# A Sub-Routines used in Algorithm 1

Here, we provide the pseudo code, related to initial rank estimation and the communication protocol used in Algorithm 1.

**Algorithm 2** RANK-ESTIMATION (at agent $j$)

**Initialization**: Rank$\leftarrow N$, Flag$\leftarrow$ **FALSE**
**for** $1 \leq t \leq N-1$ **do** ▷ Rank Estimation
  **if** $t == 1$ **OR** Flag$==$ **False then**
    $I_j(t) = t$ ▷ Play arm $t$ at time $t$
    **if** Matched at time $t$, i.e., $M_t(t) = j$ **then**
      Rank$\leftarrow t$, Flag$\leftarrow$ **TRUE**
    **end if**
  **else**
    $I_j(t) =$Rank
  **end if**
**end for**
**return** Rank

**Algorithm 3** DOMINATED-ARM-DETECTION ($\mathcal{O}, i$, Rank) (at agent $j$)

**Input** $\mathcal{O} \in [K]$ - Arm to communicate, $i \in \mathbb{N}$ - the phase and Rank - the rank of the agent
$\mathcal{C} \leftarrow \emptyset$
$\widehat{S}_i \leftarrow S_i + 2^{i-1} + 1$
**for** $\widehat{S}_i \leq t < \widehat{S}_i + K \max(0,$Rank-2$)$ **do** ▷ The first Rank-2 sub-blocks in the Communication block of phase $i$
  $I^{(j)}(t) = \mathcal{O}$ ▷ Play the most matched arm, which is input to this sub-routine
**end for**
**for** $\widehat{S}_i + K \max(0,$Rank-2$) < t \leq \widehat{S}_i + K(\max(0,$Rank-1$))$ **do** ▷ The Rank$-1^{th}$ sub-block in the Communication block of phase $i$
  $I^{(j)}(t) = ((t - (\widehat{S}_i + K \max(0,$Rank-2$))) \mod K) + 1$ ▷ Play arms in round robin
  **if** Collision Occurs **then** ▷ Collision in (Rank-1)th sub-block
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{I^{(j)}(t)\}$ ▷ Update set of arms to delete
  **end if**
**end for**
**return** $\mathcal{C}$

Observe that in the above algorithm, agent ranked 1, will only play its best arm, in all rounds of the communication block.

# B Analysis of the Algorithm and Proof of Theorem 1

## B.1 Overall Proof Architecture

The proof of Theorem 1 follows by plugging in the estimates from Corollary 16, into Corollary 11. The subject matter in Appendix B.4 is to prove Corollary 11 and the subject matter in Appendix B.5 is to prove Corollary 16. All the notations and definitions needed for the proof are collected in Appendix B.2.

## B.2 Notation and Definitions needed for the Proof

In order to implement the proof, we specify certain notations and definitions. For every $i \in \mathbb{N}$, denote by $S_i := N + (2^{i-1} - 1) + (i-1)NK$, to be the first time slot in the regret minimization block of phase $i$. For any phase $i \in \mathbb{N}$, agent $j \in [N]$ and arm $k \in [K]$, denote by $\widetilde{N}_i^{(j)}[k]$ to be the number of times agent $j$ was matched to arm $k$ in phase $i$. Recall the notation that for all agents $j \in [N]$, its stable match partner arm was denoted as $k_*^{(j)} \in [K]$. Similarly the set of dominated arms for any agent $j \in \{2, \cdots, N\}$ we defined as $\mathcal{D}_j^* := \{k_*^{(1)}, \cdots, k_*^{(j-1)}\}$. Recall that we had set $\mathcal{A}_*^{(j)} := [K] \setminus \mathcal{D}_*^{(j)}$. For any arm $k \in [K]$ and agent $j \in [N]$,

denote by $\Delta_k^{(j)} := \mu_{jk_*^{(j)}} - \mu_{jk}$.

Our first definition is whether a given phase is good for a particular agent or not. We call a phase $i \in \mathbb{N}$ **Good for Agent** $j$ if

1. $\mathcal{A}_i^{(j)} = \mathcal{A}_j^*$, i.e., $\mathcal{A}_i^{(j)} = [K] \setminus \{k_*^{(1)}, \cdots, k_*^{(j-1)}\}$.
2. The number of times each arm $k \in \mathcal{A}_*^{(j)} \setminus \{k_*^{(j)}\}$ is matched to agent $j$ in the regret minimization block of phase $i$ is less than or equal to $\frac{10\alpha i}{(\Delta_k^{(j)})^2}$. Note that by definition of $\Delta_k^{(j)}$ and the fact that arm-means are unique for an agent, for every agent $j \in [N]$ and arm $k \in \mathcal{A}_*^{(j)} \setminus \{k_*^{(j)}\}$, $\Delta_k^{(j)} > 0$.
3. The arm that is most matched in the regret minimization block of phase $i$ is $k_*^{(j)}$.

We denote by the event $\chi_i^{(j)}$ to be the indicator random variable, i.e.,

$$\chi_i^{(j)} = \mathbf{1}_{\text{Phase } i \text{ is } \textbf{Good} \text{ for Agent } j}.$$

For every agent $j \in [N]$, denote by the random time $\tau^{(j)}$ to be the first phase index, such that all phases larger than $\tau^{(j)}$ is Good for agent $j$. Formally,

$$\tau^{(j)} := \inf \left\{ i \in \mathbb{N} : \left( \prod_{l \geq i} \chi_l^{(j)} \right) = 1 \right\},$$
$$\widetilde{\tau}^{(j)} := \max(\tau^{(1)}, \cdots, \tau^{(j)}).$$

Notice that after phase $i \geq \widetilde{\tau}^{(j)}$, for all agents $j' \leq j$, $\mathcal{A}_i^{(j')} = \mathcal{A}_*^{(j')}$. In other words, the set of active arms of all agents ranked $j$ and lower are 'frozen' after phase $\widetilde{\tau}^{(j)}$ to the 'correct' set of arms.

We now, describe certain set of events. For any agent $j \in [N]$ and arm $k \in [K] \setminus \{k_*^{(1)}, \cdots, k_*^{(j)}\}$, denote by the event $\mathcal{E}_k^{(j)}$ as

$$\mathcal{E}_k^{(j)} := \left\{ N_k^{(j)}(T) - N_k^{(j)}(S_{\widetilde{\tau}^{(j)}}) \geq \left\lceil \frac{4\alpha \log(T)}{(\Delta_k^{(j)})^2} \right\rceil \right\} \cap \{\widetilde{\tau}^{(N)} < \infty\}. \tag{2}$$

Denote by the event $\mathcal{E}$ as the union, i.e.,

$$\mathcal{E} := \bigcap_{j=1}^{N} \bigcap_{k \in [K] \setminus \{k_*^{(1)}, \cdots, k_*^{(j)}\}} \mathcal{E}_k^{(j)}. \tag{3}$$

Recall that we had defined $\Delta$ to be the smallest arm-gap, namely

$$\Delta := \min_{j \in [N]} \min_{k \in \mathcal{A}_*^{(j)} \setminus \{k_*^{(j)}\}} \Delta_k^{(j)},$$

and that $i^* \in \mathbb{N}$ was defined as

$$i^* := \min \left\{ i \in \mathbb{N} : \frac{20NK\alpha i}{\Delta^2} \leq 2^{i-1} \right\}. \tag{4}$$

### B.3 Technical Preliminaries

In order to be precise in our calculations, we will need to explicitly specify a probability space. Let $\mathbb{Y} \coloneqq (Y_k^{(j)}(t))_{1\le k\le K, 1\le j\le N, t\ge 1}$, be a family of iid random variables, with each being defined uniformly in the interval $[0,1]$. The interpretation being that when agent $j$, gets matched with arm $k$, for the $t$th time, it receives a binary reward equal to $\mathbf{1}(Y_k^{(j)}(t) \le \mu_k^{(j)})$. Thus, the dynamics of the algorithm can be constructed as a deterministic (measurable) function of the family of random variables $\mathbb{Y}$.

### B.4 Regret Decomposition

**Lemma 9.** *The regret of any agent $j \in [N]$, at time $T$ can be decomposed as*

$$\mathbb{E}\left[R_T^{(j)}\right] \le \mathbb{E}[S_{\widetilde{\tau}(j)}] + \underbrace{2(j-1)(\log_2(T)+1)}_{\textit{Regret due to Communication}} + \underbrace{\mathbb{E}\left[\sum_{j'=1}^{j-1}\sum_{k\in\mathcal{A}_j^*}(N_k^{(j')}(T) - N_k^{(j')}(S_{\widetilde{\tau}(j)}))\mu_{jk_*^{(j)}}\right]}_{\textit{Regret due to Collision}} + \underbrace{\mathbb{E}\left[\sum_{k\in\mathcal{A}_j^*\setminus\{k_*^{(j)}\}}(N_k^{(j)}(T) - N_k^{(j)}(S_{\widetilde{\tau}(j)}))\Delta_k^{(j)}\right]}_{\textit{Regret due to Sub-optimal arm pull}}.$$

*Proof.* From the definition of regret, we have

$$\begin{aligned}
\mathbb{E}[R_T(j)] &= \mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}_{I^{(j)}(t)\neq k_*^{(j)}}(\mu_{jI^{(j)}(t)} - \mu_{jk_*^{(j)}})\right],\\
&= \mathbb{E}\left[\sum_{t=1}^{S_{\widetilde{\tau}(j)}}\mathbf{1}_{I^{(j)}(t)\neq k_*^{(j)}}(\mu_{jI^{(j)}(t)} - \mu_{jk_*^{(j)}})\right] + \mathbb{E}\left[\sum_{t=S_{\widetilde{\tau}(j)}}^{T}\mathbf{1}_{I^{(j)}(t)\neq k_*^{(j)}+1}(\mu_{jI^{(j)}(t)} - \mu_{jk_*^{(j)}})\right],\\
&\le \mathbb{E}[S_{\widetilde{\tau}(j)}] + \mathbb{E}\left[\sum_{t=S_{\widetilde{\tau}(j)}}^{T}\mathbf{1}_{I^{(j)}(t)\neq k_*^{(j)}}(\mu_{jI^{(j)}(t)} - \mu_{jk_*^{(j)}})\right].
\end{aligned}$$

The inequality follows from the assumption that, for all $j \in [N]$ and arm $k \in [K]$, $\mu_{jk} \in [0,1]$. Now, we decompose the second term as follows

$$\mathbb{E}\left[\sum_{t=S_{\widetilde{\tau}(j)}}^{T}\mathbf{1}_{I^{(j)}(t)\neq k_*^{(j)}}(\mu_{jI^{(j)}(t)} - \mu_{jk_*^{(j)}})\right] \le \underbrace{2(j-1)\mathbb{E}\left[\sum_{i\ge\widetilde{\tau}^{(j)}}\mathbf{1}_{2^i\le T}\right]}_{\text{Communication}} + \underbrace{\mathbb{E}\left[\sum_{j'=1}^{j-1}\sum_{k\in\mathcal{A}_j^*}(N_k^{(j')}(T) - N_k^{(j')}(S_{\widetilde{\tau}(j)}))\mu_{jk_*^{(j)}}\right]}_{\text{Collision}} + \underbrace{\mathbb{E}\left[\sum_{k\in\mathcal{A}_j^*\setminus\{k_*^{(j)}\}}(N_k^{(j)}(T) - N_k^{(j)}(S_{\widetilde{\tau}(j)}))\Delta_k^{(j)}\right]}_{\text{Sub-optimal arm pull}}.$$

This in-equality follows from the following facts

- During communication, agent ranked $j$ will face exactly $j-1$ collisions, in its round-robin arm search. Additionally, during each of the top-ranked $j-1$ agent's round robin arm-search, agent $j$ experiences one collision.
- Agent $j$ experiences a collision at an arm $k$, if and only if, exactly one agent ranked 1 through $j-1$ is matched to arm $k$, at the same time. This then gives that the total upper bound on the number of collisions is the number of times agents 1 through $j-1$, are matched to any arm in $\mathcal{A}_*^{(j)}$. This gives the regret due to collisions.

- Finally, each sub-optimal arm match incurs regret, which is captured in the last term.

□

It thus remains to bound the expected number of times, any agent $j^{'} \in [N]$, plays arm $k \in \mathcal{A}_*^{(j^{'})}$.

**Proposition 10.** *For any agent $j \in [N]$, arm $k \in \mathcal{A}_*^{(j)} \setminus \{k_*^{(j)}\}$,*

$$\mathbb{E}[N_k^{(j)}(T) - N_k^{(j)}(S_{\widetilde{\tau}^{(j)}})] \leq \frac{32\alpha \log(T)}{(\Delta_k^{(j)})^2} + 1 + \frac{8T^{-4\alpha}}{(\Delta_k^{(j)})^2} + T^{2-4\alpha}.$$

*Proof.* For any agent $j \in [N]$, arm $k \in [K]$, and time $t \geq 2$, denote by $\mathrm{UCB}_k^{(j)}(t) = \widehat{\mu}_{jk}(t-1) + \sqrt{\frac{2\alpha \log(T)}{N_k^{(j)}(t-1)}}$, to be the UCB index of arm $k$, at time $t$, by agent $j$. Let $\varepsilon := \frac{\Delta_k^{(j)}}{2}$.

$$\begin{aligned}
\mathbb{E}[N_k^{(j)}(T) - N_k^{(j)}(S_{\widetilde{\tau}^{(j)}})] &= \mathbb{E}\left[\sum_{t=S_{\widetilde{\tau}^{(j)}}+1}^{T} \mathbf{1}_{I^{(j)}(t)=k}\right], \\
&\leq \mathbb{E}\left[\sum_{t=S_{\widetilde{\tau}^{(j)}}+1}^{T} \mathbf{1}_{\mathrm{UCB}_k^{(j)}(t) \geq \mathrm{UCB}_{k_j^*}^{(j)}(t)}\right], \\
&\leq \mathbb{E}\left[\sum_{t=S_{\widetilde{\tau}^{(j)}}+1}^{T} \left(\mathbf{1}_{\mathrm{UCB}_k^{(j)}(t) \geq \mu_{jk_*^{(j)}} - \varepsilon, I^{(j)}(t)=k} + \mathbf{1}_{\mathrm{UCB}_{k_*^{(j)}}^{(j)}(t) \leq \mu_{jk_*^{(j)}} - \varepsilon}\right)\right].
\end{aligned}$$

The first inequality follows by the definition of the algorithm. The second inequality follows from a standard argument (c.f. Theorem 8.1 of [Lattimore and Szepesvári, 2018]). By linearity of expectation, we can rewrite the above as,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=S_{\widetilde{\tau}^{(j)}}+1}^{T} \left(\mathbf{1}_{\mathrm{UCB}_k^{(j)}(t) \geq \mu_{jk_*^{(j)}} - \varepsilon, I^{(j)}(t)=k} + \mathbf{1}_{\mathrm{UCB}_{k_*^{(j)}}^{(j)}(t) \leq \mu_{jk_*^{(j)}} - \varepsilon}\right)\right] = \\
&\sum_{t=1}^{T} \mathbb{P}\left[\mathrm{UCB}_k^{(j)}(t) \geq \mu_{jk_*^{(j)}} - \varepsilon, I^{(j)}(t) = k, t \geq S_{\widetilde{\tau}^{(j)}} + 1\right] + \sum_{t=1}^{T} \mathbb{P}\left[\mathrm{UCB}_{k_*^{(j)}}^{(j)}(t) \leq \mu_{jk_*^{(j)}} - \varepsilon, t \geq S_{\widetilde{\tau}^{(j)}} + 1\right].
\end{aligned}$$

Each of the two terms can be computed in a standard fashion, as outlined in Chapter 8 of [Lattimore and Szepesvári, 2018]. We reproduce them here for completeness. For brevity, we are quite loose with the constants and have not optimized them.

$$\begin{aligned}
\mathbb{P}\left[\mathrm{UCB}_k^{(j)}(t) \geq \mu_{jk_*^{(j)}} - \varepsilon, I^{(j)}(t) = k\right] &\leq \mathbb{P}\left[\bigcup_{s=1}^{t} \left\{N_k^{(j)}(t) = s, \widehat{\mu}_{jk}(t-1) + \sqrt{\frac{2\alpha \log_2(t)}{s}} \geq \mu_{jk_*^{(j)}} - \varepsilon\right\}\right], \\
&\leq \mathbb{P}\left[\bigcup_{s=1}^{T} \left\{N_k^{(j)}(t) = s, \widehat{\mu}_{jk}(t-1) + \sqrt{\frac{2\alpha \log(T)}{s}} \geq \mu_{jk_*^{(j)}} - \varepsilon\right\}\right], \\
&= \mathbb{P}\left[\bigcup_{s=1}^{T} \left\{\widehat{\mu}_{jk,s} + \sqrt{\frac{2\alpha \log(T)}{s}} \geq \mu_{jk_*^{(j)}} - \varepsilon\right\}\right], \\
&\leq \sum_{s=1}^{T} \mathbb{P}\left[\widehat{\mu}_{jk,s} + \sqrt{\frac{2\alpha \log(T)}{s}} \geq \mu_{jk_*^{(j)}} - \varepsilon\right], \\
&= \sum_{s=1}^{T} \mathbb{P}\left[\widehat{\mu}_{jk,s} + \sqrt{\frac{2\alpha \log(T)}{s}} \geq \mu_{jk} + \Delta_k^{(j)} - \varepsilon\right],
\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{=} \sum_{s=1}^{T} \mathbb{P}\left[\widehat{\mu}_{jk,s} - \mu_{jk} \geq -\sqrt{\frac{2\alpha\log_2(T)}{s}} + \frac{\Delta_k^{(j)}}{2}\right], \\
&\leq \frac{9\alpha\log(T)}{(\Delta_k^{(j)})^2} + 1 + \sum_{s=\lfloor \frac{9\alpha\log(T)}{(\Delta_k^{(j)})^2} \rfloor}^{T} \mathbb{P}\left[\widehat{\mu}_{jk,s} - \mu_{jk} \geq -\sqrt{\frac{2\alpha\log(T)}{s}} + 1 + \frac{\Delta_k^{(j)}}{2}\right], \\
&\overset{(b)}{\leq} \frac{9\alpha\log(T)}{(\Delta_k^{(j)})^2} + 1 + \sum_{s=\lceil \frac{9\alpha\log(T)}{(\Delta_k^{(j)})^2} \rceil}^{T} \mathbb{P}\left[\widehat{\mu}_{jk,s} - \mu_{jk} \geq 0.025\Delta_k^{(j)}\right], \\
&\overset{(c)}{\leq} \frac{9\alpha\log(T)}{(\Delta_k^{(j)})^2} + 1 + \sum_{s=\lceil \frac{9\alpha\log(T)}{(\Delta_k^{(j)})^2} \rceil}^{\infty} \exp\left(-s(\Delta_k^{(j)})^2 \frac{1}{1600}\right), \\
&\leq \frac{9\alpha\log(T)}{(\Delta_k^{(j)})^2} + 1 + \frac{178T^{-4\alpha}}{(\Delta_k^{(j)})^2}.
\end{aligned}
$$

Step $(a)$ follows from the definition that $\varepsilon = \frac{\Delta_k^{(j)}}{2}$. In Step $(b)$, we use that fact that for $s \geq \lceil \frac{9\alpha\log(T)}{(\Delta_k^{(j)})^2} \rceil$, $\sqrt{\frac{2\alpha\log_2(T)}{s}} \leq \frac{\sqrt{2}\Delta_k^{(j)}}{3}$ and $0.5 - \frac{\sqrt{2}}{3} \geq 0.025$. In step $(c)$, we use Hoeffding's inequality. Similarly, we bound the other inequality as

$$
\begin{aligned}
\sum_{t=1}^{T} \mathbb{P}\left[\mathrm{UCB}_{k_*^{(j)}}^{(j)}(t) \leq \mu_{jk_*^{(j)}} - \varepsilon, t \geq S_{\widetilde{\tau}^{(j)}} + 1\right] &\leq \sum_{t=1}^{T} \mathbb{P}\left[\mathrm{UCB}_{k_*^{(j)}}^{(j)}(t) \leq \mu_{jk_*^{(j)}} - \varepsilon\right], \\
&= \sum_{t=1}^{T} \mathbb{P}\left[\widehat{\mu}_{jk_*^{(j)}}(t) + \sqrt{\frac{2\alpha\log(T)}{N_{k_*^{(j)}}^{(j)}(t)}} \leq \mu_{jk_*^{(j)}} - \varepsilon\right], \\
&\leq \sum_{t=1}^{T}\sum_{s=1}^{t} \mathbb{P}\left[\widehat{\mu}_{jk_*^{(j)},s} + \sqrt{\frac{2\alpha\log(T)}{s}} \leq \mu_{jk_*^{(j)}} - \varepsilon\right], \\
&\leq \sum_{t=1}^{T}\sum_{s=1}^{t} \exp\left(-2s\left(\sqrt{\frac{2\alpha\log(T)}{s}} + \varepsilon\right)^2\right), \\
&\leq T^{2-4\alpha}.
\end{aligned}
$$

□

From the above two propositions, we obtain the following corollary

**Corollary 11.** *The regret of any agent $j \in [N]$, at time $T$ can be bounded s*

$$
\begin{aligned}
\mathbb{E}\left[R_T^{(j)}\right] \leq \mathbb{E}[S_{\widetilde{\tau}^{(j)}}] + \underbrace{2(j-1)(\log_2(T)+1)}_{\textit{Regret due to Communication}} &+ \underbrace{\sum_{j'=1}^{j-1}\sum_{k\in\mathcal{A}_j^*}\left(\frac{9\alpha\log(T)}{(\Delta_k^{(j')})^2} + 1 + \frac{178T^{-4\alpha}}{(\Delta_k^{(j')})^2} + T^{2-4\alpha}\right)\mu_{jk_*^{(j)}}}_{\textit{Regret due to Collision}} \\
&+ \underbrace{\sum_{k\in\mathcal{A}_j^*\setminus\{k_*^{(j)}\}}\left(\frac{9\alpha\log(T)}{(\Delta_k^{(j)})^2} + 1 + \frac{178T^{-4\alpha}}{(\Delta_k^{(j)})^2} + T^{2-4\alpha}\right)\Delta_k^{(j)}}_{\textit{Regret due to Sub-optimal arm pull}}.
\end{aligned}
$$

*Proof.* The Corollary follows from the following facts

- For every $j' < j$, $\widetilde{S}^{(j')} \leq \widetilde{S}^{(j)}$ almost-surely.

- For every $j^{'} < j$, every arm in set $\mathcal{A}_*^{(j)}$ is sub-optimal.
- Plugging in the estimates from Proposition 10 in Lemma 9.

□

Thus, it remains to bound $\mathbb{E}[S_{\widetilde{\tau}^{(j)}}]$, which is the subject of the next section.

### B.5 Bound on Mean and Exponential Moment of $\widetilde{\tau}^{(j)}$

Since for all $i \in \mathbb{N}$, $S_i = N + (2^{i-1} - 1) + (i-1)NK$, it suffices to bound the exponential moment $\mathbb{E}[2^{\widetilde{\tau}^{(j)}}]$ and the mean $\mathbb{E}[\widetilde{\tau}^{(j)}]$ to complete the regret guarantee. In order to do so, we first start by analyzing the probability that a phase is bad for an agent and then use that to bound the exponential moment of $\tau^{(j)}$. We shall now bound the probability that a phase is bad for a particular agent.

**Lemma 12.** *For any phase $i > i^*$, any agent $j$ and arm $k \in \mathcal{A}_*^{(j)} \setminus \{k_*^{(j)}\}$, we have*

$$\mathbb{P}[\chi_i^{(j)} = 0, i \geq \widetilde{\tau}^{(j-1)}] \leq 2N^2 \left(\frac{2}{e^5}\right)^i.$$

*Proof.* The proof follows from the basic properties of the UCB algorithm, which we can bound as follows. Recall the notation that for any agent $j \in [N]$, phase $i \in \mathbb{N}$ and arm $k \in [K]$, the quantity $N_k^{(j)}[i]$ denotes the number of times agent $j$ was matched to arm $k$ in the regret minimization blocks upto and including phase $i$. For any phase $i \geq i^*$, agent $j$ and arm $k \in \mathcal{A}_*^{(j)} \setminus \{k_*^{(j)}\}$, we have

$$\mathbb{P}\left[N_k^{(j)}[i] - N_k^{(j)}[i-1] > \left\lceil \frac{10\alpha i}{(\Delta_k^{(j)})^2} \right\rceil, i \geq \widetilde{\tau}^{(j-1)}\right] \leq \mathbb{P}\left[\bigcup_{t \geq S_{i-1} + \lceil \frac{10\alpha i}{(\Delta_k^{(j)})^2} \rceil}^{S_i} N_k^{(j)}(t) = \left\lceil \frac{10\alpha i}{(\Delta_k^{(j)})^2} \right\rceil + N_k^{(j)}[i-1], I^{(j)}(t) = k, i \geq \widetilde{\tau}^{(j-1)}\right].$$

Since $N_k^{(j)}[i-1] \geq 0$, the above can be simplified to

$$\mathbb{P}\left[N_k^{(j)}[i] - N_k^{(j)}[i-1] > \left\lceil \frac{10\alpha i}{(\Delta_k^{(j)})^2} \right\rceil, i \geq \widetilde{\tau}^{(j-1)}\right] \leq \mathbb{P}\left[\bigcup_{t \geq S_{i-1} + \lceil \frac{10\alpha i}{(\Delta_k^{(j)})^2} \rceil}^{S_i} N_k^{(j)}(t) \geq \left\lceil \frac{10\alpha i}{(\Delta_k^{(j)})^2} \right\rceil, i \geq \widetilde{\tau}^{(j-1)}\right]. \quad (5)$$

Now, by applying an union bound to the RHS, we obtain from the preceding display that

$$\mathbb{P}\left[\bigcup_{t \geq S_{i-1} + \lceil \frac{10\alpha i}{(\Delta_k^{(j)})^2} \rceil}^{S_i} N_k^{(j)}(t) \geq \left\lceil \frac{10\alpha i}{(\Delta_k^{(j)})^2} \right\rceil, I^{(j)}(t) = k, i \geq \widetilde{\tau}^{(j-1)}\right] \leq \sum_{t=S_{i-1}}^{S_i} \mathbb{P}\left[N_k^{(j)}(t) \geq \left\lceil \frac{10\alpha i}{(\Delta_k^{(j)})^2} \right\rceil, I^{(j)}(t) = k, i \geq \widetilde{\tau}^{(j-1)}\right]. \quad (6)$$

The classical large-deviation estimate for UCB from [Auer et al., 2002] gives that

$$\mathbb{P}\left[N_k^{(j)}(t) \geq \left\lceil \frac{10\alpha i}{(\Delta_k^{(j)})^2} \right\rceil, I^{(j)}(t) = k, i \geq \widetilde{\tau}^{(j-1)}\right] \leq 2e^{-5i}, \tag{7}$$

for all times $t \in \{S_{i-1}, \cdots, S_i\}$. In words, this estimate gives that UCB will play a sub-optimal arm with very low probability, on the event that the sub-optimal arm has been played sufficiently many times. We can use that estimate, since on the event that $i \geq \widetilde{\tau}^{(j-1)}$, we have that the set of active arms of agent $j$ in phase $i$, denoted by $\mathcal{A}_i^{(j)} = \mathcal{A}_*^{(j)}$. Thus, arm $k_*^{(j)}$ is the best arm for agent $j$ in phase $i$. Now, combining Equations (5),(6),(7), we get

$$\mathbb{P}\left[N_k^{(j)}[i] - N_k^{(j)}[i-1] > \left\lceil \frac{10\alpha i}{(\Delta_k^{(j)})^2} \right\rceil, i \geq \widetilde{\tau}^{(j-1)}\right] \leq \sum_{t=S_{i-1}}^{S_i} 2e^{-5i} \leq 2\left(\frac{2}{e^5}\right)^i. \tag{8}$$

To conclude the proof, notice the following fact.

**Proposition 13.** *For every $i \geq i^*$,*

$$\{\chi_i^{(j)} = 0, i \geq \widetilde{\tau}^{(j-1)}\} \subseteq \bigcup_{j'=1}^{j} \bigcup_{k \in \mathcal{A}_*^{(j')} \setminus \{k_*^{(j')}\}} \left\{[N_k^{(j')}[i] - N_k^{(j')}[i-1] \geq \left\lceil \frac{10\alpha i}{(\Delta_k^{(j')})^2} \right\rceil, i \geq \widetilde{\tau}^{(j-1)}\right\},$$

*where $i^*$ is defined in Equation (4).*

*Proof.* It suffices to establish that

$$\bigcap_{j'=1}^{j} \bigcap_{k \in \mathcal{A}_*^{(j')} \setminus \{k_*^{(j')}\}} \left\{[N_k^{(j')}[i] - N_k^{(j')}[i-1] \leq \left\lceil \frac{10\alpha i}{(\Delta_k^{(j')})^2} \right\rceil, i \geq \widetilde{\tau}^{(j-1)}\right\} \subseteq \{\chi_i^{(j)} = 1, i \geq \widetilde{\tau}^{(j-1)}\}.$$

This follows as, the phase $i^*$ is such that $\frac{10\alpha i}{\Delta^2} NK < 2^{i-1}$. Suppose all events on the LHS hold. Then, agent $j$ is matched at-most $\frac{10\alpha i}{(\Delta_k^{(j)})^2}$ times to any arm $\mathcal{A}_*^{(j)} \setminus \{k_*^{(j)}\}$ (the sub-optimal arms). This in turn is upper bounded by $\frac{10\alpha i}{\Delta^2}$. Since there are $K - j$ arms in the set $\mathcal{A}_*^{(j)} \setminus \{k_*^{(j)}\}$, the total number of sub-optimal arm pulls is at-most $(K-j)\frac{10\alpha i}{\Delta^2}$.

In order to bound the total number of collisions agent $j$ will face, we make use of two simple observations. First is that, under the events on the LHS, every agent $j' < j$ will match no more than $\frac{10\alpha i}{(\Delta_k^{(j')})^2}$ times with arm $k$, for all $k \in \mathcal{A}_*^{(j)}$. Second, is that if agent $j$ faces a collision at an arm $k \in \mathcal{A}_*^{(j)}$, then it must be the case that exactly one agent ranked 1 through to $j-1$ must have been matched to arm $k$ in the same slot. These two observations give that the total number of times agent $j$ will face a collision at an arm $k \in \mathcal{A}_*^{(j)}$ is at-most the sum of times arm $k$ is matched to agents ranked 1 through to $j-1$, which in turn is upper bounded by $(j-1)\frac{10\alpha i}{\Delta^2}$. Since therer are exactly $K - j + 1$ arms in $\mathcal{A}_*^{(j)}$, the total number of collisions incurred by agent $j$ in phase $i$, when the events on the LHS hold is upper bounded by $(K-j+1)(j-1)\frac{10\alpha i}{\Delta^2}$, which in turn is upper bounded by $NK\frac{10\alpha i}{\Delta^2}$.

However, as there is at-least $20NK\frac{\alpha i}{\Delta^2}$ time slots in phase $i$, the preceding argument yields that agent $j$ must be matched to arm $k_*^{(j)}$ at-least $20NK\frac{\alpha i}{\Delta^2} - NK\frac{10\alpha i}{\Delta^2} - K\frac{10\alpha i}{\Delta^2} \geq NK\frac{10\alpha i}{\Delta^2}$ times. Thus, agent $j$ is matched to arm $k_*^{(j)}$ the most number of times in phase $i$, i.e., $\chi_i^{(j)} = 1$.

□

Thus, from Proposition 13, and applying an union bound using Equation (8), we get

$$\mathbb{P}[\chi_i^{(j)} = 0, i \geq \widetilde{\tau}^{(j-1)}] \leq 2jK\left(\frac{2}{e}\right)^i,$$

$$\leq 2NK\left(\frac{2}{e}\right)^{i}.$$

□

We use this to now compute the mean of $\widetilde{\tau}^{(j)}$.

**Proposition 14.** *For every $j \in [N]$, we have*

$$\mathbb{E}[\widetilde{\tau}^{(j)}] \leq j(i^* + 3NK),$$

*where $i^*$ is defined in Equation (4).*

*Proof.*

$$\begin{aligned}
\mathbb{E}[\widetilde{\tau}^{(j)}] &= \sum_{x\geq 1}\mathbb{P}[\widetilde{\tau}^{(j)} \geq x], \\
&= \sum_{x\geq 1}\mathbb{P}[\widetilde{\tau}^{(j)} \geq x, \widetilde{\tau}^{(j-1)} > x] + \sum_{x\geq 1}\mathbb{P}[\widetilde{\tau}^{(j)} \geq x, \widetilde{\tau}^{(j-1)} \leq x], \\
&= \sum_{x\geq 1}\mathbb{P}[\widetilde{\tau}^{(j-1)} > x] + \sum_{x\geq 1}\mathbb{P}[\widetilde{\tau}^{(j)} \geq x, \widetilde{\tau}^{(j-1)} \leq x], \\
&\leq \mathbb{E}[\widetilde{\tau}^{(j-1)}] + i^* + \sum_{x\geq i^*}\mathbb{P}[\widetilde{\tau}^{(j)} \geq x, \widetilde{\tau}^{(j-1)} \leq x], \\
&\overset{(a)}{\leq} \mathbb{E}[\widetilde{\tau}^{(j-1)}] + i^* + \sum_{x\geq i^*} 2N^2\left(\frac{2}{e^5}\right)^{x}, \\
&\leq \mathbb{E}[\widetilde{\tau}^{(j-1)}] + i^* + 3NK, \\
&\overset{(b)}{\leq} j(i^* + 3NK).
\end{aligned}$$

Step $(a)$ follows Lemma 12 and step $(b)$ follows from the fact that $\widetilde{\tau}^{(0)} = 0$ almost-surely. □

Similarly, we can compute the exponential moment of $\widetilde{\tau}^{(j)}$.

**Proposition 15.** *For every $j \in [N]$, we have*

$$\mathbb{E}[2^{\widetilde{\tau}^{(j)}}] \leq 1 + j(2^{i^*} + 4NK),$$

*where $i^*$ is defined in Equation (4).*

*Proof.*

$$\begin{aligned}
\mathbb{E}[2^{\widetilde{\tau}^{(j)}}] &= \sum_{x\geq 1}\mathbb{P}[2^{\widetilde{\tau}^{(j)}} \geq x], \\
&= \sum_{x\geq 1}\mathbb{P}[\widetilde{\tau}^{(j)} \geq \log_2(x)], \\
&= \sum_{x\geq 1}\mathbb{P}[\widetilde{\tau}^{(j)} \geq \log_2(x), \widetilde{\tau}^{(j-1)} > \log_2(x)] + \sum_{x\geq 1}\mathbb{P}[\widetilde{\tau}^{(j)} \geq \log_2(x), \widetilde{\tau}^{(j-1)} \leq \log_2(x)], \\
&= \sum_{x\geq 1}\mathbb{P}[\widetilde{\tau}^{(j-1)} > \log_2(x)] + \sum_{x\geq 1}\mathbb{P}[\widetilde{\tau}^{(j)} \geq \log_2(x), \widetilde{\tau}^{(j-1)} \leq \log_2(x)], \\
&= \mathbb{E}[2^{\widetilde{\tau}^{(j-1)}}] + \sum_{x\geq 1}\mathbb{P}[\widetilde{\tau}^{(j)} \geq \log_2(x), \widetilde{\tau}^{(j-1)} \leq \log_2(x)], \\
&\overset{(a)}{\leq} \mathbb{E}[2^{\widetilde{\tau}^{(j-1)}}] + 2^{i^*} + \sum_{x\geq 2^{i^*}} 2NK\left(\frac{2}{e^5}\right)^{\log_2 x},
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}[2^{\widetilde{\tau}^{(j-1)}}] + 2^{i^*} + \sum_{x \geq 2^{i^*}} 2NKx^{1-\frac{5}{\ln(2)}}, \\
&\leq \mathbb{E}[2^{\widetilde{\tau}^{(j-1)}}] + 2^{i^*} + 4NK, \\
&\overset{(b)}{\leq} 1 + j(2^{i^*} + 4NK).
\end{aligned}
$$

Step $(a)$ follows from Lemma 12 and step $(b)$ follows from the fact that $\widetilde{\tau}^{(0)} = 0$ almost-surely. □

**Corollary 16.** *For every $j \in [N]$, we have*

$$
\begin{aligned}
\mathbb{E}[S_{\widetilde{\tau}^{(j)}}] &\leq N + j(2^{i^*} + 4NK) + jNK(i^* + 3NK), \\
&= N + j(2^{i^*} + i^* + 4(NK)^2 + 3NK),
\end{aligned}
$$

*where $i^*$ is defined in Equation (4).*

*Proof.* We know that for any $i \in \mathbb{N}$, $S_i := N + (2^{i-1} - 1) + (i-1)NK$. The result then follows from Propositions 14 and 15. □

## C Incentive Compatibility - Proof of Proposition 4

We restate Proposition 4 for the reader's convenience.

**Proposition 17.** *The UCB-D3 algorithm profile is $\varepsilon : (\varepsilon_j)_{j=1}^N$ stable where, for all $j \in [N]$, $\varepsilon_j = \sum_{l=1}^{j-1} \mathbf{1}_{(\mu_{jl} > \mu_{jk_j^*})} \frac{\mu_{jl}}{\mu^{(l)}} \mathbb{E}[R_T^{(l)}] + \mathbb{E}[R_T^{(j)}]$, where for all $j' \in [N]$, $\mathbb{E}[R_T^{(j')}]$ is given in Equation (1).*

This proposition gives that for agent ranked $j$, $\varepsilon_j = O\left(j \frac{\mu_{\max}^{(j)}}{\mu_{\min}^{(j)}} \mathbb{E}[R_T^{((j))}]\right)$, where $\mu_{\max}^{(j)}$ ($\mu_{\min}^{(j)}$) is the maximum (minimum) arm-mean for agent $j$. This establishes that UCB-D3 is approximately incentive compatible, namely, even if an agent deviates from the UCB-D3 algorithm, the possible improvement in reward is $O(\log(T))$.

*Proof of Proposition 4.* We bound the equilibrium property of UCB-D3; as follows. Observe that agent ranked $j$ will only collide with agents ranked 1 through $j-1$. Now, if all agents 1 through to $j-1$ are all playing arms $k_*^{(1)}, \cdots k_*^{(j-1)}$ respectively (their individual best arms), then the best arm (by definition) for agent $j$ to play will be arm $k_*^{(j)}$. On the other hand, when any agent $j' \leq j-1$ does not play arm $k_*^{(j')}$, the maximum expected reward collected by agent $j$ can be at-most $\max(\mu_{jk_*^{(j')}}, \mu_{jk_*^{(j)}})$. Under the UCB-D3; strategy profile, the expected number of times any agent $j \in [N]$, plays an arm in the set $[K] \setminus \{k_*^{(1)}, \cdots, k_*^{(j)}\}$ is at-most $\frac{1}{\mu^{(j)}} \mathbb{E}[R_T^{(j)}]$, where $\mu^{(j)} := \min_{k \in [K]} \mu_{jk}$ is the smallest arm-gap. Notice that for all agents $j$, $\mu^{(j)} > 0$, by model assumptions. This then gives us the following decomposition

$$
\sup_{s'} \mathbb{E}[\mathrm{Rew}_T^{(j)}(s_{-j}, s')] \leq \sum_{l=1}^{j-1} \mathbf{1}_{\mu_{jl} > \mu_{jk_*^{(j)}}} \frac{\mu_{jl}}{\mu^{(l)}} \mathbb{E}[R_T^{(l)}] + \mu_{jk_*^{(j)}} T, \tag{9}
$$

where $\mathbb{E}[R_T^{(l)}]$ is given in Theorem 1. Similarly, from the definition of regret, we have

$$
\mathbb{E}[\mathrm{Rew}_T^{(j)}(s)] \geq \mu_{jk_*^{(j)}} T - \mathbb{E}[R_T^{(j)}], \tag{10}
$$

where $\mathbb{E}[R_T^{(j)}]$ is given in Theorem 1. Thus, from Equations (9) and (10), we get that

$$
\sup_{s'} \left( \mathbb{E}[\mathrm{Rew}_T^{(j)}(s_{-j}, s')] - \mathbb{E}[\mathrm{Rew}_T^{(j)}(s)] \right) \leq \sum_{l=1}^{j-1} \mathbf{1}_{\mu_{jl} > \mu_{jk_*^{(j)}}} \frac{\mu_{jl}}{\mu^{(l)}} \mathbb{E}[R_T^{(l)}] + \mathbb{E}[R_T^{(j)}].
$$

□

# D Proof of Regret Lower Bound

We will use the following notations throughout the proof of the lower bound.

1. Can assume without loss of generality (W.l.o.g.) that the rank of any agent $i \in [N]$ is $i$.

2. Any agent related symbol is a superscript. Arm related is a sub-script. Thus, for any time $t$, the number of times arm $k \in [K]$ is played by agent $j$ is $N_k^{(j)}(t)$. The number of time the agent $j$ is blocked up to time $t$ is given as $C^{(j)}(t)$.

3. Distribution of agent $i \in [N]$ and arm $k \in [K]$ is given by $\nu_{jk}$, which has mean $\mu_{jk}$. W.l.o.g. let us assume $\max_k \mu_{jk} > 0$.

4. The stable match partner of any agent $j \in [N]$ is given by $k_*^{(j)} \in [K]$. The set of dominated arms for the agent $j$ is given as $\mathcal{D}_*^{(j)} = \{k_*^{(j)} : 1 \leq j' \leq j-1\}$, the set of non-dominated arms is given as $\mathcal{A}_*^{(j)} = [K] \setminus \mathcal{D}_*^{(j)}$.

5. For any agent $j \in [N]$, arm $k \in [K]$, $\Delta_k^{(j)} := \mu_{jk_*^{(i)}} - \mu_{jk}$, the arm-gap. This can be negative.

### D.1 Divergence Decomposition

We need to setup a few notations for the proof of divergence decomposition lemma. The proof generalizes the framework in Chapter 15 of [Lattimore and Szepesvári, 2018] for the multi-agent framework.[8]

**Canonical multi-agent bandit model:** We now define the ($N$-agent, $K$-arm, T-horizon) bandit models. The canonical bandit model ($N$-agent, $K$-arm, T-horizon) lies in a measurable space $\{\Omega, \mathcal{F}\}$. Let $K^{(j)}(t)$ denotes the arm chosen by the $j$-th agent on time $t$, and $X^{(j)}(t)$ denotes the rejection or reward obtained from that arm for agent $j$ in round $t$. We denote the rejection by the symbol $\emptyset$. Therefore, $K^{(j)}(t) \in [K]$, and $X^{(j)}(t) \in [0,1] \cup \{\emptyset\}$ for all $j \in [N]$ and $t \in [T]$. Also, $K^{(j)}(t)$, and $X^{(j)}(t)$ for all $j \in [N]$ and $t \in [T]$ are measurable with respect to $\mathcal{F}$. Let $H(t) = (K^{(j)}(t'), X^{(j)}(t') \forall j \in [N], \forall t' \leq t)$ be the random variable representing the history of actions taken and rewards seen up to and including round $t$. We have $H(t) \in \mathcal{H}(t) \equiv \left([K]^N \times ([0,1] \cup \{\emptyset\})^N\right)^t$. We may set $\Omega \equiv \mathcal{H}(T)$ and the sigma algebra generated by the history as $\mathcal{F} \equiv \sigma(H(T))$.

**Environment:** The bandit environment is specified by $\nu = (\nu_{jk}, \forall j \in [N], k \in [K])$ where $\nu_{jk}$ is the distribution of rewards obtained when arm $k$ is matched to agent $j$ in this environment.

**Policy:** A policy is a sequence of distribution of possible request to the arms from the agents (which can assimilate any coordination among the agents) conditioned on the past events. More formally, the policy $\boldsymbol{\pi} = \{\boldsymbol{\pi}_t(\cdot) : t \in [T]\}$ where $\boldsymbol{\pi}_t(\cdot) \equiv \{\pi_t(k, j|\cdot) : \forall k \in [K], j \in [N]\}$ is the function that maps the history upto time $t-1$ to the action $K^{(j)}(t), \forall j \in [N]$. Further, $\pi_t(k, j|\cdot) : \mathcal{H}(t-1) \to [0,1]$ denotes the probability, as a function of history $H(t-1)$ of agent $j$ playing arm $k$.

**Probability Measure:** Each environment $\boldsymbol{\nu}$ and policy $\pi$ jointly induces a probability distribution over the measurable space $\{\Omega, \mathcal{F}\}$ denoted by $\mathbb{P}_{\nu,\pi}$. Let $\mathbb{E}_{\nu,\pi}$ denote the expectation induced. The density of a particular history up to time $T$, under an environment $\boldsymbol{\nu}$ and a policy $\pi$, can be defined as

$$\begin{aligned} &d\mathbb{P}_{\nu,\pi}\left(\mathbf{k}(t), \mathbf{x}(t) : t \in [T]\right) \\ &= \prod_{t=1}^{T} \pi_t(\mathbf{k}(t)|h(t-1)) p_\nu(\mathbf{x}(t)|\mathbf{k}(t)) d\lambda(\mathbf{x}(t); \boldsymbol{\nu}) d\rho(\mathbf{k}(t)). \end{aligned}$$

[8] See, [Besson and Kaufmann, 2017] for a related approach for regret lower bound proof in the colliding bandit models [Avner and Mannor, 2014].

Here, $\lambda(\mathbf{x};\boldsymbol{\nu}) = \prod_{j=1}^{N} \lambda_j(x^{(j)})$ is the dominating measure over the rewards with $\lambda_j(x^{(j)}) = \delta_\emptyset + \sum_k \nu_{jk}$.[9] Also, $\rho(\mathbf{k})$ is the counting measure on the collective action of the agents.

**Lemma 18** (Divergence Decomposition)**.** *For two bandit instances $\boldsymbol{\nu} = \{\nu_{jk} : j \in [N], k \in [K]\}$, and $\boldsymbol{\nu}' = \{\nu_{jk} : j \in [N], k \in [K]\}$, and any admissible policy $\pi$ the following divergence decomposition is true*

$$D(\mathbb{P}_{\nu,\pi}, \mathbb{P}_{\nu',\pi}) = \sum_{j=1}^{N}\sum_{k=1}^{K} \mathbb{E}_{\nu,\pi}[N_k^{(j)}(T)] D(\nu_{jk}, \nu'_{jk}).$$

*Proof.* The divergence between two measures, which correspond to two different environments under a policy $\pi$, $\mathbb{P}_{\nu,\pi}$ and $\mathbb{P}_{\nu',\pi}$ can be expressed as

$$\begin{aligned}
&D(\mathbb{P}_{\nu,\pi}, \mathbb{P}_{\nu',\pi}) \\
&= \mathbb{E}_{\nu,\pi}\left[\sum_{t=1}^{T} \log\left(\frac{d\mathbb{P}_{\nu,\pi}}{d\mathbb{P}_{\nu',\pi}}\right)\right] \\
&\overset{(i)}{=} \mathbb{E}_{\nu,\pi}\left[\log\left(\frac{p_\nu(\mathbf{x}(t)|\mathbf{k}(t))}{p_{\nu'}(\mathbf{x}(t)|\mathbf{k}(t))}\right)\right] \\
&\overset{(ii)}{=} \mathbb{E}_{\nu,\pi}\left[\sum_{t=1}^{T} \log\left(\frac{\prod_{j:x^{(j)}(t)\neq\emptyset} p_\nu(x^{(j)}(t)|\mathbf{k}(t))}{\prod_{j:x^{(j)}(t)\neq\emptyset} p_{\nu'}(x^{(j)}(t)|\mathbf{k}(t))}\right)\right] \\
&= \mathbb{E}_{\nu,\pi}\left[\sum_{t=1}^{T} \sum_{j:x^{(j)}(t)\neq\emptyset} \mathbb{E}_\nu\left[\log\left(\frac{p_\nu(x^{(j)}(t)|\mathbf{k}(t))}{p_{\nu'}(x^{(j)}(t)|\mathbf{k}(t))}\right)\Big|\mathbf{k}(t)\right]\right] \\
&\overset{(iii)}{=} \mathbb{E}_{\nu,\pi}\left[\sum_{t=1}^{T} \sum_{j:x^{(j)}(t)\neq\emptyset} D\left(\nu_{jk^{(j)}(t)}, \nu'_{jk^{(j)}(t)}\right)\right] \\
&= \sum_{j=1}^{N}\sum_{k=1}^{K} \mathbb{E}_{\nu,\pi}\left[\sum_{t=1}^{T} \mathbf{1}_{(k^{(j)}(t)=k, x^{(j)}(t)\neq\emptyset)} D\left(\nu_{jk}, \nu'_{jk}\right)\right] \\
&\overset{(iv)}{=} \sum_{j=1}^{N}\sum_{k=1}^{K} \mathbb{E}_{\nu,\pi}\left[N_k^{(j)}(T)\right] D\left(\nu_{jk}, \nu'_{jk}\right).
\end{aligned}$$

In the above series equation (i) is true because the density of the policy cancels out for the two different environments. Equation (ii) holds because if for some set of actions $\mathbf{k}(t)$ agent $j$ observes $x^{(j)}(t) = \emptyset$ that indicates agent $j$ is rejected on that round. This is independent of the environment. In particular, we have $p_\nu(x^{(j)}(t)|\mathbf{k}(t)) = p_{\nu'}(x^{(j)}(t)|\mathbf{k}(t))$ if $x^{(j)}(t) = \emptyset$ for any $\mathbf{k}(t)$. In deriving equation (iii) we make use of the definition of divergence. Equation (iv) uses the definition of $N_k^{(j)}(t)$ the total number of times agent $j$ successfully plays arm $k$ up to time $T$. □

### D.2 Proof of Regret Decomposition (Lemma 6)

*Proof.* We fix any agent $j \in [N]$ for the rest of the proof. We have the expected regret for the agent $j$, under a policy $\pi$ and any bandit instance $\boldsymbol{\nu}$ as

$$R_T^{(j)}(\boldsymbol{\nu}, \pi) = \sum_{k=1}^{K} \Delta_k^{(j)} \mathbb{E}_{\nu,\pi}[N_k^{(j)}(T)] + \sum_{k=1}^{K} \mu_{jk_*^{(j)}} \mathbb{E}_{\nu,\pi}[C^{(j)}(T)].$$

[9] Here $\delta_\emptyset$ is the dirac measure on $\emptyset$ denoting the rejection event. For multiple a pair environments we can define a dominating measure as $\lambda(\mathbf{x};\boldsymbol{\nu}_1,\boldsymbol{\nu}_2) = \sum_{i=1,2}\lambda(\mathbf{x};\boldsymbol{\nu}_i)$. This is used in the proof of Lemma 18.

This is true as for each collision the agent $j$ obtains $\mu_{jk_*^{(j)}}$ regret (0 reward) in expectation, and for each successful play of arm $k$ it obtains $\Delta_k^{(j)}$ regret. Therefore, a trivial regret lower bound is

$$R_T^{(j)}(\boldsymbol{\nu}, \pi) \geq \sum_{k=1}^{K} \Delta_k^{(j)} \mathbb{E}_{\nu,\pi}[N_k^{(j)}(T)].$$

For an OSB instance, we know that the number of times the agents 1 to $(j-1)$ plays arm $k_*^{(j)}$ successfully, the agent $j$ should either move to a sub-optimal arm (as the arm $k_*^{(j)}$ is the optimal arm for agent $j$ in an OSB instance) or it is blocked. In the best possible scenario, the agent $j$ successfully plays its second best arm, in each of these instances. This holds as $\Delta_{\min}^{(j)} \leq \mu_{jk_*^{(j)}}$ for non-negative rewards. Therefore, the regret from the events when agents 1 to $(j-1)$ plays arm $k_*^{(j)}$ successfully, is lower bounded by

$$R_T^{(j)}(\boldsymbol{\nu}, \pi) \geq \sum_{j'=1}^{j-1} \Delta_{\min}^{(j)} \mathbb{E}_{\nu,\pi}[N_{k_*^{(j)}}^{(j')}(T)].$$

Therefore, the combined regret lower bound is given as

$$R_T^{(j)}(\boldsymbol{\nu}, \pi) \geq \max\left\{ \sum_{k=1}^{K} \Delta_k^{(j)} \mathbb{E}_{\nu,\pi}[N_k^{(j)}(T)], \sum_{j'=1}^{j-1} \Delta_{\min}^{(j)} \mathbb{E}_{\nu,\pi}[N_{k_*^{(j)}}^{(j')}(T)] \right\}$$

□

### D.3 Proof of Regret Lower Bound (Theorem 7)

*Proof.* We consider any instance in the class of OSB $\boldsymbol{\nu}$, universally consistent policy $\pi$, agent $j \in [N]$, and arm $k \in [K] \setminus \{k_*^{(j')} : 1 \leq j' \leq j\}$. Let us consider the instance $\boldsymbol{\nu'}$ (which is specific to the $j$ and $k$ pair) where $\nu'_{j'k'} = \nu_{j'k'}$ for all $j' \neq j, k' \neq k$, $\nu'_{jk}$ such that $D(\nu_{jk}, \nu'_{jk}) \leq D_{\inf}(\nu_{jk}, \mu_{jk_*^{(j)}}, \mathcal{P}) + \epsilon$ and $\mu'_{jk} \equiv \mu(\nu'_{jk}) > \mu_{jk_*^{(j)}}$ for some $\epsilon > 0$. Note, for $\mu_{jk_*^{(j)}} < 1$ and $\Delta_k^{(j)} > 0$, which holds by assumption, the distribution $\nu'_{jk}$ exists by definition of $D_{\inf}(\cdot)$. In short, for the $j$-th agent we make the $k$-th arm optimal. The optimal arm for agent $j$ in the instance $\boldsymbol{\nu'}$ is the arm $k$.

For any event $A$ (and its complement $A^c$), due to Pinsker's inequality we have

$$D(\mathbb{P}_{\nu,\pi}, \mathbb{P}_{\nu',\pi}) \geq \log\left(\frac{1}{2(\mathbb{P}_{\nu,\pi}(A) + \mathbb{P}_{\nu,\pi}(A^c))}\right). \tag{11}$$

Let us now consider the event $A = \{N_k^{(j)}(T) \geq T/2\}$. Therefore, due to the regret decomposition lemma 6, we have the regrets:

1. In instance $\boldsymbol{\nu}$ as $R_T(\boldsymbol{\nu}, \pi) \geq R_T^{(j)}(\boldsymbol{\nu}, \pi) \geq \Delta_k^{(j)} \frac{T}{2} \mathbb{P}_{\nu,\pi}\left(\{N_k^{(j)}(T) \geq T/2\}\right)$.
2. In instance $\boldsymbol{\nu'}$ as $R_T(\boldsymbol{\nu'}, \pi) \geq R_T^{(j)}(\boldsymbol{\nu'}, \pi) \geq (\mu'_{jk} - \mu_{jk_j^*}) \frac{T}{2} \mathbb{P}_{\nu,\pi}\left(\{N_k^{(j)}(T) < T/2\}\right)$.

As the only change in reward distribution happens in agent $j$, arm $k$ pair, we have from Lemma 18:

$$D(\mathbb{P}_{\nu,\pi}, \mathbb{P}_{\nu',\pi}) = D(\nu_{jk}, \nu'_{jk}) \mathbb{E}_{\nu,\pi}[N_k^{(j)}(T)] \leq \left(\epsilon + D_{\inf}(\nu_{jk}, \mu_{jk_*^{(j)}}, \mathcal{P})\right) \mathbb{E}_{\nu,\pi}[N_k^{(j)}(T)].$$

The last inequality holds true by construction of $\nu'_{jk}$.

Substituting the above three relations in Equation (11) we obtain for any $\epsilon > 0$.

$$\left(\epsilon + D_{\inf}(\nu_{jk}, \mu_{jk_*^{(j)}}, \mathcal{P})\right) \mathbb{E}_{\nu,\pi}[N_k^{(j)}(T)] \geq \log\left(\frac{1}{2(\mathbb{P}_{\nu,\pi}(A) + \mathbb{P}_{\nu,\pi}(A^c))}\right) \geq \log\left(\frac{T \min((\mu'_{jk} - \mu_{jk_*^{(j)}}), \Delta_k^{(j)})}{4(R_T^{(j)}(\boldsymbol{\nu}, \pi) + R_T^{(j)}(\boldsymbol{\nu'}, \pi))}\right).$$

Here, the final inequality hold as the policy $\pi$ is assumed to be universally consistent. Therefore, taking the following holds after taking the

$$\lim_{\epsilon\to 0}\liminf_{T\to\infty}\frac{\mathbb{E}_{\nu,\pi}[N_k^{(j)}(T)]}{\log T}\geq\lim_{\epsilon\to 0}\frac{1}{\epsilon+D_{\inf}(\nu_{jk},\mu_{jk_*^{(j)}},\mathcal{P})}=\frac{1}{D_{\inf}(\nu_{jk},\mu_{jk_*^{(j)}},\mathcal{P})}.$$

As the above bound is true for any uniformly consistent policy $\pi$, and for agent $j$, and arm $k\in[K]\setminus\{k_*^{(j')}:1\leq j'\leq j\}$. We use the regret decomposition lemma (Lemma 6) to obtain the final aysmptotic regret lower bound for any agent $j\in[N]$ as

$$\liminf_{T\to\infty}\frac{R_T^{(j)}(\boldsymbol{\nu})}{\log T}\geq\max\left\{\sum_{j'=1}^{j-1}\frac{\Delta_{\min}^{(j)}}{D_{\inf}(\nu_{j'k_*^{(j)}},\mu_{j'k_*^{(j')}},\mathcal{P})},\sum_{k\notin\mathcal{A}_*^{(j)}\setminus k_*^{(j)}}\frac{\Delta_k^{(j)}}{D_{\inf}(\nu_{jk},\mu_{jk_*^{(j)}},\mathcal{P})}\right\}$$

Here, we use the fact that $k_*^{(j)}\notin\mathcal{D}_*^{(j')}\cup\{k_*^{(j')}\}$ for all $j'<j$. Also note, $\sum_{j'=1}^{0}(\cdot)=0$ and $\mathcal{D}_*^{(1)}=\emptyset$ for the highest ranked arm. □

### D.4 Proof of Corollary 8

The above corollary follows readily from Theorem 7. Let for agent $j'$ from 1 to $j-1$ the optimal arm be $j'$ with mean $1/2$ and all the other arms have mean $1/2-\Delta$, where $\Delta>0$ is small enough. Also, let the $j$-th agent have the arm means between $1/2$ for the $j$-th arm and $1/4$ for any other arm. For $\mathcal{P}$ the class of Bernoulli rewards, we have $D_{\inf}(\nu_{j'k_*^{(j)}},\mu_{j'k_*^{(j')}},\mathcal{P})\leq\Delta^2/4$ for all $j'\leq j-1$, and $\Delta_{\min}^{(j)}=1/4$. Therefore, the regret of the $j$-th agent is lower bounded as $\frac{(j-1)\log(T)}{16\Delta^2}$.

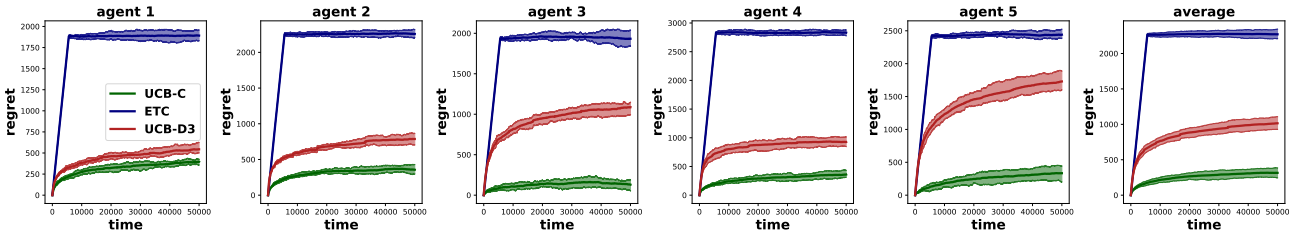
## E Additional Simulations

In this section, we compare our algorithm to both, the ETC based decentralized algorithm and the centralized UCB. The main conclusion is that, in both small and large systems, our algorithm outperforms the prior decentralized ETC algorithm [Liu et al., 2019] and is comparable to the centralized UCB algorithm of [Liu et al., 2020].

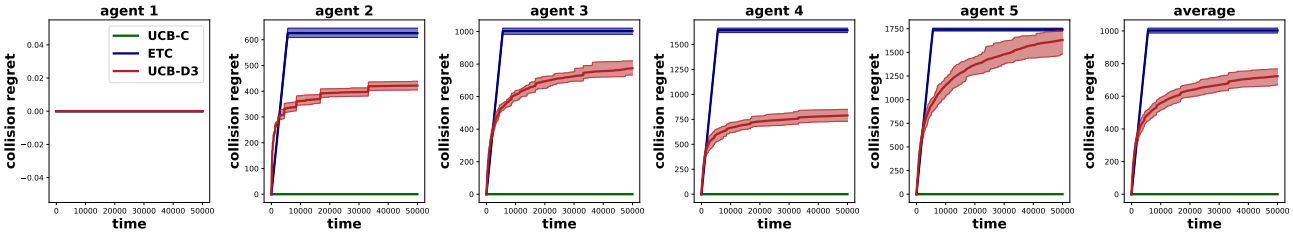
**Simulation Setup** We consider 4 systems – the first two systems are the OSB systems with 10 agents and 10 arms (Figure 4), and 10 agents, 15 arms (Figure 5). In both these systems, a random permutation $\sigma$ was first chosen and the arm-mean of arm $\sigma(i)$ for agent $i$ was set to 0.9. All other arm-means was chosen randomly and uniformly in $[0,0.8]$. We then consider two non OSB systems with 5 agents and 7 arms (Figure 3) and 10 agents and 15 arms (Figure 6). In these two systems, every agent $j$, uniformly spaces the arm-means between 0.1 and 0.9, with each agent having a random permutations over the arms to arrange the arm-means. All plots are averaged over 30 trials with confidence intervals of 95%. For brevity, Figures 4,5 and 6 are in the Appendix.

**Comparison with other Algorithms** - In Figures 4a, 5a, 3a and 6a, we plot the regret of all agents, for the three algorithms. We observe that UCB-D3 outperforms ETC and is slightly poorer compared to the centralized UCB algorithm. The centralized UCB has no collisions as a central arbiter matches agents and arms in the centralized UCB, and thus the regret is expected to be lower than any decentralized algorithm, which incurs some collisions. Figures 4b, 5b, 3b, 6b highlight this, where we plot the regret incurred by all algorithms only on account of collisions. The collisions incurred in our algorithm are lower compared to the decentralized ETC algorithm, thereby incurring lower regret compared to ETC. Although for a few high ranked agents, ETC has lower collisions (Fig. 5b), the overall regret of agents is lower with UCB-D3 algorithm as opposed to ETC. The deletion of dominant arms plays a key role, which enables our algorithm to have reduced collisions and thus lower regret.

**Equilibrium Freezing of UCB-D3** - In Figures 4c, 5c, 3c and 6c, we plot a 'heatmap' of the arms recommended by the agents over the 13 phases. The darker the shade, the higher the frequency (over the different simulation runs), that a particular agent recommended a particular arm in a particular phase. We observe from Figure 5c that after a random phase, all agent always recommend their stable match partner arm. Moreover, the time for an agent to settle into the 'equilibrium' of always recommending their estimated stable match partner arm is larger for lower ranked agents. Nevertheless, Figure 5c shows that after a random time, the agents delete their dominated arms thereby "freezing the system into an equilibrium".



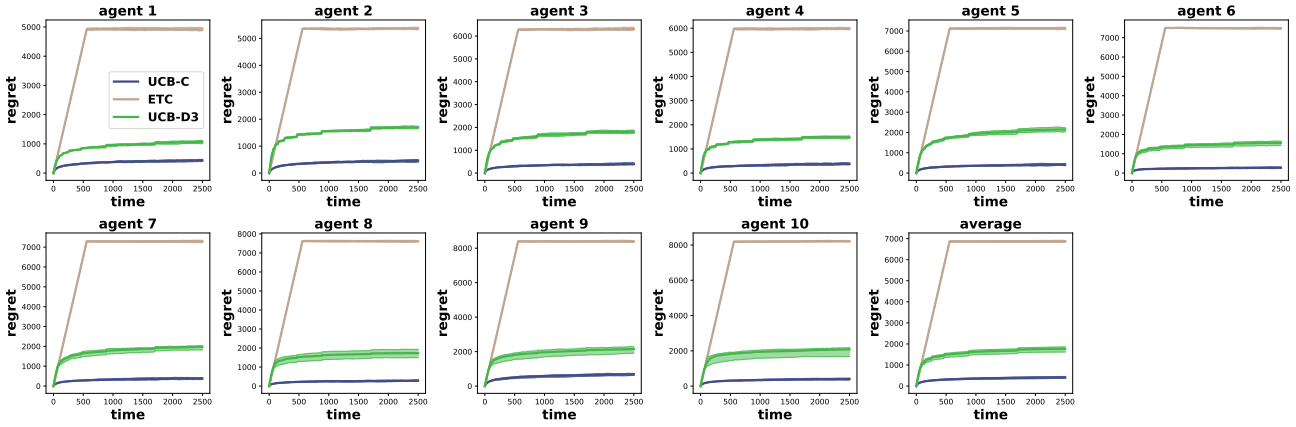

(a) Regret plots of all agents



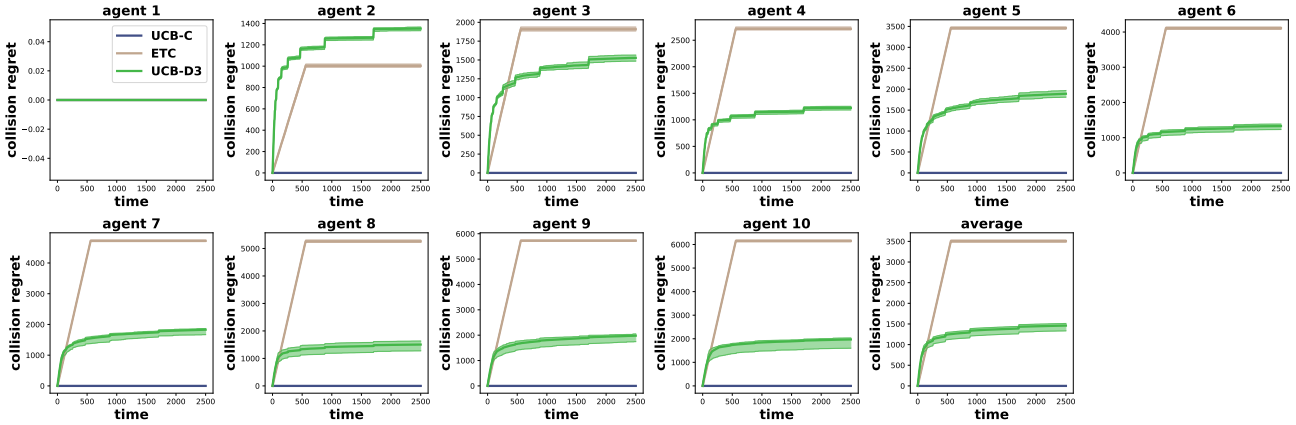

(b) A plot showing the cumulative regret only due to collisions. The centralized UCB ensures that agents never collide and thus do not lose out on regret.


(c) Arms recommended by the agents across phases over different runs of the algorithm.

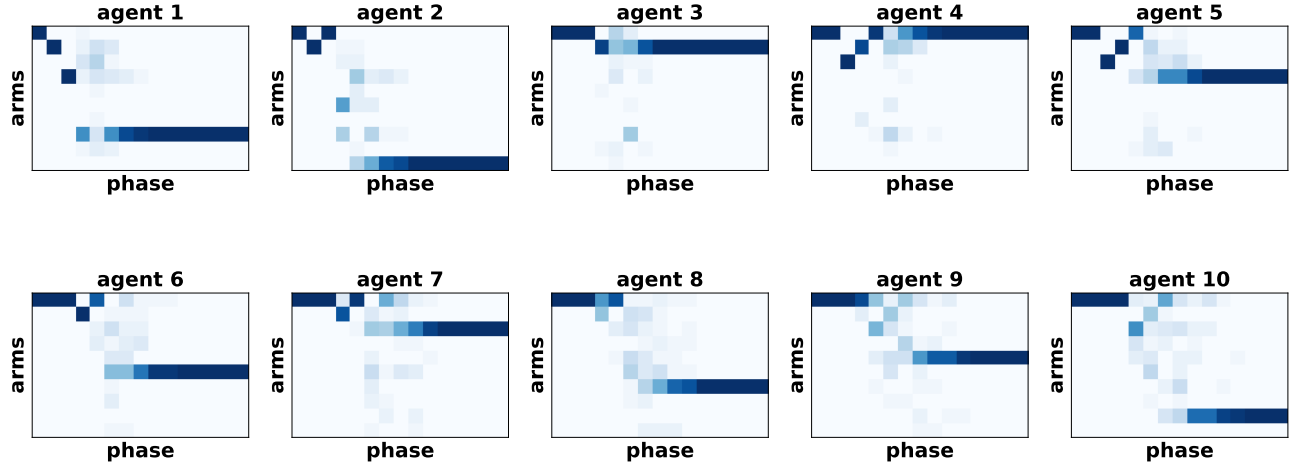
Figure 3: Simulations on a system with 5 agents and 7 arms. For each agent $i \in [5]$, a permutation over the arms $\sigma_i$ was chosen, and the arm-means are equally spaced among the 7 arms from 0.1 to 0.9 in the increasing order of permutation. This is thus not a OSB instance. The rewards are binary. The value of $H = 801$ was used for ETC.




(a) Regret plots of all agents.

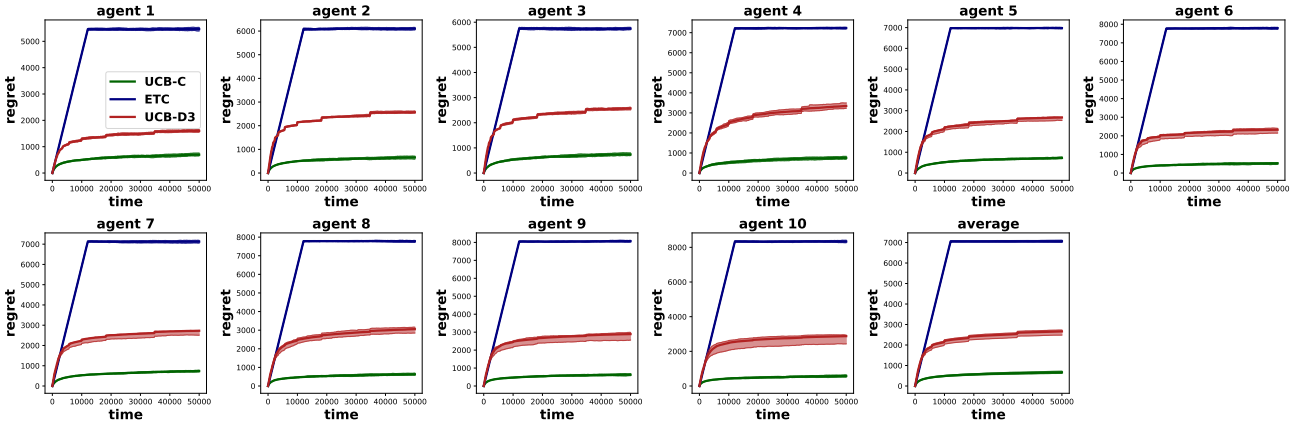



(b) A plot showing the cumulative regret only due to collisions. The centralized UCB ensures that agents never collide and thus do not lose out on regret.

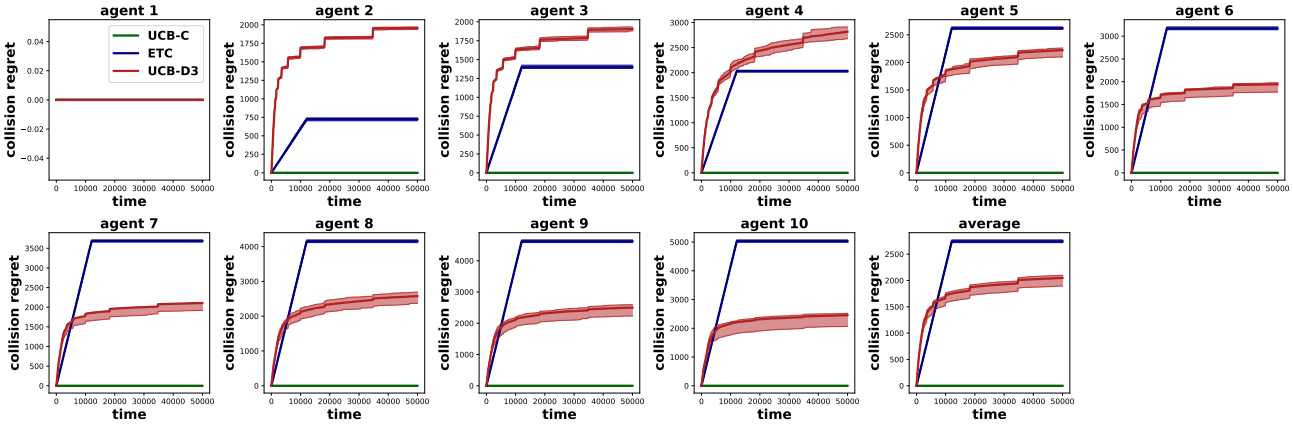



(c) Arms recommended by the agents across phases over different runs of the algorithm.

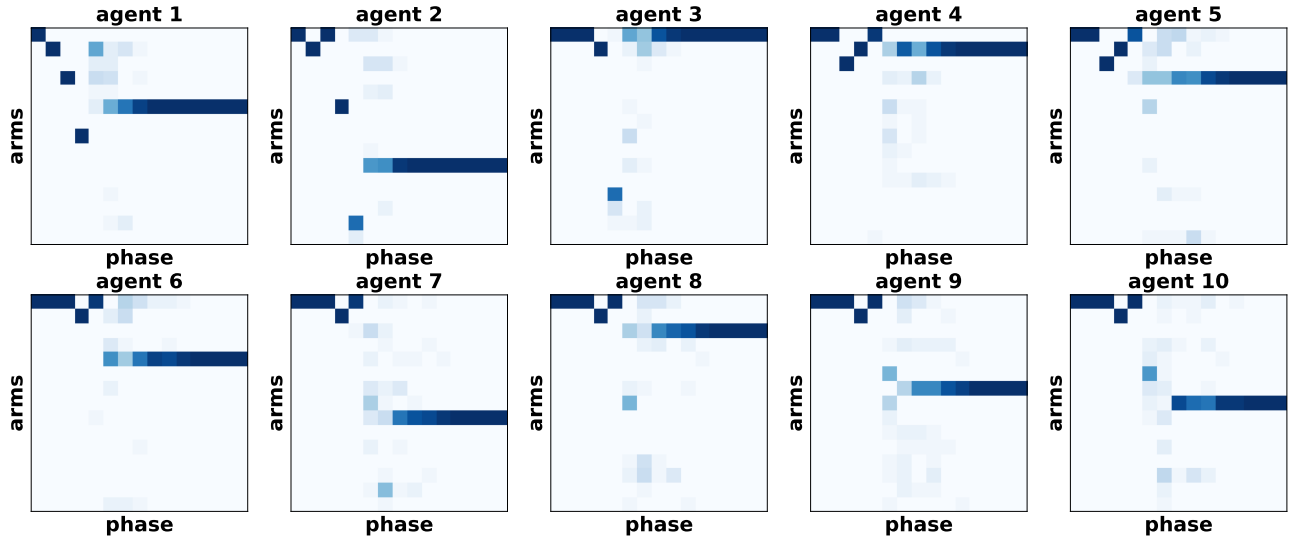
Figure 4: Simulations on a system with 10 agents and 10 arms. The arm-means for sub-optimal arms for each agent are chosen i.i.d. uniformly over $[0, 0.8]$, while the arm-mean of agent $i \in [10]$ for arm $\sigma(i)$ (its optimal stable match arm) was set to 0.9. The rewards are binary. Here, $\sigma(\cdot)$ denotes a permutation. This is thus a OSB instance. The value of $H = 1117$ used for ETC.




(a) Regret plots of all agents

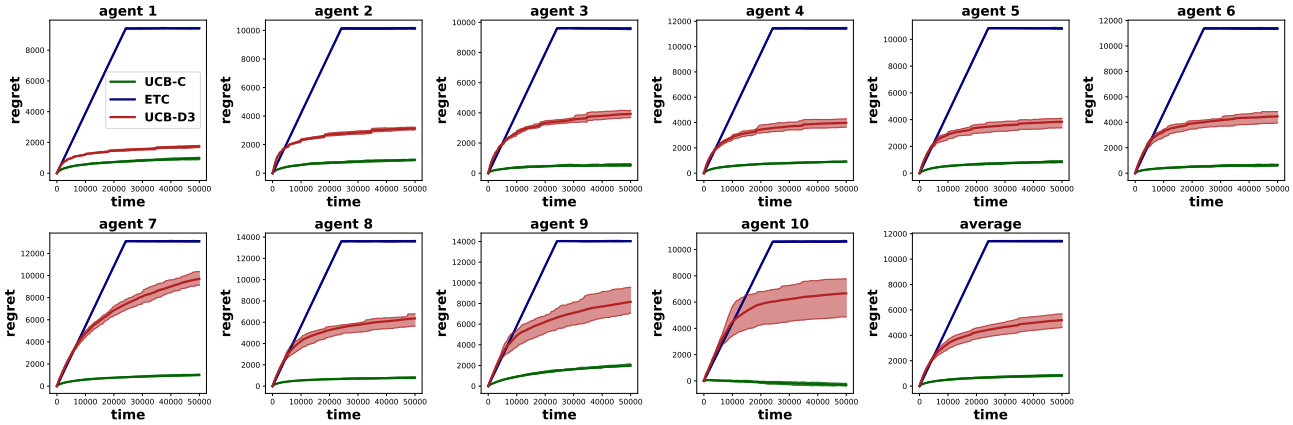



(b) A plot showing the cumulative regret only due to collisions. The centralized UCB ensures that agents never collide and thus do not lose out on regret.

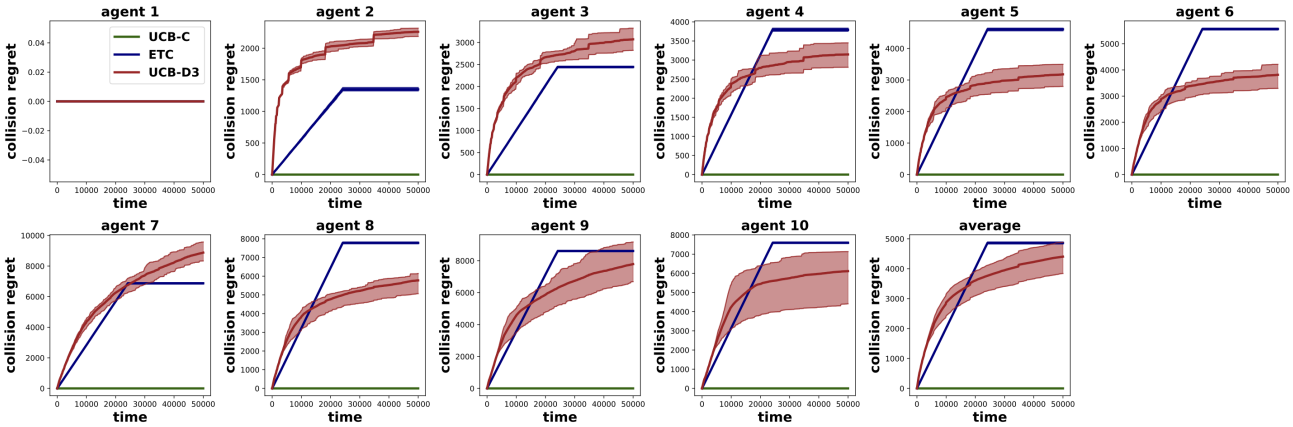



(c) Arms recommended by the agents across phases over different runs of the algorithm.

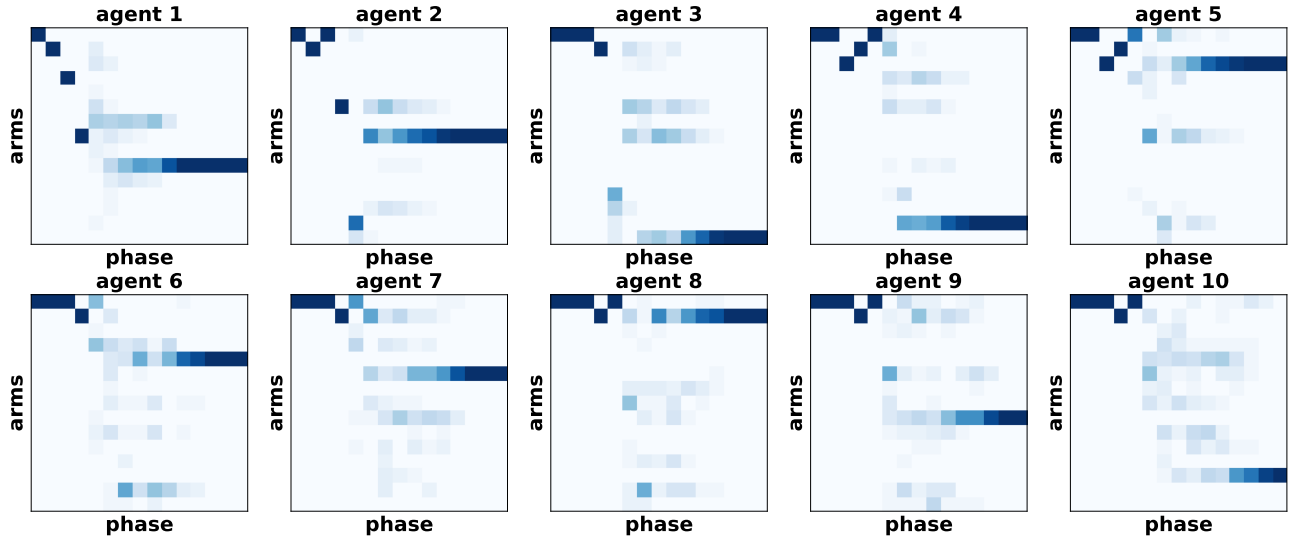
Figure 5: Simulations on a system with 10 agents and 15 arms. The arm-means for sub-optimal arms for each agent are chosen i.i.d. uniformly over $[0, 0.8]$, while the arm-mean for agent $i \in [10]$ and arm $\sigma(i)$ (stable match partner arm) was set to 0.9. The rewards are binary. Here, $\sigma(\cdot)$ denotes a permutation. This is thus a OSB instance. The value of $H = 805$ was used for ETC.




(a) Regret plots of all agents




(b) A plot showing the cumulative regret only due to collisions. The centralized UCB ensures that agents never collide and thus do not lose out on regret.




(c) Arms recommended by the agents across phases over different runs of the algorithm.

Figure 6: Simulations on a system with 10 agents and 15 arms. For each agent $i \in [10]$, a permutation over the arms $\sigma_i$ was chosen, and the arm-means are equally spaced among the 7 arms from 0.1 to 0.9 in the increasing order of permutation. This is thus not a OSB instance. The rewards are binary. The value of $H = 1610$ was used for ETC.