Beyond $\log^2(T)$ Regret for Decentralized Bandits in Matching Markets

Soumya Basu [1] **Karthik Abinav Sankararaman** [2] **Abishek Sankararaman** [3]

Abstract

We design decentralized algorithms for regret minimization in two sided matching markets with one-sided bandit feedback that significantly improves upon the prior works (Liu et al., 2020a; Sankararaman et al., 2021; Liu et al., 2020b). First, for general markets, for any $\varepsilon>0$, we design an algorithm that achieves a $O(\log^{1+\varepsilon}(T))$ regret to the agent-optimal stable matching, with unknown time horizon T, improving upon the $O(\log^2(T))$ regret achieved in (Liu et al., 2020b). Second, we provide the optimal $\Theta(\log(T))$ regret for markets satisfying *uniqueness consistency* – markets where leaving participants don't alter the original stable matching. Previously, $\Theta(\log(T))$ regret was achievable (Sankararaman et al., 2021; Liu et al., 2020b) in the much restricted *serial dictatorship* setting, when all arms have the same preference over the agents. We propose a phase based algorithm, where in each phase, besides deleting the globally communicated dominated arms, the agents locally delete arms with which they collide often. This *local deletion* is pivotal in breaking deadlocks arising from rank heterogeneity of agents across arms. We further demonstrate superiority of our algorithm over existing works through simulations.

1. Introduction

Decentralized decision making by competing agents under uncertainty, each one motivated by one's own objective, is a key feature in online market places, e.g. TaskRabbit, UpWork, DoorDash. An emerging line of research (Aridor et al., 2020; Johari et al., 2021; Liu et al., 2020a; Sankararaman et al., 2021; Liu et al., 2020b) in the field of multi-agent bandits is dedicated to understanding algorithmic principles in the interplay of competition, learning and regret minimization. The two-sided matching market (Gale & Shapley, 1962) is one such thread, where regret minimization is first studied in (Liu et al., 2020a) with a centralized arbiter, and in (Sankararaman et al., 2021; Liu et al., 2020b) at different levels of decentralization.

We study the fully decentralized two-sided matching market, comprising of N demand-side entities (a.k.a. agents) and K supply-side entities (a.k.a. arms). Each entity has a separate preference ranking of the opposite side agents, i.e. each agent over K arms, and each arm over N agents, and aims to match with the most preferred entity. In the bandit setting, the agents lack prior knowledge of their respective preferences, thus need to learn the preferences only through their *own past interactions*, while the arms know their preferences. In each round, every agent simultaneously chooses an arm of their choice, and are either *matched* to their arm of choice and receive a stochastic reward (reward defines agents' preference); or are *blocked*, are notified of this, and receive no reward.

A matching between agents and arms is *stable* if there exist no pair of agents and arms that are not matched with each other under the matching, prefer each other over their current partners. When each participant knows its preference and the agents propose to the arms across multiple rounds, the system admits any *stable matching* as a Nash equilibrium. Further, the *agent-optimal* stable matching is the one, which yields the highest reward among multiple possible stable matchings, to all agents (Gale & Shapley, 1962). Our objective is to design a uniform protocol for the agents, which allows each agent to quickly find and match with its agent-optimal arm, thus maximizing their cumulative reward.

Under a restrictive special case, known as serial dictatorship, where all arms have the same preference (Sankararaman et al., 2021) shows it is possible for each agent to attain $O(NK\log(T))$ cumulative regret (the gap between optimal and achieved reward) in T rounds. They rely on an Upper confidence bound (UCB) based algorithm, called UCB-D3, that uses strategic deletion of arms done through global communication alongside UCB based explore-exploit. When the decentralization is relaxed, and each agent observes the

[1]Google, Mountain View, CA, USA [2]Facebook, Menlo Park, CA, USA [3]Amazon, Palo Alto, CA, USA. Correspondence to: Soumya Basu <basusoumya@utexas.edu>, Karthik Abinav Sankararaman <karthikabinavs@gmail.com>, Abishek Sankararaman <abishek.90@gmail.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

response of all the arms per round, (Liu et al., 2020b) designs Collision Avoidance UCB (CA-UCB), which avoids collision by arm deletion and reduced switching of arms. This provides $O(\exp(N^4)\log^2(T))$ regret for each agent with respect to an agent-pessimal stable matching. Under agent-pessimal stable matching, all the agents obtain the minimum reward among all stable matching, thus it is less desirable than the agent-optimal stable matching. Moreover, the relaxation of decentralization comes at the cost of privacy, and the possible lack of truthfulness of agents which can hurt the performance, and discourage participation.

Our first contribution is to show through a simple phase based explore then commit protocol each agent can achieve $O(K\log^{1+\varepsilon}(T))$ regret for any $\varepsilon > 0$, for sufficiently large time horizon T. We further focus on two-sided markets that satisfy *uniqueness consistency* (`UnqC`) (Karpov, 2019), where the stable match is *robust* —namely if any subset of matched pairs of a stable match leave, the remaining matches are still stable. This also relaxes the stringent requirement of homogeneous preferences imposed by the serial dictatorship model. Arguably, robustness to matches departing the system and heterogeneity of preferences is a desirable property in large dynamic markets such as TaskRabbit, InstaCart etc. Under `UnqC`, where UCB-D3 fails due to heterogeneous arm ranking, we design UCB-D4 (UCB-D3 augmented with local deletions) that allows each agent to achieve $O(NK\log(T))$ regret. In local deletion, an agent aggressively and locally removes the arms to which it collides above a well designed threshold.

1.1. Main Contributions

1. General markets. The best existing regret bound for general stable matching due to (Liu et al., 2020b) obtains a $O(\exp(N^4)\log^2(T))$ *agent-pessimal* regret under *partial decentralized* feedback in time horizon T. The CA-UCB algorithm is used where agents switch with very low frequency to the arm with highest UCB index in a carefully chosen subset. The slow switching allows for collision avoidance but at a cost of high regret. Therefore, we design, phased ETC, a simple phase based algorithm, used uniformly by agents, which sets up a protocol that allows collision free exploration at the beginning of each phase. Then using the Gale-Shapley (Gale & Shapley, 1962) algorithm commits to a stable matching with their individual estimated preference lists. We prove this algorithm achieves a $O(K\log^{1+\varepsilon}(T))$ *agent-optimal* regret under *fully decentralized* feedback for any $\varepsilon > 0$, by setting the exploration duration based on ε. Although our proposed algorithm beats the state-of-the-art CA-UCB guarantees in many dimensions – fully vs partially decentralized, agent optimal vs pessimal regret, $O(K\log^{1+\varepsilon}(T))$ vs $O(\exp(N^4)\log^2(T))$ – it suffers from cold-start, i.e. it works for $T = \Omega(\exp(N/\Delta^2))$, where Δ is the minimum reward gap across all arms and agents. This leaves open the quest for an optimal $O(\log(T))$ regret without the curse of cold-start.

2. Markets with uniqueness consistency. We next focus on markets with uniqueness consistency (or `UnqC` in short), where there is a unique and robust stable matching. In this setting, the best known result (excluding serial dictatorship) is $O(\exp(N^4)\log^2(T))$, achieved by CA-UCB. For serial dictatorship, a special case of `UnqC` where all arms have the same preference order, the best result, $O((j-1)K\log(T))$ for agent ranked j for all arms, is obtained by UCB-D3. UCB-D3, a phase based algorithm however, cannot incur sub-linear regret when the preferences are heterogeneous (empirically shown in the Section 6). We introduce UCB-D4, a generalization of UCB-D3 that handles heterogeneous arm preferences under uniqueness consistency, and achieves the coveted $O(\log(T))$ regret for any $T = \Omega(N/\Delta^2)$, i.e. without the cold-start problem. Our key algorithmic and theoretical insights behind this result are as follows.

Algorithmic. UCB-D4 augments UCB-D3 with a *local deletion*, where an agent in each phase deletes an arm locally if it experiences collision more than a ($\beta\times$phase length) times, for an appropriate $\beta > 0$. The local deletion plays an important role in eliminating deadlocks that can be created under UCB-D3. In particular, consider the case when uniqueness consistency holds. Here arms have heterogeneous preferences, and agent a and agent b can block each other from exploring their non-stable matched arms. Interestingly, we discover that such deadlocks do not occur when playing the stable matched arms. Due to the specific nature of the deadlock, if an agent 'frequently' collides with an arm it is safe to delete that arm locally. The key is to carefully set the threshold of local deletion. A small threshold can remove the stable matched arm with constant probability due to the stochastic feedback, thus incur linear regret. Whereas, a large threshold deter the agent from getting out of the deadlock fast enough to achieve $O(\log(T))$ regret. UCB-D4 with $\beta < 1/K$ strikes the correct balance.

Theoretical. We introduce a dual induction proof-technique linked with α-condition. The α-condition bestows two orders: one among agents – left order, and one among arms – right order. In the left order, the agents have their stable matched arm as best arm once the stable matched arms for higher order agents are removed. For the right order same holds with the roles of agents and arms swapped. Our dual induction uses these two orders. We show local deletion ensures the arms inductively, following the right order, become 'available' for their respective stable matched agents. This allows the agents to inductively, following the left order, identify and broadcast their respective stable matched arm. The second induction, depends on first, and is driven by global deletion of dominated arms, alongside deadlock resolution due to local deletion.

3. Empirical. We compare the performance of our algorithms against CA-UCB, an algorithm for matching markets under *partial decentralization*: the response of all arms per round is visible. Our proposed phased ETC and UCB-D4 both work under *full decentralization*: only sees response of the proposed arm. Extensive simulations show, despite the restrictive feedback, phased ETC outperforms CA-UCB in general instances, while UCB-D4 does so under uniqueness consistency. We also empirically validate that UCB-D3 produces linear regret under uniqueness consistency.

1.2. Related Work

The field of bandit learning has a vast literature with multiple textbooks on this subject (Lattimore & Szepesvári, 2020; Bubeck & Cesa-Bianchi, 2012; Slivkins, 2019). Our work falls in the multi-agent bandits setting, which is effective in modelling applications like wireless networks (Avner & Mannor, 2016; Darak & Hanawal, 2019), online advertising (Hillel et al., 2013), data-centers and search scheduling (Sankararaman et al., 2019) where multiple decision makers interact causing an interesting interplay of competition, learning and consensus (Boursier & Perchet, 2020; Brânzei & Peres, 2019). The study of multi-agent bandits in two-sided matching markets is initiated by (Liu et al., 2020a). They study the centralized setting, when the agents pass on their estimated preference to a single decision maker who then proposes to the arms. In this setting, the authors design a UCB based protocol that completely avoids collision (due to centralized proposals) and attains a $O(NK\log(T))$ regret with respect to the agent-pessimal stable matching. The papers of (Sankararaman et al., 2021) and (Liu et al., 2020b) are the closest to our work as both of them studies decentralized (the latter only partially) two sided matching markets. We have already mentioned their relation to this work, and how we improve upon them. Two sided markets under full information, commonly known as the *stable marriage problem* was introduced in the seminal work of (Gale & Shapley, 1962) where they established the notion of stability and provided the optimal algorithm to obtain a stable matching. Our results rely on the recent combinatorial characterization, namely α-condition, of UnqC in stable marriage problem by (Karpov, 2019). Further, in the economics literature, uncertainty in two sided matching have been studied recently in (Johari et al., 2021; Ashlagi et al., 2019) in directions tangential to our work. We provide a detailed related work in Appendix B.

2. System Model

Agents and arms. We have N agents and $K \geq N$ arms. When agent $j \in [N]$ is matched to arm $k \in [K]$ it receives a reward sampled (independent of everything) from a latent distribution with support $[0,1]$, and mean $\mu_{jk} \in [0,1]$. We assume that $\{\mu_{jk}\}_{j,k}$ are all distinct. For each agent, the arm means impose a preference order over the arms, with higher means preferred over lower means. Similarly, every arm $k \in [K]$ has a total preference order over the arms $>_{\texttt{arm}(k)}$; for $j, j' \in [N]$, if $\texttt{agent}(j) >_{\texttt{arm}(k)} \texttt{agent}(j')$ then arm k prefers agent j over j'. When context is clear, we use j for $\texttt{agent}(j)$, and k for $\texttt{arm}(k)$. For any subset of agents and arms, the *preference profile* is the preference order of the agents restricted to the given set of arms, and vice versa. The game proceeds in T rounds (value of T unknown to the agents) where every agent simultaneously *plays* one arm, and is either matched to that arm or is notified that it was not matched. In every round, each arm is matched to the most preferred agent playing that arm in that round (if any).

Stable Matching. Given the preference order for agents and arms, consider a matching denoted by the set $\mathbf{k}^*, \mathbf{j}^*$, with k_j^* denoting the matched arm for agent j, and j_k^* denoting the matched agent for arm k. Under a *stable-matching* there exist no two pairs (j,k) and (j',k') such that $k = k_j^*$ and $k' = k_{j'}^*$, but $\mu_{jk} < \mu_{jk'}$ and $\texttt{agent}(j) >_{\texttt{arm}(k')} \texttt{agent}(j')$ – if agent j matches with k' both improve their position. An agent-optimal stable match is unique and is one where all agents obtain their respective best possible match among all possible stable matchings (see, (Gale & Shapley, 1962)).

Decentralized bandit with no information. All agents have common information, that we have N total agents and K arms each labeled 1 through K. In each round, every agent observes only the outcome of its action and cannot observe the other agents' play/outcome. Specifically, when it is rejected by the arm it observes the collision signal, otherwise it observes the reward obtained from the arm it proposed to. Thus, our feedback structure is *fully decentralized* where in each round, an agent's decision to play an arm to play can depend only on common information before the start, its past actions and outcomes. Agents can however agree to a common protocol that map their observed outcomes to arms in every round.

Agent optimal regret. Total reward obtained by any agent is compared against that obtained by playing the agent optimal stable match in all rounds. Let $P_j(t) \in [K]$ be the arm played by agent j in round t and $M_j(t) \in \{0,1\}$ be the indicator random variable denoting whether agent j was matched to $P_j(t)$. Let k_j^* be the agent-optimal stable match of agent j. The T-round individual regret for an agent is

$$R_j(T) = T\mu_{jk_j^*} - \mathbb{E}\left[\sum_{t=1}^{T} M_j(t)\mu_{jP_j(t)}\right].$$

The goal is to design a protocol that all agents follow to minimize their individual regret.

We make a few important remarks on the model.

1. Feedback structure. Our feedback structure is same

as (Sankararaman et al., 2021), and more restrictive than the one proposed very recently in (Liu et al., 2020b) called decentralized bandit with partial information (see also, Section 6). In the latter, all the agents can observe the matching between arms and agents in each round, alongside it's own reward or collision information. Further, the knowledge of preference order of the arms is explicitly assumed.

2. Why agent optimal regret? Our rationale for this is to compare against an oracle in which the arm-preferences of the agents are common information known to all agents, and each agent plays to maximize its individual total reward. This corresponds to a repeated game, in which, the set of pure strategy Nash equilibria corresponds to all agents playing a particular stable match in all rounds. The optimal equilibria that maximizes the reward for all agents simultaneously, corresponds to the agent-optimal stable matching.

3. Model generalizations. We focus on $N \leq K$, as for $N > K$ some agents will remain unmatched under full information, and have 0 regret by definition. Our algorithms, can be modified easily to handle such cases. Further, we can admit any sub-Gaussian reward distribution with finite mean in place of rewards supported on $[0, 1]$, with minor changes in our proof.

3. Achieving $O(\log^{(1+\varepsilon)}(T))$ Regret in Matching Markets for $\varepsilon > 0$

In this section, we give a simple Phased Explore then Commit (ETC) style algorithm that obtains an asymptotic per-agent regret of $O(\log^{1+\varepsilon}(T))$ for any given $\varepsilon > 0$. The algorithm proceeds in phases of exponentially growing lengths, with phase i lasting 2^i rounds. In each phase $i \geq 0$, all agents explore for the first i^ε rounds, and then subsequently exploit by converging to a stable matching through the Gale-Shapley strategy. The explorations are organized so that the agents avoid collision, which can be done in a simple decentralized fashion, where each agent gets assigned an unique id in the range $\{1, \cdots, N\}$ (Algorithm 3 in Appendix A). The pseudo-code for phased ETC is given in Algorithm 1.

The minimum reward gap across all arms and agents is $\Delta = \min\{|\mu_{jk} - \mu_{jk'}| : j \in [N], k, k' \in [K]\}$.

Theorem 1. *If every agent runs Algorithm 1 with input parameter $\varepsilon > 0$, then the regret of any agent j after T rounds satisfies*

$$R_T^{(j)} \leq \frac{K(\log_2(T)+2)^{1+\varepsilon}}{1+\varepsilon} + (N^2+K)(\log_2(T)+2) + 2^{\left((8/\Delta^2)^{1/\varepsilon} 4^{(1+\varepsilon)/\varepsilon}+1\right)} + \frac{e}{e-2}.$$

The proof is given in Appendix C. Several remarks are in order now on the regret upper bound of phased-ETC.

Algorithm 1 Phased ETC Algorithm

```
Index ← INDEX-ESTIMATION()
for N + 1 ≤ t ≤ T do
    i ← ⌊log2(t − 1)⌋
    if t − 2^i + 1 ≤ K⌊i^ε⌋ then
        Play arm (t + Index − 2^i + 1) mod K
    else
        Play GALE-SHAPLEY-MATCHING (Gale & Shapley, 1962) of Algorithm 5 with arm-preference ordered by empirical means computed using all the explore samples thus far.
    end if
end for
```

1. Comparison with CA-UCB. From an asymptotic viewpoint (when $T \to \infty$), our result improves upon the recent result of (Liu et al., 2020b). Their proposed algorithm CA-UCB, achieves a regret of $O(\log^2(T))$ with respect to the *agent-pessimal* stable matching. Theorem 1 shows that, even under our (restrictive) *decentralized bandit* model, the phased ETC algorithm achieves $O(\log^{1+\varepsilon}(T))$ regret with respect to the *agent-optimal* stable matching.

2. Exponential dependence on gap. The constant in the regret bound has an exponential dependence on Δ^{-2}, which limits it's applicability to practical values of T. We note that, the CA-UCB algorithm proposed in (Liu et al., 2020b), has an exponential dependence on N^4 that is multiplied with the $\log^2(T)$ term in their regret bound. It is an interesting open problem, to obtain an algorithm, that obtains $O(\log^{1+\varepsilon}(T))$ regret for the general matching bandit case, without any exponential dependence.

3. Comparison with Single Phase ETC. An ETC algorithm was proposed in (Liu et al., 2020a), which requires the knowledge of minimum reward gap (Δ) and the time horizon (T) to compute the necessary exploration. We remove the dependence on the reward gap and time horizon by our interleaved exploration and exploitation in the phased-ETC.

In the following section, we present UCB-D4 — UCB-D3 with Local Deletion, a decentralized algorithm that achieves regret scaling as $O\left(\frac{\log(T)}{\Delta^2}\right)$ without any exponential dependence on problem parameters, whenever the underlying system satisfies *uniqueness consistency*.

4. Achieving $O(\log(T))$ Regret under Uniqueness Consistency

We first introduce the uniqueness consistency in stable matching systems, and provide known combinatorial characterization of such systems, before presenting our algorithm.

Uniqueness Consistency. The uniqueness consistency is an

important subclass of graphs with unique stable-matching, and is defined as below after (Karpov, 2019).

Definition 1 ((Karpov, 2019))**.** *A preference profile satisfies* uniqueness consistency *iff (i) there exists a unique stable matching* $(\mathbf{k}^*, \mathbf{j}^*)$*; (ii) for any subset of arms or agents, the restriction of the preference profile on this subset with their stable-matched pair according to* $(\mathbf{k}^*, \mathbf{j}^*)$ *has a unique stable matching.*

The uniqueness consistency implies that even if an arbitrary subset of agent leave the system with their respective stable matched arms, we are left with a system with unique stable matching among the rest of agents and arms. This allows for any algorithm, which is able to identify at least one stable pair in a unique stable-matching system, to iteratively lead the system to the global unique stable matching.

Combinatorial characterization. In (Karpov, 2019), a necessary and sufficient condition for the uniqueness consistency of the stable matching system is established in the form of α-condition (defined shortly). However, in order to connect α-condition with the serial dictatorship studied in (Sankararaman et al., 2021), we find it instructive to first present the sequential preference condition (SPC) (Eeckhout, 2000), a generalization of serial dictatorship, before defining the α-condition in full generality. The definition of SPC (Eeckhout, 2000) (which is stated in terms of men and women preferences) restated in our system where we have agents and arms is as follows.

Let σ denote a pair of permutations of $[N]$ and $[K]$. Then $[K]_\sigma = \{a_1^{(\sigma)}, \ldots, a_K^{(\sigma)}\}$ and $[N]_\sigma = \{A_1^{(\sigma)}, \ldots, A_N^{(\sigma)}\}$ denote permutations of the ordered sets $[K]$ and $[N]$, respectively. The k-th arm in $[K]_\sigma$ is the $a_k^{(\sigma)}$-th arm in $[K]$, and the j-th agent in $[N]_\sigma$ is the $A_j^{(\sigma)}$-th agent in $[N]$.

Definition 2. Sequential preference condition *(SPC) is satisfied iff there is an order of agents and arms so that*

$$\forall j \in [N]_{\rm spc}, \forall k \in [K]_{\rm spc}, k > j, : \mu_{jj} > \mu_{jk}, \text{ and}$$
$$\forall k \leq N, \forall j \in [N]_{\rm spc}, j > k : \texttt{agent}\, j_k^* >_{\texttt{arm}\, k} \texttt{agent}\, j.$$

We next introduce a generalization of the SPC condition known as α-condition. This was first introduced in (Karpov, 2019) recently, and shown to be the *weakest sufficient condition* for a market to admit an unique stable matching. We again restate the definition from (Karpov, 2019) in our scenario.

Definition 3. *The* $\boldsymbol{\alpha}$***-condition** is satisfied iff there is a stable matching* $(\mathbf{k}^*, \mathbf{j}^*)$*, a left-order of agents and arms*

$$\text{s.t. } \forall j \in [N]_l, \forall k > j, k \in [K]_l : \mu_{jk_j^*} > \mu_{jk},$$

and a (possibly different) right-order of agents and arms

$$\text{s.t. } \forall k < j \leq N, a_k \in [K]_r, A_j \in [N]_r :$$
$$\texttt{agent}\, A_{j_{a_k}^*} >_{\texttt{arm}\, a_k} \texttt{agent}\, A_j.$$

The following theorem in (Karpov, 2019) provides the characterization of uniqueness consistency, where unacceptable mates are absent (an unacceptable pair (j, k) means agent j does not accept arm k, or the vice versa).[1]

Theorem 2 ((Karpov, 2019))**.** *If there is no unacceptable mates, then the* α*-condition is a necessary and sufficient condition for the uniqueness consistency.*

Without loss of generality, henceforth we consider the SPC order to be identity, i.e. $[N]_{\rm spc} = [N]$ and $[K]_{\rm spc} = [K]$. Again, without loss of generality we consider the left order in the α-condition to be identity, i.e. $[N]_l = [N]$ and $[K]_l = [K]$. Therefore, for the rest of the paper we only deal with arm order $a_k^{(r)} = a_k$ and agent order $A_j^{(r)} = A_j$, for arm $k \in [K]$ and agent $j \in [N]$.

We end with a few remarks on these three systems.

1. Serial Dictatorship, SPC, and* $\boldsymbol{\alpha}$*-condition. As mentioned earlier, the α-condition generalizes SPC. SPC is satisfied when the left order is identical to the right order in α-condition. Further, SPC generalizes serial dictatorship, as the unique rank of the agents in the latter and their respective stable matched arms creates an SPC order. We present examples of the three systems.

1. This system is Serial dictatorship, SPC, and α-condition

$$\texttt{agent} : a : 1 > 2 > 3, b : 1 > 2 > 3, c : 2 > 1 > 3,$$
$$\texttt{arm} : 1 : a > b > c, 2 : a > b > c, 3 : a > b > c.$$

2. This system is not Serial dictatorship as arms do not have a unique rank. But it satisfies SPC and α-condition, with a valid SPC order $\{(a, 1), (b, 2), (c, 3)\}$.

$$\texttt{agent} : a : 1 > 2 > 3, b : 1 > 2 > 3, c : 2 > 1 > 3,$$
$$\texttt{arm} : 1 : a > b > c, 2 : a > b > c, 3 : a > c > b.$$

3. The third system is not Serial dictatorship or SPC as there is no SPC order. But it satisfies α-condition as a valid left order is $\{(a, 1), (b, 2), (c, 3)\}$, and the corresponding right order is $\{(b, 2), (c, 3), (1, a)\}$.

$$\texttt{agent} : a : 1 > 2 > 3, b : 1 > 2 > 3, c : 2 > 1 > 3,$$
$$\texttt{arm} : 1 : b > c > a, 2 : b > a > c, 3 : b > c > a.$$

Currently, the α-condition is the weakest sufficient condition for uniqueness of stable matching. Necessary condition for the uniqueness of stable matching remains elusive and is a long standing open problem in combinatorics.

2. Stable matching under SPC and* $\boldsymbol{\alpha}$*-condition. The definition of SPC implies that the unique stable matching is

[1] (Karpov, 2019) state the theorem for a system with same number of men and women. However, the theorem readily extends to our system where we have K arms and N agents with $N \leq K$. Indeed, the unmatched $(N - K)$ arms do not influence the stable matching under any arbitrary restriction of the preference profile.

obtained when, under the SPC order, for every $i \leq N$, the agent i is matched to the arm i. Furthermore, under α-condition for any $j \in [N]$, the agent j is matched with arm j, and the agent A_j is matched with the arm a_j in the stable matching. We present a proof in the Appendix E.

4.1. UCB-D4: UCB Decentralized Dominate-Delete with Local Deletions

In this section, we describe our main algorithm. The algorithm applies a novel technique called local deletions that interleaves with the phased global deletion strategy of UCB-D3 algorithm proposed in (Sankararaman et al., 2021). Algorithm 2 describes this algorithm in detail.

Algorithm 2 UCB-D4 algorithm (for an agent j)

Input: Parameters $\beta \in (0, 1/K)$, and $\gamma > 1$
Set $\mathsf{Index}(j) \leftarrow$ `INDEX-ESTIMATION()`
Global deletion $G_j[0] = \phi,\ \forall j \in [N]$
for phase $i = 1, 2, \ldots$ **do**
  Reset the collision counters $C_{jk}[i] = 0, \forall k \in [K]$
  Delete dominated arms, $\mathcal{A}_j[i] \leftarrow [K] \setminus G_j[i-1]$
  if $t < 2^i + N + NK(i-1)$ **then**
    Local deletion $L_j[i] \leftarrow \{k : C_{jk}[i] \geq \lceil \beta 2^i \rceil\}$
    Play an arm $P_j(t) \in \arg\max_{k \in \mathcal{A}_j[i] \setminus L_j[i]} \left(\widehat{\mu}_{k,j}(t-1) + \sqrt{\frac{2\gamma \log(t)}{N_{k,j}(t-1)}} \right)$
    if Arm $k = P_j(t)$ is matched **then**
      Update estimate $\widehat{\mu}_{k,j}$, and matching count $N_{k,j}$
    else
      Increase collision counter $C_{jk}[i] \leftarrow C_{jk}[i] + 1$
    end if
  else if $t = 2^i + N + NK(i-1)$ **then**
    $\mathcal{O}_j[i] \leftarrow$ most matched arm so far in phase i
    $G_j[i] \leftarrow$ `COMMUNICATION(`i, $\mathcal{O}_j[i]$`)`
  end if
end for

The index of each agent j, namely $\mathsf{Index}(j)$, is set as the rank of this agent for the arm 1, which can be learned accurately in N rounds and in a distributed way (details in Algorithm 3 in Appendix A).

At a high-level, the algorithm proceeds in phases as follows. At each phase i, every agent j first updates its set of active arms for the phase given by $\mathcal{A}_j[i]$ by removing all the arms in the global dominated set $G_j[i-1]$. Then for the next 2^i time-steps, every agent plays the arm with the highest UCB from its active set $\mathcal{A}_j[i]$. Here, $\hat{\mu}_{jk}(t)$ denotes the estimated mean reward for, and $N_{jk}(t)$ the number of matches with, arm k by agent j at the end of round t. Whenever the agent collides with an arm at least $\lceil \beta 2^i \rceil$ times (tracked by collision counters $C_{jk}[i]$), it removes this arm from the active set.

Finally, in the last NK steps of this phase, each agent participates in a communication protocol (Algorithm 4 in Appendix A) to update the global dominated sets $G_j[i]$. In the rounds from $(K \times (\mathsf{Index}(j) - 1))$ to $(K \times \mathsf{Index}(j) - 1)$, agent j proposes to the K arms in round robin order, and adds the rejected arms to its globally dominated set $G_j[i]$. In the remaining rounds, it proposes to its most played arm in phase i, *i.e.,* arm $\mathcal{O}_j[i]$. We show that the globally dominated set $G_j[i] = \{\mathcal{O}_{j'}[i] : j' >_{\mathcal{O}_{j'}[i]} j\}$.

5. Main Results

We now present our main result in Theorem 3 which show that UCB-D4 attains near optimal logarithmic regret when the stable matching satisfies uniqueness consistency.

System parameters. We introduce the following definitions first that will be used in the regret upper bound:
1. The *blocking agents* for agent $j \in [N]$ and arm $k \in [K]$, $\mathcal{B}_{jk} := \{j' : \texttt{agent}(j') >_{\texttt{arm}(k)} \texttt{agent}(j)\}$.

2. The *dominated arms* for agent $j \in [N]$, $\mathcal{D}_j := \{k^*_{j'} : j' \leq j-1\}$.

3. The *blocked non-dominated arms* for agent $j \in [N]$, $\mathcal{H}_j = \{k : \exists j' \in \mathcal{B}_{jk} : k \notin \mathcal{D}_j \cup \mathcal{D}_{j'}\}$.

4. The *max blocking agent* for agent $j \in [N]$ $J_{\max}(j) = \max\left(j+1, \{j' : \exists k \in \mathcal{H}_j, j' \in \mathcal{B}_{jk}\}\right)$.

5. The *right-order mapping* for α-condition for agent $j \in [N]$ is $lr(j)$ so that $A_{lr(j)} = j$ with A_j as defined in Definition 3. We define the $lr_{\max}(j) = \max\{lr(j') : j' \leq j\}$.

6. The gaps of each agent j, is given as $\Delta_{jk} = (\mu_{jk^*_j} - \mu_{jk})$ which can be negative for some arms k. We define $\Delta_{\min} = \min\{\Delta_{jk} : j \in [N], k \in [K], \Delta_{jk} > 0\}$ as the minimum positive gap across arms and agents.

Some comments are due on some of the above definitions. The *dominated arms* is defined similar to (Sankararaman et al., 2021). These arms may have higher mean, hence higher long term UCB, than the agent's stable matched arm. If not removed, the agent will incur linear regret due to collision from these arms which are played by blocking agents in steady state. The *blocked non-dominated arms* are absent in serial dictatorship, but in the SPC and α-condition they emerge due to heterogeneity of arm preference orders. These arms are not necessarily removed during the global deletion (i.e. through the set $G_j[i]$ in UCB-D4), and may create a deadlock for an agent where the agent keeps playing these arms without exploring them due to collisions.

Regret bounds. We present the $O(\log(T))$ regret bound for the systems following uniqueness consistency in Theorem 3.

Theorem 3. *For a stable matching instance satisfying α-condition (Definition 3), suppose each agent follows* UCB-D4 *(Algorithm 2) with $\gamma > 1$ and $\beta \in (0, 1/K)$.*

Further, let

$$f_\alpha(j) = lr_{\max}(j) + j,\ i^* = \max\{8, i_1, i_2\},$$
$$i_1 = \min\{i : (N-1)\tfrac{10\gamma i}{\Delta^2_{min}} < \beta 2^{(i-1)}\},$$
$$\text{and } i_2 = \min\{i : (N - 1 + NK(i-1)) \le 2^{i+1}\}.$$

Then the regret for an agent $j \in [N]$ is upper bounded by

$$\mathbb{E}[R_j(T)] \le \underbrace{\sum_{k \notin \mathcal{D}_j \cup k^*_j} \tfrac{8\gamma}{\Delta_{jk}} \left(\log(T) + \sqrt{\tfrac{\pi}{\gamma}\log(T)}\right)}_{\textit{sub-optimal match}}$$
$$+ \underbrace{\sum_{k \notin \mathcal{D}_j} \sum_{j' \in \mathcal{B}_{jk} : k \notin \mathcal{D}_{j'}} \tfrac{8\gamma\mu_{k^*_j}}{\Delta^2_{j'k}} \left(\log(T) + \sqrt{\tfrac{\pi}{\gamma}\log(T)}\right)}_{\textit{collision}}$$
$$+ \underbrace{(K - 1 + |\mathcal{B}_{jk^*_j}|)\log_2(T)}_{\textit{communication}}$$
$$+ \underbrace{O\left(\tfrac{N^2K^2}{\Delta^2_{\min}} + (\beta|\mathcal{H}_j|f_\alpha(J_{\max}(j)) + f_\alpha(j) - 2)2^{i^*}\right)}_{\textit{transient phase, independent of } T}.$$

Furthermore, if the system satisfies SPC (Definition 2) the T independent term in the above regret bound can be replaced by $O\left(\frac{N^2K^2}{\Delta^2_{\min}} + (\beta|\mathcal{H}_j|J_{\max}(j) + j - 1)2^{i^*}\right)$.

We conclude with some remarks about our main result.

Scaling of regret bounds. Our regret bounds in Theorem 3 are desirable for the following reasons.
1. *Dependence on the arm gaps and horizon.* Our regret upper bound is near optimal in the arm gaps and horizon, as it matches with the regret lower bound of $\Omega\left(\frac{\log(T)}{\Delta^2_{\min}}\right)$ in Corollary 6 of (Sankararaman et al., 2021).
2. *Polynomial dependence.* The constant term associated with the regret bound is polynomial in all the system parameters, including the minimum gap, as

$$2^{i^*} = O\left(\max\left\{\tfrac{N}{\beta\Delta^2_{min}}\log(\tfrac{N}{\beta\Delta^2_{min}}), NK\log(NK)\right\}\right).$$

Hence, UCB-D4 regret bounds, under uniqueness consistency, has no exponential dependence on system parameters.
3. *Dependence on preference profile.* The preference profile influences the regret mainly in three ways. First, each agent deletes the dominated arms $\mathcal{D}_j$ so incurs no long term regret for those arms. The first term in the regret bound scales as $(K - j)$. Second, for any agent and any non-dominated arms (which are not globally deleted) the blocking agents during exploration creates collision leading to the second term scaling as at most $|\mathcal{H}_j \setminus \mathcal{D}_j|$. Finally, the constant in the regret bound scales linearly with the order of the agent and all of its blocking agents ($j + J_{\max}(j)$) in SPC. For the uniqueness consistency, the constant depends on both the left and right order ($f_\alpha(j) = j + lr_{\max}(j)$) of the agent and all of its blocking agents ($f_\alpha(j) + f_\alpha(J_{\max}(j))$).

Local deletion threshold. The local deletion threshold is set as $\beta \times$ phase length $+ \Theta(1)$ with $\beta < 1/K$. Increasing the threshold leads to higher regret until local deletion vanishes. This happens as more collision is allowed until an arm is deleted. But higher threshold allows for quick detection of the stable matched arm. However, decreasing the threshold leads to a more aggressive deletion leading to lower regret from collision per phase, at a cost of longer detection time for the stable matched arm. In particular, if instead we set local deletion threshold as $\Theta(\text{polylog}(\text{phase length}))$ then the constant 2^{i^*} becomes exponential, i.e. $\exp(N/\Delta^2_{\min})$. At one extreme, for $\beta \ge 1/K$, the global deletion may stop freezing as the stable matched arm for an agent is not guaranteed to emerge as the most matched arm. In the other extreme, for a threshold of $\Theta(\log(\text{phase length}))$ with large enough probability the stable matched arm may get locally deleted. In both extremes, the UCB-D4 ceases to work.

5.1. Key Insights into the Proof of Theorem 3

The full proof of Theorem 3 is presented in the Appendix E. We first present why UCB-D3 (Sankararaman et al., 2021) that works for the serial dictatorship setting fails under SPC.

Deadlocks beyond serial dictatorship. Under SPC while running UCB-D3 (Sankararaman et al., 2021) the *blocked non-dominated arms* cause deadlock where two agents are unable to sample their respective best stable matched arm leading to linear regret. This is best explained by an example. Let us consider the 3 agent (a, b, and c), and 3 arm (1, 2, and 3) system with the preference lists given as

$$\texttt{agent} : a : 1 > 2 > 3, b : 2 > 1 > 3, c : 3 > 1 > 2,$$
$$\texttt{arm} : 1 : a > b > c, 2 : a > b > c, 3 : a > c > b.$$

This system satisfies SPC with the agent optimal stable matching $\{(a,1),(b,2),(c,3)\}$, but it is not a serial dictatorship. If we run the UCB-D3 algorithm in (Sankararaman et al., 2021) then the agent a matches with arm 1,and agent b and c both delete arm 1 using global deletion. However, without additional coordination, with non negligible probability, agent b may not match with arm 3 (collision with agent c), and agent c may not match with arm 2 (collision with agent b). No further global deletion is guaranteed resulting in a deadlock, hence linear regret (see, Section 6).

We first focus on the proof for SPC condition before tackling α-condition, as the latter builds on the former.

Proof Sketch for SPC. There are two main components of the proof, *inductive freezing of stable matching pairs*, and *vanishing of local deletion*, both in expected constant time.

Inductive freezing of stable matching pairs. Local deletion

of arms with more than ($\beta \times$ phase length) collisions, for $\beta < 1/K$, ensures the agent ranked 1 (under SPC) removes all the *blocked non-dominated arms*, thus it never gets stuck in a deadlock. This gives agent 1 opportunity to detect arm 1 as its most matched arm in expected constant time, as in the presence of arm 1, agent 1 matches with any non-deleted sub-optimal arm $O(\log(\text{phase length}))$ times with high probability under UCB dynamics. Note that agent 1 never deletes arm 1 (either globally or locally) due to the SPC condition. Agent 2 next deletes dominated arm 1, and for similar reasons as above matches most with arm 2. The proof is completed using an induction over the agents following the SPC order, where in agent j freezing happens in $O(ji^*)$ phases.

Constant time vanishing of local deletion. We establish that the local deletion vanishes in expected constant time. Indeed, when all the agents settle to their respective stable matched arms, the sub-optimal play is limited to $O(log(\text{phase length}))$, hence total collision to blocked non-dominated arm is also logarithmic in phase length. This leads to the vanishing of local deletion, which requires ($\beta \times$ phase length) collisions, with an extra $O(\beta|\mathcal{H}_j|J_{\max}(j)2^{i^*})$ regret, where $J_{\max}(j)$ is the maximum blocking agent.

Stable regime. The terms that grow with T as $O(\log(T))$ are accounted for by keeping count of expected (1) number of sub-optimal matches, (2) number of collisions from blocking agents, and (3) the regret due to communication sub-phases. □

Uniqueness consistency - A tale of two orders. When moving from SPC instances to uniqueness consistency (equivalently α-condition due to Theorem 2) we no longer have the simple inductive structure that we leverage in the proof sketch of SPC. Under α-condition we have two orders instead. The left order states when arm 1 to arm $(j-1)$ are removed agent j has arm j as the most preferred arm. Whereas, the right order states when agent A_1 to agent $A_{(k-1)}$ (recall $\{A_1, \ldots, A_N\}$ is a separate permutation of the agents) are removed arm a_k prefers agent A_k. Global deletion here must follow left order. However, unlike SPC the set of blocking agents for agent 1 and arm 1 (i.e. $\mathcal{B}_{11}$) is nonempty. Thus, agent 1 cannot get matched with arm 1 majority of time unless the agents in $\mathcal{B}_{11}$ stop playing arm 1. We next show how this is resolved.

Proof Sketch of α-condition. We show due to local deletion an *inductive warmup of stable matching pairs* precedes the two phases mentioned in the proof sketch of SPC.

Inductive warmup of stable matching pairs. We begin with an observation that holds for every stable matching: for any two stable matching pair (j,k), (j',k'), the arm k is either sub-optimal (has mean reward lesser than k') for agent j', or arm k prefers agent j over agent j'. Hence, in our system, for any $2 \le j \le N$, either (a) arm a_j (stable matched arm for agent A_j) is suboptimal for agent A_1, or (b) arm a_j prefers A_j over A_1. In case (b) arm A_1 never causes collision in a_j for agent A_j. Now suppose case (a) holds. Due to α-condition, the agent A_1 is the most preferred for arm a_1. Therefore, arm a_1 is always available to agent A_1, and a_1 has higher mean than a_j. This implies the UCB algorithm plays arm a_j only $O(\log(\text{phase length}))$ times with high probability in any phase. Once this happens, arm a_2 is almost always available to A_2, and the induction sets in. This is used to prove A_j is warmed up in $O(ji^*)$ phases, or agent j is warmed up in $O(lr(j)i^*)$ phases. Once agent 1 is warmed up the inductive freezing starts as it matches with arm 1. The rest closely follows the proof of SPC. □

6. Numerical Simulations

In this section, we present our numerical simulations with 5 agents and 6 arms. Additional results with larger instances are deferred to the Appendix F.

Baselines. We use two baselines with their own feedback.
1. *Centralized UCB (UCB-C)* is proposed in (Liu et al., 2020a) for the centralized feedback setting, where each agent in every round submits it's preference order (based on UCB indices) to a centralized agent who assigns the agent optimal matching under this preference, and the rewards are observed locally.
2. *Collision Avoidance UCB (CA-UCB)* is proposed in (Liu et al., 2020b) for the *decentralized with partial information* setting where at each round all agents observe the player matched to each arm, but rewards are observed locally.

The centralized setting is the most relaxed where no collision happens, followed by the partial decentralized setting where each round the matching of arms can be learned without any collision. Both are relaxed compared to our *decentralized with no information* setting where any global information can be obtained only through collisions.

Results. We generate random instances to compare the performance of UCB-C, CA-UCB and UCB-D4 in their respective settings. Since (Liu et al., 2020b) does not mention how to set their hyper-parameter λ, we report the best result by running a grid search over λ. For UCB-D4, we use $\beta = 1/2K$ and $\gamma = 2$. We simulate all the algorithms on the same sample paths, for a total 50 sample paths and report mean, 75% and 25% agent-optimal regret.

Figure 1 shows that in a general instance phased ETC outperforms CA-UCB even with a restricted feedback, whereas, as expected, UCB-C outperforms phased ETC. Figure 2 shows that when uniqueness condition holds UCB-D4, despite the restricted feedback, outperforms CA-UCB, while it is comparable to the centralized UCB-C.

Linear Regret of UCB-D3 in SPC. Figure 3 shows that even when the SPC condition holds, UCB-D3 may result in linear regret, emphasizing the importance of UCB-D4.

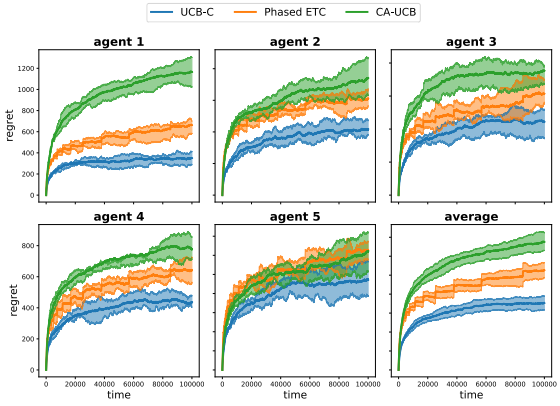

Figure 1. A general instance with $N = 5$, and $K = 6$. Regret: UCB-C (blue) $<$ Phased-ETC (orange) $<$ CA-UCB (green).

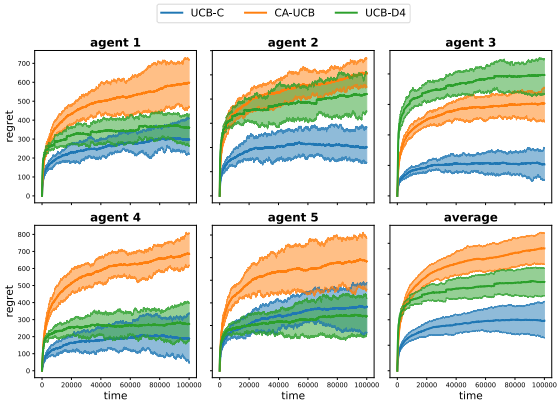

Figure 2. An α-condition instance with $N = 5$, and $K = 6$. Regret: UCB-C (blue) $<$ UCB-D4 (green) $<$ CA-UCB (orange).

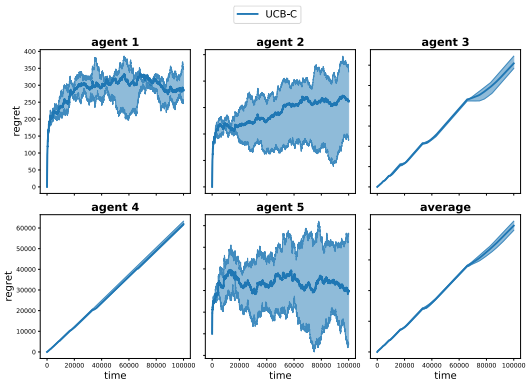

Figure 3. Linear regret of UCB-D3 for an SPC instance.

References

Aridor, G., Mansour, Y., Slivkins, A., and Wu, Z. S. Competing bandits: The perils of exploration under competition. *arXiv preprint arXiv:2007.10144*, 2020.

Ashlagi, I., Krishnaswamy, A. K., Makhijani, R., Saban, D., and Shiragur, K. Assortment planning for two-sided sequential matching markets. *arXiv preprint arXiv:1907.04485*, 2019.

Avner, O. and Mannor, S. Multi-user lax communications: a multi-armed bandit approach. In *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*, pp. 1–9. IEEE, 2016.

Bistritz, I. and Leshem, A. Game of thrones: Fully distributed learning for multiplayer bandits. *Mathematics of Operations Research*, 2020.

Boursier, E. and Perchet, V. SIC-MMAB: synchronisation involves communication in multiplayer multi-armed bandits. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 12048–12057, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c4127b9194fe8562c64dc0f5bf2c93bc-Abstract.html.

Boursier, E. and Perchet, V. Selfish robustness and equilibria in multi-player bandits. In Abernethy, J. D. and Agarwal, S. (eds.), *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pp. 530–581. PMLR, 2020. URL http://proceedings.mlr.press/v125/boursier20a.html.

Brânzei, S. and Peres, Y. Multiplayer bandit learning, from competition to cooperation. *arXiv preprint arXiv:1908.01135*, 2019.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations & Trends in Machine Learning*, 2012.

Buccapatnam, S., Tan, J., and Zhang, L. Information sharing in distributed stochastic bandits. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pp. 2605–2613. IEEE, 2015.

Chawla, R., Sankararaman, A., Ganesh, A., and Shakkottai, S. The gossiping insert-eliminate algorithm for multi-agent bandits. In Chiappa, S. and Calandra, R. (eds.), *The*

23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy], volume 108 of *Proceedings of Machine Learning Research*, pp. 3471–3481. PMLR, 2020. URL http://proceedings.mlr.press/v108/chawla20a.html.

Dai, X. and Jordan, M. I. Learning strategies in decentralized matching markets under uncertain preferences. *arXiv preprint arXiv:2011.00159*, 2020.

Darak, S. J. and Hanawal, M. K. Multi-player multi-armed bandits for stable allocation in heterogeneous ad-hoc networks. *IEEE Journal on Selected Areas in Communications*, 37(10):2350–2363, 2019.

Dubey, A. and Pentland, A. S. Cooperative multi-agent bandits with heavy tails. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2730–2739. PMLR, 2020. URL http://proceedings.mlr.press/v119/dubey20a.html.

Eeckhout, J. On the uniqueness of stable marriage matchings. *Economics Letters*, 69(1):1–8, 2000.

Gale, D. and Shapley, L. S. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

Hillel, E., Karnin, Z. S., Koren, T., Lempel, R., and Somekh, O. Distributed exploration in multi-armed bandits. In Burges, C. J. C., Bottou, L., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 854–862, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/598b3e71ec378bd83e0a727608b5db01-Abstract.html.

Johari, R., Kamble, V., and Kanoria, Y. Matching while learning. *Operations Research*, 2021.

Kalathil, D., Nayyar, N., and Jain, R. Decentralized learning for multiplayer multiarmed bandits. *IEEE Transactions on Information Theory*, 60(4):2331–2345, 2014.

Karpov, A. A necessary and sufficient condition for uniqueness consistency in the stable marriage matching problem. *Economics Letters*, 178:63–65, 2019.

Kolla, R. K., Jagannathan, K., and Gopalan, A. Collaborative learning of stochastic bandits over a social network. *IEEE/ACM Transactions on Networking*, 26(4): 1782–1795, 2018.

Landgren, P., Srivastava, V., and Leonard, N. E. Distributed cooperative decision making in multi-agent multi-armed bandits. *Automatica*, 125:109445, 2021.

Larrnaaga, M., Ayesta, U., and Verloop, I. M. Dynamic control of birth-and-death restless bandits: Application to resource-allocation problems. *IEEE/ACM Transactions on Networking*, 24(6):3812–3825, 2016.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Liu, L. T., Mania, H., and Jordan, M. I. Competing bandits in matching markets. In Chiappa, S. and Calandra, R. (eds.), *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1618–1628. PMLR, 2020a. URL http://proceedings.mlr.press/v108/liu20c.html.

Liu, L. T., Ruan, F., Mania, H., and Jordan, M. I. Bandit learning in decentralized matching markets. *arXiv preprint arXiv:2012.07348*, 2020b.

Mehrabian, A., Boursier, E., Kaufmann, E., and Perchet, V. A practical algorithm for multiplayer bandits when arm means vary among players. In Chiappa, S. and Calandra, R. (eds.), *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1211–1221. PMLR, 2020. URL http://proceedings.mlr.press/v108/mehrabian20a.html.

Rosenski, J., Shamir, O., and Szlak, L. Multi-player bandits - a musical chairs approach. In Balcan, M. and Weinberger, K. Q. (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 155–163. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/rosenski16.html.

Sankararaman, A., Ganesh, A., and Shakkottai, S. Social learning in multi agent multi armed bandits. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(3):1–35, 2019.

Sankararaman, A., Basu, S., and Sankararaman, K. A. Dominate or delete: Decentralized competing bandits in serial dictatorship. In Banerjee, A. and Fukumizu, K. (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings*

of Machine Learning Research, pp. 1252–1260. PMLR, 2021. URL http://proceedings.mlr.press/v130/sankararaman21a.html.

Slivkins, A. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.