Fair Algorithms for Multi-Agent Multi-Armed Bandits

Safwan Hossain
University of Toronto
safwan.hossain@mail.utoronto.ca

Evi Micha
University of Toronto
emicha@cs.toronto.edu

Nisarg Shah
University of Toronto
nisarg@cs.toronto.edu

Abstract

We propose a multi-agent variant of the classical multi-armed bandit problem, in which there are N agents and K arms, and pulling an arm generates a (possibly different) stochastic reward for each agent. Unlike the classical multi-armed bandit problem, the goal is not to learn the "best arm"; indeed, each agent may perceive a different arm to be the best for her personally. Instead, we seek to learn a fair distribution over the arms. Drawing on a long line of research in economics and computer science, we use the *Nash social welfare* as our notion of fairness. We design multi-agent variants of three classic multi-armed bandit algorithms and show that they achieve sublinear regret, which is now measured in terms of the lost Nash social welfare.

1 Introduction

In the classical (stochastic) multi-armed bandit (MAB) problem, a principal has access to K arms and pulling arm j generates a stochastic reward for the principal from an unknown distribution D_j with an unknown mean μ_j^*. If the mean rewards were known a priori, the principal could just repeatedly pull the *best arm* given by $\arg\max_j \mu_j^*$. However, the principal has no apriori knowledge of the quality of the arms. Hence, she uses a learning algorithm which operates in rounds, pulls arm j^t in round t, observes the stochastic reward generated, and uses that information to learn the best arm over time. The performance of such an algorithm is measured in terms of its cumulative regret up to a horizon T, defined as $\sum_{t=1}^{T}(\max_j \mu_j^* - \mu_{j^t}^*)$. Note that this is the difference between the total mean reward that would have been achieved if the best arm was pulled repeatedly and the total mean reward of the arms pulled by the learning algorithm up to round T.

This problem can model situations where the principal is deliberating a policy decision and the arms correspond to the different alternatives she can implement. However, in many real-life scenarios, making a policy decision affects not one, but several agents. For example, imagine a company making a decision that affects all its employees, or a conference deciding the structure of its review process, which affects various research communities. This can be modeled by a multi-agent variant of the multi-armed bandit (MA-MAB) problem, in which there are N agents and pulling arm j generates a (possibly different) stochastic reward for each agent i from an unknown distribution $D_{i,j}$ with an unknown mean $\mu_{i,j}^*$.

Before pondering about learning the "best arm" over time, we must ask what the best arm even means in this context. Indeed, the "best arm" for one agent may not be the best for another. A first attempt may be to associate some "aggregate quality" to each arm; for example, the quality of arm j may be defined as the total mean reward it gives to all agents, i.e., $\sum_i \mu_{i,j}^*$. This would

nicely reduce our problem to the classic multi-armed bandit problem, for which we have an armory of available solutions [30]. However, this approach suffers from the *tyranny of the majority* [28]. For example, imagine a scenario with ten agents, two arms, and deterministic rewards. Suppose four agents derive a reward of 1 from the first arm but 0 from the second, while the remaining six derive a reward of 1 from the second arm but 0 from the first. The aforementioned approach will deem the second arm as the best and a classical MAB algorithm will converge to *repeatedly* pulling the second arm, thus unfairly treating the first four agents (a minority). A solution which treats each group in a "proportionally fair" [28] manner should ideally converge to pulling the first arm 40% of the time and the second 60% of the time. Alternatively, we can allow the learning algorithm to "pull" a probability distribution over the arms and seek an algorithm that converges to placing probability 0.4 on the first arm and 0.6 on the second.

This problem of making a fair collective decision when the available alternatives — in this case, probability distributions over the arms — affect multiple agents is well-studied in computational social choice [7]. The literature offers a compelling fairness notion called the *Nash social welfare*, named after John Nash. According to this criterion, the fairest distribution maximizes the *product* of the expected utilities (rewards) to the agents. A distribution p that places probability p_j on each arm j gives expected utility $\sum_j p_j \cdot \mu^*_{i,j}$ to agent i. Hence, the goal is to maximize $\text{NSW}(p, \mu^*) = \prod_{i=1}^{N}(\sum_{j=1}^{K} p_j \cdot \mu^*_{i,j})$ over p. One can verify that this approach on the aforementioned example indeed yields probability 0.4 on the first arm and 0.6 on the second, as desired. It is also interesting to point out that with a single agent ($N = 1$), the distribution maximizing the Nash social welfare puts probability 1 on the best arm, thus effectively reducing the problem to the classical multi-armed bandit problem (albeit with subtle differences which we highlight in Section 6).

Maximizing the Nash social welfare is often seen as a middle ground between maximizing the utilitarian social welfare (sum of utilities to the agents), which is unfair to minorities (as we observed), and maximizing the egalitarian social welfare (minimum utility to any agent), which is considered too extreme [28]. The solution maximizing the Nash social welfare is also known to satisfy other qualitative fairness desiderata across a wide variety of settings [1, 4, 6, 8, 12, 16, 17, 18]. For example, a folklore result shows that in our setting such a solution will always lie in *the core*; we refer the reader to the work of Fain et al. [17] for a formal definition of the core as well as a short derivation of this fact using the first-order optimality condition. For further discussion on this, see Sections 1.2 and 7.

When *exactly* maximizing the Nash social welfare is not possible (either due to a lack of complete information, as in our case, or due to computational difficulty), researchers have sought to achieve approximate fairness by approximately maximizing this objective [2, 10, 11, 19, 20, 26]. Following this approach in our problem, we define the (cumulative) regret of an algorithm at horizon T as $\sum_{t=1}^{T}(\max_p \text{NSW}(p, \mu^*) - \text{NSW}(p^t, \mu^*))$, where p^t is the distribution selected in round t. Our goal in this paper is to design algorithms whose regret is sublinear in T.

1.1 Our Results

We consider three classic algorithms for the multi-armed bandit problem: Explore-First, Epsilon-Greedy, and UCB [30]. All three algorithms attempt to balance exploration (pulling arms only to learn their rewards) and exploitation (using the information learned so far to pull "good" arms). Explore-First performs exploration for a number of rounds optimized as a function of T followed by exploitation in the remaining rounds to achieve regret $\widetilde{\mathcal{O}}\left(K^{1/3}T^{2/3}\right)$. Epsilon-Greedy flips a coin in each round to decide whether to perform exploration or exploitation and achieves the same regret bound. Its key advantage over Explore-First is that it does not need to know the horizon T upfront. UCB merges exploration and exploitation to achieve a regret bound of $\widetilde{\mathcal{O}}\left(K^{1/2}T^{1/2}\right)$.

Here, $\widetilde{\mathcal{O}}$ hides log factors. Traditionally, the focus is on optimizing the exponent of T rather than that of K as the horizon T is often much larger than the number of arms K. It is known that the dependence of UCB's regret on T is optimal: no algorithm can achieve *instance-independent* $o(T^{1/2})$ regret [3].[1]

We propose natural multi-agent variants of these three algorithms. Our variants take the Nash social welfare objective into account and select a distribution over the arms in each round instead of a single arm. For Explore-First, we derive $\widetilde{\mathcal{O}}\left(N^{2/3}K^{1/3}T^{2/3}\right)$ regret bound, which recovers the aforementioned single-agent bound with an additional factor of $N^{2/3}$. We also show that changing a parameter of the algorithm yields a regret bound of $\widetilde{\mathcal{O}}\left(N^{1/3}K^{2/3}T^{2/3}\right)$, which offers a different tradeoff between the dependence on N and K. For Epsilon-Greedy, we recover the same two regret bounds, although the analysis becomes much more intricate. This is because, as mentioned above, Epsilon-Greedy is a horizon-independent algorithm (i.e. it does not require apriori knowledge of T), unlike Explore-First. For UCB, we derive $\widetilde{\mathcal{O}}\left(NKT^{1/2}\right)$ and $\widetilde{\mathcal{O}}\left(N^{1/2}K^{\frac{3}{2}}T^{1/2}\right)$ regret bounds; our dependence on K worsens compared to the classical single-agent case, but importantly, we recover the same $\sqrt{T}$ dependence. Finally, we note that even for $N = 1$, a learning algorithm is slightly more powerful in our setting than in the classical setting since it can choose a distribution over the arms as opposed to a deterministic arm. Nonetheless, we derive an $\Omega(\sqrt{T})$ instance-independent lower bound on the regret of any algorithm in our setting, establishing the asymptotic optimality of our UCB variant.

Deriving these regret bounds for the multi-agent case requires overcoming two key difficulties that do not appear in the single-agent case. First, our goal is to optimize a complicated function, the Nash social welfare, rather than simply selecting the best arm. This requires a Lipschitz-continuity analysis of the Nash social welfare function and the use of new tools such as the McDiarmid's inequality which are not needed in the standard analysis. Second, the optimization is over an infinite space (the set of distributions over arms) rather than over a finite space (the set of arms). Thus, certain tricks such as a simple union bound no longer work; we use the concept of δ-covering, used heavily in the Lipschitz bandit framework [24], in order to address this.

1.2 Related Work

Since the multi-armed bandit problem was introduced by Thompson [31], many variants of it have been proposed, such as sleeping bandit [23], contextual bandit [33], dueling bandit [34], Lipschitz bandit [24], etc. However, all these variants involve a single agent who is affected by the decisions. We note that other multi-agent variants of the multi-armed bandit problem have been explored recently [5, 9]. However, they still involve a common reward like in the classical multi-armed bandit problem. Their focus is on getting the agents to cooperate to maximize this common reward.

Another key aspect of our framework is the focus on fairness. Recently, several papers have focused on fairness in the multi-armed bandit problem. For instance, Joseph et al. [22] design a UCB variant which guarantees what they refer to as meritocratic fairness to the arms, i.e., that a worse arm is never preferred to a better arm regardless of the algorithm's confidence intervals for them. Liu et al. [27] require that similar arms be treated similarly, i.e., two arms with similar mean rewards be selected with similar probabilities. Gillen et al. [21] focus on satisfying fairness with respect to an unknown fairness metric. Finally, Patil et al. [29] assume that there are external constraints requiring that each arm be pulled in at least a certain fraction of the rounds and design

[1] In instance-independent bounds, the constant inside the big-Oh notation is not allowed to depend on the (unknown) distributions in the given instance. UCB also achieves an $O(\log T)$ instance-dependent regret bound, which is also known to be asymptotically optimal [25]. For further discussion, see Section 7.

algorithms that achieve low regret subject to this constraint. All these papers seek to achieve fairness *with respect to the arms*. In contrast, in our work, the arms are "inanimate" (e.g. policy decisions) and we seek fairness *with respect to the agents*, who are separate from the arms.

More broadly, the problem of making a fair decision given the (possibly conflicting) preferences of multiple agents is well-studied in computational social choice [7] in a variety of contexts. For example, one can consider our problem as that of randomized voting (alternatively known as *fair mixing* [4]) by viewing the agents as voters and the arms are candidates. The goal is then to pick a fair lottery over the candidates given the voters' preferences. This is also a special case of other more complex models studied in the literature such as fair public decision-making [12] and fair allocation of public goods [17]. However, in computational social choice, voters typically have fixed preferences over the candidates. In contrast, rewards observed by the agents in our framework are stochastic. From this viewpoint, our work provides algorithms for maximizing the Nash social welfare when noisy information can be queried regarding agent preferences.

2 Preliminaries

For $n \in \mathbb{N}$, define $[n] = \{1, \ldots, n\}$. Let $N, K \in \mathbb{N}$. In the *multi-agent multi-armed bandit* (MA-MAB) problem, there is a set of *agents* $[N]$ and a set of *arms* $[K]$. For each agent $i \in [N]$ and arm $j \in [K]$, there is a *reward distribution* $D_{i,j}$ with mean $\mu^*_{i,j}$ and support $[0,1]$;[2] when arm j is pulled, each agent i observes an independent *reward* sampled from $D_{i,j}$. Let us refer to $\mu^* = (\mu^*_{i,j})_{i\in[N],j\in[K]} \in [0,1]^{N\times K}$ as the (true) reward matrix.

Policies: As mentioned in the introduction, pulling an arm deterministically may be favorable to one agent, but disastrous to another. Hence, we are interested in *probability distributions* over arms, which we refer to as *policies*. The K-simplex, denoted Δ^K, is the set of all policies. For a policy $p \in \Delta^K$, p_j denotes the probability with which arm j is pulled. Note that due to linearity of expectation, the expected reward to agent i under policy p is $\sum_{j=1}^K p_j \cdot \mu^*_{i,j}$.

Nash social welfare: The Nash social welfare is defined the product of (expected) rewards to the agents. Given $\mu = (\mu_{i,j})_{i\in[N],j\in[K]}$, and policy $p \in \Delta^K$, define $\mathrm{NSW}(p,\mu) = \prod_{i=1}^N \left(\sum_{j=1}^K p_j \cdot \mu_{i,j}\right)$. Thus, the (true) Nash social welfare under policy p is $\mathrm{NSW}(p,\mu^*)$. Hence, if we knew μ^*, we would pick an *optimal policy* $p^* \in \arg\max_{p\in\Delta^K} \mathrm{NSW}(p,\mu^*)$. However, because we do not know μ^* in advance, our algorithms will often produce an estimate $\widehat{\mu}$, and use it to choose a policy; the quantity $\mathrm{NSW}(p,\widehat{\mu})$ will play a key role in our algorithms and their analysis.

Algorithms: An algorithm for the MA-MAB problem chooses a policy p^t in each round $t \in \mathbb{N}$. Then, an arm a^t is sampled according to policy p^t, and for each agent $i \in [N]$, a reward X^t_{i,a^t} is sampled independently from distribution D_{i,a^t}. At the end of round t, the algorithm learns the sampled arm a^t and the reward vector $(X^t_{i,a^t})_{i\in[N]}$, which it can use to choose policies in the later rounds.

Reward estimates: All our algorithms maintain an estimate of the mean reward matrix μ^* at every round. For round t and arm $j \in [K]$, let $n^t_j = \sum_{s=1}^{t-1} \mathbb{1}[a^s = j]$ denote the number of times arm j is pulled at the beginning of round t, and let $\widehat{\mu}^t_{i,j} = \frac{1}{n^t_j} \sum_{s\in[t-1]:a^s=j} X^s_{i,j}$ denote the average

[2] We need the support of the distribution to be non-negative and bounded, but the upper bound of 1 is without loss of generality. All our bounds scale linearly with the upper bound on the support.

reward experienced by agent i from the n_j^t pulls of arm j thus far. Our algorithms treat these as an estimate of $\mu_{i,j}^*$ available at the beginning of round t. Let $\widehat{\mu}^t = (\widehat{\mu}_{i,j}^t)_{i\in[N],j\in[K]}$.

Regret: Recall that p^* is an optimal policy that has the highest Nash social welfare. The *instantaneous regret* in round t due to an algorithm choosing p^t is $r^t = \mathrm{NSW}(p^*, \mu^*) - \mathrm{NSW}(p^t, \mu^*)$. The (cumulative) *regret* in round T due to an algorithm choosing $p^1, \ldots, p^T$ is $R^T = \sum_{t=1}^T r^t$. We note that R^T and r^t are defined for a specific algorithm, which will be clear from the context. We are interested in bounding the *expected regret* $\mathbb{E}[R^T]$ of an algorithm at round T, where the expectation is over the randomness involved in sampling the arms a^t and the agent rewards $(X_{i,a^t}^t)_{i\in[N]}$ for $t \in [T]$.[3] We say that an algorithm is *horizon-dependent* if it needs to know T in advance in order to yield bounded regret at round T, and *horizon-independent* if it yields such a bound without knowing T in advance.

δ-Covering: Given a metric space (X, d) and $\delta > 0$, a set $S \subseteq X$ is called a *δ-cover* if for each $x \in X$, there exists $s \in S$ with $d(x, s) \leq \delta$. That is, from each point in the metric space, there is a point in the δ-cover that is no more than δ distance away. We will heavily use the fact that there exists a δ-cover of $(\Delta^K, \|\cdot\|_1)$ (i.e. the K-simplex under the L_1 distance) with size at most $(1 + 2/\delta)^K$ [32, p. 126], which follows from a simple discretization of the simplex.

3 Explore-First

ALGORITHM 1: Explore-First

Input: Number of agents N, number of arms K, horizon T

Parameters : Exploration period L

`// Pull each arm` L `times`

for $t = 1, \ldots, K \cdot L$ **do** `// Exploration`

  $j \leftarrow \lceil t/L \rceil$

  $p^t \leftarrow$ policy that puts probability 1 on arm j `// Pull arm` j `deterministically`

end

Compute the estimated reward matrix $\widehat{\mu} \triangleq \widehat{\mu}^{K\cdot L+1}$ of the rewards observed so far

Compute $\widehat{p} \in \arg\max_{p\in\Delta^K} \mathrm{NSW}(p, \widehat{\mu})$

for $t = K \cdot L + 1, \ldots, T$ **do** `// Exploitation`

  $p^t \leftarrow \widehat{p}$

end

Perhaps the simplest algorithm (with a sublinear regret bound) in the classic single-agent MAB framework is Explore-First. It is composed of two distinct stages. The first stage is *exploration*, during which the algorithm pulls each arm L times. At the end of this stage, the algorithm computes the arm $\widehat{a}$ with the best estimated mean reward, and in the subsequent *exploitation* stage, pulls arm $\widehat{a}$ in every round. The algorithm is horizon-dependent, i.e., it takes the horizon T as input and sets L as a function of T. Setting $L = \Theta\left(K^{-\frac{2}{3}} T^{\frac{2}{3}} \log^{\frac{1}{3}}(T)\right)$ yields regret bound $\mathbb{E}[R^T] = \mathcal{O}\left(K^{\frac{1}{3}} T^{\frac{2}{3}} \log^{\frac{1}{3}}(T)\right)$ [30].

[3] The algorithms we study do not introduce any further randomness in choosing the policies.

In our multi-agent variant, presented as Algorithm 1, the exploration stage pulls each arm L times as before. However, at the end of this stage, the algorithm computes, not an arm $\widehat{a}$, but a policy $\widehat{p}$ with the best estimated Nash social welfare. During exploitation, it then uses policy $\widehat{p}$ in every round. With an almost identical analysis as in the single-agent setting, we recover the aforementioned regret bound with an additional $N^{2/3}$ factor for N agents.

Using a novel and more intricate argument, we show that a different tradeoff between the exponents of N and K can be obtained, where $N^{2/3}$ is reduced to $N^{1/3}$ at the expense of increasing $K^{1/3}$ to $K^{2/3}$ (and adding a logarithmic term). We later use this approach again in our analysis of more sophisticated algorithms.

Before we proceed to the proof, we remark that Algorithm 1 can be implemented efficiently. The only non-trivial step is to compute the optimal policy given an estimated reward matrix, i.e., $\widehat{p} \in \arg\max_{p \in \Delta^K} \mathrm{NSW}(p, \widehat{\mu})$. Since the Nash social welfare is known to be log-concave, this can be solved in polynomial time [15].

We begin by presenting a few elementary lemmas regarding the behavior of the Nash social welfare function $\mathrm{NSW}(p, \mu)$. We are mainly interested in how much the function can change when its arguments change. To that end, the following folklore result translates the difference in a product to a sum of point-wise differences that are easier to deal with.

Lemma 1. *Let $a_i, b_i \in [0,1]$ for $i \in [N]$. Then, $\left|\prod_{i=1}^N a_i - \prod_{i=1}^N b_i\right| \leq \sum_{i=1}^N |a_i - b_i|$.*

Proof. We prove this using induction on N. For $N = 1$, the lemma trivially holds. Suppose it holds for $N = n$. For $N = n+1$, we have

$$\begin{aligned}
\left|\prod_{i=1}^{n+1} a_i - \prod_{i=1}^{n+1} b_i\right| &= \left|\prod_{i=1}^{n+1} a_i - b_{n+1}\prod_{i=1}^{n} a_i + b_{n+1}\prod_{i=1}^{n} a_i - \prod_{i=1}^{n+1} b_i\right| \\
&\leq \left(\prod_{i=1}^{n} a_i\right)|a_{n+1} - b_{n+1}| + b_{n+1} \cdot \left|\prod_{i=1}^{n} a_i - \prod_{i=1}^{n} b_i\right| \\
&\leq |a_{n+1} - b_{n+1}| + \sum_{i=1}^{n} |a_i - b_i| = \sum_{i=1}^{n+1} |a_i - b_i|\,,
\end{aligned}$$

where the second transition is due to the triangle inequality, and the third transition holds due to the induction hypothesis and because $a_i, b_i \in [0,1]$ for each i. □

Using Lemma 1, we can easily analyze Lipschitz-continuity of $\mathrm{NSW}(p, \mu)$ when either p or μ changes and the other is fixed. First, we consider change in p with μ fixed.

Lemma 2. *Given a reward matrix $\mu \in [0,1]^{N \times K}$ and policies $p^1, p^2 \in \Delta^K$, we have*

$$\left|\mathrm{NSW}(p^1, \mu) - \mathrm{NSW}(p^2, \mu)\right| \leq N \cdot \left\|p^1 - p^2\right\|_1 = N \cdot \sum_{j \in [K]} \left|p_j^1 - p_j^2\right|.$$

Proof. Using Lemma 1, we have

$$\left|\mathrm{NSW}(p^1, \mu) - \mathrm{NSW}(p^2, \mu)\right| \leq \sum_{i \in [N]} \left|\sum_{j \in [K]} (p_j^1 - p_j^2) \cdot \mu_{i,j}\right| \leq N \cdot \sum_{j \in [K]} \left|p_j^1 - p_j^2\right|,$$

where the final transition is due to the triangle inequality and because $\mu_{i,j} \in [0,1]$ for each i, j. □

Next, we consider change in μ with p fixed.

Lemma 3. *Given a policy $p \in \Delta^K$, and reward matrices $\mu^1, \mu^2 \in [0,1]^{N\times K}$, we have*

$$\left|\mathrm{NSW}(p,\mu^1) - \mathrm{NSW}(p,\mu^2)\right| \le \sum_{i\in[N]}\sum_{j\in[K]} p_j \cdot \left|\mu^1_{i,j} - \mu^2_{i,j}\right|.$$

Proof. Again, using Lemma 1, we have

$$\left|\mathrm{NSW}(p,\mu^1) - \mathrm{NSW}(p,\mu^2)\right| \le \sum_{i\in[N]} \left|\sum_{j\in[K]} p_j \cdot (\mu^1_{i,j} - \mu^2_{i,j})\right| \le \sum_{i\in[N],j\in[K]} p_j \cdot \left|\mu^1_{i,j} - \mu^2_{i,j}\right|,$$

where the last transition is due to the triangle inequality. □

We are now ready to derive the regret bounds for Explore-First.

Theorem 1. *Explore-First is horizon-dependent and has the following expected regret at round T.*

- *When* $L = \Theta\left(N^{\frac{2}{3}}K^{-\frac{2}{3}}T^{\frac{2}{3}}\log^{\frac{1}{3}}(NKT)\right)$, $\mathbb{E}[R^T] = \mathcal{O}\left(N^{\frac{2}{3}}K^{\frac{1}{3}}T^{\frac{2}{3}}\log^{\frac{1}{3}}(NKT)\right)$.
- *When* $L = \Theta\left(N^{\frac{1}{3}}K^{-\frac{1}{3}}T^{\frac{2}{3}}\log^{\frac{2}{3}}(NKT)\right)$, $\mathbb{E}[R^T] = \mathcal{O}\left(N^{\frac{1}{3}}K^{\frac{2}{3}}T^{\frac{2}{3}}\log^{\frac{2}{3}}(NKT)\right)$.

Proof. Note that the instantaneous regret $r^t(p^t)$ in any round t can be at most 1 because $\mathrm{NSW}(p,\mu^*) \in [0,1]$ for every policy p. Thus,

$$\mathbb{E}[R^T] = \sum_{t=1}^{T}\mathbb{E}[r^t] \le KL\cdot 1 + (T-KL)\cdot\mathbb{E}[\mathrm{NSW}(p^*,\mu^*) - \mathrm{NSW}(\widehat{p},\mu^*)]. \tag{1}$$

Thus, our goal is to bound $\mathbb{E}[\mathrm{NSW}(p^*,\mu^*) - \mathrm{NSW}(\widehat{p},\mu^*)]$. We bound this in two ways.

First approach: We present this approach briefly since it largely mimics the classical analysis with an application of Lemma 3. Here, we bound how much $\widehat{\mu}$ can deviate from μ^*. Specifically, we let $\epsilon = \sqrt{\frac{\log(NKT)}{L}}$ and define the event $\mathcal{E} \triangleq \forall i\in[N], \forall j\in[K] : \left|\widehat{\mu}_{i,j} - \mu^*_{i,j}\right| \le \epsilon$. Since L is fixed, we have $\mathbb{E}[\widehat{\mu}_{i,j}] = \mu^*_{i,j}$. Hence, we can apply Hoeffding's inequality followed by the union bound to derive $\Pr[\mathcal{E}] \ge 1 - 2/T^2$. Conditioned on $\mathcal{E}$, from Lemma 3 we have $\mathrm{NSW}(p,\mu^*) - \mathrm{NSW}(p,\widehat{\mu}) \le N\epsilon$ for every policy p, which implies

$$\mathrm{NSW}(p^*,\mu^*) \le \mathrm{NSW}(p^*,\widehat{\mu}) + N\epsilon \le \mathrm{NSW}(\widehat{p},\widehat{\mu}) + N\epsilon \le \mathrm{NSW}(\widehat{p},\mu^*) + 2N\epsilon,$$

where the second transition is because $\widehat{p} \in \arg\max_{p\in\Delta^K}\mathrm{NSW}(p,\widehat{\mu})$. Substituting this into Equation (1), using the fact that $\mathbb{E}[\mathrm{NSW}(p^*,\mu^*) - \mathrm{NSW}(\widehat{p},\mu^*)] \le 1\cdot\mathbb{E}[\mathrm{NSW}(p^*,\mu^*) - \mathrm{NSW}(\widehat{p},\mu^*)|\mathcal{E}] + \Pr[\neg\mathcal{E}]\cdot 1$, and setting $L = \Theta\left(N^{\frac{2}{3}}K^{-\frac{2}{3}}T^{\frac{2}{3}}\log^{\frac{1}{3}}(NKT)\right)$ yields the first regret bound.

Second approach: We now focus on another approach for bounding $\mathbb{E}[\mathrm{NSW}(p^*,\mu^*) - \mathrm{NSW}(\widehat{p},\mu^*)]$, which is more intricate and offers a different tradeoff between the dependence on N and K. Notice that for a given p, $\mathbb{E}[\mathrm{NSW}(p,\widehat{\mu})] = \mathrm{NSW}(p,\mu^*)$ because all $\widehat{\mu}_{i,j}$-s are independent and expectation decomposes over sums and products of independent random variables. Thus, we can use McDiarmid's inequality to bound $|\mathrm{NSW}(p,\widehat{\mu}) - \mathrm{NSW}(p,\mu^*)|$ at a given p.

Fix a δ-cover $\mathcal{P}$ of $(\Delta^K, \|\cdot\|_1)$ with $|\mathcal{P}| \le (1+2/\delta)^K$. Fix $p\in\mathcal{P}$. Notice that $\widehat{\mu}_{i,j} = (1/L)\cdot\sum_{s=1}^{L} X^s_{i,j}$, where $X^s_{i,j}$ is the reward to agent i from the s-th pull of arm j during the exploration phase.

We thus decompose $\widehat{\mu}$ into $N\cdot L$ random variables: for each $i\in[N]$ and $s\in[L]$, we let $X^s_i = (X^s_{i,j})_{j\in[K]}$. To apply McDiarmid's inequality, we need to analyze the maximum amount c^s_i

by which changing X_i^s can change $\text{NSW}(p, \widehat{\mu})$. Using Lemma 3, it is easy to see that $c_i^s \leq 1/L$ for each $i \in [N]$ and $s \in [L]$. Now, applying McDiarmid's inequality, we have

$$\Pr\left[|\text{NSW}(p, \widehat{\mu}) - \text{NSW}(p, \mu^*)| \leq \epsilon\right] \leq 2e^{\frac{-2\epsilon^2}{\sum_{i\in[N], s\in[L]}(c_i^s)^2}} = 2e^{\frac{-2L\epsilon^2}{N}}.$$

Setting $\epsilon = \sqrt{\frac{N\log(|\mathcal{P}|T)}{2L}}$, we have that for each $p \in \mathcal{P}$,

$$\Pr\left[|\text{NSW}(p, \widehat{\mu}) - \text{NSW}(p, \mu^*)| \leq \sqrt{\frac{N\log(|\mathcal{P}|T)}{2L}}\right] \leq \frac{2}{|\mathcal{P}|T}.$$

Using the union bound, we have that

$$\Pr\left[\forall p \in \mathcal{P} : |\text{NSW}(p, \widehat{\mu}) - \text{NSW}(p, \mu^*)| \leq \sqrt{\frac{N\log(|\mathcal{P}|T)}{2L}}\right] \geq 1 - \frac{2}{T}.$$

For $p \in \Delta^K$, let $\overline{p} \in \arg\min_{p'\in\mathcal{P}} \|p - p'\|_1$. Then, since $\mathcal{P}$ is a δ-cover, we have $\|p - \overline{p}\|_1 \leq \delta$. Thus, due to Lemma 2, we have

$$\begin{aligned}|\text{NSW}(p, \widehat{\mu}) - \text{NSW}(p, \mu^*)| &\leq \sum_{\mu\in\{\widehat{\mu},\mu^*\}} |\text{NSW}(p, \mu) - \text{NSW}(\overline{p}, \mu)| + |\text{NSW}(\overline{p}, \widehat{\mu}) - \text{NSW}(\overline{p}, \mu^*)| \\ &\leq 2N\delta + |\text{NSW}(\overline{p}, \widehat{\mu}) - \text{NSW}(\overline{p}, \mu^*)|.\end{aligned}$$

Setting $\delta = \frac{1}{NT}$, we have

$$\Pr\left[\forall p \in \Delta^K : |\text{NSW}(p, \widehat{\mu}) - \text{NSW}(p, \mu^*)| \leq \frac{2}{T} + \sqrt{\frac{N\log(|\mathcal{P}|T)}{2L}}\right] \geq 1 - \frac{2}{T}.$$

Next, we use the fact that

$$\text{NSW}(p^*, \mu^*) - \text{NSW}(\widehat{p}, \mu^*) \leq \sum_{p\in\{p^*,\widehat{p}\}} |\text{NSW}(p, \widehat{\mu}) - \text{NSW}(p, \mu^*)|.$$

Hence,

$$\Pr\left[|\text{NSW}(p^*, \mu^*) - \text{NSW}(\widehat{p}, \mu^*)| \leq \frac{4}{T} + \sqrt{\frac{2N\log(|\mathcal{P}|T)}{L}}\right] \geq 1 - \frac{2}{T}.$$

Next, we substitute $|\mathcal{P}| \leq (1+2/\delta)^K \leq (3/\delta)^K$, $\delta = \frac{1}{NT}$, and $L = \Theta\left(N^{\frac{1}{3}}K^{-\frac{1}{3}}T^{\frac{2}{3}}\log^{\frac{2}{3}}(NKT)\right)$, and then substitute the derived bound in Equation (1) to get the second regret bound. □

4 Epsilon-Greedy

A slightly more sophisticated algorithm than Explore-First is Epsilon-Greedy, which is presented as Algorithm 2. It spreads out exploration instead of performing it all at the beginning. Specifically, at each round t, it performs exploration with probability ϵ^t, and exploitation otherwise. Exploration cycles through the arms in a round-robin fashion, while exploitation uses the policy p^t with the highest Nash social welfare under the current estimated reward matrix (rather than choosing a single estimated best arm as in the classical algorithm).

ALGORITHM 2: ϵ^t-Greedy

Input: Number of agents N, number of arms K

Parameters : Exploration probabilities ϵ^t for $t \in \mathbb{N}$

$curr \leftarrow 1$ `// Next arm to pull during exploration`

for $t = 1, 2, \ldots,$ **do**

  Toss a coin with success probability ϵ^t

  if *success* **then** `// Exploration`

    `// Round-robin among arms during exploration`

    $p^t \leftarrow$ policy that puts probability 1 on arm $curr$ `// Pull it deterministically`

    $curr \leftarrow curr + 1$ `// When` $curr$ `becomes` $K+1$`, reset to 1`

  else `// Exploitation`

    Compute the estimated reward matrix $\widehat{\mu}^t$ from the rewards observed so far

    $p^t \leftarrow \arg\max_{p \in \Delta^K} \mathrm{NSW}(p, \widehat{\mu}^t)$

  end

end

We remark that, like Explore-First, Epsilon-Greedy can also be implemented efficiently. The only non-trivial step is to compute $\widehat{p} \in \arg\max_{p \in \Delta^K} \mathrm{NSW}(p, \widehat{\mu})$, which, as we mentioned before, can be done in polynomial time.

The key advantage of Epsilon-Greedy over Explore-First is that it is horizon-independent. However, in the $\widehat{\mu}$ computed in Explore-First at the end of exploration, each $\widehat{\mu}_{i,j}$ is the average of L iid samples, where L is fixed. In contrast, in the $\widehat{\mu}^t$ computed in Epsilon-Greedy in round t, each $\widehat{\mu}^t_{i,j}$ is the average of n^t_j iid samples. The fact that n^t_j is itself a random variable and the $\widehat{\mu}^t_{i,j}$-s are correlated through the n^t_j-s prevents a direct application of certain statistical inequalities, thus complicating the analysis of Epsilon-Greedy. To address this, we first present a sequence of useful lemmas that apply to *any* algorithm, and then use them to prove the regret bounds of Epsilon-Greedy and later UCB.

4.1 Useful Lemmas

Recall that μ^* and $\widehat{\mu}^t$ denote the true reward matrix and the estimated reward matrix at the beginning of round t, respectively. Our goal is to find an upper bound on the quantity $\left|\mathrm{NSW}(p, \mu^*) - \mathrm{NSW}(p, \widehat{\mu}^t)\right|$ that, with high probability, holds at every $p \in \Delta^K$ simultaneously. To that end, we first need to show that $\widehat{\mu}^t$ will be close to μ^* with high probability.

Recall that random variable n^t_j denotes the number of times arm j is pulled by an algorithm before round t, and $\widehat{\mu}^t_{i,j}$ is an average over n^t_j independent samples. Hence, we cannot directly apply Hoeffding's inequality, but we can nonetheless use standard tricks from the literature.

Lemma 4. *Define* $r^t_j = \sqrt{\frac{2\log(NKt)}{n^t_j}}$*, and event*

$$\mathcal{E}^t \triangleq \forall i \in [N], j \in [K] : \left|\widehat{\mu}^t_{i,j} - \mu^*_{i,j}\right| \leq r^t_j.$$

Then, for any algorithm and any t*, we have* $\Pr[\mathcal{E}^t] \geq 1 - \frac{2}{t^3}$.

Proof. Fix t. For $i \in [N]$, $j \in [K]$, and $\ell \in [t]$, let $\overline{v}^\ell_{i,j}$ denote the average reward to agent i from

the first ℓ pulls of arm j, and define $\overline{r}_j^\ell = \sqrt{\frac{2\log(NKt)}{\ell}}$. Then, by Hoeffding's inequality, we have

$$\forall i \in [N], j \in [K], \ell \in [t] : \Pr\left[\left|\overline{v}_{i,j}^\ell - \mu_{i,j}\right| > \overline{r}_j^\ell\right] \leq \frac{2}{(NKt)^4}.$$

By the union bound, we get

$$\Pr\left[\forall i \in [N], j \in [K], \ell \in [t] : \left|\overline{v}_{i,j}^\ell - \mu_{i,j}\right| \leq \overline{r}_j^\ell\right] \geq 1 - \frac{2}{(NKt)^3}.$$

Because $n_j^t \in [t]$ for each $j \in [K]$, the above event implies our desired event $\mathcal{E}^t$. Hence, we have that $\Pr[\mathcal{E}^t] \geq 1 - 2/(NKt)^3 \geq 1 - 2/t^3$. □

Conditioned on $\mathcal{E}^t$, we wish to bound $\left|\mathrm{NSW}(p, \mu^*) - \mathrm{NSW}(p, \widehat{\mu}^t)\right|$ simultaneously at all $p \in \Delta^K$. We provide two such (incomparable) bounds, which will form the crux of our regret bound analysis. The first bound is a direct application of the Lipschitz-continuity analysis from Lemma 3.

Lemma 5. *Conditioned on $\mathcal{E}^t$, we have that*

$$\forall p \in \Delta^K : \left|\mathrm{NSW}(p, \widehat{\mu}^t) - \mathrm{NSW}(p, \mu^*)\right| \leq N \cdot \sum_{j \in [K]} p_j \cdot r_j^t.$$

Proof. Conditioned on $\mathcal{E}^t$, we have $\left|\widehat{\mu}_{i,j}^t - \mu_{i,j}^*\right| \leq r_j^t$ for each $j \in [K]$. In that case, it is easy to see that the upper bound from Lemma 3 becomes $N \cdot \sum_{j \in [K]} p_j \cdot r_j^t$. □

The factor of N in Lemma 5 stems from analyzing how much $\widehat{\mu}^t$ may deviate from μ^* conditioned on $\mathcal{E}^t$, in the worst case. However, even after conditioning on $\mathcal{E}^t$, $\widehat{\mu}^t$ remains a random variable. Hence, one may expect that its deviation, and thus the difference $\left|\mathrm{NSW}(p, \widehat{\mu}^t) - \mathrm{NSW}(p, \mu^*)\right|$, may be smaller in expectation. Thus, to derive a different bound than in Lemma 5, we wish to apply McDiarmid's inequality. However, there are two issues in doing so directly.

- McDiarmid's inequality bounds the deviation of $\mathrm{NSW}(p, \widehat{\mu}^t)$ from its expected value. If $\widehat{\mu}^t$ consisted of independent random variables, like in Explore-First, this would be equal to $\mathrm{NSW}(p, \mu^*)$. However, in general, these variables may be correlated through n_j^t. We use a conditioning trick to address this issue.
- We cannot hope to apply McDiarmid's inequality at each $p \in \Delta^K$ separately and use the union bound because Δ^K is infinite. So we apply it at each p in a δ-cover of Δ^K, apply the union bound, and then translate the guarantee to nearby $p \in \Delta^K$ using the Lipschitz-continuity analysis from Lemma 2.

The next result is one of the key technical contributions of our work with a rather long proof.

Lemma 6. *Define the event*

$$\mathcal{H}^t \triangleq \forall p \in \Delta^K : \left|\mathrm{NSW}(p, \widehat{\mu}^t) - \mathrm{NSW}(p, \mu^*)\right| \leq \sqrt{12NK\log(NKt)} \cdot \sum_{j \in [K]} p_j \cdot r_j^t + \frac{4}{t}.$$

Then, for any algorithm and any t, we have $\Pr[\mathcal{H}^t | \mathcal{E}^t] \geq 1 - 2/t^3$.

Proof. Fix $p \in \Delta^K$. Fix $\delta > 0$, and let $\mathcal{P}$ be a δ-cover of the policy simplex Δ^K with $|\mathcal{P}| \leq (1 + 2/\delta)^K$ [32, p. 126].

Conditioned on $\mathcal{E}^t$ (i.e. $\left|\widehat{\mu}_{i,j}^t - \mu_{i,j}^*\right| \leq r_j^t = \sqrt{\frac{2\log(NKt)}{n_j^t}}, \forall i \in [N], j \in [K]$), we wish to derive a high probability bound on $\left|\mathrm{NSW}(p, \widehat{\mu}^t) - \mathrm{NSW}(p, \mu^*)\right|$. We can bound the deviation of $\mathrm{NSW}(p, \widehat{\mu}^t)$ from its expected value. However, unlike in the case of Explore-First, we cannot directly claim that the expected value is $\mathrm{NSW}(p, \mu^*)$ because, as we mentioned above, $\widehat{\mu}^t$ consists of random variables that may be correlated through the random varaible $n^t = (n_1^t, \ldots, n_K^t)$ taking values in $[t]^K$. Thus, we need a more careful argument.

For $i \in [N]$, $j \in [K]$, and $\ell_j \in [t]$, let $\overline{v}_{i,j}^{\ell_j}$ denote the average reward to agent i from the first ℓ_j pulls of arm j, and define $\overline{r}_j^{\ell_j} = \sqrt{\frac{2\log(NKt)}{\ell_j}}$. Let $\ell = (\ell_1, \ldots, \ell_K) \in [t]^K$ and $\overline{v}^\ell = (\overline{v}_{i,j}^{\ell_j})_{i\in[N], j\in[K]}$. Each $\overline{v}_{i,j}^{\ell_j}$ is independent and satisfies $\mathbb{E}[\overline{v}_{i,j}^{\ell_j}] = \mu_{i,j}^*$. Since expectation decomposes over sums and products of independent random variables, we have $\mathbb{E}[\mathrm{NSW}(p, \overline{v}^\ell)] = \mathrm{NSW}(p, \mu^*)$.

Evaluating conditional expectation: We next argue that further conditioning on the high probability event $\mathcal{E}^t$ does not change the expectation by much. Formally,

$$\begin{aligned}
&\left|\mathrm{NSW}(p, \mu^*) - \mathbb{E}[\mathrm{NSW}(p, \overline{v}^\ell)|\mathcal{E}^t]\right| \\
&= \left|\mathbb{E}[\mathrm{NSW}(p, \overline{v}^\ell)] - \mathbb{E}\left[\mathrm{NSW}(p, \overline{v}^\ell)|\mathcal{E}^t\right]\right| \\
&= \Pr[\neg\mathcal{E}^t] \cdot \left|\mathbb{E}\left[\mathrm{NSW}(p, \overline{v}^\ell)|\neg\mathcal{E}^t\right] - \mathbb{E}\left[\mathrm{NSW}(p, \overline{v}^\ell)|\mathcal{E}^t\right]\right| \\
&\leq \Pr[\neg\mathcal{E}^t] \leq \frac{2}{t^3} \leq \frac{2}{t},
\end{aligned} \tag{2}$$

where the penultimate transition holds because NSW is bounded in $[0, 1]$, and the final transition is due to Lemma 4.

Applying McDiarmid's inequality: We first decompose $\overline{v}^\ell$ into N random variables: for each $i \in [N]$, let $\overline{v}_i^\ell = (\overline{v}_{i,j}^\ell)_{j\in[K]}$. To apply McDiarmid's inequality, we need to analyze the maximum amount c_i by which changing $\overline{v}_i^\ell$ can change $\mathrm{NSW}(p, \overline{v}^\ell)$. Fix $i \in [N]$, and fix all the variables except $\overline{v}_i^\ell$. Conditioned on $\mathcal{E}^t$, each $\overline{v}_{i,j}^\ell$ can change by at most $2\overline{r}_j^{\ell_j}$. Hence, using Lemma 3, we have that $c_i \leq 2\sum_{j\in[K]} p_j \cdot \overline{r}_j^{\ell_j}$. Now, applying McDiarmid's inequality, we have

$$\forall \ell \in [t]^K : \Pr\left[\left|\mathrm{NSW}(p, \overline{v}^\ell) - \mathbb{E}\left[\mathrm{NSW}(p, \overline{v}^\ell)|\mathcal{E}^t\right]\right| \geq \epsilon \mid \mathcal{E}^t\right] \leq 2e^{\frac{-2\epsilon^2}{\sum_{i\in[N]} c_i^2}} \leq 2e^{\frac{-2\epsilon^2}{4N\cdot\left(\sum_{j\in[K]} p_j \cdot \overline{r}_j^{\ell_j}\right)^2}}.$$

Using Equation (2), and setting $\epsilon = \sqrt{2N\log(|\mathcal{P}|t^{K+3})} \cdot \sum_{j\in[K]} p_j \cdot \overline{r}_j^{\ell_j}$, we have that

$$\forall \ell \in [t]^K : \Pr\left[\left|\mathrm{NSW}(p, \overline{v}^\ell) - \mathrm{NSW}(p, \mu^*)\right| \geq \sqrt{2N\log(|\mathcal{P}|t^{K+3})} \cdot \sum_{j\in[K]} p_j \cdot \overline{r}_j^{\ell_j} + \frac{2}{t} \;\middle|\; \mathcal{E}^t\right] \leq \frac{2}{|\mathcal{P}|t^{K+3}}.$$

Next, by union bound, we get

$$\Pr\left[\forall \ell \in [t]^K : \left|\mathrm{NSW}(p, \overline{v}^\ell) - \mathrm{NSW}(p, \mu^*)\right| \geq \sqrt{2N\log(|\mathcal{P}|t^{K+3})} \cdot \sum_{j\in[K]} p_j \cdot \overline{r}_j^{\ell_j} + \frac{2}{t} \;\middle|\; \mathcal{E}^t\right] \leq \frac{2}{|\mathcal{P}|t^3}.$$

Because $n_j^t \in [t]$ for each $j \in [K]$, we have

$$\Pr\left[\left|\mathrm{NSW}(p,\widehat{\mu}^t) - \mathrm{NSW}(p,\mu^*)\right| \geq \sqrt{2N\log(|\mathcal{P}|t^{K+3})}\cdot\sum_{j\in[K]} p_j\cdot r_j^t + \frac{2}{t} \;\middle|\; \mathcal{E}^t\right] \leq \frac{2}{|\mathcal{P}|t^3}.$$

Extending to all policies in $\mathcal{P}$: Using the union bound, we have that

$$\Pr\left[\forall p\in\mathcal{P}: \left|\mathrm{NSW}(p,\widehat{\mu}^t) - \mathrm{NSW}(p,\mu^*)\right| \leq \sqrt{2N\log(|\mathcal{P}|t^{K+3})}\cdot\sum_{j\in[K]} p_j\cdot r_j^t + \frac{2}{t} \;\middle|\; \mathcal{E}^t\right] \geq 1-\frac{2}{t^3}.$$

Extending to all policies in $\mathbf{\Delta^K}$: For $p\in\Delta^K$, let $\overline{p}\in\arg\min_{p'\in\mathcal{P}}\|p-p'\|_1$. Then, since $\mathcal{P}$ is a δ-cover, we have $\|p-\overline{p}\|_1\leq\delta$. Thus, due to Lemma 2, we have

$$\begin{aligned}\left|\mathrm{NSW}(p,\widehat{\mu}^t) - \mathrm{NSW}(p,\mu^*)\right| &\leq \sum_{\mu\in\{\widehat{\mu}^t,\mu^*\}} \left|\mathrm{NSW}(p,\mu) - \mathrm{NSW}(\overline{p},\mu)\right| \\ &\quad + \left|\mathrm{NSW}(\overline{p},\widehat{\mu}^t) - \mathrm{NSW}(\overline{p},\mu^*)\right| \\ &\leq 2N\delta + \left|\mathrm{NSW}(\overline{p},\widehat{\mu}^t) - \mathrm{NSW}(\overline{p},\mu^*)\right|.\end{aligned}$$

Setting $\delta=\frac{1}{Nt}$, we have

$$\Pr\left[\forall p\in\Delta^K: \left|\mathrm{NSW}(p,\widehat{\mu}^t) - \mathrm{NSW}(p,\mu^*)\right| \leq \sqrt{2N\log(|\mathcal{P}|t^{K+3})}\cdot\sum_{j\in[K]} p_j\cdot r_j^t + \frac{4}{t} \;\middle|\; \mathcal{E}^t\right] \geq 1-\frac{2}{t^3}.$$

Substituting $|\mathcal{P}|\leq(1+2/\delta)^K\leq(3/\delta)^K$ with $\delta=\frac{1}{Nt}$ yields the desired bound. □

Finally, we use the following simple observation in deriving our asymptotic bounds.

Proposition 1. *For constant $p\in\mathbb{R}$, $\sum_{t=1}^T t^p$ is $\Theta(\log T)$ if $p=-1$ and $\Theta\left(T^{p+1}\right)$ otherwise.*

4.2 Analysis of Epsilon-Greedy

We can now use these lemmas to derive the regret bounds for Epsilon-Greedy.

Theorem 2. *Epsilon-Greedy is horizon-independent, and has the following expected regret at any round T.*

- *If $\epsilon^t=\Theta\left(N^{\frac{2}{3}}K^{\frac{1}{3}}t^{-\frac{1}{3}}\log^{\frac{1}{3}}(NKt)\right)$ for all t, $\mathbb{E}[R^T]=\mathcal{O}\left(N^{\frac{2}{3}}K^{\frac{1}{3}}T^{\frac{2}{3}}\log^{\frac{1}{3}}(NKT)\right)$.*
- *If $\epsilon^t=\Theta\left(N^{\frac{1}{3}}K^{\frac{2}{3}}t^{-\frac{1}{3}}\log^{\frac{2}{3}}(NKt)\right)$ for all t, $\mathbb{E}[R^T]=\mathcal{O}\left(N^{\frac{1}{3}}K^{\frac{2}{3}}T^{\frac{2}{3}}\log^{\frac{2}{3}}(NKT)\right)$.*

Proof. Fix $t\in[T]$. Let b^t denote the number of times Epsilon-Greedy performs exploration up to round t. Note that $\mathbb{E}[b^t]=\sum_{s=1}^t\epsilon^s\geq t\epsilon^t$, where the last step follows from the fact that ϵ^t is monotonically decreasing in both cases of the theorem. Let $\theta>0$ be a constant such that $\epsilon^t\geq\theta\cdot t^{-1/3}$ in both cases of the theorem.

Define the event $\mathcal{B}^t\triangleq b^t\geq\gamma\cdot t\epsilon^t$, where $\gamma=1-1/\theta$. Then, by Hoeffding's inequality, we have

$$\Pr[\neg\mathcal{B}^t]\leq e^{-2(1-\gamma)^2\theta^2t^{1/3}}=e^{-2t^{1/3}}\leq e^{-\log t}=\frac{1}{t}. \tag{3}$$

Because the algorithm performs round-robin during exploration, conditioned on $\mathcal{B}^t$, we have that $n_j^t \geq \frac{b^t}{K} \geq \frac{\gamma \cdot t\epsilon^t}{K}$ for each arm j,[4] which implies $r_j^t \leq \sqrt{\frac{2K\log(NKt)}{\gamma \cdot t\epsilon^t}}$ for each j. Thus, conditioned on $\mathcal{B}^t$, we have

$$\forall p \in \Delta^K : \sum_{j\in[K]} p_j \cdot r_j^t \leq \max_{j\in[K]} r_j^t \leq \sqrt{\frac{2K\log(NKt)}{\gamma \cdot t\epsilon^t}}. \tag{4}$$

We are now ready to use the bounds from Lemmas 5 and 6. We focus on the event

$$\mathcal{C}_\alpha^t \triangleq \forall p \in \Delta^K : \left|\mathrm{NSW}(p,\mu^*) - \mathrm{NSW}(p,\widehat{\mu}^t)\right| \leq \alpha^t \cdot \sum_{j\in[K]} p_j \cdot r_j^t + \frac{4}{t}.$$

Conditioned on $\mathcal{E}^t \wedge \mathcal{H}^t$, note that $\mathcal{C}_\alpha^t$ holds for $\alpha^t = N$ due to Lemma 5, and for $\alpha^t = \sqrt{12NK\log NKt}$ due to Lemma 6.

Let $\widehat{p}^t \in \arg\max_{p\in\Delta^K} \mathrm{NSW}(p,\widehat{\mu}^t)$. We wish to bound the regret $\mathrm{NSW}(p^*,\mu^*) - \mathrm{NSW}(\widehat{p}^t,\mu^*)$ that Epsilon-Greedy incurs when performing exploitation in round t by choosing policy $\widehat{p}^t$. Conditioned on $\mathcal{E}^t \wedge \mathcal{H}^t \wedge \mathcal{B}^t$, we have

$$\begin{aligned}
&\mathrm{NSW}(p^*,\mu^*) - \mathrm{NSW}(\widehat{p}^t,\mu^*) \\
&= \left(\mathrm{NSW}(p^*,\mu^*) - \mathrm{NSW}(p^*,\widehat{\mu}^t)\right) + \left(\mathrm{NSW}(p^*,\widehat{\mu}^t) - \mathrm{NSW}(\widehat{p}^t,\widehat{\mu}^t)\right) + \left(\mathrm{NSW}(\widehat{p}^t,\widehat{\mu}^t) - \mathrm{NSW}(\widehat{p}^t,\mu^*)\right) \\
&\leq \sum_{p\in\{p^*,\widehat{p}^t\}} \left|\mathrm{NSW}(p,\mu^*) - \mathrm{NSW}(p,\widehat{\mu}^t)\right| \leq 2\alpha^t\sqrt{\frac{2K\log(NKt)}{\gamma \cdot t\epsilon^t}} + \frac{8}{t},
\end{aligned} \tag{5}$$

where the penultimate transition holds because $\widehat{p}^t$ is the optimal policy under $\widehat{\mu}^t$, so $\mathrm{NSW}(p^*,\widehat{\mu}^t) \leq \mathrm{NSW}(\widehat{p}^t,\widehat{\mu}^t)$, and the final transition follows from Equation (4) and the fact that $\mathcal{E}^t \wedge \mathcal{H}^t$ imply $\mathcal{C}_\alpha^t$.

We are now ready to analyze the expected regret of Epsilon-Greedy at round T. We have

$$\begin{aligned}
\mathbb{E}[R^T] &= \sum_{t=1}^T \mathbb{E}[r^t] \leq \sum_{t=1}^T \mathbb{E}\left[\epsilon^t \cdot 1 + (1-\epsilon^t)\cdot\left(\mathrm{NSW}(p^*,\mu^*) - \mathrm{NSW}(\widehat{p}^t,\mu^*)\right)\right] \\
&\leq \sum_{t=1}^T \Bigg(\epsilon^t + \Pr\left[\mathcal{E}^t \wedge \mathcal{H}^t \wedge \mathcal{B}^t\right] \cdot \mathbb{E}\left[\mathrm{NSW}(p^*,\mu^*) - \mathrm{NSW}(\widehat{p}^t,\mu^*) \;\middle|\; \mathcal{E}^t \wedge \mathcal{H}^t \wedge \mathcal{C}_\alpha^t\right] \\
&\qquad\qquad + \Pr\left[\neg\mathcal{E}^t \vee \neg\mathcal{H}^t \vee \neg\mathcal{B}^t\right] \cdot 1\Bigg) \\
&\leq \sum_{t=1}^T \left(\epsilon^t + 2\alpha^t\sqrt{\frac{2K\log(NKt)}{\gamma \cdot t\epsilon^t}} + \frac{8}{t} + \frac{4}{t^3} + \frac{1}{t}\right),
\end{aligned}$$

where the final transition holds due to Equation (5), Lemma 4, Lemma 6, and Equation (3). Notice that we are using the fact that

$$\Pr[\mathcal{E}^t \wedge \mathcal{H}^t] = \Pr[\mathcal{E}^t] \cdot \Pr[\mathcal{H}^t|\mathcal{E}^t] \geq (1-2/t^3)\cdot(1-2/t^3) \geq 1 - 4/t^3.$$

To obtain the first regret bound, we set $\epsilon^t = \Theta\left(N^{\frac{2}{3}}K^{\frac{1}{3}}t^{-\frac{1}{3}}\log^{\frac{1}{3}}(NKt)\right)$ and $\alpha^t = N$, and obtain

$$\mathbb{E}[R^T] = \mathcal{O}\left(N^{\frac{2}{3}}K^{\frac{1}{3}}\log^{\frac{1}{3}}(NKT)\sum_{t=1}^T t^{-\frac{1}{3}}\right) = \mathcal{O}\left(N^{\frac{2}{3}}K^{\frac{1}{3}}T^{\frac{2}{3}}\log^{\frac{1}{3}}(NKT)\right).$$

[4]Technically, $n_j^t \geq \lfloor \frac{b^t}{K} \rfloor$ for each arm j, but we omit the floor for the ease of presentation.

For the second regret bound, we set $\epsilon^t = \Theta\left(N^{\frac{1}{3}} K^{\frac{2}{3}} t^{-\frac{1}{3}} \log^{\frac{2}{3}}(NKt)\right)$ and $\alpha^t = \sqrt{12NK\log(NKt)}$, and obtain

$$\mathbb{E}[R^T] = \mathcal{O}\left(N^{\frac{1}{3}} K^{\frac{2}{3}} \log^{\frac{2}{3}}(NKT) \sum_{t=1}^{T} t^{-1/3}\right) = \mathcal{O}\left(N^{\frac{1}{3}} K^{\frac{2}{3}} T^{\frac{2}{3}} \log^{\frac{2}{3}}(NKT)\right).$$

Note that in both cases, we omit the $\mathcal{O}(1/t)$ and $\mathcal{O}\left(1/t^3\right)$ terms because they are dominated by the $\mathcal{O}\left(1/t^{1/3}\right)$ term. In both cases, we use Proposition 1 at the end. □

5 UCB

ALGORITHM 3: UCB

Input: Number of agents N, number of arms K

Parameters : Confidence parameter α^t for each $t \in \mathbb{N}$

`// Pull each arm once`

for $t = 1, \ldots, K$ **do**

  $p^t \leftarrow$ policy that puts probability 1 on arm t `// Pull arm` t `deterministically`

end

for $t = K+1, \ldots$ **do**

  Compute the estimated reward matrix $\widehat{\mu}^t$

  $p^t \leftarrow \arg\max_{p \in \Delta^K} \mathrm{NSW}(p, \widehat{\mu}^t) + \alpha^t \sum_{j \in [K]} p_j \cdot r_j^t$, where $r_j^t \triangleq \sqrt{\frac{\log(NKt)}{n_j^t}}$.

end

In the classical multi-armed bandit setting, UCB first pulls each arm once. Afterwards, it merges exploration and exploitation cleverly by pulling, in each round, an arm maximizing the sum of its estimated reward and a confidence interval term similar to r_j^t in Algorithm 3. Our multi-agent variant similarly selects a policy that maximizes the estimated Nash social welfare plus a confidence term for a policy, which simply takes a linear combination of the confidence intervals of the arms.

Unlike Explore-First and Epsilon-Greedy, which can be implemented efficiently, it is not clear if our UCB variant admits an efficient implementation due to this step of computing $\arg\max_{p \in \Delta^K} \mathrm{NSW}(p, \widehat{\mu}) + \alpha^t \sum_{j \in [K]} p_j r_j^t$. Due to the added linear term, the objective is no longer log-concave. This remains a challenging open problem. However, we notice that this can also be viewed as the problem of optimizing a polynomial over a simplex, which, while NP-hard in general, is known to admit a PTAS when the degree is a constant [13, 14]. Hence, in our case, when the number of agents N is a constant, this step can be computed approximately, but it remains to be seen how this approximation translates to the final regret bounds.

We show that UCB achieves the desired $\sqrt{T}$ dependence on the horizon (up to logarithmic factors). In Section 6, we show that this is optimal.

Theorem 3. *UCB is horizon-independent, and has the following expected regret at any round T.*

- *If $\alpha^t = N$ for all t, $\mathbb{E}[R^T] = \mathcal{O}\left(NKT^{\frac{1}{2}} \log(NKT)\right)$.*
- *If $\alpha^t = \sqrt{12NK\log(NKt)}$ for all t, $\mathbb{E}[R^T] = \mathcal{O}\left(N^{\frac{1}{2}} K^{\frac{3}{2}} T^{\frac{1}{2}} \log^{\frac{3}{2}}(NKT)\right)$.*

Proof. Fix one of two parameter choices:

1. $\alpha^t = N$ for all t and $c = N$.
2. $\alpha^t = \sqrt{12NK\log(NKt)}$ for all t and $c = \sqrt{12NK\log(NKT)}$.

Note that in both cases, $\alpha^t \le c$ for all t. Hence, c serves as an upper bound on α^t that does not depend on t. We show that in both cases, running UCB with the α^t parameter value yields a regret bound of $\mathbb{E}[R^T] = O(cK\sqrt{T}\log(NKT))$. Substituting the two choices of c then yields the two regret bounds.

Let us again focus on the event

$$\mathcal{C}_\alpha^t \triangleq \forall p \in \Delta^K : \left|\mathrm{NSW}(p,\mu^*) - \mathrm{NSW}(p,\widehat{\mu}^t)\right| \le \alpha^t \cdot \sum_{j\in[K]} p_j \cdot r_j^t + \frac{4}{t}.$$

Recall the clean events $\mathcal{E}^t$ and $\mathcal{H}^t$ defined in Lemmas 4 and 6. As argued in the proof of Theorem 2, we have that $\mathcal{E}^t \wedge \mathcal{H}^t$ implies $\mathcal{C}_\alpha^t$ for both choices of α^t; for $\alpha^t = N$, it follows from Lemma 5, and for $\alpha^t = \sqrt{12NK\log(NKt)}$, it follows from Lemma 6. Using Lemmas 4 and 6 as well as the union bound, we have that $\Pr[\neg\mathcal{C}_\alpha^t] \le 1 - \Pr[\mathcal{E}^t \wedge \mathcal{H}^t] = 1 - \Pr[\mathcal{E}^t]\cdot\Pr[\mathcal{H}^t|\mathcal{E}^t] \le 1 - (1-2/t^3)\cdot(1-2/t^3) \le 4/t^3$.

Define a clean event $\mathcal{C}_\alpha^* \triangleq \bigwedge_{t\ge\sqrt{T}} \mathcal{C}_\alpha^t$. Here, we do not care about the first $\sqrt{T}$ rounds because the maximum regret from these rounds is $\mathcal{O}\left(\sqrt{T}\right)$, which is permissible given our desired regret bounds. By the union bound, we have $\Pr[\neg\mathcal{C}_\alpha^*] \le T \cdot 4/(\sqrt{T})^3 = 4/\sqrt{T}$. Thus, $\mathcal{C}_\alpha^*$ is a high-probability event. In what follows, we derive an upper bound on the expected regret conditioned on $\mathcal{C}_\alpha^*$, i.e., $\mathbb{E}[R^T|\mathcal{C}_\alpha^*]$. Since conditioning on a high-probability event does not affect the expected value significantly, the desired regret bound will then follow.

For any $t \in [T]$, conditioned on $\mathcal{C}_\alpha^t$ we have that

$$\begin{aligned}
\mathrm{NSW}(p^*,\mu^*) &\le \mathrm{NSW}(p^*,\widehat{\mu}^t) + \alpha^t \sum_{j\in[K]} p_j^* \cdot r_j^t + \frac{4}{t} \\
&\le \mathrm{NSW}(p^t,\widehat{\mu}^t) + \alpha^t \sum_{j\in[K]} p_j^t \cdot r_j^t + \frac{4}{t} \\
&\le \mathrm{NSW}(p^t,\mu^*) + 2\alpha^t \sum_{j\in[K]} p_j^t \cdot r_j^t + \frac{8}{t},
\end{aligned}$$

where the first and the last transition are from conditioning on $\mathcal{C}_\alpha^t$, and the second transition is because $p = p^t$ maximizes the quantity $\mathrm{NSW}(p,\widehat{\mu}^t) + \alpha^t \sum_{j\in[K]} p_j \cdot r_j^t$ in the UCB algorithm.

Let us write $p^{[T]} = (p^1,\ldots,p^T)$ for the random variable denoting the policies used by the algorithm, and $\overline{p}^{[T]} = (\overline{p}^1,\ldots,\overline{p}^T)$ to denote a specific value in $(\Delta^K)^T$ taken by the random variable.

Instead of analyzing $\mathbb{E}[R^T|\mathcal{C}_\alpha^*]$ directly, we further condition on UCB choosing a specific sequence of policies $\overline{p}^{[T]}$. That is, we are interested in deriving an upper bound on $\mathbb{E}[R^T|\mathcal{C}_\alpha^* \wedge p^{[T]} = \overline{p}^{[T]}]$.[5] Interestingly, we show that this quantity is $\mathcal{O}\left(cK\sqrt{T}\log(NKT)\right)$ for *every* possible $\overline{p}^{[T]}$.

Fix an arbitrary $\overline{p}^{[T]}$. For $t \in [T]$ and $j \in [K]$, define $q_j^t = \sum_{s=1}^t \overline{p}_j^s$. Then, $\mathbb{E}[n_j^t|p^{[T]} = \overline{p}^{[T]}] = q_j^t$. For each $j \in [K]$, let T_j be the smallest t for which $q_j^t \ge 2\sqrt{T\log(NKT)}$ (if no such t exists, let $T_j = T$); note that given $\overline{p}^{[T]}$, T_j is fixed and not a random variable. Also, we have that $q_j^{T_j} = \Theta\left(\sqrt{T\log(NKT)}\right)$ for each $j \in [K]$.

[5] Note that even after conditioning on $p^{[T]} = \overline{p}^{[T]}$, there is still randomness left in sampling actions from the policies and sampling the rewards of those actions.

Let us define a clean event $\mathcal{B} \triangleq \forall j \in [K], n_j^{T_j} \geq \sqrt{T \log(NKT)}$. We first show that this is a high probability event. Indeed, using Hoeffding's inequality, we have that for each $j \in [K]$,

$$\begin{aligned}\Pr\left[n_j^{T_j} < \sqrt{T\log(NKT)} \;\middle|\; \mathcal{C}_\alpha^* \wedge p^{[T]} = \overline{p}^{[T]}\right] &\leq \Pr\left[n_j^{T_j} < s_j^{T_j} - \sqrt{T\log(NKT)} \;\middle|\; \mathcal{C}_\alpha^* \wedge p^{[T]} = \overline{p}^{[T]}\right] \\ &\leq \frac{1}{N^2K^2T^2}.\end{aligned}$$

Taking union bound over $j \in [K]$, we have that $\Pr\left[\neg\mathcal{B} \;\middle|\; \mathcal{C}_\alpha^* \wedge p^{[T]} = \overline{p}^{[T]}\right] \leq \frac{1}{N^2KT^2}$.

Next, we bound $\mathbb{E}[R^T | \mathcal{C}_\alpha^* \wedge p^{[T]} = \overline{p}^{[T]}]$ by using event $\mathcal{B}$.

$$\begin{aligned}&\mathbb{E}\left[R^T \;\middle|\; \mathcal{C}_\alpha^* \wedge p^{[T]} = \overline{p}^{[T]}\right] \\ &= \sum_{t=1}^{T} \mathbb{E}\left[\mathrm{NSW}(p^*, \mu^*) - \mathrm{NSW}(p^t, \mu^*) \;\middle|\; \mathcal{C}_\alpha^* \wedge p^{[T]} = \overline{p}^{[T]}\right] \\ &\leq \max(K, \sqrt{T}) + \sum_{t=\max(K,\sqrt{T})+1}^{T} \Bigg(1 \cdot \mathbb{E}\left[\mathrm{NSW}(p^*, \mu^*) - \mathrm{NSW}(p^t, \mu^*) \;\middle|\; \mathcal{C}_\alpha^* \wedge p^{[T]} = \overline{p}^{[T]} \wedge \mathcal{B}\right] \\ &\qquad\qquad + \Pr\left[\neg\mathcal{B} \;\middle|\; \mathcal{C}_\alpha^* \wedge p^{[T]} = \overline{p}^{[T]}\right] \cdot 1\Bigg) \\ &\leq \max(K, \sqrt{T}) + \sum_{t=\max(K,\sqrt{T})+1}^{T} \mathbb{E}\left[2\alpha^t \sum_{j\in[K]} \overline{p}_j^t \cdot r_j^t + \frac{8}{t} \;\middle|\; \mathcal{C}_\alpha^* \wedge p^{[T]} = \overline{p}^{[T]} \wedge \mathcal{B}\right] \\ &\qquad + T \cdot \Pr\left[\neg\mathcal{B} \;\middle|\; \mathcal{C}_\alpha^* \wedge p^{[T]} = \overline{p}^{[T]}\right] \qquad (6) \\ &\leq \max(K, \sqrt{T}) + 1 + 2c\sqrt{2\log(NKT)} \sum_{t=1}^{T} \sum_{j\in[K]} \frac{\overline{p}_j^t}{\sqrt{c_j}}. \qquad (7)\end{aligned}$$

The final transition holds because $\alpha^t \leq c$ for all t, $r_j^t = \sqrt{\frac{2\log(NKT)}{n_j^t}}$, and conditioned on $\mathcal{B}$, $n_j^t \geq c_j$ for each $j \in [K]$ and $t \in [T]$, where $c_j = 1$ if $t < T_j$, and $c_j = \sqrt{T\log(NKT)}$ if $t \geq T_j$. Hence,

$$\begin{aligned}\mathbb{E}\left[R^T \;\middle|\; \mathcal{C}_\alpha^* \wedge p^{[T]} = \overline{p}^{[T]}\right] &\leq \max(K, \sqrt{T}) + 1 + 2c\sqrt{2\log(NKT)} \sum_{j\in[K]} \sum_{t=1}^{T} \frac{\overline{p}_j^t}{\sqrt{c_j}} \\ &= \max(K, \sqrt{T}) + 1 + 2c\sqrt{2\log(NKT)} \sum_{j\in[K]} \left(\sum_{t=1}^{T_j-1} \frac{\overline{p}_j^t}{1} + \sum_{t=T_j}^{T} \frac{\overline{p}_j^t}{\sqrt{T\log(NKT)}}\right) \\ &\leq \max(K, \sqrt{T}) + 1 + 2c\sqrt{2\log(NKT)} \sum_{j\in[K]} \left(q_j^{T_j} + \frac{T}{\sqrt{T\log(NKT)}}\right) \\ &= \mathcal{O}\left(cK\sqrt{T}\log(NKT)\right).\end{aligned}$$

Because this bound holds for every possible $\overline{p}^{[T]}$, we also have that $\mathbb{E}[R^T | \mathcal{C}_\alpha^*] = \mathcal{O}\left(cK\sqrt{T}\log(NKT)\right)$. Finally, we can see that

$$\mathbb{E}[R^T] = \Pr[\mathcal{C}_\alpha^*] \cdot \mathbb{E}[R^T | \mathcal{C}_\alpha^*] + \Pr[\neg\mathcal{C}_\alpha^*] \cdot \mathbb{E}[R^T | \neg\mathcal{C}_\alpha^*]$$

$$\leq 1 \cdot \mathcal{O}\left(cK\sqrt{T}\log(NKT)\right) + \frac{4}{\sqrt{T}} \cdot 1 = \mathcal{O}\left(cK\sqrt{T}\log(NKT)\right).$$

Recall that substituting $c = N$ and $c = \sqrt{12NK\log(NKT)}$ yields the two regret bounds. □

We emphasize that our analysis of the multi-agent UCB differs significantly from the analysis of the classical (single-agent) UCB. For example, the use of clean event $\mathcal{C}_\alpha^*$ is unique to our analysis. More importantly, the expression in Equation (7) is also unique to our setting in which the algorithm can "pull" a probability distribution over the arms. The corresponding expression in case of the classical UCB turns out to be much simpler and straightforward to bound. In contrast, we need to use additional tricks to derive the bound of $\mathcal{O}\left(cK\sqrt{T}\log(NKT)\right)$.

Finally, in the proof presented above, we showed that, assuming the clean event $\mathcal{C}_\alpha^*$, the expected regret is small conditioned on *any* sequence of policies that the UCB algorithm might use. At the first glance, this may seem surprising. However, a keen reader can observe that the clean event $\mathcal{C}_\alpha^*$ can only occur when the UCB algorithm uses a "good" sequence of policies that leads to low expected regret. A similar phenomenon is observed in the analysis of the classical (single-agent) UCB algorithm as well (see, e.g., [30]): assuming a different clean event, the classical UCB algorithm is guaranteed to not pull suboptimal arms too many times.

6 Lower Bound

We lastly turn our focus to proving lower bounds on the expected regret of any algorithm for our multi-agent multi-armed bandit (MA-MAB) problem. In the classical multi-armed bandit problem, it is known that no algorithm can achieve a regret bound of $E[R^T] = o(\sqrt{KT})$, when the constant inside the little-Oh notation is required to be independent of the distributions in the given instance [3]. For further discussion on bounds where the constant is allowed to depend on the distributions in the given instance, we refer the reader to Section 7. Our goal in this section is to reproduce this lower bound for our multi-agent multi-armed bandit problem. This would establish that the $\sqrt{T}$-dependence of the expected regret of our UCB variant on the horizon T from Theorem 3 is optimal. Note that our focus is solely on the dependence of the expected regret on T as T is typically much larger than both the number of agents N and the number of arms K. We leave it to future work to optimize the dependence on N and K.

First, we notice that any lower bound derived for the case of a single agent also holds when there are $N > 1$ agents. This is because one can consider instances in which all but one of the agents derive a fixed reward of 1 from every arm. Note that the contribution of such agents to the product in the Nash social welfare expression is always 1 regardless of the policy chosen. Hence, the Nash social welfare reduces to simply the expected utility of the remaining agent, i.e., the Nash social welfare in an instance with only this one agent. Therefore, any lower bound on the expected regret that holds for MA-MAB with a single agent also holds for MA-MAB with $N > 1$ agents.

Next, let us focus on the MA-MAB problem with $N = 1$ agent. At the first glance, this may look almost identical to the classical multi-armed bandit problem. After all, if there is but one agent, the policy maximizing the Nash social welfare places probability 1 on the arm j^* that gives the highest mean reward to the agent. Thus, like in the classical problem, our goal would be to converge to pulling arm j^* repeatedly and our regret would also be measured with respect to the best policy which deterministically pulls arm j^* in every round. However, there are two subtle differences which prevent us from directly borrowing the classical lower bound.

1. In our MA-MAB problem, an algorithm is allowed to "pull" a *distribution over the arms* p^t in round t and learn the stochastically generated rewards for a *random* arm j^t sampled from

this distribution. This makes the algorithm slightly more powerful than an algorithm in the classical MAB problem which must deterministically choose an arm to pull.

2. In our MA-MAB problem, the regret in round t is computed as the difference between the mean reward of the best arm and the *expected* mean reward of an arm j^t sampled according to the distribution p^t used by the algorithm. In the classical problem, one would replace the latter term with the mean reward of the arm actually pulled in round t.

The latter distinction is not particularly troublesome because our focus is on the *expected* regret of an algorithm anyway. However, the first distinction makes it impossible to directly borrow lower bounds from the classical MAB problem.

One might wonder if there is still a way to reduce the MA-MAB problem with $N = 1$ agent to the classical MAB problem. For example, given an algorithm A for MA-MAB with $N = 1$, what if we construct an algorithm $\widehat{A}$ for the classical MAB and use the lower bound on the expected regret of $\widehat{A}$ to derive a lower bound on the expected regret of A? The problem with such reduction is that once A chooses a distribution p^t, we have no control over which arm will be sampled. This choice is crucial as it will determine what information the algorithm gets to learn. We cannot mimic this learning process in our deterministic algorithm $\widehat{A}$. Upon careful consideration, it also seems difficult to express the expected regret of A as the convex combination of the expected regret of several deterministic algorithms for the classical MAB.

Instead of aiming to find a black-box reduction to the classical problem, we therefore investigate in detail the proof of the $\Omega\left(\sqrt{KT}\right)$ lower bound for the classical MAB due to Auer et al. [3, Theorem 5.1] and observe that their argument goes through for our MA-MAB problem as well. Instead of repeating their proof, we survey the key steps of their proof in which they assume the algorithm to be deterministically pulling an arm and highlight why the argument holds even when this is not the case.

- In the proof of their Lemma A.1, in the explanation of their Equation (30), they cite the assumption that given the rewards observed in the first $t-1$ rounds (they denote this by the vector $\mathbf{r}^{t-1}$), the algorithm pulls a fixed arm i_t in round t. They refer to the distribution $\mathbf{P}_i\{r^t|\mathbf{r}^{t-1}\}$ of the reward in round t given $\mathbf{r}^{t-1}$. In their case, the randomness in r^t is solely due to stochasticity of the rewards since the arm pulled (i_t) is fixed. However, in our case, one can think of $\mathbf{P}_i\{r^t|\mathbf{r}^{t-1}\}$ as containing randomness both due to the random choice of i_t and due to the stochasticity of the rewards, and their equations still go through.

- In the same equation, they consider $\mathbf{P}_{unif}\{i_t = i\}$, the probability that arm i is pulled in round t. In their case, the only randomness is due to $\mathbf{r}^{t-1}$. In our case, there is additional randomness due to the sampling of an arm in round t from a distribution p^t. However, this does not affect their calculations.

- Finally, in the proof of their Theorem A.2, they again consider the probability $\mathbf{P}_i\{i_t = i\}$ and the same argument as above ensures that their proof continues to hold in our setting.

Thus, we have the following lower bound.

Proposition 2. *For any algorithm for the MA-MAB problem, there exists a problem instance such that* $\mathbb{E}[R^T] = \Omega\left(\sqrt{KT}\right)$.

7 Discussion

We introduce a multi-agent variant of the multi-armed bandit problem in which different agents have different preferences over the arms and we seek to learn a tradeoff between the arms that is fair with respect to the agents, where the Nash social welfare is used as the fairness notion. Our work leaves several open questions and directions for future work.

Computation. As we observed in the paper, our Explore-First and Epsilon-Greedy variants can be implemented in polynomial time. However, for our UCB variant, it is not clear if the step of computing $\arg\max_{p \in \Delta^K} \mathrm{NSW}(p, \widehat{\mu}) + \alpha^t \sum_{j \in [K]} p_j r_j^t$ can be computed in polynomial time due to the added linear term at the end. As we mentioned in Section 5, there exists a PTAS for this step when the number of agents N is constant, but the complexity in the general case remains open. One might wonder if it helps to take the logarithm of the Nash social welfare, i.e., solve $\arg\max_{p \in \Delta^K} \log \mathrm{NSW}(p, \widehat{\mu}) + \alpha^t \sum_{j \in [K]} p_j r_j^t$. Indeed, since $\log \mathrm{NSW}$ is a concave function, this can be solved efficiently. However, our key lemmas use bounds on the NSW function that do not hold for $\log \mathrm{NSW}$ function. Further, such an approach would yield a regret bound where the regret is in terms of $\log \mathrm{NSW}$, which cannot be easily converted into regret in terms of NSW.

Logarithmic regret bound for UCB. In the classical stochastic multi-armed bandit setting, UCB has two known regret bounds with optimal dependence on T. There is an *instance-independent* bound that grows roughly as $\sqrt{T}$ (where the constants depend only on K and not on the unknown mean rewards in the given instance) and an *instance-dependent* bound that grows roughly as $\log T$ (where the constants may depend on the unknown mean rewards in the given instance in addition to K). While we recover the former bound in our multi-agent setting, we were not able to derive an instance-dependent logarithmic regret bound. This remains a major challenging open problem.

Improved lower bounds. In Section 6, we observe that the instance-independent $\Omega(\sqrt{KT})$ lower bound from the classical setting also holds in our multi-agent setting. Given that our upper bounds increase with N, it would be interesting to see if we can derive lower bounds that also increase with N. Deriving instance-dependent lower bounds in our setting would also be interesting.

Fairness. While maximizing the Nash social welfare is often seen as a fairness guarantee of its own, as discussed in the introduction, the policy with the highest Nash social welfare is also known to satisfy other fairness guarantees. However, it is not clear if the additive regret bounds we derive in terms of the Nash social welfare also translate to bounds on the amount by which these other fairness guarantees are violated. Considering other fairness guarantees and bounding their total violation is also an interesting direction for the future.

Multi-agent extensions. More broadly, our work opens up the possibility of designing multi-agent extensions of other multi-armed bandit problems. For example, one can consider a multi-agent dueling bandit problem, in which an algorithm asks an agent (or all agents) to compare two arms rather than report their reward for a single arm. Meaningfully defining the regret for such frameworks and designing algorithms that bound it is an exciting future direction.

References

[1] G. Amanatidis, G. Birmpas, A. Filos-Ratsikas, A. Hollender, and A. A. Voudouris. 2020. Maximum Nash Welfare and Other Stories About EFX. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*. 24–30.

[2] N. Anari, T. Mai, S. O. Gharan, and V. V. Vazirani. 2018. Nash Social Welfare for Indivisible Items Under Separable, Piecewise-Linear Concave Utilities. In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 2274–2290.

[3] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. 2002. The nonstochastic multi-armed bandit problem. *SIAM J. Comput.* 32, 1 (2002), 48–77.

[4] H. Aziz, A. Bogomolnaia, and H. Moulin. 2019. Fair Mixing: The Case of Dichotomous Preferences. In *Proceedings of the 20th ACM Conference on Economics and Computation (EC)*. 753–781.

[5] E. Bargiacchi, T. Verstraeten, D. Roijers, A. Nowé, and H. Hasselt. 2018. Learning to coordinate with coordination graphs in repeated single-stage multi-agent decision problems. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*. 482–490.

[6] F. Brandl, F. Brandt, D. Peters, C. Stricker, and W. Suksompong. 2020. Funding Public Projects: A Case for the Nash Product Rule. Manuscript.

[7] F. Brandt, V. Conitzer, U. Endress, J. Lang, and A. D. Procaccia (Eds.). 2016. *Handbook of Computational Social Choice*. Cambridge University Press.

[8] I. Caragiannis, D. Kurokawa, H. Moulin, A. D. Procaccia, N. Shah, and J. Wang. 2019. The Unreasonable Fairness of Maximum Nash Welfare. *ACM Transactions on Economics and Computation (TEAC)* 7, 3 (2019), 1–32.

[9] M. Chakraborty, K. Y. P. Chua, S. Das, and B. Juba. 2017. Coordinated Versus Decentralized Exploration In Multi-Agent Multi-Armed Bandits.. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. 164–170.

[10] R. Cole, N. Devanur, V. Gkatzelis, K. Jain, T. Mai, V. V. Vazirani, and S. Yazdanbod. 2017. Convex Program Duality, Fisher Markets, and Nash Social Welfare. In *Proceedings of the 18th ACM Conference on Economics and Computation (EC)*. 459–460.

[11] R. Cole and V. Gkatzelis. 2018. Approximating the Nash Social Welfare with Indivisible Items. *SIAM J. Comput.* 47, 3 (2018), 1211–1236.

[12] V. Conitzer, R. Freeman, and N. Shah. 2017. Fair Public Decision Making. In *Proceedings of the 18th ACM Conference on Economics and Computation (EC)*. 629–646.

[13] E. de Klerk, M. Laurent, and P. A. Parrilo. 2006. A PTAS for the Minimization of Polynomials of Fixed Degree Over the Simplex. *Theoretical Computer Science* 361, 2-3 (2006), 210–225.

[14] E. de Klerk, M. Laurent, and Z. Sun. 2015. An Alternative Proof of a PTAS for Fixed-Degree Polynomial Optimization Over the Simplex. *Mathematical Programming* 151, 2 (2015), 433–457.

[15] E. Eisenberg and D. Gale. 1959. Consensus of subjective probabilities: The pari-mutuel method. *The Annals of Mathematical Statistics* 30, 1 (1959), 165–168.

[16] B. Fain, A. Goel, and K. Munagala. 2016. The Core of the Participatory Budgeting Problem. In *Proceedings of the 12th Conference on Web and Internet Economics (WINE)*. 384–399.

[17] B. Fain, K. Munagala, and N. Shah. 2018. Fair Allocation of Indivisible Public Goods. In *Proceedings of the 19th ACM Conference on Economics and Computation (EC)*. 575–592.

[18] R. Freeman, S. M. Zahedi, and V. Conitzer. 2017. Fair and Efficient Social Choice in Dynamic Settings. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. 4580–4587.

[19] J. Garg, M. Hoefer, and K. Mehlhorn. 2018. Approximating the Nash Social Welfare with Budget-Additive Valuations. In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 2326–2340.

[20] J. Garg and P. McGlaughlin. 2019. Improving Nash Social Welfare Approximations. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*. 294–300.

[21] S. Gillen, C. Jung, M. Kearns, and A. Roth. 2018. Online learning with an unknown fairness metric. In *Proceedings of the 32nd Annual Conference on Neural Information Processing Systems (NIPS)*. 2600–2609.

[22] M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth. 2016. Fairness in learning: Classic and contextual bandits. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NIPS)*. 325–333.

[23] R. Kleinberg, A. Niculescu-Mizil, and Y. Sharma. 2010. Regret bounds for sleeping experts and bandits. *Machine learning* 80, 2-3 (2010), 245–272.

[24] R. Kleinberg, A. Slivkins, and E. Upfal. 2008. Multi-Armed Bandits in Metric Spaces. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC)*. 681–690.

[25] T. L. Lai and H. Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* 6, 1 (1985), 4–22.

[26] E. Lee. 2017. APX-Hardness of Maximizing Nash Social Welfare with Indivisible Items. *Inform. Process. Lett.* 122 (2017), 17–20.

[27] Y. Liu, G. Radanovic, C. Dimitrakakis, D. Mandal, and D. C. Parkes. 2017. Calibrated fairness in bandits. arXiv preprint arXiv:1707.01875.

[28] H. Moulin. 2003. *Fair Division and Collective Welfare*. MIT Press.

[29] V. Patil, G. Ghalme, V. Nair, and Y. Narahari. 2020. Achieving Fairness in Stochastic Multi-Armed Bandits. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*. 5379–5386.

[30] A. Slivkins. 2019. Introduction to Multi-Armed Bandits. *Foundations and Trends in Machine Learning* 12, 1-2 (2019), 1–286.

[31] W. R. Thompson. 1933. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika* 25, 3/4 (1933), 285–294.

[32] M. J. Wainwright. 2019. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press.

[33] M. Woodroofe. 1979. A One-Armed Bandit Problem with a Concomitant Variable. *J. Amer. Statist. Assoc.* 74, 368 (1979), 799–806.

[34] Y. Yue, J. Broder, R. Kleinberg, and T. Joachims. 2012. The K-Armed Dueling Bandits Problem. *J. Comput. System Sci.* 78, 5 (2012), 1538–1556.