A. Proofs

We will use the following standard concentration bounds:

Let $X_1, ..., X_m$ be independent random variables such that X_i always lies in the interval $[0, 1]$. Define $\overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$ and $\mu = \mathbb{E}[\overline{X}]$.

Theorem 5. *(Chernoff Bound) For any* $0 < \delta < 1$,

$$\Pr\left(\sum_{j=1}^{m} X_j \geq (1-\delta)\cdot \mathbb{E}[\sum_{j=1}^{m} X_j]\right) \leq \exp\left(-\frac{\mathbb{E}[\sum_{j=1}^{m} X_j]\delta^2}{2}\right)$$

Theorem 6. *(Hoeffding's Inequality) For any* $\delta > 0$,

$$\Pr\left(|\overline{X} - \mu| \geq \delta\right) \leq 2 \cdot \exp\left(-2 \cdot m \cdot \delta^2\right)$$

A.1. MC algorithm proofs

We prove the following lemmas from which Theorem 1 follows.

Lemma 2. $\forall \epsilon > 0$ *and* $0 < \delta < 1$, *then after* $T_0 = \left\lceil \frac{16\cdot K}{\epsilon^2} \cdot \ln\left(\frac{4\cdot K^2}{\delta}\right)\right\rceil$ *rounds of random exploration, all players have an* ϵ*-correct ranking of the arms w.p.* $\geq 1 - \delta$

Lemma 3. *Let* $\delta \in [0, 1]$. *For* $\epsilon_1 = \frac{0.1}{K}$, *if the number of rounds* T_0 *used to estimate* N *is at least* $\left\lceil \frac{\log\left(\frac{2}{\delta}\right)}{2{\epsilon_1}^2}\right\rceil$, *then w.p.* $\geq 1 - \delta$ *we have that* $N^* = N$.

Lemma 4. *Denote by* R^F *the regret accumulated due to players running the musical chairs subroutine. Conditioned on the event that all* N *players learned an* ϵ*-correct ranking and that* $N^* = N$, *it holds that the expected value of* R^F *is at most* $2 \cdot \exp(2) \cdot N^2$.

A.1.1. Proof of Lemma 2

At a high level, the proof proceeds as follows: We find the value of the number of observations of each arm, denoted by C, required for ensuring all players learn an ϵ-correct ranking w.h.p. We use a union bound over all bad events of learning an inaccurate estimate of the expected reward of each one of the arms, and Hoeffding's inequality to upper bound the probability for each estimate to be far from its true mean given that the player has seen at least C observations of each arm. We then use a union bound over the N players, bounding the probability of the event that some player did not learn an ϵ-correct ranking. The last step is finding the value of T_0, the number of rounds of random exploration, required to obtain at least C observations of each arm for each player w.h.p.

We now turn to the formal proof. Given δ we define $\delta_1 := \frac{\delta}{2}$ and $\delta_2 := \frac{\delta}{2}$. Note that if for player i it is true that $\forall j \in \{1, ..., K\}$ $|\tilde{\mu}_j - \mu^j| \leq \frac{\epsilon}{2}$ then player i must have an ϵ-correct ranking.
We now calculate what is the required number of observations, C, of each arm in order to get

$$\Pr\left(\text{player } i \text{ doesn't have an } \epsilon\text{-correct ranking} \mid \text{player } i \text{ viewed} \geq C \text{ observations of each arm}\right) < \frac{\delta_1}{N}.$$

Specifically, we have the following:

$$\begin{aligned}
&\Pr\left(\text{player } i \text{ does not have an } \epsilon\text{-correct ranking} \mid \text{player } i \text{ viewed} \geq C \text{ observations of each arm}\right)\\
&\leq \Pr\left(\exists j \text{ s.t. } |\tilde{\mu}_j - \mu^j| > \frac{\epsilon}{2} \mid \text{player } i \text{ viewed} \geq C \text{ observations of each arm}\right)\\
&\underbrace{\leq}_{\text{union bound}} \sum_{j=1}^{K} \Pr\left(|\tilde{\mu}_j - \mu^j| > \frac{\epsilon}{2} \mid \text{player } i \text{ viewed} \geq C \text{ observations of each arm}\right)\\
&= \sum_{j=1}^{K}\sum_{n=C}^{\infty} \Pr\left(|\tilde{\mu}_j - \mu^j| > \frac{\epsilon}{2} | \text{ \# of views } = n\right) \Pr\left(\text{ viewed } n | n >= C\right)
\end{aligned}$$

Using Hoeffding's inequality, this is at most

$$\underbrace{\leq}_{\text{Hoeffding's Inequality}} \sum_{j=1}^{K}\sum_{n=C}^{\infty} 2\cdot\exp\left(\left(\frac{-n\cdot\epsilon^2}{2}\right)\right)\Pr\left(\text{ viewed } n|n>=C\right)$$

$$\leq \sum_{j=1}^{K} 2\cdot\exp\left(\left(\frac{-C\cdot\epsilon^2}{2}\right)\right)\sum_{n=C}^{\infty}\Pr\left(\text{ viewed } n|n>=C\right)$$

$$= K\cdot 2\cdot\exp\left(\left(\frac{-C\cdot\epsilon^2}{2}\right)\right)$$

Notice that we can apply Hoeffding's inequality here since each observation of the reward of an arm is sampled independent of the number of times we view it. This is true since every player is sampling uniformly at random at every round of learning (and independent of all previous rounds).
In order for this to be $< \frac{\delta_1}{N}$ we need:

$$2\cdot K\cdot\exp\left(-C\cdot\frac{\epsilon^2}{2}\right) < \frac{\delta_1}{N}$$

$$\implies C > \ln\left(\frac{2\cdot K\cdot N}{\delta_1}\right)\cdot\frac{2}{\epsilon^2}$$

Now we show that if all players have at least $C > \ln\left(\frac{2\cdot K\cdot N}{\delta_1}\right)\cdot\frac{2}{\epsilon^2}$ observations of each arm then w.p. $\geq 1-\delta_1$ all players have an ϵ-correct ranking:
We start by defining the following events:

- A will denote the event that all players have an ϵ-correct ranking ($\overline{A}$ will denote A complement)
- A_i will denote the event that player i has an ϵ-correct ranking
- B will denote the event that all players have observed each arm at least C times ($\overline{B}$ will denote B complement)
- B_i will denote the event that player i has observed each arm at least C times

$$\Pr\left(A|B\right)$$

$$\geq 1-\Pr\left(\bigvee_i \overline{A}_i|B_i\right)$$

$$\underbrace{\geq}_{\text{union bound}} 1-\sum_{i=1}^{N}\Pr\left(\overline{A}_i|B_i\right)$$

$$\geq 1-N\cdot\frac{\delta_1}{N} = 1-\delta_1$$

Now we show that there exists a T_0 large enough so that all players have $> C$ observations of each arm w.p. $\geq 1-\delta_2$.
We define $A_{i,j}(t) = I\left\{\text{player } i \text{ observed arm } j \text{ at round } t\right\}$.
Note that for any round t and any i,j we have that $\Pr\left(A_{i,j}(t)=1\right) = \frac{1}{K}\cdot\left(1-\frac{1}{K}\right)^{N-1} \implies \mathbb{E}\left[A_{i,j}(t)\right] = \frac{1}{K}\cdot\left(1-\frac{1}{K}\right)^{N-1}$.

So for any i, j we have that

$$\begin{aligned}
&\Pr\left(\text{player } i \text{ has} \le \tfrac{1}{2}\cdot T_0 \cdot \mathbb{E}\left[A_{i,j}(t)\right] \text{ observations}\right)\\
&= \Pr\left(\sum_{t=1}^{T_0} A_{i,j}\left(t\right) \le \frac{1}{2}\cdot T_0 \cdot \mathbb{E}\left[A_{i,j}(t)\right]\right)\\
&\underbrace{\le}_{\text{Chernoff bound}} \exp\left(\frac{-\frac{1}{4}\cdot T_0 \cdot \mathbb{E}\left[A_{i,j}(t)\right]}{2}\right)
\end{aligned}$$

Note that we can apply Chernoff bound here since for any i,j, $A_{i,j}$ are i.i.d across t, since all players are employing random sampling at every round of learning.
Using a union bound we get that:

$$\begin{aligned}
&\Pr\left(\exists i, j s.t. \sum_{t=1}^{T_0} A_{i,j}\left(t\right) \le \frac{1}{2}\cdot T_0 \cdot \mathbb{E}\left[A_{i,j}(t)\right]\right)\\
&\le N \cdot K \cdot \exp\left(\frac{-\frac{1}{4}\cdot T_0 \cdot \mathbb{E}\left[A_{i,j}(t)\right]}{2}\right)
\end{aligned}$$

In order for this probability to be upper bounded by δ_2 we need:

$$\begin{aligned}
&N \cdot K \cdot \exp\left(\frac{-\frac{1}{4}\cdot T_0 \cdot \mathbb{E}\left[A_{i,j}(t)\right]}{2}\right) < \delta_2\\
&\implies T_0 > \frac{1}{8\cdot \mathbb{E}\left[A_{i,j}(t)\right]} \cdot \ln\left(\frac{N\cdot K}{\delta_2}\right)
\end{aligned}$$

We have shown that if $T_0 > \frac{1}{8\cdot\mathbb{E}[A_{i,j}(t)]}\cdot \ln\left(\frac{N\cdot K}{\delta_2}\right)$ then w.p. $\ge 1-\delta_2$ we have $\forall i, j$ the number of observations player i has of arm j, $\sum_{t=1}^{T_0} A_{i,j}\left(t\right)$, $> \frac{1}{2}\cdot T_0 \cdot \mathbb{E}\left[A_{i,j}(t)\right]$.
We also need the total number of observations each player has of each arm to be at least C,
i.e.

$$\begin{aligned}
&\sum_{t=1}^{T_0} A_{i,j}\left(t\right) > \frac{1}{2}\cdot T_0 \cdot \mathbb{E}\left[A_{i,j}(t)\right] \ge C > \ln\left(\frac{2\cdot K \cdot N}{\delta_1}\right)\cdot\frac{2}{\epsilon^2}\\
&\implies T_0 \ge 2\cdot\frac{1}{\mathbb{E}\left[A_{i,j}(t)\right]}\cdot \ln\left(\frac{2\cdot K\cdot N}{\delta_1}\right)\cdot\frac{2}{\epsilon^2}
\end{aligned}$$

So we have two constraints on T_0, thus we take:
$T_0 = \left\lceil \max\left\{\frac{1}{8\cdot\mathbb{E}[A_{i,j}(t)]}\cdot\ln\left(\frac{N\cdot K}{\delta_2}\right), 2\cdot\frac{1}{\mathbb{E}[A_{i,j}(t)]}\cdot\ln\left(\frac{2\cdot K\cdot N}{\delta_1}\right)\cdot\frac{2}{\epsilon^2}\right\}\right\rceil$.
We remind the reader that $\mathbb{E}\left[A_{i,j}(t)\right] = \frac{1}{K}\cdot\left(1-\frac{1}{K}\right)^{N-1} \ge \frac{1}{K}\cdot\frac{1}{4}$ for all $K > 1$.
So we take $T_0 = \left\lceil \max\left\{\frac{K}{2}\cdot\ln\left(\frac{N\cdot K}{\delta_2}\right), \frac{16\cdot K}{\epsilon^2}\cdot\ln\left(\frac{2\cdot N\cdot K}{\delta_1}\right)\right\}\right\rceil$ and the result holds. Using the events A and B as defined above, we get that

$$\begin{aligned}
&\Pr\left(A\right) = 1 - \Pr\left(\overline{A}\right)\\
&= 1 - \left(\Pr\left(\overline{A}|B\right)\cdot\Pr\left(B\right) + \Pr\left(\overline{A}|\overline{B}\right)\cdot\Pr\left(\overline{B}\right)\right)\\
&\ge 1 - \left(\Pr\left(\overline{A}|B\right) + \Pr\left(\overline{B}\right)\right)\\
&\ge 1 - \left(\delta_1 + \delta_2\right) \ge 1 - \delta
\end{aligned}$$

Notice that letting $T_0 = \left\lceil \max\left\{\frac{K}{2}\cdot\ln\left(\frac{N\cdot K}{\delta_2}\right), \frac{16\cdot K}{\epsilon^2}\cdot\ln\left(\frac{2\cdot N\cdot K}{\delta_1}\right)\right\}\right\rceil$ is only possible if one knows N. If N is unknown, then one can increase T_0 and set it to

$$T_0 = \left\lceil \max\left\{\frac{K}{2}\cdot\ln\left(\frac{K^2}{\delta_2}\right), \frac{16\cdot K}{\epsilon^2}\cdot\ln\left(\frac{2\cdot K^2}{\delta_1}\right)\right\}\right\rceil,$$

and the lemma would still hold since we assume that $N < K$.

Since $\forall i \in [1, K] : \mu^i \in [0, 1]$, we get that $\epsilon^2 \leq 1$. In addition, as defined initially: $\delta_1 = \delta_2 = \frac{\delta}{2}$. Thus the second term in this max expression always dominates the first, thus we get:

$$T_0 = \left\lceil \frac{16 \cdot K}{\epsilon^2} \cdot \ln\left(\frac{4 \cdot K^2}{\delta}\right) \right\rceil,$$

A.1.2. PROOF OF LEMMA 3

We start by outlining the main points of the proof. We use the empirical estimate of the collision probability, $\hat{p}$, generated by counting the number of collisions occurred during t rounds divided by t, in order to extract the empirical estimate of the number of players, N^*. We then show that in order for the estimated number of players to be correct, the difference between the empirical estimate of the collision probability, $\hat{p}$, and the true probability, p, needs to be smaller than some value, which we denote ϵ_1. We then use Hoeffding's inequality to upper bound the event that $\hat{p}$ is farther than ϵ_1 from p after t rounds of exploration, and find the value of t that is ensures this probability is smaller than δ. Taking the number of exploration rounds, T_0, to be greater or equal to this t, will give us the correct estimate of the number of players. The last segment only simplifies the expression of ϵ_1 to a more readable term.

We now turn to the detailed proof. Fix some player i, and let C_t be the number of collisions observed by the player until time t. Also, let p be the true probability of a collision, when N players are choosing arms uniformly at random among K arms. The probability of a player not experiencing a collision is

$$\begin{aligned}\Pr(\text{no collision}) &= \sum_{j=1}^{K} \Pr(\text{choose arm j}) \cdot \Pr(\text{no other player chooses arm j}) \\ &= \sum_{j=1}^{K} \frac{1}{K} \cdot \left(1 - \frac{1}{K}\right)^{N-1} \\ &= \frac{1}{K} \cdot K \cdot \left(1 - \frac{1}{K}\right)^{N-1} = \left(1 - \frac{1}{K}\right)^{N-1}\end{aligned}$$

Thus the probability of a collision at any round of learning is : $p = 1 - \left(1 - \frac{1}{K}\right)^{N-1}$. Note that $p < 1$ for any $N, K > 0$. Inverting this equation, we get

$$N = \frac{\log(1-p)}{\log\left(1 - \frac{1}{K}\right)} + 1.$$

Therefore, if we let $\hat{p}_t := \frac{C_t}{t}$ be the empirical estimate of the collision probability after t rounds, it is natural to take the estimator defined as

$$N^* \; = \; \text{round}\left(\frac{\log\left(1 - \hat{p}_t\right)}{\log\left(1 - \frac{1}{K}\right)} + 1\right) \; = \; \text{round}\left(\frac{\log\left(\frac{t - C_t}{t}\right)}{\log\left(1 - \frac{1}{K}\right)} + 1\right)$$

Our goal will be to show that when t is sufficiently large, $N^* = N$ with arbitrarily high probability. Specifically, we will upper bound the probability of the estimator $\frac{\log\left(\frac{t-C_t}{t}\right)}{\log\left(1-\frac{1}{K}\right)} + 1$ being far from the true value N (which also includes the unlikely case $C_t = t$, in which case the estimator is infinite).

Recalling that $N = \frac{\log(1-p)}{\log\left(1-\frac{1}{K}\right)} + 1$, and the definition of N^*, to ensure that $N^* = N$ it is enough to require

$$\left| \frac{\log\left(1 - \hat{p}_t\right)}{\log\left(1 - \frac{1}{K}\right)} - \frac{\log(1-p)}{\log\left(1 - \frac{1}{K}\right)} \right| \leq \gamma$$

for some $\gamma < 1/2$, which is equivalent to requiring

$$\left| \frac{\log\left(\frac{1-\hat{p}_t}{1-p}\right)}{\log\left(1 - \frac{1}{K}\right)} \right| \leq \gamma.$$

Let β denote the actual difference between $\hat{p}_t$ and p, so that $\hat{p}_t = p + \beta$. Therefore, the above is equivalent to

$$\begin{aligned} -\gamma \le \frac{\log\left(\frac{1-p-\beta}{1-p}\right)}{\log\left(1-\frac{1}{K}\right)} \le \gamma \\ \iff \quad \gamma \log\left(1-\frac{1}{K}\right) \le \log\left(\frac{1-p-\beta}{1-p}\right) \le -\gamma \log\left(1-\frac{1}{K}\right) \\ \iff \quad \left(1-\frac{1}{K}\right)^{\gamma} \le \frac{1-p-\beta}{1-p} \le \left(1-\frac{1}{K}\right)^{-\gamma} \\ \iff \quad (1-p)\left(1-\frac{1}{K}\right)^{\gamma} \le 1-p-\beta \le (1-p)\left(1-\frac{1}{K}\right)^{-\gamma} \\ \iff \quad -1+p+(1-p)\left(1-\frac{1}{K}\right)^{\gamma} \le -\beta \le -1+p+(1-p)\left(1-\frac{1}{K}\right)^{-\gamma} \\ \iff \quad 1-p-(1-p)\left(1-\frac{1}{K}\right)^{-\gamma} \le \beta \le 1-p-(1-p)\left(1-\frac{1}{K}\right)^{\gamma} \\ \iff \quad (1-p)\cdot\left(1-\left(1-\frac{1}{K}\right)^{-\gamma}\right) \le \beta \le (1-p)\cdot\left(1-\left(1-\frac{1}{K}\right)^{\gamma}\right) \end{aligned}$$

Therefore, if we can ensure that $|\hat{p}_t - p| \le \epsilon_1$, where

$$\epsilon_1 = \min\left\{\left|(1-p)\cdot\left(1-\left(1-\frac{1}{K}\right)^{-\gamma}\right)\right|, \left|(1-p)\cdot\left(1-\left(1-\frac{1}{K}\right)^{\gamma}\right)\right|\right\}$$

for some $\gamma < 1/2$ (say 0.49), we get that $N^* = N$ as required. If t is sufficiently large, this can be done using Hoeffding's inequality: $\hat{p}_t$ is an average of t i.i.d. random variables with expectation p, hence with probability at least $1-\delta$, $|\hat{p}_t - p| \le \epsilon_1$ provided that $t \ge \frac{\log(2/\delta)}{2\epsilon_1^2}$.

We now replace the expression of ϵ_1 above, which is a bit unwieldy, with a simpler lower bound (where we also take $\gamma = 0.49$). First, plugging in the expression for p, we get

$$\epsilon_1 = \min\left\{\left|\left(\left(1-\frac{1}{K}\right)^{N-1}\cdot\left(1-\left(1-\frac{1}{K}\right)^{-0.49}\right)\right)\right|, \left|\left(\left(1-\frac{1}{K}\right)^{N-1}\cdot\left(1-\left(1-\frac{1}{K}\right)^{0.49}\right)\right)\right|\right\}$$

We first lower bound the first expression:

$$\begin{aligned} \left|\left(\left(1-\frac{1}{K}\right)^{N-1}\cdot\left(1-\left(1-\frac{1}{K}\right)^{-0.49}\right)\right)\right| = -\left(\left(1-\frac{1}{K}\right)^{N-1}\cdot\left(1-\left(1-\frac{1}{K}\right)^{-0.49}\right)\right) \\ \ge \left(1-\frac{1}{K}\right)^{K-1}\cdot\left(-\left(1-\left(1-\frac{1}{K}\right)^{-0.49}\right)\right) \ge \frac{1}{\exp(1)}\cdot\left(-\left(1-\left(1-\frac{1}{K}\right)^{-0.49}\right)\right) \end{aligned}$$

We use a Taylor expansion to lower bound $\left(-1+\left(1-\frac{1}{K}\right)^{-0.49}\right)$: Considering $f(x) = \left(-1+(1-x)^{-0.49}\right)$, the first derivative is $f'(x) = 0.49\cdot(1-x)^{-1.49}$ and the second derivative of $f(x)$ is $f''(x) = 0.49\cdot 1.49\cdot(1-x)^{-2.49}$, which is non-negative for any $x \in [0,1]$. Therefore, $f(x) \ge f(0) + f'(0)\cdot x = 0.49\cdot x$ for any $x \in [0,1]$, and replacing x with $1/K$ we get that $-\left(1-\left(1-\frac{1}{K}\right)^{-0.49}\right) \ge \frac{0.49}{K}$

Similarly, we lower bound the second expression:

$$\left|\left(\left(1-\frac{1}{K}\right)^{N-1}\cdot\left(1-\left(1-\frac{1}{K}\right)^{0.49}\right)\right)\right| = \left(\left(1-\frac{1}{K}\right)^{N-1}\cdot\left(1-\left(1-\frac{1}{K}\right)^{0.49}\right)\right)$$
$$\geq \left(1-\frac{1}{K}\right)^{K-1}\cdot\left(1-\left(1-\frac{1}{K}\right)^{0.49}\right) \geq \frac{1}{\exp(1)}\cdot\left(1-\left(1-\frac{1}{K}\right)^{0.49}\right)$$

We use Taylor expansion again to lower bound $\left(1-\left(1-\frac{1}{K}\right)^{0.49}\right)$. We look at the function: $f(x) = 1-(1-x)^{0.49}$. The first derivative of $f(x)$ is $f'(x) = 0.49\cdot(1-x)^{-0.51}$ and the second derivative of $f(x)$ is $f''(x) = 0.49\cdot 0.51\cdot(1-x)^{-1.51}$. Note that $\forall x \in [0,1] : f''(x) \geq 0$. Thus we get: $f(x) \geq f(0) + f'(0)\cdot x = 0.49\cdot x$. Thus the lower bound is again: $\left|\left(\left(1-\frac{1}{K}\right)^{N-1}\cdot\left(1-\left(1-\frac{1}{K}\right)^{0.49}\right)\right)\right| \geq \frac{0.49}{K\cdot\exp(1)}$

Combining the above, we showed that

$$\epsilon_1 \geq \frac{0.49}{\exp(1)\cdot K} \geq \frac{0.1}{K}.$$

Taking this value for ϵ_1, we get that if we run the learning phase for $\left\lceil\frac{\log\left(\frac{2}{\delta_2}\right)}{2{\epsilon_1}^2}\right\rceil$ rounds, then w.p. $\geq 1-\delta_2$, we have that $N^* = N$.

A.1.3. Proof of Lemma 4

We begin with stating the main ideas in the proof: We calculate a lower bound on the probability for any player to become fixed on an arm at any round after the beginning of the musical chairs subroutine. We use this probability to get a bound on the expected time it takes any player to become fixed. We then calculate the total expected regret accumulated by players during this time.

The proof proceeds as follows: we remind the reader that the musical chairs phase is when a set of N players who, with high probability, have learned an ϵ-correct ranking each choose an arm uniformly at random from the best N arms and stay 'fixed‘ on that arm until the end of the epoch or game. Thus once a player has 'fixed‘, the only case in which she can contribute regret is if another non-fixed player collides with her.

We will denote by N to be the number of players starting the musical chairs phase. let r be the number of rounds since the start of the musical chairs phase (i.e. when the musical chairs phase starts, $r = 0$).

Denote by T_f the time it takes for one player running the musical chairs subroutine to become fixed. We will first bound T_f.

N_m will denote the maximum number of players (if the game has a dynamic player setting rather than static).

We start by fixing some player who is running the musical chairs subroutine. We will denote by v_t the number of players that entered late and are not running the musical chairs subroutine, rather they are choosing arms uniformly at random. For any round t after the musical chairs phase begins the probability for this player to become fixed is at least:

$$\sum_{\text{all unfixed arms}} \frac{1}{N}\cdot\left(1-\frac{1}{N}\right)^{N-1}\cdot\left(1-\frac{1}{K}\right)^{v_t} \geq \frac{1}{N}\cdot\left(1-\frac{1}{N}\right)^{N-1}\cdot\left(1-\frac{1}{K}\right)^{N_m-N}$$

In the above expression the first term is the probability that the player we are considering chooses the specific arm in the summation. The second term is a lower bound on the probability that all players who are running the musical chairs subroutine and are unfixed do not choose that specific arm since the actual probability is $(1-\frac{1}{N})^{N_r}$, where N_r is the number of player remaining unfixed, which is greater than or equal to $(1-\frac{1}{N})^{N}$. The third term is a lower bound on the probability that all players who entered late in the epoch, who choose arms from all K arms uniformly at random, do not choose the specified arm. In the static setting this term would be 1.

We use the convention that $(1-\frac{1}{N})^{N-1}$ equals 1 when N is 1 since in this case the probability that the specified player becomes fixed is 1 times the probability that no new player (who is exploring randomly) chooses the last unavailable arm.

We continue bounding the probability that at any round t some player running the musical chairs subroutine, who has not fixed already, becomes fixed:

$$\frac{1}{N} \cdot \left(1 - \frac{1}{N}\right)^{N-1} \cdot \left(1 - \frac{1}{K}\right)^{N_m - N} \geq \frac{1}{N} \cdot \left(1 - \frac{1}{N}\right)^{N-1} \cdot \left(1 - \frac{1}{N_m}\right)^{N_m - 1}$$

In the above inequality we use the fact that $N_m \leq K$.

Now we use the fact that $(1 - \frac{1}{x})^{x-1} \geq \frac{1}{\exp(1)}$ for $x \geq 1$ to get that

$$\frac{1}{N} \cdot \left(1 - \frac{1}{N}\right)^{N-1} \cdot \left(1 - \frac{1}{N_m}\right)^{N_m - 1} \geq \frac{1}{\exp{(2)}} \cdot \frac{1}{N}$$

Now we have established that for any player running the musical chairs subroutine, who has not become 'fixed' yet, the probability, at any round t, to become fixed in the next round is at least $\frac{1}{\exp(2)} \cdot \frac{1}{N}$. So for any player, from the moment they begin running the musical chairs subroutine the expected time it takes to become fixed is at most the expected number of times flipping a biased coin with probability $\frac{1}{\exp(2)} \cdot \frac{1}{N}$, which is $\exp{(2)} \cdot N$.

Thus the expected time it takes any player to fix is at most $\exp{(2)} \cdot N$.

Notice that after the learning phase in every collision there is at most one fixed player. The regret, at any round after the learning phase, is bounded by two times the number of unfixed players. Therefore the total regret accumulated due to any player running the musical chairs subroutine is bounded by two times the time is takes her to 'fix'. Denote by T_f^i the time it takes player i to 'fix' on an arm. The total expected regret accumulated by players running the musical chairs subroutine is: $\leq \mathbb{E}\left[\sum_{i=1}^{N} 2 \cdot T_f^i\right]$ Each player running the musical chairs subroutine can only contribute at most 2 to the

Let $r_{i,t}$ be an indicator variable which equals 1 if player i incurred regret at round t and 0 otherwise. We will denote by $F \subseteq [N]$ the set of players who started running the musical chairs subroutine and fixed on an arm and by $U \subseteq [N]$ the set of players who have not fixed yet.

We will now bound the expected regret due to players running the musical chairs subroutine. Notice that we do not include any regret due to players who entered late and are randomly exploring.

$$\begin{aligned}
&\leq \mathbb{E}\left[\sum_{i=1}^{N} \sum_{t=T_0+1}^{T} r_{i,t}\right] \leq \mathbb{E}\left[\sum_{t=T_0+1}^{T} \sum_{i \in U} r_{i,t} + \sum_{j \in F} r_{j,t}\right] \\
&\leq \mathbb{E}\left[\sum_{t=T_0+1}^{T} \sum_{i \in U} r_{i,t} + \sum_{i \in U} r_{i,t}\right] \leq \mathbb{E}\left[\sum_{t=T_0+1}^{T} 2 \cdot \sum_{i \in U} r_{i,t}\right] \\
&\leq \mathbb{E}\left[\sum_{t=T_0+1}^{T} 2 \cdot \sum_{i \in U} I_{\text{player } i \text{ not fixed at round } t}\right] \leq \mathbb{E}\left[\sum_{t=T_0+1}^{T} 2 \cdot N \cdot I_{\text{player } i \text{ not fixed at round } t}\right] \\
&\leq 2 \cdot N \cdot T_f
\end{aligned}$$

where the third inequality ($\sum_{j \in F} r_{j,t} \leq \sum_{i \in U} r_{i,t}$) is due to the fact that $\sum_{j \in F} r_{j,t}$ is upper bounded by the total number of collisions at round t (since there can only be one fixed player per collision) and the total number of collisions is upper bounded by $\sum_{i \in U} r_{i,t}$.

In conclusion, we have that the expected regret due to players who run the musical chairs subroutine is $\leq 2 \cdot T_f \cdot N = 2 \cdot \exp{(2)} \cdot N^2$

A.2. Analysis of DMC algorithm

Let T_1 denote the epoch length, $T_f = \mathbb{E}[F] \leq \exp{(2)} \cdot N_m$ be the expected time it takes any player to fix on an arm during the 'Musical Chairs' period, $C = T_0 + T_f$, and N_m be the maximum number of active players at any given round.

A.2.1. Proof of Theorem 2

At a high level, the proof of this theorem proceeds as follows: We take the bound we get from Lemma 1, and find the optimal epoch length T_1. This optimal value balances the regret due to random exploration that occurs in each new epoch, and the regret due to entering/leaving players that cause regret during the epoch, before restarting the MC algorithm.

Using lemma 1 we will now compute the optimal epoch length, T_1.

We have that the expected regret is

$$\begin{aligned} &\leq \frac{T}{T_1} \cdot (N_m \cdot (T_0 + 2 \cdot T_f)) + e \cdot N_m \left(T_1 - T_0\right) + l \left(T_1 - T_0\right) \\ &\leq \frac{T}{T_1} \cdot (K \cdot (T_0 + 2 \cdot T_f)) + e \cdot K \left(T_1 - T_0\right) + l \left(T_1 - T_0\right) \end{aligned}$$

We differentiate with respect to T_1 and set equal to zero to find the optimal value of T_1 (up to rounding):
$T_1 = \left\lceil \sqrt{\frac{T \cdot K \cdot (T_0 + 2 \cdot T_f)}{K \cdot e + l}} \right\rceil$ and since $e + l \leq x$ we take $T_1 = \left\lceil \sqrt{\frac{T \cdot (T_0 + 2 \cdot T_f)}{x}} \right\rceil$
Note that we require T_1 to be greater than T_0 (the epoch length must be at least as long as the learning period) . Thus if the optimal T_1 is less than T_0, we set $T_1 = T_0$.
Plugging T_1 into the regret bounds yields that the expected regret is $\leq O\left(\frac{T}{T_1} \cdot T_0 + x \cdot T_1\right)$ where
$T_1 = O\left(\sqrt{\frac{T \log(T)}{x}}\right)$ and $T_0 = O\left(\log\left(T\right)\right)$.

Thus we get that the expected regret is $\leq O\left(\frac{T}{\sqrt{\frac{T\log(T)}{x}}} \cdot \log(T) + x \cdot \sqrt{\frac{T \log(T)}{x}}\right)$
$= O\left(\sqrt{\frac{T \cdot x}{\log(T)}} \cdot \log(T) + x \cdot \sqrt{\frac{T\log(T)}{x}}\right) = O\left(\sqrt{Tx} \cdot \frac{\log(T)}{\sqrt{\log(T)}} + \sqrt{Tx} \cdot \sqrt{\log(T)}\right)$
$= \tilde{O}\left(\sqrt{Tx}\right)$ where the $\tilde{O}$ hides logarithmic factors in x, T, and x is an upper bound on the number of players entering and exiting.

A.2.2. Proof of Lemma 1

We start with an outline of the proof: we sum all sources of regret: regret due to random exploration and fixing stage in each epoch, regret due to leaving players that might leave high ranking arms unexploited, and regret due to entering players that explore and cause collisions. We then use the lemmas used in the proof of the MC algorithm, with a slight change in the probability of error, to account for the fact that we want to be successful in all epochs. This gives us the bound on the total expected regret.

We now turn to the detailed proof. We start by noting that the number of epochs is at most $\lceil \frac{T}{T_1} \rceil$, and we compute a bound on the expected regret per epoch and sum over the epochs using the linearity of expectation.
The regret per epoch is composed of three terms:

- Regret due to learning and fixing on an arm
- Regret due to entering players
- Regret due to leaving players

We will now compute each of these terms.

Regret due to learning and fixing on an arm For each epoch, we have at most N_m players who are learning for T_0 rounds and fixing on an arm, each one taking T_f rounds, in expectation. For every given round of learning or fixing an upper bound on the expected regret is N_m since the best expected reward minus the worst expected reward is less than 1. Also, as shown in lemma 4, the expected regret due to players who run the musical chairs subroutine is at most $2 \cdot \exp(2) \cdot N_m^2 = 2 \cdot T_f \cdot N_m$. This means that the total expected regret for this term, per epoch, is $\leq N_m \cdot (T_0 + 2 \cdot T_f)$.

Regret due to entering players We remind the reader that players cannot enter during the learning period. Since a newly entered player learns until the end of the epoch she contributes at most $T_1 - T_0$ regret due to learning. During this time period the player may also collide with other players who are already fixed or collide with players who have not finished fixing after the expected time it takes to fix. A newly entering player can collide with at most 1 other fixed player. Since we already count the regret of a any round of a non-fixed player as regret we only need to add regret added due to collisions with fixed players. Thus the total regret a single newly entering player can contribute is $2 \cdot (T_1 - T_0)$ and the total regret due to all newly entering players is at most $e_i \cdot 2 \cdot (T_1 - T_0)$ per epoch, where e_i is the number of players who enter at epoch i.

Regret due to leaving players We remind the reader that players cannot leave during the learning period.
Every player that leaves can contribute at most 1 regret at each round, since in the worst case this player causes the best arm to remain unused during the fixed period. This would mean that for every player that leaves, per epoch, we have at most $T_1 - T_0$ added regret. Denote by l_i the number of players that leave during epoch i ($\sum_{i=1}^{\frac{T}{T_1}} l_i = l$). For all players that leave in epoch i, this amounts to having an added regret of $l_i \cdot (T_1 - T_0)$.

Now summing over the all epochs we get that the total expected regret of the DMC algorithm is:

$$\begin{aligned} &\leq \sum_{i=1}^{\frac{T}{T_1}} \left(N_m \cdot (T_0 + 2T_f) + e_i \cdot N_m \left(T_1 - T_0\right) + l_i \left(T_1 - T_0\right)\right) \\ &= \frac{T}{T_1} \cdot \left(N_m \cdot (T_0 + 2T_f)\right) + e \cdot N_m \cdot (T_1 - T_0) + l \left(T_1 - T_0\right) \end{aligned}$$

For the regret bound to be correct, we need to ensure that players learn an ϵ-correct ranking and correctly estimate the number of players at every epoch. By using lemma 2 and lemma 3 with confidence parameters set to $\frac{\delta}{2 \cdot T}$, and taking the union bound over all epochs, we ensure that with high probability the players learn the true rankings and estimate the number of players correctly at each epoch, and thus we get that w.p. $\geq 1 - \delta$ we have that the regret bound holds. For this reason T_0 includes a $\log(T)$ factor, as stated above.

Notice that letting $T_0 = \left\lceil \max\left(\frac{16 \cdot K}{\epsilon^2} \cdot \ln\left(\frac{4 \cdot N_m \cdot K}{\frac{\delta}{2 \cdot T}}\right), 50 \cdot \log\left(\frac{2}{1 - \frac{\delta}{2 \cdot T}}\right)\right)\right\rceil$ is only possible if one knows N_m. If N_m is unknown, then one can increase T_0 and set it to $T_0 = \left\lceil \max\left(\frac{16 \cdot K}{\epsilon^2} \cdot \ln\left(\frac{4 \cdot K^2}{\frac{\delta}{2 \cdot T}}\right), 50 \cdot \log\left(\frac{2}{1 - \frac{\delta}{2 \cdot T}}\right)\right)\right\rceil$ and the lemma holds since we assume that $N_m < K$.

A.3. Analysis of scenarios from section 4

A.3.1. Proof of Theorem 3

We will denote by p_i player i for $i \in \{1, 2\}$ and the highest ranked arm as a_1. In a nutshell, the argument goes as follows: We show that the probability that p_2 does not ever learn the ranking of a_1 in rounds $[\lceil \frac{T}{2} \rceil, \lceil \frac{T}{2} + f \cdot T \rceil]$, and that p_1 stayed on arm a_1 during rounds $[\lfloor \frac{T}{4} \rfloor, \lceil \frac{T}{2} \rceil]$, is at least some constant b, not dependent on T. As a result, from time $\lceil \frac{T}{2} + f \cdot T \rceil$ onwards, the exploration probability of p_2 will be at least $\epsilon_2(f \cdot T) = \frac{c \cdot K^2}{d^2 \cdot (K-1) \cdot f \cdot T}$ (this is because p_2 has played for at least $f \cdot T$ rounds) . Thus the expected time for p_2 to explore a_1, given that she does not rank a_1 highly enough to exploit it, is $\geq \frac{1}{\epsilon_2(f \cdot T)} = \Omega(T)$. So the expected regret will be $\geq b \cdot \Omega(T) + (1 - b) \cdot 0 = \Omega(T)$.

Denote by A the event that p_2 never learns a_1 in rounds $[\lfloor \frac{T}{2} \rfloor, \lceil \frac{T}{2} + fT \rceil$. Denote by B the event that p_1 does not leave a_1 in rounds $[\lfloor \frac{T}{4} \rfloor, \lceil \frac{T}{2} \rceil]$. Since $\Pr(A \bigwedge B) = \Pr(A|B) \cdot \Pr(B)$ we will lower bound both of these terms, $\Pr(A|B)$ and

$\Pr\left(B\right)$. We start by lower bounding $\Pr\left(B\right)$:

$$
\begin{aligned}
&\Pr\left(p_1 \text{ leaves } a_1\text{at least once during the rounds } \lfloor\frac{T}{4}\rfloor, \lceil\frac{T}{2}\rceil\right) \leq \\
&\sum_{t=\lfloor\frac{T}{4}\rfloor}^{\lceil\frac{T}{2}\rceil} \Pr\left(p_1 \text{ explores at round } t\right) = \sum_{t=\lfloor\frac{T}{4}\rfloor}^{\lceil\frac{T}{2}\rceil} \frac{cK^2}{d^2(K-1)t} \\
&\leq \sum_{t=\lfloor\frac{T}{4}\rfloor}^{\lceil\frac{T}{2}\rceil} \frac{cK^2}{d^2(K-1)\lfloor\frac{T}{4}\rfloor} \leq \left(\lfloor\frac{T}{4}\rfloor\right) \frac{cK^2}{d^2(K-1)\lfloor\frac{T}{4}\rfloor} = \frac{cK^2}{d^2(K-1)}
\end{aligned}
$$

Thus $\Pr\left(B\right) \geq 1 - \frac{cK^2}{d^2(K-1)}$.

We now compute a lower bound on $\Pr\left(A|B\right)$, the probability that p_2 does not learn the ranking of a_1 in rounds $[\left\lceil\frac{T}{2}\right\rceil, \left\lceil\frac{T}{2} + f \cdot T\right\rceil]$ given that p_1 never left a_1 during the rounds $[\lfloor\frac{T}{4}\rfloor, \lceil\frac{T}{2}\rceil]$:

$$
\begin{aligned}
\Pr\left(\overline{A}|B\right) &= \Pr\left(p_2 \text{ learns } a_1 \text{ in rounds } [\left\lceil\tfrac{T}{2}\right\rceil, \left\lceil\tfrac{T}{2} + f \cdot T\right\rceil]|B\right) \\
&\leq \Pr\left(p_2 \text{ has one successful sample of } a_1|B\right) \\
&= \Pr\left(p_2 \text{ samples } a_1 \text{ and } p_1 \text{ is absent}|B\right) \\
&\leq \Pr\left(p_1 \text{ is absent at some round in } [\left\lceil\tfrac{T}{2}\right\rceil, \left\lceil\tfrac{T}{2} + f \cdot T\right\rceil]|B\right) \\
&= \sum_{n=\frac{T}{2}}^{\frac{T}{2}+f\cdot T} \Pr\left(p_1 \text{ left } a_1 \text{ for the first time at round n, and any number of times afterwards}|B\right)
\end{aligned}
$$

Denote by A_n the event that p_1 leaves a_1 for the first time since round $\lceil\frac{T}{2}\rceil$ at round n and by B_n the event that p_1 has not left a_1 from round $\lceil\frac{T}{2}\rceil$ until round $n-1$.

So we get:

$$\Pr\left(p_2 \text{ learns } a_1 \text{ in rounds } [\lceil \tfrac{T}{2} \rceil, \lceil \tfrac{T}{2} + f \cdot T \rceil] | B\right) \leq \sum_{n=\lceil \frac{T}{2} \rceil}^{\lceil \frac{T}{2}+f\cdot T \rceil} \Pr\left(A_n | B\right)$$

$$= \sum_{n=\lceil \frac{T}{2} \rceil}^{\lceil \frac{T}{2}+f\cdot T \rceil} \Pr\left(p_1 \text{ leaves due to exploration at round } n | B_n, B\right)$$

$$\cdot \Pr\left(B_n | B\right) + \Pr\left(p_1 \text{ leaves due to a collision at round } n | B_n, B\right) \cdot \Pr\left(B_n | B\right)$$

$$\leq \sum_{n=\lceil \frac{T}{2} \rceil}^{\lceil \frac{T}{2}+f\cdot T \rceil} \Pr\left(p_1 \text{ leaves due to exploration at round } n | B_n, B\right)$$

$$+ \Pr\left(p_1 \text{ leaves due to a collision at round } n | B_n, B\right)$$

$$\underbrace{\leq}_{*} \sum_{n=\lceil \frac{T}{2} \rceil}^{\lceil \frac{T}{2}+f\cdot T \rceil} \frac{c \cdot K^2}{d^2 \cdot (K-1) \cdot n} + \left(1 - \left(1 - \alpha^{\lfloor \frac{T}{4} \rfloor}\right)\right)$$

$$= \sum_{n=\lceil \frac{T}{2} \rceil}^{\lceil \frac{T}{2}+f\cdot T \rceil} \frac{c \cdot K^2}{d^2 \cdot (K-1) \cdot n} + \alpha^{\lfloor \frac{T}{4} \rfloor} \leq \sum_{n=\lceil \frac{T}{2} \rceil}^{\lceil \frac{T}{2}+f\cdot T \rceil} \frac{c \cdot K^2}{d^2 \cdot (K-1) \cdot \frac{T}{2}} + \alpha^{\lfloor \frac{T}{4} \rfloor}$$

$$= g \cdot \left(\frac{T}{2} + f \cdot T - \frac{T}{2}\right) \cdot \left(\frac{c \cdot K^2}{d^2 \cdot (K-1) \cdot \frac{T}{2}} + \alpha^{\lfloor \frac{T}{4} \rfloor}\right)$$

$$= g \cdot \frac{c \cdot K^2 \cdot 2 \cdot f}{d^2 \cdot (K-1)} + g \cdot f \cdot T \cdot \alpha^{\lfloor \frac{T}{4} \rfloor}$$

g is a constant that when multiplied by $\left(\frac{T}{2} + f \cdot T - \frac{T}{2}\right)$ ensures that these numbers are whole numbers.
The inequality labeled $*$ is due to the fact that since it is given that p_1 always exploited arm 1 from rounds $\lfloor \frac{T}{4} \rfloor$ to $\lceil \frac{T}{2} \rceil$ then she would have a persistence probability $\geq 1 - \alpha^{\lfloor \frac{T}{4} \rfloor}$.
This fact can be seen from the following calculation:

$$p_{t+1} = p_t \cdot \alpha + (1-\alpha) = (p_{t-1} \cdot \alpha + (1-\alpha)) \cdot \alpha + (1-\alpha)$$

$$= p_{t-1} \cdot \alpha^2 + (1-\alpha)(\alpha+1) = (p_{t-2} \cdot \alpha + (1-\alpha)) \cdot \alpha^2 + (1-\alpha)(\alpha+1)$$

$$= p_{t-2} \cdot \alpha^3 + (1-\alpha) \cdot \left(\alpha^2 + \alpha + 1\right)$$

Which implies that if p_1 did not leave a_1 from round $\lfloor \frac{T}{4} \rfloor$ until round $\lceil \frac{T}{2} \rceil$ then the persistence parameter of p_1 at round $\lceil \frac{T}{2} \rceil$ will be $p_{\lceil \frac{T}{2} \rceil} = p_{\lfloor \frac{T}{4} \rfloor} \cdot \alpha^{\lfloor \frac{T}{4} \rfloor} + (1-\alpha) \cdot \left(1 + \alpha + \alpha^2 + ... + \alpha^{\lfloor \frac{T}{4} \rfloor - 1}\right) = p_{\lfloor \frac{T}{4} \rfloor} \cdot \alpha^{\lfloor \frac{T}{4} \rfloor} + (1-\alpha) \cdot \sum_{r=0}^{\lfloor \frac{T}{4} \rfloor - 1} \alpha^r = p_{\lfloor \frac{T}{4} \rfloor} \cdot \alpha^{\lfloor \frac{T}{4} \rfloor} + (1-\alpha) \cdot \frac{1-\alpha^{\lfloor \frac{T}{4} \rfloor}}{1-\alpha} \geq 1 - \alpha^{\lfloor \frac{T}{4} \rfloor}$

Thus we see that the probability that p_2 does not learn a_1 in rounds $[\lceil \frac{T}{2} \rceil, \lceil \frac{T}{2} + f \cdot T \rceil]$ given that p_1 did not leave a_1 in rounds $[\lfloor \frac{T}{4} \rfloor, \lceil \frac{T}{2} \rceil]$ is $\geq 1 - \left(\frac{c \cdot K^2 \cdot 2 \cdot f}{d^2 \cdot (K-1)} + f \cdot T \cdot \alpha^{\lfloor \frac{T}{4} \rfloor}\right)$.

We will now show that this is lower bounded by some constant, not dependent on T, by showing that:
$\Pr\left(p_2 \text{ learns } a_1 \text{ in rounds } [\lceil \frac{T}{2} \rceil, \lceil \frac{T}{2} + f \cdot T \rceil]\right) \leq \frac{1}{2}$ and thus:
$\Pr\left(p_2 \text{ does not learn } a_1 \text{ in rounds } [\lceil \frac{T}{2} \rceil, \lceil \frac{T}{2} + f \cdot T \rceil]\right) \geq \frac{1}{2}$

We will now show that either the following two statements are true, in which case we have shown that the probability of p_2 not learning a_1 is greater than some constant, or they are not both true and the MEGA algorithm will have linear regret due to other reasons, inherent to the algorithm:

- $\frac{c \cdot K^2 \cdot 2 \cdot f}{d^2 \cdot (K-1)} \leq \frac{1}{4}$

- $f \cdot T \cdot \alpha^{\frac{T}{4}} \leq \frac{1}{4}$

These two statements depend on the choice of parameters of MEGA: c, d, α. For the first statement we show that if these parameters are chosen in a way which contradicts the inequality, then the MEGA algorithm results in linear regret, thus satisfying the main claim in any case. The second statement is true by assumption.
We start with the first statement:

$\frac{c \cdot K^2 \cdot 2 \cdot f}{d^2 \cdot (K-1)} \leq \frac{1}{4}$
For this to be true we need: $f \leq \frac{d^2 \cdot (K-1)}{8 \cdot c \cdot K^2}$. At the same time we want f large enough such that $f \cdot T \geq X$ for some constant $X \geq \frac{c \cdot K^2}{d^2 \cdot (K-1)}$ to ensure that the exploration coefficient for p_2 will be less than 1 when p_1 leaves.
If we have an f such that: $\frac{X}{T} \leq f \leq \frac{d^2 \cdot (K-1)}{8 \cdot c \cdot K^2}$ then we are done. If we cannot find such an f then we are in the case where: $\frac{d^2 \cdot (K-1)}{8 \cdot c \cdot K^2} \leq \frac{X}{T}$.
This happens if: $d^2 \leq \frac{8 \cdot c \cdot K^2 \cdot X}{(K-1) \cdot T}$ or $c \geq \frac{T \cdot d^2 \cdot (K-1)}{8 \cdot K^2 \cdot X}$.
We note that on those two cases the regret will be linear due to exploration, since:
If $d^2 \leq \frac{8 \cdot c \cdot K^2 \cdot X}{(K-1) \cdot T}$ then $\forall t \leq \frac{T}{8 \cdot X} : \epsilon_t = \frac{c \cdot K^2}{d^2 \cdot (K-1) \cdot t} \geq \frac{c \cdot K^2}{\frac{X \cdot 8 \cdot c \cdot K^2}{(K-1) \cdot T} \cdot (K-1) \cdot t} = \frac{T}{8 \cdot X \cdot t} \geq \frac{T}{8 \cdot X \cdot \frac{T}{8 \cdot X}} = 1$.

If $c \geq \frac{T \cdot d^2 \cdot (K-1)}{8 \cdot K^2 \cdot X}$ then $\forall t \leq \frac{T}{8 \cdot X} : \epsilon_t = \frac{c \cdot K^2}{d^2 \cdot (K-1) \cdot t} \geq \frac{\frac{T \cdot d^2 \cdot (K-1)}{8 \cdot K^2 \cdot X} \cdot K^2}{d^2 \cdot (K-1) \cdot t} = \frac{T}{8 \cdot X \cdot t} \geq \frac{T}{8 \cdot X \cdot \frac{T}{8 \cdot X}} = 1$.

Thus, in both cases, players will be exploring for at least $\frac{T}{8 \cdot X}$ rounds, resulting in linear regret.

We now consider the second statement, i.e. : $f \cdot T \cdot \alpha^{\frac{T}{4}} \leq \frac{1}{4}$
Note that for this to be true, we need: $\alpha^{\frac{T}{4}} \leq \frac{1}{4 \cdot f \cdot T} \implies \alpha \leq \left(\frac{1}{4 \cdot f \cdot T}\right)^{\frac{4}{T}}$. This value is very close to 1, and approaches 1 as $T \to \infty$. Thus for any reasonable choice of α, the expected regret is $\Omega\left(T\right)$ and when α is not in this range, i.e. very close to 1, then p will hardly deviate from p_0 which means that during a collision event both players have the same chance of not persisting and making an arm unavailable. This will make the unavailability mechanism not functional, and moreover, it will cause more collisions by forcing the players onto the second arm. For clarity we use a weaker bound on α: $\alpha \leq 1 - \frac{4 \log(4fT)}{T}$. We show that this is weaker:

$$\frac{1}{4fT}^{\frac{4}{T}} = \exp\left(\log(\frac{1}{4fT}) \cdot \frac{4}{T}\right) = \exp\left(-\log(4fT) \cdot \frac{4}{T}\right) \geq 1 - \log(4fT) \cdot \frac{4}{T}$$

Thus, by assumption, the second statement is true as well.
Thus we have shown that $\Pr\left(A \bigwedge B\right) \geq \frac{1}{2} \cdot \left(1 - \frac{cK^2}{d^2(K-1)}\right)$. So the expected regret of the MEGA algorithm is $\Omega(T)$ w.p. $\geq \frac{1}{2} \cdot \left(1 - \frac{cK^2}{d^2(K-1)}\right)$. This is true since it would take $\Omega\left(T\right)$ rounds, in expectation, for player 2 to learn the true ranking of arm 1 after player 1 has left the game. The upper bound for the DMC algorithm calculated in the dynamic setting is also an upper bound for this scenario taking $x = 2$ which yields an expected regret $\leq O\left(\sqrt{T}\right)$.

A.3.2. PROOF OF THEOREM 4

We will first give a sketch of the proof: we use the scenario described in (Avner & Mannor, 2014), where a player leaves after the system has 'stabilized'. They provide an upper bound on the added regret caused by the exiting player, on top of the regret caused by other sources. We use this bound in order to calculate the added regret cause by a stream of exiting player, each leaving after the system has 'stabilized'. We compare this bound to the bound of the total expected regret of the DMC algorithm in that scenario.

We now continue with the formal proof. Since one player leaves every $X = (2 \cdot r) \cdot \lceil T^\lambda \rceil$ rounds, the rounds at which a player leaves are: $0 \cdot \lceil T^\lambda \rceil, 2 \cdot \lceil T^\lambda \rceil, 4 \cdot \lceil T^\lambda \rceil, ..., \left(\lceil \frac{T}{\lceil T^\lambda \rceil} \rceil - 1\right) \cdot \lceil T^\lambda \rceil$. (Avner & Mannor, 2014) present a case in which only one player leaves at round t, and causes an added regret of t^β. They show that for the scenario of one player leaving, this is a worst case regret bound. We sum over half the total rounds to avoid concerns of counting regret of the last round. Using this bound of t^{β} we get that the total added regret due to all of the players who left is at least:

$$\begin{aligned}
&\sum_{r=0}^{\frac{1}{2}\cdot\lceil\frac{T}{2\cdot T^{\lambda}}\rceil} \left(2\cdot r\cdot\lceil T^{\lambda}\rceil\right)^{\beta} \geq \sum_{r=0}^{\frac{1}{2}\cdot\frac{T}{2\cdot T^{\lambda}}} \left(2\cdot r\cdot T^{\lambda}\right)^{\beta} = T^{\lambda\cdot\beta}\cdot 2^{\beta}\cdot \sum_{r=0}^{\frac{1}{2}\cdot\frac{T}{2\cdot T^{\lambda}}} r^{\beta} \\
&\geq T^{\lambda\cdot\beta}\cdot 2^{\beta}\cdot \int_{r=0}^{\frac{1}{2}\cdot\frac{T}{2\cdot T^{\lambda}}} r^{\beta} dr = T^{\lambda\cdot\beta}\cdot 2^{\beta}\cdot\frac{1}{\beta+1}\cdot\frac{1}{4^{\beta+1}}\cdot\frac{T^{\beta+1}}{T^{\lambda\cdot(\beta+1)}} \\
&= \frac{2^{\beta}}{(\beta+1)\cdot 4^{\beta+1}}\cdot T^{\lambda\cdot\beta+\beta+1-\lambda\cdot\beta-\lambda} = \frac{1}{(\beta+1)\cdot 2^{\beta+2}}\cdot T^{1-(\lambda-\beta)}
\end{aligned}$$

So we get that:

$$R_{mega}\left(T\right) \geq \frac{1}{(\beta+1)\cdot 2^{\beta+2}}\cdot T^{1-(\lambda-\beta)}$$

Thus the expected regret of the MEGA algorithm in this scenario is at least $\min\{T, \frac{1}{(\beta+1)\cdot 2^{\beta+2}}\cdot T^{1-(\lambda-\beta)}\}$ since regret cannot be larger than linear. Notice that this bound does not include any regret due to learning, collisions, and other sources of regret.

We will compare this bound with an upper bound of the DMC algorithm for this scenario.

In the analysis of the DMC algorithm in the dynamic setting we calculated regret when at most x players enter or leave. In this case we have $\frac{T}{T^{\lambda}}$ players leaving or entering. Thus the regret bound is $O\left(\sqrt{xT}\right) = O\left(\sqrt{T\frac{T}{T^{\lambda}}}\right) = O\left(T^{1-\frac{\lambda}{2}}\right)$.

B. Supplementary Figures

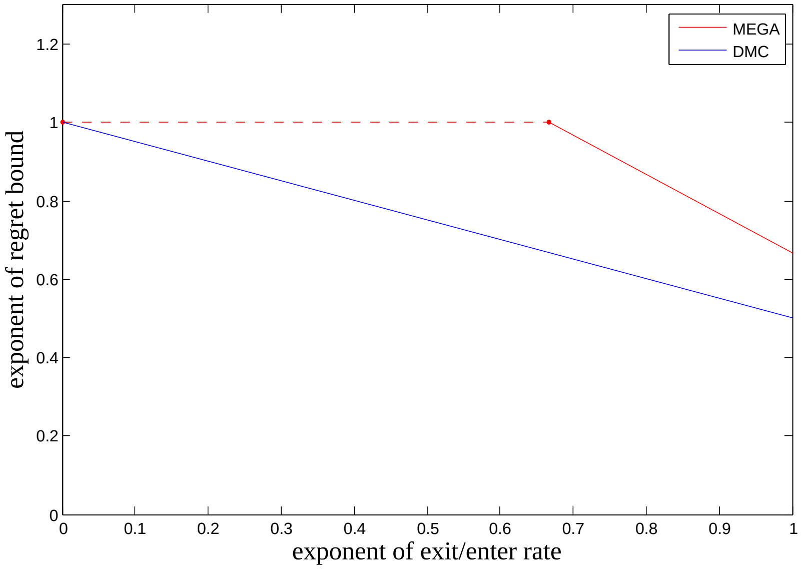

Figure 4. A graphical illustration of theorem 4: comparison of regret bound. The red line represents the exponent $\min\{1, 1-(\lambda-\beta)\}$ in the regret upper bound for the MEGA algorithm, as a function of λ, where we use $\beta = \frac{2}{3}$, the value recommended in (Avner & Mannor, 2014). Note that the dashed red line is a region not covered in the MEGA algorithm analysis, but the regret is trivially at most T. The blue line represents the regret exponent $1 - \frac{\lambda}{2}$ of the DMC algorithm, as a function of λ.

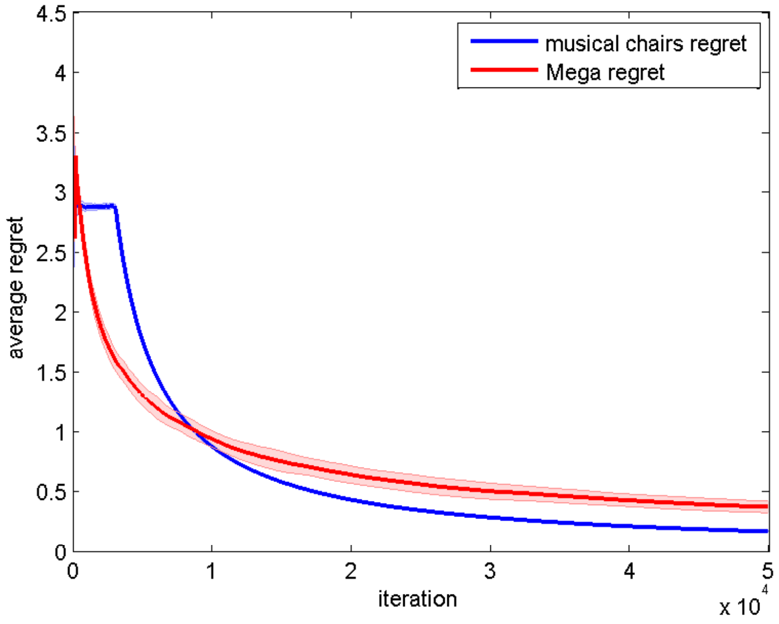

Figure 5. Static setting: average regret after 50,000 iterations

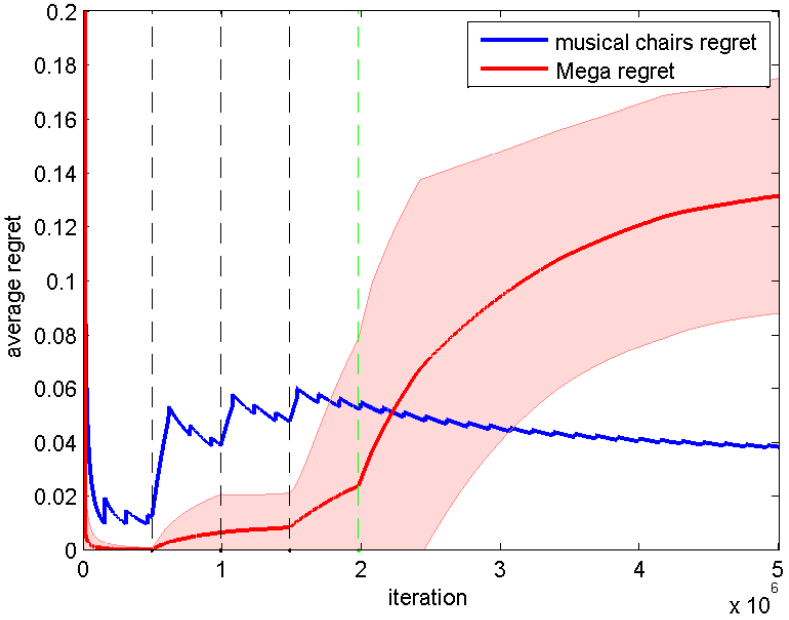

Figure 6. Generalized case 1: A generalization of the scenario presented in theorem 3, section 4, for multiple players. The game starts with one player and every $T^{0.84}$ rounds another player enters, until the number of players reaches 4. Then after another $T^{0.84}$ rounds, the first player leaves. Rewards are chosen deterministically. There are 10 arms with a gap of 0.7 between the expected reward of the N^{th} and $N+1^{th}$ best arm, and T_1 was set to $167,845$. The black dashed lines represents players entering and the green dashed line shows when the first player exits.