Decentralized Scheduling with QoS Constraints: Achieving O(1) QoS Regret of Multi-Player Bandits

Qingsong Liu
Tsinghua University, Beijing, China

Zhixuan Fang*
Tsinghua University, Beijing, China

ABSTRACT

We consider a decentralized multi-player multi-armed bandit (MP-MAB) problem where players cannot observe the actions and rewards of other players and no explicit communication or coordination between players is possible. Prior studies mostly focus on maximizing the sum of rewards of the players over time. However, the total reward maximization learning may lead to imbalanced reward among players, leading to poor Quality of Service (QoS) for some players. In contrast, our objective is to let each player n achieve a predetermined average reward over time, i.e., achieving a predetermined level of QoS. We develop a novel decentralized algorithm to accomplish this objective by leveraging the methodology of randomized matching, which ensures that all players have an $O(1)$ QoS regret. We reveal an analog between our MP-MAB model and the online wireless queuing systems, which builds a connection between QoS in MP-MAB learning and stability in queuing theory.

ACM Reference Format:
Qingsong Liu and Zhixuan Fang. 2024. Decentralized Scheduling with QoS Constraints: Achieving O(1) QoS Regret of Multi-Player Bandits . In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

This section introduces our studied problem. We study a stochastic multi-player game played by a set of N players $\mathcal{N} = \{1, ..., N\}$ over a finite time horizon T. Each player cannot communicate with other players and faces with a common set of K arms denoted by $\mathcal{K} = \{1, ..., K\}$. We assume that $K \geq N$, since otherwise we can simply add dummy arms with 0 reward. At each round t, all players simultaneously pick one arm to play. We denote by $a_n(t)$ the arm that player n chooses at round t, and the action profile (vector of arms selected) at round t is $\boldsymbol{a}(t) \in [K]^N$. Players do not know which arms the other players chose, and need not even know the number of players N.

Rewards setting. We assume that, when multiple players choose the same arm, none of them can obtain a reward due to collision. We denote by $\eta_i(\boldsymbol{a})$ the no-collision indicator of arm i with respect to the action profile $\boldsymbol{a}$ such that $\eta_i(\boldsymbol{a}) = 0$ if $|\mathcal{N}_i(\boldsymbol{a})| > 1$, and $\eta_i(\boldsymbol{a}) = 1$ otherwise, where $\mathcal{N}_i(\boldsymbol{a}) = \{n \mid a_n = i\}$ is the set of players that chose arm i in action profile $\boldsymbol{a}$. Then for each player n, the instantaneous reward of hers at round t is $S_t^n = r_{n,a_n(t)}(t) \cdot \eta_{a_n(t)}(\boldsymbol{a}(t))$, where $r_{n,a_n(t)}(t)$ is a random reward that has a continuous distribution on $[0, 1]$. The reward sequence of arm i for player n, $\{r_{n,i}(t)\}_{t=1}^T$, is i.i.d. with an unknown expectation of $\mu_{n,i}$. We consider the heterogeneous setting where $\mu_{n,i}$ may not equal $\mu_{m,i}$ when $m \neq n$. An immediate example for the above collision reward model is wireless channel allocation, where the transmission of one user creates interference for other users on the same channel and causes all transmissions to fail.

Feedback setting. At each round t, each player n can observe her reward S_t^n together with the collision indicator $\eta_{a_n(t)}(\boldsymbol{a}(t))$. This makes sense in the context of cellular networks, as the transmitter can receive an ACK/NACK signal after each transmission, which can be used to determine if a collision has occurred.

Objective. Unlike most literature on decentralized MP-MAB [1–3, 7–10] that aims to maximize the total reward (See the latest survey [4] for reference), our goal is to let every player n achieve at least a target QoS value γ_n, i.e.,

$$\mathbb{E}[S_t^n] \geq \gamma_n, \ \forall t \in [T], \ n \in [N], \tag{1}$$

where the expectation is over the randomness of rewards and policy. We emphasize that our model is fully decentralized, i.e., every player cannot communicate with others and use extra information made by others to make her decisions, and players do not know each other's QoS values. Regarding the objective (1), we adopt QoS regret as our performance metric that defined as follows:

$$R(T) = \sum_{t=1}^{T} \max_n \left(\gamma_n - \mathbb{E}[S_t^n]\right)^+, \tag{2}$$

where $(x)^+$ denotes $\max\{x, 0\}$. A meaningful policy should produce at least sublinear QoS regret performance, i.e., $R(T)/T \to 0$, and would be ideal if $R(T) = O(1)$, i.e., *bounded* QoS regret. Of course, we cannot hope the bounded QoS regret is possible unless there exists a centralized algorithm that can make such a guarantee. We thus first understand what conditions the QoS requirements $\boldsymbol{\gamma}$ should satisfy for the players to guarantee their QoS requirements under centralized coordination. This motivates defining the capacity of our MP-MAB game as follows

$$\Delta = \max_{P \in \Phi} \min_n \Big(\sum_{i=1}^{K} P_{n,i} \cdot \mu_{n,i} - \gamma_n \Big) > 0. \tag{3}$$

If $\Delta < 0$, it means that no matter what the central controller's policy is, there exists at least one player whose QoS requirement γ_n is larger than her effective reward rate, and her QoS regret will grow over time. Indeed, even if $\Delta = 0$, i.e., there exists an n such that $\gamma_n = \sum_{i=1}^K P_{n,i}\mu_{n,i}$, the QoS regret still grows over time due to stochastic fluctuations. Hence, we require $\Delta > 0$ in our model.

1.1 The corresponding queuing systems

Here we reveal that our MP-MAB game behaves like an online queuing system in wireless networks. The system consists of N sources competing for K channels to transmit packets to a common Base Station (BS). At each (discrete) time $t = 0, 1, \ldots$, the following occurs: (a) A new data packet will arrive at the source n's queue with a fixed, time-independent probability γ_i. We model the arrival

*Corresponding author: Zhixuan Fang (zfang@mail.tsinghua.edu.cn).
The full paper has been accepted to AAAI 2024 [5].

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

event of time t as A_t^n and we have $\mathbb{P}(A_t^n = 1) = \gamma_n,\ \forall t$. (b) Each source n chooses one channel $a_n(t) \in [K]$ to transmit the first data packet in her queue. If the queue is empty, she would send a null/hello packet on her chosen channel (The BS will examine all received packets and discard null/hello packets). (c) Each channel i is unreliable and experiences i.i.d. ON-OFF channel fading. The probability that the channel i between source n and BS is ON is $\mu_{n,i}$ at any time. Here $\mu_{n,i}$ is heterogeneous w.r.t the source n, as the quality of each channel is often different for different sources in the cognitive radio context. When more than one source transmits the packet (including null/hello packet) on the same channel, their transmission would fail due to the interference. (d) If a data packet fails to transmit, the source would transmit it again in the next time until it succeeds. Each source not only receives feedback on whether her packet is transmitted successfully at her chosen channel, but also whether there exists other players choosing the same channel as she does via ACK/NACK signals, i.e., collision sensing.

We denote $Q^n(t)$ as the number of untransmitted data packets of source n at the beginning of time t (before new packet arrives). Formally, if S_t^n is the event indicator that source n clears a packet at time t and A_t^n is again the event indicator source n received a new packet at time t, then the dynamics of $Q^n(t)$ is as follows:

$$Q^n(t+1) = \max\{Q^n(t) + A_t^n - S_t^n, 0\},\ Q^n(0) = 0. \tag{4}$$

Since coordinating a large number of sources in a centralized manner is infeasible, decentralized scheduling policies are desirable in practice. Hence, a common objective is to design a fully-decentralized algorithm to guarantee the stability [6] of this queuing system. The following theorem reveals an analog between our MP-MAB model and this queuing system, bridging QoS in MP-MAB learning and stability in queuing theory.

Theorem 1. *For our MP-MAB problem, any algorithm that achieves sublinear QoS regret can also stabilize the corresponding queuing system, wherein all sources follow this algorithm by replacing the arm pulling with channel selecting.*

We remark that the opposite direction of Theorem 1 does not hold, i.e., the algorithm that achieves stability for the queuing system may not guarantee a sublinear QoS regret for our MP-MAB model. This helps strengthen the significance of our work.

2 ALGORITHMS AND MAIN RESULTS

This section presents the proposed decentralized algorithms, accompanied with its performance bounds. We design our decentralized algorithms by using randomized matching to allocate arms to players (a round-robin manner), which is a non-standard algorithm in MP-MAB literature. Here we remark that any randomized matching policy between players and arms can be characterized by a doubly stochastic matrix. We denote by $\mathbb{B}_K$ the set of doubly stochastic matrices that belongs to $[0,1]^{K\times K}$. To formalize our decentralized algorithm design, we need some definitions below.

Definition 1. *A **dominant mapping** is a function $\phi : \mathbb{R}^N \times \mathbb{R}^K \to \mathbb{B}_K$ which takes $(\boldsymbol{\gamma}, \boldsymbol{\mu})$ as input and returns a feasible doubly stochastic matrix P if it exists (and the identity matrix otherwise). We design the dominant mapping as follows,*

$$\phi(\boldsymbol{\gamma}, \boldsymbol{\mu}) = \arg\min_{P\in\mathbb{B}_K} \max_{n\in[N]} -\ln\Big(\sum_{i=1}^{K} P_{n,i}\cdot\mu_{n,i} - \gamma_n\Big) + \frac{1}{2K}\|P\|_2^2. \tag{5}$$

Our designed dominant mapping ensures that once the estimation error for $(\boldsymbol{\mu}, \boldsymbol{\gamma})$ is below a threshold, the returned doubly stochastic matrix $\tilde{P}$ can strictly satisfy the QoS requirements of all players, with a margin of order Δ.

Definition 2. *A **permutation matrix** $P \in [0,1]^{K\times K}$ is a square binary matrix that has exactly one entry of 1 in each row and each column and 0s elsewhere. Each such matrix represents an one-to-one matching between players and arms (where we pad with some virtual players since $K \le N$).*

Definition 3. *A **BvN** (Birkhoff von Neumann) **decomposition** is a function $\psi : \mathbb{B}_K \to \mathbb{P}(\mathbb{B}_K)$ that associates to any doubly stochastic matrix P a random variable $\psi(P)$ such that $\mathbb{E}[\psi(P)] = P$; stated otherwise, it expresses P as a convex combination of permutation matrices, i.e., there exist $\theta_1, ..., \theta_m \ge 0$, $\sum_{i=1}^m \theta_i = 1$, $m = (K-1)^2$ and permutation matrices $P_1, ..., P_m$ such that $\psi(P) = \sum_{i=1}^m \theta_i P_i$.*

Informally speaking, those definitions describe the policies players would follow in the decentralized case: a dominant mapping gives adequate marginals that ensure zero QoS regret (since each player n obtains in expectation a reward of $\sum_{i=1}^K P_{n,i}\cdot\mu_{n,i}$ at each round, which is larger than γ_n by definition). And a BvN decomposition describes the associated coupling to avoid collisions while maintaining marginals. Explicitly, given a common $\phi(\boldsymbol{\gamma}, \boldsymbol{\mu})$, the joint decentralized strategy for each player is to draw a **shared random variable** $\omega \in \mathbb{R}$ and then choose arms according to the permutation $\psi(\phi(\boldsymbol{\gamma}, \boldsymbol{\mu}))(\omega)$. We can verify that such strategy can ensure that players select arms in a collision-free round-robin manner while satisfying all players' QoS requirements.

Based on this intuition, our decentralized QoS guaranteeing algorithm, **AdeQoS**, proceeds in epochs comprising three phases: exploration, (implicit) communication, and consensus. The exploration phase allows players to obtain enough samples of each arm to estimate their expected rewards. During the communication phase, players attempt to infer information about the arm statistics and others' QoS values through forced collisions. After going through these two phases, players independently converge to an identical doubly stochastic matrix, obviating the need for a central entity. This allows for collision-free arm selection using BvN decomposition and shared randomness, making the decentralized problem resemble the centralized one. The remaining challenge is to verify if the current doubly stochastic matrix meets all players' QoS requirements without a central entity, which necessitates a consensus phase. The consensus phase allows the players, that no longer believe the current doubly stochastic matrix being played can satisfy her QoS requirement, signal other players to agree on a new one via forced collision. Due to the space limit, the complete pseudocode of AdeQoS and corresponding analysis are provided in the full paper [5]. Theorem 2 below provides the performance guarantees for AdeQoS.

Theorem 2. *For any $\Delta > 0$, if each player runs AdeQoS then the QoS regret is **bounded**:*

$$R(T) \le O(K^3 + N^3K^2 + N^4Ke^{16/\Delta^2}/\Delta^8).$$

Equipped with Theorem 1, Theorem 2 also implies that the AdeQoS can stabilize the corresponding queuing system.

REFERENCES

[1] Ilai Bistritz and Amir Leshem. Distributed multi-player bandits-a game of thrones approach. In *Advances in Neural Information Processing Systems*, pages 7222–7232, 2018.

[2] Etienne Boursier, Emilie Kaufmann, Abbas Mehrabian, and Vianney Perchet. A practical algorithm for multiplayer bandits when arm means vary among players. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, Palermo, Sicily, Italy, 2020.

[3] Etienne Boursier and Vianney Perchet. Sic-mmab: synchronisation involves communication in multiplayer multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 12071–12080, 2019.

[4] Etienne Boursier and Vianney Perchet. A survey on multi-player bandits. *arXiv preprint arXiv:2211.16275*, 2022.

[5] Qingsong Liu and Zhixuan Fang. Decentralized scheduling with qos constraints: Achieving o (1) qos regret of multi-player bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13981–13989, 2024.

[6] Michael Neely. *Stochastic network optimization with application to communication and queueing systems*. Springer Nature, 2022.

[7] Chengshuai Shi, Wei Xiong, Cong Shen, and Jing Yang. Heterogeneous multi-player multi-armed bandits: Closing the gap and generalization. *Advances in Neural Information Processing Systems*, 34:22392–22404, 2021.

[8] Harshvardhan Tibrewal, Sravan Patchala, Manjesh K Hanawal, and Sumit J Darak. Multiplayer multi-armed bandits for optimal assignment in heterogeneous networks. *arXiv preprint arXiv:1901.03868*, 2019.

[9] Po-An Wang, Alexandre Proutiere, Kaito Ariu, Yassir Jedra, and Alessio Russo. Optimal algorithms for multiplayer multi-armed bandits. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, Palermo, Sicily, Italy, 2020.

[10] Po-An Wang, Alexandre Proutiere, Kaito Ariu, Yassir Jedra, and Alessio Russo. Optimal algorithms for multiplayer multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 4120–4129. PMLR, 2020.