',1;\lambda)-\lambda, & \text{if } a=1, \\ V_{n,t+1}(c,\delta+1;\lambda), & \text{if } a=0. \end{cases} \tag{21}$$

The value function $V_{n,t}(c,\delta;\lambda)$ can be calculated using the backward induction method [12] from time $t = T, T-1,\ldots,1$:

$$V_{n,t}(c,\delta;\lambda) = \max\Big\{P_n(1|c,\delta) + \sum_{c'\in\{0,1\}} P_n(c'|c,\delta)V_{n,t+1}(c',1;\lambda)-\lambda, \; V_{n,t+1}(c,\delta+1;\lambda)\Big\}, \tag{22}$$

where $V_{n,T+1}(c,\delta;\lambda)=0$ for all (c,δ).

If the action-value function associated with the active action $Q_{n,t}(c,\delta,1;\lambda)$ is higher than the action-value function associated with the passive action $Q_{n,t}(c,\delta,0;\lambda)$, i.e, $Q_{n,t}(c,\delta,1;\lambda) > Q_{n,t}(c,\delta,0;\lambda)$, then the optimal decision to problem (18) is to schedule arm n.

An equivalent policy is to employ the gain index, which is defined by [11], [27], [34]

$$\alpha_{n,t}(c,\delta) = Q_{n,t}(c,\delta,1;\lambda) - Q_{n,t}(c,\delta,0;\lambda). \tag{23}$$

The gain index $\alpha_{n,t}(c,\delta)$ is the difference between two action-value functions. Following the policy with the gain index $\alpha_{n,t}(c,\delta)$, arm n is scheduled if the gain index is positive, i.e., $\alpha_{n,t}(c,\delta)>0$. To that end, we develop a Maximum Gain First (MGF) policy. This policy selects at most M arms with the highest gain.

Remark 1. *Analysis of the action-value function* (21) *yields useful insights to design our online learning algorithm. From* (21)*, if the action is passive (i.e., $a=0$), it is evident from* (5) *that the AoCSI transition will be from δ to $\delta+1$ and the CSI c will remain in the same state c. Consequently, the actual state transition is known to the BS with probability* 1 *when a passive action is taken. Therefore, to design our online algorithm, we do not need to compute any probabilities associated with the passive action.*

V. Problem Statement for Online Setting

In our online learning environment, the BS interacts with N arms through K episodes, where the length of the time horizon in each episode is H time-slots. A key challenge in this setting is that the decision maker has no prior knowledge of the actual transition probabilities $\mathbf{P} = (\mathbb{P}_n)_{n=1}^N$. In every episode k, the BS starts from an initial state $(\mathbf{c}(1),\boldsymbol{\delta}(1)) = \{(c_1(1),\delta_1(1)),(c_2(1),\delta_2(1)),\ldots,(c_N(1),\delta_N(1))\}$ and implements a policy π^k. Subsequently, it utilizes the observations up to H time-slots to iteratively estimate the underlying transition probabilities.

Finding the optimal policy for RMAB problems is generally intractable and PSPACE-hard [42]. Therefore, it is significantly challenging to evaluate the performance of an online learning policy with respect to the optimal policy of RMAB. The authors in [15], [17] employ Lagrangian dual problem to evaluate the performance of the developed online policies. Similar to [15], [17], we utilize Lagrangian dual problem to evaluate the performance of our online policy π^k in episode k. Through simulation, we demonstrate that the Online-MGF policy converges to the MGF policy with known channel statistics within a very few episodes (see Section VII).

Definition 1. Regret. Given the actual transition probabilities $\mathbf{P}$, the regret of the policy π^k in episode k is given by

$$\text{Reg}(k) = L(\pi^*,\mathbf{P},\lambda^*,\mathbf{c}(1),\boldsymbol{\delta}(1)) - L(\pi^k,\mathbf{P},\lambda^k,\mathbf{c}(1),\boldsymbol{\delta}(1)), \tag{24}$$

where π^* is the optimal policy for the Lagrangian dual problem (17) and λ^* minimizes the Lagrangian $L(\pi^*,\mathbf{P},\lambda,\mathbf{c}_1,\delta_1)$ defined in (17). Using (24), we get the total regret as follows:

$$\text{Reg}(K) = \sum_{k=1}^{K} \text{Reg}(k). \tag{25}$$

VI. Design of Online Policy for Solving (9)-(10)

We solve the online learning problem in two steps: (i) Utilizing a UCB-based approach, we first develop an Online-MGF policy in Section VI-A that does not need to satisfy any indexability condition; (ii) Next, we analyze the regret and find a regret bound which illustrates that the developed algorithm achieves sublinear regret in Section VI-B.

A. Online-MGF Policy

We utilize UCB-based online learning to provide an algorithm for the Online-MGF policy. To compute the gain index in (23), we need to know the action-value functions, and, hence, the transition probabilities $\mathbf{P}$. Because $\mathbf{P}$ is unknown in our model, we develop an online learning method in Algorithm 2.

1) Confidence Bound for Transition Probabilities: For every arm n and every episode k, we maintain the number of visits $B_n^k(c,\delta,c')$, which denotes the number of transitions from CSI $C_n(t-\Delta_n(t))=c$ and AoCSI $\Delta_n(t)=\delta$ to the CSI $C_n(t)=c'$ under the active action $a_n(t)=1$ observed by the last k episodes. Because our learning algorithm only needs to estimate probabilities associated with the active actions, we need to observe the state transitions under the active actions only. For a small given constant $\mu>0$, we define the confidence radius as follows:

$$\rho_n^k(c,\delta) = \sqrt{\frac{4\log(4\tau N\frac{k^2}{\mu})}{\max(1,B_n^k(c,\delta))}}, \tag{26}$$

where $B_n^k(c,\delta) = \sum_{c'\in\{0,1\}} B_n^k(c,\delta,c')$ is the number of visits to state (c,δ) for arm n under active action by episode k, τ is the maximum value in the truncated AoCSI space, and N is the number of users.

The empirical transition probability is given by

$$\hat{P}_n^k(c'|c,\delta) = \frac{B_n^k(c,\delta,c')}{B_n^k(c,\delta)}. \tag{27}$$

Then, the confidence ball of possible transition probabilities is given by

$$\mathcal{B}^k = \Bigg\{ P_n^k | \sum_{c' \in \{0,1\}} \Big| P_n^k(c'|c,\delta) - \hat{P}_n^k(c'|c,\delta) \Big| \leq \rho_n^k(c,\delta), \forall n, c, \delta \Bigg\}. \tag{28}$$

2) Algorithm for Online-MGF Policy: Using an optimistic approach, we estimate the transition probabilities [14], [15]. We choose the optimistic transition probability $P_n^k(1|c,\delta)$ for each arm n in episode k that maximizes the value function within the confidence bound $\rho_n^k(c,\delta)$. The optimization problem at episode k in time-slot t is given by

$$\max_{P_n^k \in \mathcal{B}^k} V_{n,t}^{P_n^k}(c,\delta;\lambda^k), \tag{29}$$

$$\text{s.t. } V_{n,t}^{P_n^k}(c,\delta;\lambda^k) = \max_{a \in \mathcal{A}} Q_{n,t}^{P_n^k}(c,\delta,a;\lambda^k), \tag{30}$$

where $Q_{n,t}^{P_n^k}(c,\delta,a;\lambda^k)$ is the action-value function given by

$$Q_{n,t}^{P_n^k}(c,\delta,a;\lambda^k) = \begin{cases} P_n^k(1|c,\delta) + \sum\limits_{c' \in \{0,1\}} P_n^k(c'|c,\delta) V_{n,t}^{P_n^k}(c',1;\lambda^k) - \lambda^k, & \text{if } a = 1, \\ V_{n,t}^{P_n^k}(c,\delta+1;\lambda^k) - \lambda^k, & \text{if } a = 0. \end{cases} \tag{31}$$

After computing the optimistic transition probabilities and optimistic value functions, we utilize the optimistic action-value functions for each arm n in episode k to compute the gain index as follows:

$$\alpha_{n,t}^k(c,\delta;\lambda^k) = Q_{n,t}^{P_n^k}(c,\delta,1;\lambda^k) - Q_{n,t}^{P_n^k}(c,\delta,0;\lambda^k). \tag{32}$$

The algorithm to compute the Online-MGF policy is provided in Algorithm 2. At every episode k, this policy selects M arms with the highest positive gain index (32).

Remark 2. A Lagrangian-based online Whittle index policy was studied in [15], which requires solving two sequential optimization problems: first for optimistic transition probabilities, then for the Whittle index. Unlike [15], our Algorithm 2 only requires solving one optimization problem (29)-(30), where we obtain the optimistic transition probability $P_n^k(1|c,\delta)$ for the CSI $C_n(t) = 1$ at time-slot t given outdated CSI value $C_n(t - \Delta_n(t)) = c$ and AoCSI $\Delta_n(t) = \delta$. This also yields $P_n^k(0|c,\delta) = 1 - P_n^k(1|c,\delta)$. Therefore, we do not need to compute the entire transition probabilities from state (c,δ) to next state (c',δ') since whenever a source n is scheduled AoCSI becomes 1, and if source n is not scheduled AoCSI increases by 1. After getting the optimistic probability $P_n^k(1|c,\delta)$, the gain index in (32) is directly computed using optimistic action-value functions (31). Hence, the optimization problem in our algorithm is much simpler compared to the two sequential optimization problems in [15].

Algorithm 1: Online Maximum Gain First (Online-MGF) Policy

Initialize N users, constraint M, episode length H.
Initialize $B_n^1(c,\delta,c') = 0$ for all c,δ,c'.
Initialize $\lambda^{(1)}$.
for *episode* $k = 1, 2, \ldots$ **do**
  Reset $t = 1$ and initial state $(\mathbf{c},\boldsymbol{\delta}) = (\mathbf{c}(1),\boldsymbol{\delta}(1))$.
  Compute transition probabilities for each arm by solving the optimization problem (29)-(30).
  Compute gain indices from (32) for each arm at every time-slot t.
  Schedule M users with highest gain indices at every time-slot t.
  Observe transitions and update visits $B_n^k(c,\delta,c')$, empirical transition probability $\hat{\mathbf{P}}^k$, confidence region $\mathcal{B}^k$.
  Update λ^{k+1} from (33).

Algorithm 2: Online Maximum Gain First (Online-MGF) Policy

Initialize N users, constraint M, episode length H.
Initialize $B_n^1(c,\delta,c') = 0$ for all c,δ,c'.
Initialize $\lambda^{(1)}$.
for *episode* $k = 1, 2, \ldots$ **do**
  Reset $t = 1$ and initial state $(\mathbf{c},\boldsymbol{\delta}) = (\mathbf{c}(1),\boldsymbol{\delta}(1))$.
  Compute $P_n^k(1|c,\delta)$ for each arm n and for all (c,δ) by solving (29)-(30).
  Compute gain indices from (32) for each arm at every time-slot t.
  Schedule M users with highest gain indices at every time-slot t.
  Observe transitions and update visits $B_n^k(c,\delta,c')$, empirical transition probability $\hat{\mathbf{P}}^k$, confidence region $\mathcal{B}^k$.
  Update λ^{k+1} from (33).

The Lagrange multiplier λ^k is updated in every k-th episode. The update rule is given by

$$\lambda^{k+1} = \max\left\{ \lambda^k + \frac{\eta}{kH}\left(\sum_{n=1}^{N} \sum_{t=1}^{H} a_n^k(t) - M \right), 0 \right\}, \tag{33}$$

where $\eta/(kH)$ is the learning parameter, and action $a_n^k(t)$ is obtained by solving the following decoupled problem:

$$a_n^k(t) = \arg\max_a Q_{n,t}^{P_n^k}(c,\delta,a;\lambda_k). \tag{34}$$

B. Regret Bound for Algorithm 2

In this section, we analyze the regret of Algorithm 2. To bound the regret, we first show that with high probability the true transition $\mathbf{P}$ lies within the confidence ball $\mathcal{B}^k$ in (28).

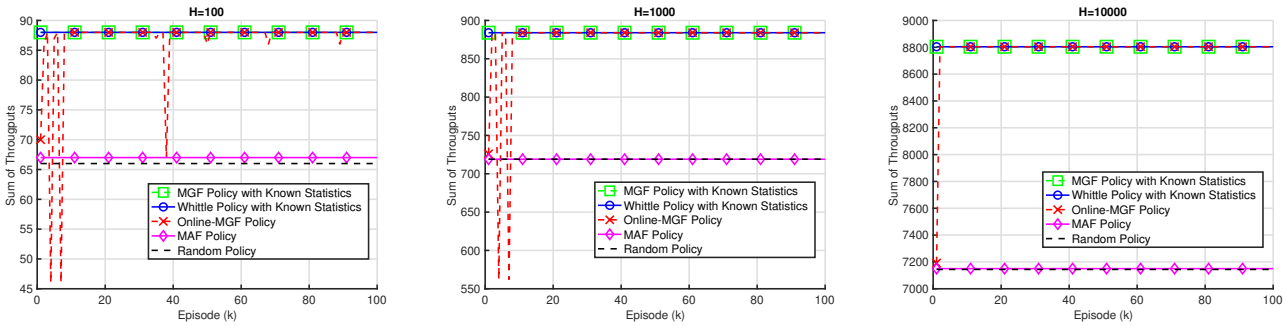

Fig. 3: Sum of Throughputs vs the number of episodes (k) with different time-horizon length in each episode (H). The other parameters are $p_1 = 0.6$, $r_1 = 0.5$, $p_2 = 0.9$, $r_2 = 0.75$, $N = 4$, and $M = 1$.

Lemma 1. *Given $\mu > 0$ and $k \geq 1$, we get that* $\Pr(\mathbf{P} \in \mathcal{B}^k) \geq 1 - \frac{\mu}{k^2}$.

Proof sketch. Leveraging [44, Theorem 2.1], we establish a bound for the L_1-deviation of the actual transition and the empirical transition. □

Lemma 1 implies that for each episode k, a confidence bound can be obtained within which the actual transition lies with high probability. Utilizing Lemma 1, we bound the regret for each episode k in the following lemma.

Lemma 2. *For given initial state $((c_1(1), \delta_1(1)), (c_2(1), \delta_2(1)), \ldots, (c_N(1), \delta_N(1)))$, the following holds with probability $1 - \mu$:*

$$\begin{aligned}&\mathrm{Reg}(k)\\ &\leq \sum_{n=1}^{N} V_{n,1}^{P_n^k}(c_n(1), \delta_n(1); \lambda^k) - V_{n,1}^{P_n}(c_n(1), \delta_n(1); \lambda^k). \quad (35)\end{aligned}$$

Proof sketch. If the confidence bound holds, i.e., $\mathbf{P} \in \mathcal{B}_k$, we can write

$$\begin{aligned}&\mathrm{Reg}(k)\\ &= L(\pi^*, \mathbf{P}, \lambda^*, \mathbf{c}(1), \boldsymbol{\delta}(1)) - L(\pi^k, \mathbf{P}, \lambda^k, \mathbf{c}(1), \boldsymbol{\delta}(1))\\ &= \sum_{n=1}^{N} V_{n,1}^{P_n}(c_n(1), \delta_n(1); \lambda^*) - V_{n,1}^{P_n}(c_n(1), \delta_n(1); \lambda^k)\\ &\leq \sum_{n=1}^{N} V_{n,1}^{P_n}(c_n(1), \delta_n(1); \lambda^k) - V_{n,1}^{P_n}(c_n(1), \delta_n(1); \lambda^k), \quad (36)\end{aligned}$$

where (36) holds because the Lagrange multiplier λ^* minimizes the Lagrangian $L(\pi^*, \mathbf{P}, \lambda)$ in (17). Next, in order to prove (35), we first show that

$$V_{n,t}^{P_n}(c, \delta; \lambda^k) \leq V_{n,t}^{P_n^k}(c, \delta; \lambda^k), \quad (37)$$

which holds by backward induction for any $t = H, H-1, \ldots, 1$ as $V_{n,H+1}(c, \delta) = 0$ for all (c, δ). When the confidence bound holds, using (37) and the fact that P_n^k solves the optimization problem (29)-(30), we can show that

$$\mathrm{Reg}(k) \leq \sum_{n=1}^{N} V_{n,1}^{P_n^k}(c_n(1), \delta_n(1); \lambda^k) - V_{n,1}^{P_n}(c_n(1), \lambda_n(1); \lambda^k). \quad (38)$$

□

Next, we utilize Lemma 2 to bound the regret for K episodes. In this sequel, we have the following theorem.

Theorem 2. *With probability $1 - \eta$, the cumulative regret of Algorithm 2 in all K episodes is given by*

$$\mathrm{Reg}(K) \leq O\left(V_{\max} |\mathcal{D}| N H \sqrt{K \log K}\right), \quad (39)$$

where $V_{\max}$ is the maximum value-function, $|\mathcal{D}|$ is the size of the state space of AoCSI, N is the number of users, H is the time-horizon length in each episode, and K is the total number of episodes.

Proof sketch. When the confidence bound holds, we get

$$\begin{aligned}&\mathrm{Reg}(k)\\ &\leq \sum_{n=1}^{N} \mathbb{E}_{P_n, \pi_n^k}\Bigg[\sum_{t=1}^{H} \rho_n^k(c_n(t), \delta_n(t), 1) +\\ &\sum_{t=1}^{H} \left| P_n^k(\cdot | c_n(t), \delta_n(t), 1) - P_n(\cdot | c_n(t), \delta_n(t), 1) \right| V_{\max}\Bigg], \quad (40)\end{aligned}$$

where π_n^k is the policy in episode k for arm n. The inequality (40) is obtained by using Lemma 2 and applying Lemma 10 of [45]. Next, we utilize the confidence radius $\rho_n^k(c, \delta, 1)$ in (26) and obtain the following bound

$$\begin{aligned}\mathrm{Reg}(k) \leq \sum_{n=1}^{N} &\frac{1}{\mu}(1 + V_{\max})\sqrt{4 \log(4\tau N K)}\\ &\mathbb{E}_{P_n, \pi_n^k}\left[\sum_{(c,\delta) \in \{0,1\} \times \mathcal{D}} \frac{\gamma_n^k(c, \delta)}{\max(1, B_n^k(c_n(1), \delta_n(1)))}\right], \quad (41)\end{aligned}$$

where $\gamma_n^k(c, \delta)$ is a random variable, which represents the number of visits to state $(c, \delta) \in \{0, 1\} \times \mathcal{D}$ in episode k for

arm n. Because Algorithm 2 needs to estimate the probabilities associated with the active actions only, we do not need to consider actions in (41). Next, applying [15, Lemma E.3], and taking the sum over all K episodes, we get the cumulative regret in (39).

Theorem 2 illustrates that Algorithm 2 achieves sub-linear regret with respect to the number of episodes K. This regret bound aligns with existing online learning algorithms in literature [15], [17], [37], [38], [40], [46]. In section VII, our numerical results illustrate that the Online-MGF policy converges to both the MGF and Whittle policies with known channel statistics within a small number of episodes.

VII. Simulation Results

In this section, we evaluate the performance of the following policies:

- **Random Policy**: This policy randomly selects users following a uniform distribution. In the simulation, we use MATLAB built-in function $\mathrm{randperm}(N, M)$.
- **MGF Policy with Known Channel Statistic**: This policy computes the gain index using the known channel statistics. Then, this policy schedules M users with the highest gain indices in every time-slot.
- **Whittle Index Policy with Known Channel Statistic**: This policy computes the Whittle index using the known channel statistics. The closed-form expression for the Whittle index with known channel statistics of this problem is provided in [2], [7]. Then, the policy schedules M users with the highest Whittle indices in every time-slot t.
- **Online-MGF Policy**: The policy is provided in Algorithm 2. The Online-MGF policy does not know the channel statistics. This policy first learns the channel statistics, then computes gain indices at the beginning of each episode, and schedule M users with the highest gain indices.
- **Maximum Age First (MAF) policy**: This policy schedules M users with the highest AoCSI values at every time-slot.

In the simulation, we consider two classes of users, namely class 1 and class 2. Users in the same class have the same channel statistics. We consider that half of the users belong to class 1 and another half belong to class 2. We denote $P_n(C_n(t) = 1|C_n(t-1) = 1)$ of user n by p_1 and $P_n(C_n(t) = 1|C_n(t-1) = 0)$ by r_1 if the user n belongs to class 1; otherwise we denote $P_n(C_n(t) = 1|C_n(t-1) = 1)$ of user n by p_2 and $P_n(C_n(t) = 1|C_n(t-1) = 0)$ by r_2 if user n belongs to class 2.

Figure 3 illustrates the sum of throughput vs the number of episodes (k) with different time horizon length in each episode $H = 100, 1000, 10000$. In this figure, we set $p_1 = 0.6$, $r_1 = 0.5$, $p_2 = 0.9$, $r_2 = 0.75$, $N = 4$, and $M = 1$. According to Figure 3, MGF policy with known channel statistics shows similar performance as Whittle index policy with known channel statistics. The figure also depicts that the proposed Online-MGF policy converges faster to the MGF policy with known channel statistics as the number of time horizon H in each episode increases. This is because for higher values of H, Online-MGF observes more state transitions in each episode, which yields accurate learning of channel statistics. Moreover, Online-MGF performs better compared to the other baselines, such as MAF and random policies. This is because MAF policy only considers AoCSI and ignores most recently observed CSI, whereas Random policy schedules the users randomly.

Figure 4 illustrates the sum of throughput vs number of episodes (k) with different configurations of N and M. In this figure, we set $H = 10000$ and the other parameters are the same as Figure 3. This figure shows that for all configurations, the Online-MGF policy converges faster to the MGF and the Whittle index policies with known channel statistics. As expected, the proposed Online-MGF policy performs better compared to MAF and random policies.

Figure 5 depicts the sum of throughput vs the number of episodes (k) with different values of $p_1 = 0.1, 0.5, 0.8$. In this figure, we set $N = 10$, $M = 4$ and other parameters are the same as Figure 4. Interestingly, when $p_1 = 0.8$, MGF policy with known channel statistics, and Online-MGF policy outperforms Whittle index policy with known channel statistics.

In Figure 6, we plot the sum of throughput vs the number of channels (M). In this figure, the sum of throughput for each point is averaged over 100 episodes. For this figure, we set $N = 100$ and other parameters are the same as Figure 4. As the number of channels increases, the sum of throughput increases. The figure illustrates that the performance gain of Online-MGF policy with MAF and Random policies first increases with M up to $M = 50$. After $M = 50$, the performance gain decreases because the number of available channels for allocating to $N = 100$ users increases.

Our simulation illustrates that the performance of our proposed Online-MGF policy aligns with the MGF policy and Whittle index policy with known channel statistics. Furthermore, Online-MGF policy outperforms MAF and random policy as the number of episodes increases in every settings. Notably, for $p_1 = 0.8$, the Online-MGF policy outperforms the Whittle index policy with known channel statistics.

VIII. Conclusion

We study an online wireless scheduling problem with imperfect CSI and unknown channel statistics. The formulated problem is an RMAB. By leveraging the sufficient statistic of the history, we are able to significantly reduce the state space of the underlying RMAB. To that end, we develop an Online-MGF policy to solve the RMAB that does not need to satisfy any indexability condition and achieves sub-linear regret with the number of episodes. Through simulation, we validate the effectiveness of the proposed online-MGF policy.

Appendix A
Proof of Theorem 1

We prove Theorem 1 in three steps: (i) First, we show that $C_n(t)$ is conditionally independent of the $(C_n(\tau -$

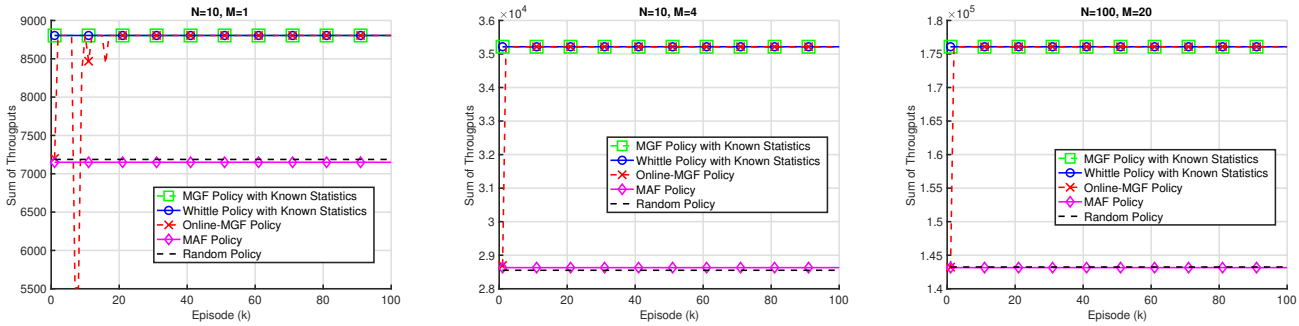

Fig. 4: Sum of Throughputs vs the number of episodes (k) with different configurations of N and M. The other parameters are $p_1 = 0.6$, $r_1 = 0.5$, $p_2 = 0.9$, $r_2 = 0.75$, and $H = 10000$.

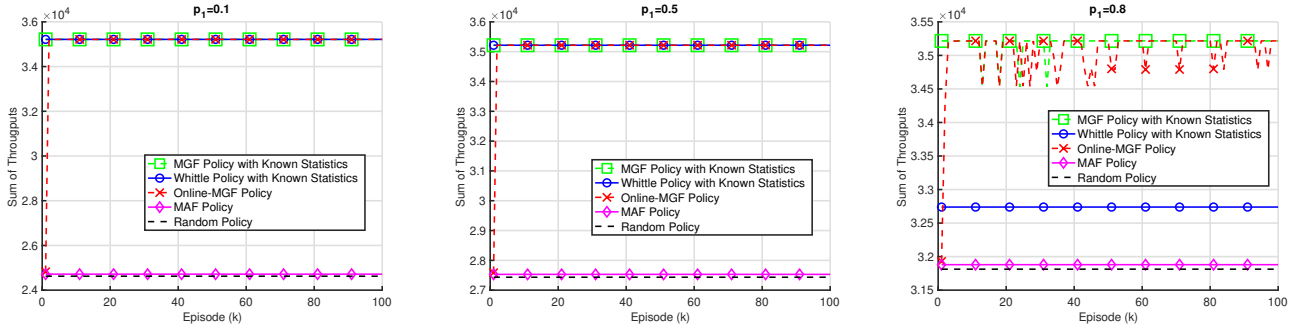

Fig. 5: Sum of Throughputs vs the number of episodes (k) with different channel statistics. The other parameters are $H = 10000, N = 10$, and $M = 4$.

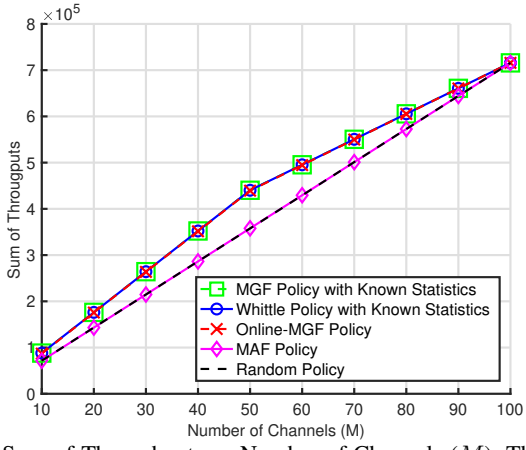

Fig. 6: Sum of Throughputs vs Number of Channels (M). The other parameters are $p_1 = 0.6$, $r_1 = 0.5$, $p_2 = 0.9$, $r_2 = 0.75$, $H = 10000$, and $N = 100$.

$\Delta_n(\tau)), \Delta_n(\tau))_{\tau=1}^{t-1}$ given the last observed CSI $C_n(t-\Delta_n(t))$ and its AoCSI $\Delta_n(t)$; (ii) Next, we show that $C_n(t)$ is conditionally independent of the scheduling decisions $(a_n(\tau))_{\tau=1}^{t-1}$ and the delivery indicators $(\beta_n(\tau))_{\tau=1}^{t-1}$ given the AoCSI history $(\Delta_n(\tau))_{\tau=1}^{t}$; (iii) Finally, we show that the value function of the finite horizon MDP given the history $\mathcal{H}_n(t)$ is equal to the value function of the finite horizon MDP given $(C_n(t-\Delta_n(t)), \Delta_n(t))$.

Due to the Markov property and time homogeneity of the CSI $C_n(t)$, we have the following Markov chain given $\Delta_n(t)$: $C_n(t) \leftrightarrow C_n(t-\Delta_n(t)) \leftrightarrow C_n((t-1)-(\Delta_n(t-1)) \leftrightarrow \ldots$. Therefore, $C_n(t)$ is conditionally independent of $(C_n(\tau-\Delta_n(\tau)), \Delta_n(\tau))_{\tau=1}^{t-1}$ given $(C_n(t-\Delta_n(t)), \Delta_n(t))$.

Moreover, $C_n(t)$ depends on $(C_n(t-\Delta_n(t)), \Delta_n(t))$ and $(\beta_n(\tau))_{\tau=1}^{t-1}$. To compute $\Delta_n(\tau)$, we require scheduling decisions $(a_n(\tau))_{\tau=0}^{t-1}$ and $(\Delta_n(\tau))_{\tau=1}^{t-1}$. Hence, $C_n(t)$ depends on $(a_n(\tau))_{\tau=0}^{t-1}$ and $(\Delta_n(\tau))_{\tau=1}^{t-1}$ through $\Delta_n(t)$. However, if the AoCSI history $(\Delta_n(\tau))_{\tau=1}^{t}$ is given, then the CSI $C_n(t)$ does not depend on the scheduling decisions $(a_n(\tau))_{\tau=1}^{t-1}$ and the delivery indicators $(\beta_n(\tau))_{\tau=1}^{t-1}$. Hence, $C_n(t)$ is conditionally independent of the scheduling decisions $(a_n(\tau))_{\tau=0}^{t-1}$ and the delivery indicators $(\beta_n(\tau))_{\tau=0}^{t-1}$ given the AoCSI history $(\Delta_n(\tau))_{\tau=1}^{t}$.

Next, from (6), we can write

$$\mathbb{E}\left[\sum_{n=1}^{N} C_n(t)a_n(t)\middle|(\mathcal{H}_n(t))_{n=1}^{N}\right]$$
$$=\sum_{n=1}^{N}\mathbb{E}\left[C_n(t)a_n(t)\middle|\mathcal{H}_n(t)\right], \quad (42)$$

where (42) holds because given $(a_n(\tau))_{\tau=1}^{t-1}$, $(\beta_n(t))_{\tau=1}^{t-1}$ are independent across both users and time-slots and the CSI $\{C_n(t), t = 0, 1, 2, \ldots, T\}$ and $\{C_m(t), t = 0, 1, 2, \ldots, T\}$ are independent for all $n \neq m$.

Next, utilizing the fact that $C_n(t)$ is conditionally indepen-

dent of $\mathcal{H}_n(t)$ given $(C_n(t-\Delta_n(t)), \Delta_n(t))$, we get that

$$\begin{aligned}&\mathbb{E}\left[C_n(t)a_n(t)\middle|\mathcal{H}_n(t)\right]\\ =&\mathbb{E}\left[C_n(t)a_n(t)\middle|C_n(t-\Delta_n(t)), \Delta_n(t)\right].\end{aligned} \tag{43}$$

Therefore, the value function of the finite horizon MDP (6)-(7) under any given policy $\pi = (a_n(1), a_n(2), \ldots, a_n(T))_{n=1}^N$ can be written as

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{n=1}^{N}\mathbb{E}\left[C_n(t)a_n(t)\middle|(\mathcal{H}_n(t))_{n=1}^N\right]\\ =&\sum_{t=1}^{T}\sum_{n=1}^{N}\mathbb{E}\left[C_n(t)a_n(t)\middle|C_n(t-\Delta_n(t)), \Delta_n(t)\right],\end{aligned} \tag{44}$$

where (44) holds from (43). Therefore, from [12, Chapter 4.3], Theorem 1 follows.

REFERENCES

[1] W. Ouyang, A. Eryilmaz, and N. B. Shroff, "Asymptotically optimal downlink scheduling over Markovian fading channels," in *IEEE INFOCOM*, 2012, pp. 1224–1232.

[2] ——, "Low-complexity optimal scheduling over time-correlated fading channels with ARQ feedback," *IEEE Trans. Mob. Comput.*, vol. 15, no. 9, pp. 2275–2289, 2016.

[3] S. H. A. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, "Optimality of myopic sensing in multichannel opportunistic access," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 4040–4050, 2009.

[4] C.-p. Li and M. J. Neely, "Exploiting channel memory for multi-user wireless scheduling without channel measurement: Capacity regions and algorithms," in *ACM MobiHoc*, 2010, pp. 50–59.

[5] Q. Zhao, B. Krishnamachari, and K. Liu, "On myopic sensing for multi-channel opportunistic access: structure, optimality, and performance," *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 5431–5440, 2008.

[6] X. Liu, E. K. P. Chong, and N. B. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 10, pp. 2053–2064, 2002.

[7] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of Whittle index for dynamic multichannel access," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5547–5567, 2010.

[8] P. Whittle, "Restless bandits: activity allocation in a changing world," *Journal of Applied Probability*, vol. 25A, pp. 287–298, 1988.

[9] P. Ansell, K. Glazebrook, J. Niño-Mora, and M. O'Keeffe, "Whittle's index policy for a multi-class queueing system with convex holding costs," *Math. Meth. Oper. Res.*, vol. 57, pp. 21–39, 04 2003.

[10] Y. Chen and A. Ephremides, "Scheduling to minimize age of incorrect information with imperfect channel state information," *Entropy*, vol. 23, no. 12, p. 1572, 2021.

[11] G. Chen and S. C. Liew, "An index policy for minimizing the uncertainty-of-information of Markov sources," *arXiv preprint arXiv:2212.02752*, 2022.

[12] D. P. Bertsekas *et al.*, "Dynamic programming and optimal control, 4th edition," *Belmont, MA: Athena Scientific*, vol. 1, 2011.

[13] M. Costa, S. Valentin, and A. Ephremides, "On the age of channel state information for non-reciprocal wireless links," in *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2015, pp. 2356–2360.

[14] S. Wang, G. Xiong, and J. Li, "Online restless multi-armed bandits with long-term fairness constraints," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 14, 2024, pp. 15 616–15 624.

[15] K. Wang, L. Xu, A. Taneja, and M. Tambe, "Optimistic Whittle index policy: Online learning for restless bandits," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 8, 2023, pp. 10 131–10 139.

[16] F. Li, J. Liu, and B. Ji, "Combinatorial sleeping bandits with fairness constraints," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 3, pp. 1799–1813, 2019.

[17] M. K. C. Shisher, V. Tripathi, M. Chiang, and C. G. Brinton, "Online learning of Whittle indices for restless bandits with non-stationary transition kernels," 2025. [Online]. Available: https://arxiv.org/abs/2506.18186

[18] L. Tassiulas, "Scheduling and performance limits of networks with constantly changing topology," *IEEE transactions on information theory*, vol. 43, no. 3, pp. 1067–1073, 2002.

[19] Y. Sun and B. Cyr, "Sampling for data freshness optimization: Non-linear age functions," *J. Commun. Netw.*, vol. 21, no. 3, pp. 204–219, 2019.

[20] G. Chen, S. C. Liew, and Y. Shao, "Uncertainty-of-information scheduling: A restless multiarmed bandit framework," *IEEE Trans. Inf. Theory*, vol. 68, no. 9, pp. 6151–6173, 2022.

[21] M. K. C. Shisher, Y. Sun, and I.-H. Hou, "Timely communications for remote inference," *IEEE/ACM Transactions on Networking*, vol. 32, no. 5, pp. 3824–3839, 2024.

[22] T. Z. Ornee and Y. Sun, "A Whittle index policy for the remote estimation of multiple continuous Gauss-Markov processes over parallel channels," in *ACM MobiHoc*, 2023, p. 91–100.

[23] G. Xiong, X. Qin, B. Li, R. Singh, and J. Li, "Index-aware reinforcement learning for adaptive video streaming at the wireless edge," in *ACM MobiHoc*, 2022, pp. 81–90.

[24] I. Kadota, A. Sinha, E. Uysal-Biyikoglu, R. Singh, and E. Modiano, "Scheduling policies for minimizing age of information in broadcast wireless networks," *IEEE/ACM Trans. Netw.*, vol. 26, no. 6, pp. 2637–2650, 2018.

[25] T. Z. Ornee and Y. Sun, "Sampling and remote estimation for the Ornstein-Uhlenbeck process through queues: Age of information and beyond," *IEEE/ACM Trans. Netw.*, vol. 29, no. 5, p. 1962–1975, oct 2021.

[26] Y. Sun, Y. Polyanskiy, and E. Uysal, "Sampling of the Wiener process for remote estimation over a channel with random delay," *IEEE Trans. Inf. Theory*, vol. 66, no. 2, pp. 1118–1135, 2020.

[27] T. Z. Ornee, M. K. C. Shisher, C. Kam, and Y. Sun, "Remote safety monitoring: Significance-aware status updating for situational awareness," 2025. [Online]. Available: https://arxiv.org/abs/2507.09833

[28] M. Costa, S. Valentin, and A. Ephremides, "On the age of channel information for a finite-state Markov model," in *IEEE ICC*, 2015, pp. 4101–4106.

[29] K. T. Truong and R. W. Heath, "Effects of channel aging in massive MIMO systems," *Journal of Communications and Networks*, vol. 15, no. 4, pp. 338–351, 2013.

[30] M. Lipski, C. Kam, S. Kompella, and T. Ephremides, "Age of channel state information for collaborative beamforming," in *ACM MobiHoc*, 2024, pp. 392–397.

[31] S. Chakraborty and Y. Sun, "Send pilot or data? Leveraging age of channel state information for throughput maximization," *arXiv preprint arXiv:2503.13866*, 2025.

[32] R. R. Weber and G. Weiss, "On an index policy for restless bandits," *Journal of applied probability*, vol. 27, no. 3, pp. 637–648, 1990.

[33] Y. Zou, K. T. Kim, X. Lin, and M. Chiang, "Minimizing age-of-information in heterogeneous multi-channel systems: A new partial-index approach," in *ACM MobiHoc*, 2021, pp. 11–20.

[34] M. K. C. Shisher, B. Ji, I.-H. Hou, and Y. Sun, "Learning and communications co-design for remote inference systems: Feature length selection and transmission scheduling," *IEEE J. Sel. Areas Inf. Theory*, vol. 4, pp. 524–538, 2023.

[35] T. Z. Ornee, M. K. C. Shisher, C. Kam, and Y. Sun, "Context-aware status updating: Wireless scheduling for maximizing situational awareness in safety-critical systems," in *IEEE MILCOM*, 2023, pp. 194–200.

[36] M. K. C. Shisher, A. Piaseczny, Y. Sun, and C. G. Brinton, "Computation and communication co-scheduling for timely multi-task inference at the wireless edge," *IEEE INFOCOM*, 2025.

[37] G. Neu and G. Bartók, "An efficient algorithm for learning with semi-bandit feedback," in *International Conference on Algorithmic Learning Theory*. Springer, 2013, pp. 234–248.

[38] D. Foster and A. Rakhlin, "Beyond UCB: Optimal and efficient contextual bandits with regression oracles," in *International conference on machine learning*. PMLR, 2020, pp. 3199–3210.

[39] R. Ortner, D. Ryabko, P. Auer, and R. Munos, "Regret bounds for restless Markov bandits," in *International conference on algorithmic learning theory*. Springer, 2012, pp. 214–228.

[40] N. Akbarzadeh and A. Mahajan, "On learning Whittle index policy for restless bandits with scalable regret," *IEEE Transactions on Control of Network Systems*, vol. 11, no. 3, pp. 1190–1202, 2023.

[41] V. Krishnamurthy, *Partially observed Markov decision processes*. Cambridge university press, 2016.

[42] C. Papadimitriou and J. Tsitsiklis, "The complexity of optimal queueing network control," in *IEEE CCC*, 1994, pp. 318–322.

[43] I. M. Verloop, "Asymptotically optimal priority policies for indexable and nonindexable restless bandits," 2016.

[44] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. J. Weinberger, "Inequalities for the L_1 deviation of the empirical distribution," *Hewlett-Packard Labs, Tech. Rep*, p. 125, 2003.

[45] P. Ju, A. Ghosh, and N. B. Shroff, "Achieving fairness in multi-agent markov decision processes using reinforcement learning," *arXiv preprint arXiv:2306.00324*, 2023.

[46] V. Tripathi and E. Modiano, "An online learning approach to optimizing time-varying costs of aoi," 2021. [Online]. Available: https://arxiv.org/abs/2105.13383