An Efficient Algorithm for Fair Multi-Agent Multi-Armed Bandit with Low Regret

Matthew Jones, Huy Nguyen, Thy Nguyen

Northeastern University
{jones.m, hu.nguyen, nguyen.thy2}@northeastern.edu

Abstract

Recently a multi-agent variant of the classical multi-armed bandit was proposed to tackle fairness issues in online learning. Inspired by a long line of work in social choice and economics, the goal is to optimize the Nash social welfare instead of the total utility. Unfortunately previous algorithms either are not efficient or achieve sub-optimal regret in terms of the number of rounds T. We propose a new efficient algorithm with lower regret than even previous inefficient ones. For N agents, K arms, and T rounds, our approach has a regret bound of $\tilde{O}(\sqrt{NKT} + NK)$. This is an improvement to the previous approach, which has regret bound of $\tilde{O}(\min(NK, \sqrt{N}K^{3/2})\sqrt{T})$. We also complement our efficient algorithm with an inefficient approach with $\tilde{O}(\sqrt{KT} + N^3K)$ regret. The experimental findings confirm the effectiveness of our efficient algorithm compared to the previous approaches.

1 Introduction

The multi-armed bandit (MAB) problem poses an identical set of choices (or actions, *arms*) at each time step to the decision-maker. When the decision maker pulls makes a choice, she receives a reward drawn from a distribution associated with that choice. The goal is to select arms to maximize the total reward in expectation, or to minimize their expected *regret* with respect to the optimal strategy (or *policy*). The MAB problem lends itself to many applications including allocation of resources in a financial portfolio (Hoffman et al. 2011; Shen et al. 2015; Huo and Fu 2017), selection of both dosages (Bastani and Bayati 2020) and treatments (Durand et al. 2018) in clinical healthcare, and a broad range of recommender systems (Zhou et al. 2017; Bouneffouf, Bouzeghoub, and Gançarski 2012, 2013).

In many scenarios, making a decision impacts not only one but multiple agents. For instance, consider a policy-maker making decisions that might influence various groups of their constituents (Zhu and Walker 2018), or a recommendation engine that picks hyperparameters for their system to serve many users. The multi-agent multi-armed bandit problem (MA-MAB), proposed in Hossain, Micha, and Shah (2021), is a variant of the MAB problem in which there are N agents and K arms, and in every round t, the algorithm selects the same arm a for every agent and receive N rewards realizations drawn from unknown distributions $\{D_{j,a}\}_{j=1}^{N}$ with unknown means $\{\mu^{\star}_{j,a}\}_{j=1}^{N}$. The reward distributions of the same arm for different agents need not be the same, and hence an optimal arm for one agent might be sub-optimal for the others. As such, the goal of picking the optimal arm in the standard MAB problem is no longer meaningful, and a different objective is needed.

In this setting, one intuitive objective would be to optimize the expected rewards over all agents. This reduces the problem to the classic MAB problem and can be solved accordingly (Slivkins 2019). In practice, this objective has a key weakness: individual rewards may be neglected in order to improve the reward of the collective. Consider an extremely divided case with two arms where just over half of agents receive reward 1 from the first arm and reward 0 from the second arm, and the other agents receive the opposite rewards. The optimal strategy in this case is to always choose the first arm: half of the agents receive reward 1, and almost half of the agents receive reward 0. While this formulation is well-studied by reduction, it unfortunately does not serve to enforce fairness among the agents. A fairer choice would be to pull each arm with roughly the same probability, which only slightly reduces the individual utilities. In other words, the agents would receive similar rewards with each other if the algorithm learns a distribution π over the arms and converges to pulling each arm almost uniformly. Thus, the optimal strategy for a MA-MAB problem corresponds not to picking the optimal arm as in the classic MAB problem, but a probability distribution over the arms.

In (Hossain, Micha, and Shah 2021), the authors proposed an objective motivated by studies on fairness in social choice: the *Nash social welfare* (NSW). This is a classical notion going back to (Nash 1950) and (Kaneko and Nakamura 1979). For n agents with expected utility $u_1, u_2, \ldots, u_n$, the Nash product $\prod_i u_i$ is sandwiched between $(\min_i u_i)^n \leq \prod_i u_i \leq \left(\frac{\sum_i u_i}{n}\right)^n$, serving as a compromise between an egalitarian approach (optimizing the minimum reward across agents) and a utilitarian approach (optimizing the sum of rewards across agents). In the context of the MA-MAB problem, the optimal fair strategy corresponds to the distribution that maximizes

the cummulative Nash product of the agents' expected rewards. Formally, $\pi^\star = \arg\max_{\pi\in\Delta^K} \mathrm{NSW}(\pi,\mu^\star) = \arg\max_{\pi\in\Delta^K} \prod_{j\in[N]} \left(\sum_{a\in[K]} \pi_a \mu^\star_{j,a}\right)$, where Δ^K is the K-simplex .

The prior work in (Hossain, Micha, and Shah 2021) proposed three algorithms for the MA-MAB problem (see table 1). Their Explore-First and ϵ-greedy algorithms have regret bounds that scale with $\Omega(T^{2/3})$. The algorithms involve computing the optimal policy given an estimated reward matrix, i.e., $\pi = \arg\max_{\pi\in\Delta^K} \mathrm{NSW}(\pi,\hat{\mu})$. Since the objective is log-concave, the optimization step can be solved in polynomial time and there thus exist efficient implementations for the algorithms. Their UCB algorithm achieves the improved regret bound of $\widetilde{O}(K\sqrt{NT\cdot\min\{N,K\}})$. Note that the dependency on T for the UCB algorithm is tight due to a reduction of the MA-MAB problem with $N=1$ to the standard multi-armed bandits problem with a lower bound of $\Omega(\sqrt{TK})$. It is not clear to the authors of (Hossain, Micha, and Shah 2021) if the UCB algorithm "admits an efficient implementation." This is due to the algorithm's step of computing $\arg\max_{\pi\in\Delta^K} \mathrm{NSW}(\pi,\hat{\mu}) + \sum_{a\in[K]} \beta_a \pi_a$, where β_a is inversely proportional to the number of times the algorithm has picked arm a. With the added linear term, the optimization program is no longer log-concave. As such, previous algorithms for this problem in (Hossain, Micha, and Shah 2021) either fail to achieve optimal dependency on T, or do not admit an efficient implementation. This leads us to investigate the following question:

Is it possible to design an algorithm that admits an efficient implementation while achieving optimal dependency on T?

Our main contribution is an efficient algorithm with $\widetilde{O}(\sqrt{NKT}+NK)$ regret. We not only not give an affirmative answer to the question above, but also achieve an improved regret bound over the UCB algorithm of (Hossain, Micha, and Shah 2021) for most regimes of N,K,T. Our algorithm preserves the efficiency of the Explore-First and ϵ-Greedy approaches in (Hossain, Micha, and Shah 2021), while achieving an improved bound over the previous state-of-the-art. We also complement our efficient algorithm with an inefficient approach with $\widetilde{O}(\sqrt{KT}+N^2K)$ regret.

2 Other related works

Many variants of the multi-armed bandit problem have been proposed and studied such as adversarial bandit (Auer et al. 2002), dueling bandit (Yue and Joachims 2009; Yue et al. 2012), Lipschitz bandit (Kleinberg 2004; Flaxman, Kalai, and McMahan 2004), contextual bandit (Hazan and Megiddo 2007), and sleeping bandit (Kleinberg, Niculescu-Mizil, and Sharma 2010). Multi-agent variants of the problem have also been investigated in (Landgren, Srivastava, and Leonard 2016; Chakraborty et al. 2017; Bargiacchi et al. 2018).

In the context of fair multi-armed bandit, (Joseph et al. 2016) proposes a framework where an arm with a higher expected reward is selected with a probability no lower than that of an arm with a lower expected reward. (Wang and Joachims 2020) requires the fair policy to sample arms with probability proportional to the value of a merit function of its mean reward. In (Liu et al. 2017) preserving fairness means the probability of selecting each arm should be similar if the two arms have a similar quality distribution. (Gillen et al. 2018) studies fairness in the linear contextual bandits setting where there are individual fairness constraints imposed by an unknown similarity metric. (Patil et al. 2020) proposes a fair MAB variant that seeks to optimize the cumulative reward while also ensures that, at any round, each arm is pulled at least a specified fraction of times.

Our objective of finding a probability distribution over arms to optimize the Nash welfare objective can also be cast as the continuum-armed bandit problem where the Nash welfare function is the objective. (Kleinberg, Slivkins, and Upfal 2019) designs an algorithm with a regret bound of $\tilde{O}\left(T^{\frac{\gamma+1}{\gamma+2}}\right)$, where γ, defined as the zooming dimension, would be $\Theta(K)$ for the MA-MAB problem. The resulting bound would be no better than $O(T^{2/3})$ and approaches $O(T)$ as K increases. It is also important to note that there is a long line of work on bandit convex optimization (Hazan and Levy 2014; Bubeck and Eldan 2016; Bubeck, Lee, and Eldan 2017; Chen, Zhang, and Karbasi 2019). One can apply the approaches to optimize the log of the Nash welfare function. However, the regret bound of the new objective does not translate to that of the original objective.

3 Preliminaries

Define $[n]=\{1,\dots,n\}$ for $n\in\mathbb{N}$. In the multi-agent multi-armed bandit problem, we have a set of agents $[N]$ and a set of arms $[K]$ for $N,K\in\mathbb{N}$. For each agent $j\in[N]$ and arm $a\in[K]$, we have a reward distribution $D_{j,a}$ with mean $\mu^\star_{j,a}$ and support in $[0,1]$. Define $\mu^\star=(\mu^\star_{j,a})_{j\in[N],a\in[K]}$ as the mean reward matrix. At each round t, a_t denotes the selected arm of round t and $r_{j,a_t,t}\sim D_{j,a_t}$ the realization of the reward distribution associated with arm a_t of each agent j.

For reward matrix $\mu=\mu_{j\in[N],a\in[K]}\in[0,1]^{N\times K}$ and policy $\pi\in\Delta^K$, we denote the Nash Social Welfare function by the product over the expected reward of each agent, $\mathrm{NSW}(\pi,\mu)=\prod_{j\in[N]}\left(\sum_{a\in[K]}\pi_a\mu_{j,a}\right)$. For a time horizon T, our goal is to choose policies $\pi_t\in\Delta^K$ at each round $t\in[T]$ to minimize the cumulative regret $R_T=\sum_{t\in[T]}\mathrm{NSW}(\pi^\star,\mu^\star)-\sum_{t\in[T]}\mathrm{NSW}(\pi_t,\mu^\star)$. We use $\pi_{a,t}$ to denote the probability of selecting arm a under policy π_t.

At each round t, our algorithm samples an arm a_t from π_t and pulls the same arm for every agent. Let $N_{a,t}=\sum_{\tau=1}^{t-1}\mathbf{1}\{a_\tau=a\}$ denote the number of times arm a has been sampled before round t, and $\hat{\mu}_{j,a,t}=\frac{1}{N_{a,t}}\sum_{\tau=1}^{t-1} r_{j,a_\tau,\tau}\mathbf{1}\{a_\tau=a\}$ the corresponding empirical mean of agent j's reward on arm a. Our main algorithm maintains a confidence bound for the mean reward matrix $\mu^\star$ at every round. Let $\hat{\mu}_t=(\hat{\mu}_{j,a,t})_{j\in[N],a\in[K]}$ denote the estimated mean reward matrix and $w_t=(w_{j,a,t})_{j\in[N],a\in[K]}$ the confidence bound matrix for round t.

Table 1: Fair algorithms for the MA-MAB problem.

Algorithm	Regret Bound	Efficient	Reference
Explore-First	$\widetilde{O}\left(\sqrt[3]{NKT^2 \cdot \min\{N,K\}}\right)$	✓	(Hossain, Micha, and Shah 2021), Theorem 1
ϵ-Greedy	$\widetilde{O}\left(\sqrt[3]{NKT^2 \cdot \min\{N,K\}}\right)$	✓	(Hossain, Micha, and Shah 2021), Theorem 2
UCB	$\widetilde{O}\left(K\sqrt{NT \cdot \min\{N,K\}}\right)$	✗	(Hossain, Micha, and Shah 2021), Theorem 3
Algorithm 1	$\widetilde{O}\left(NK + \sqrt{NKT}\right)$	✓	Theorem 4.5
Algorithm 2	$\widetilde{O}\left(N^2K + \sqrt{KT}\right)$	✗	Theorem 4.6

Algorithm 1: Fair multi-agent UCB algorithm

input: K, N, T, δ
for $t = 1$ to T **do**
  if $t \leq K$ **then**
    $\pi_t \leftarrow$ policy that puts probability 1 on arm t
  else
    $\forall j, a, \hat{\mu}_{j,a,t} = \frac{1}{N_{a,t}} \sum_{\tau=1}^{t-1} r_{j,a_\tau,\tau} \mathbf{1}\{a_\tau = a\}$
    $\forall j, a, U_{j,a,t} = \min(\hat{\mu}_{j,a,t} + w_{j,a,t}, 1)$
    $\pi_t \leftarrow \operatorname{argmax}_{\pi \in \Delta^K} \mathrm{NSW}(\pi, U_t)$
  end if
  Sample a_t from π_t
  Observe rewards $\{r_{j,a_t,t}\}_{j \in N}$
  $N_{a_t,t+1} \leftarrow N_{a_t,t} + 1$
end for

4 Algorithms

Our UCB algorithm, Algorithm 1, first selects each arm once in order to obtain an initial estimate $\hat{\mu}$ for μ. Then, it computes the upper confidence bound estimate $U_t = \{U_{j,a,t}\}_{j \in [N], a \in [K]}$ of the true mean for each arms a of N agents and finds a policy π_t that optimizes the Nash social welfare function given U_t. Due to the log-concavity of Nash social welfare, we use standard convex optimization tools to optimize $\log(\mathrm{NSW}(\pi, U_t))$ and simultaneously optimize $\mathrm{NSW}(\pi, U_t)$.

This approach differs from the UCB algorithm in Hossain, Micha, and Shah (2021) in two aspects. First, the optimization step in their UCB algorithm uses an additive regularization term in the objective rather than on the estimate $\hat{\mu}$, and is therefore not log-concave. Second, our confidence interval is defined in terms of the empirical mean and as a result our confidence interval is about a factor of $\sqrt{1 - \hat{\mu}_{j,a,t}}$ tighter than that of Hossain, Micha, and Shah (2021). As our proof involves bounding the regret by the Lipschitz continuity of the Nash social welfare, our confidence interval allows for a careful analysis of the algorithm.

Note that the UCB algorithm in Hossain, Micha, and Shah (2021) is horizon-independent. Although Algorithm 1 requires the value of the time horizon T as input, it can be easily modified to be horizon-independent. One approach is to modify the confidence interval $w_{j,a,t}$ so it becomes a function of the current time step, t , rather than the time horizon, T. Specifically, both $\ln(4NKT/\delta)$ terms of $w_{j,a,t}$ in Algorithm 1 would become $\ln(8NKt^2/\delta)$. Lemma 4.2 and Theorem 4.5 can be easily adapted to the new confidence interval. The horizon-independent variant of our algorithm would have the same regret bound of $\widetilde{O}\left(NK + \sqrt{NKT}\right)$.

Overall approach: Our objective is to bound the regret $\sum_{t \in [T]} \mathrm{NSW}(\pi^\star, \mu^\star) - \mathrm{NSW}(\pi_t, \mu^\star)$. Observe that $\mathrm{NSW}(\cdot, \cdot)$ is monotone in the second argument so $\mathrm{NSW}(\pi^\star, U_t) \geq \mathrm{NSW}(\pi^\star, \mu^\star)$. By the optimality of π_t, we also have $\mathrm{NSW}(\pi_t, U_t) \geq \mathrm{NSW}(\pi^\star, U_t)$. Thus, we can reduce the problem to bounding $\sum_{t \in [T]} \mathrm{NSW}(\pi_t, U_t) - \mathrm{NSW}(\pi_t, \mu^\star)$.

The key idea to bound the regret in a single round t is to look at the expected reward of each agent j if the mean reward was U_t. Formally, let $g_{j,t} = \sum_{a \in [K]} \pi_{a,t}(1 - U_{j,a,t}) = 1 - \sum_{a \in [K]} \pi_{a,t} U_{j,a,t}$. If there are a lot of agents with large $g_{j,t}$, then the Nash product $\mathrm{NSW}(\pi_t, U_t) = \prod_j (1 - g_{j,t})$ is small and the regret is therefore small. More precisely, we consider two cases depending on whether there exists a $p \geq 0$ such that the set of agents $\{j \in [N] : g_{j,t} \geq 2^{-p}\}$ is of size at least $3 \cdot 2^p \ln(T)$. If such a p exists, then

$$\mathrm{NSW}(\pi_t, U_t) \leq (1 - 2^{-p})^{3 \cdot 2^p \ln T} \leq \frac{1}{T^3}$$

As a result, the regret of round t is negligible.

We now need to bound the regret of rounds where no such p exists. The key idea is to show that when no p exists, the upper confidence bounds are on average very close to the true means. For intuition, suppose the following similar statement holds. Let $g'_{j,t} = \sum_{a \in [K]} \pi_{a,t}(1 - \hat{\mu}_{j,a,t})$ and for all $p \geq 0$, the set of agents $\{j \in [N] : g'_{j,t} \geq 2^{-p}\}$ is of size at most $3 \cdot 2^p \ln(T)$. Given this condition we can bound

$$\begin{aligned} \sum_{j \in [N]} g'_{j,t} &\leq \int_0^1 [\text{number of agents } j \text{ s.t. } g'_{j,t} \geq x] dx \\ &\leq 1 + 6 \ln T \log N \end{aligned}$$

Notice that our estimation error $w_{j,a,t}$ is a function of $1 - \hat{\mu}_{j,a,t}$ and we showed that $1 - \hat{\mu}_{j,a,t}$ is small on average. Thus, by making a careful averaging argument, we can show that the upper bound U_t is close to $\mu^\star$ on average. The regret bound then follows from the smoothness of the function

$\mathrm{NSW}(\pi_t, \cdot)$. The actual proof has to overcome additional technical challenges due to the difference between the desired $g'_{j,t}$ and the actual $g_{,t}$.

Analysis

We seek to bound $\sum_{t\in[T]} \mathrm{NSW}(\pi_t, U_t) - \mathrm{NSW}(\pi_t, \mu^\star)$ using the smoothness of the NSW objective. We include the missing proof of the Lipschitz-continuity property and the other lemmas in the appendix.

Lemma 4.1. *(Lemma 3, (Hossain, Micha, and Shah 2021)) Given a policy $\pi \in \Delta^k$ and reward matrices $\mu^1, \mu^2 \in [0,1]^{N\times K}$, we have*

$$\left|\mathrm{NSW}(\pi, \mu^1) - \mathrm{NSW}(\pi, \mu^2)\right| \le \sum_{j\in[N]} \sum_{a\in[K]} \pi_a \left|\mu^1_{j,a} - \mu^2_{j,a}\right|$$

Lemma 4.1 implies that a Lipschitz-continuity analysis for bounding $\sum_{t\in[T]} \mathrm{NSW}(\pi_t, U_t) - \mathrm{NSW}(\pi_t, \mu^\star)$ would benefit from a tight confidence bound on the means of the rewards. The following lemma proves a confidence interval that is about factor of $\sqrt{1-\hat{\mu}_{j,a,t}}$ tighter than that of Hossain, Micha, and Shah (2021).

Lemma 4.2. *For any $\delta \in (0,1)$, with probability at least $1-\delta/2$, $\forall\ t > K, a \in [K]$, $j \in [N]$, $\left|\mu^\star_{j,a} - \hat{\mu}_{j,a,t}\right| \le \sqrt{\frac{12(1-\hat{\mu}_{j,a,t})\ln(4NKT/\delta)}{N_{a,t}}} + \frac{12\ln(4NKT/\delta)}{N_{a,t}} = w_{j,a,t}$.*

Our confidence bound $w_{j,a,t}$ in Lemma 4.2 has a $\widetilde{O}\left(\sqrt{\frac{1-\hat{\mu}_{j,a,t}}{N_{a,t}}}\right)$ term and a $\widetilde{O}\left(\frac{1}{N_{a,t}}\right)$ term. Using Young's inequality, we can bound both the first term by $\widetilde{O}\left(\left(1-\hat{\mu}_{j,a,t}\right) + \frac{1}{N_{a,t}}\right)$, and thus we have

$$w_{j,a,t} \in \widetilde{O}\left(\left(1-\hat{\mu}_{j,a,t}\right) + \frac{1}{N_{a,t}}\right). \tag{1}$$

As mentioned above, if there are a lot of agents with large $g_{j,t}$ at round t, then the regret will be negligible. Thus, our main goal is to analyze the regret for rounds t's when there are not enough such agents. The following lemma formalizes the intuition and bounds the sum of expected empirical reward over all agents by the confidence intervals w_t. In other words, it bounds the $\left(1-\hat{\mu}_{j,a,t}\right)$ term of Equation 1 over all actions a and agents j.

Lemma 4.3. *Define $g_{j,t} = \sum_{a\in[K]} \pi_{a,t}\left(1 - U_{j,a,t}\right)$, and $S(t,p) = \left\{j \text{ for } j \in [N] : g_{j,t} \ge 2^{-p}\right\}$. If $|S(t,p)| < 2^p \cdot 3\ln T$ for all $p \ge 0$, then*

$$\sum_{j\in[N]} \sum_{a\in[K]} \pi_{a,t}(1-\hat{\mu}_{j,a,t}) \le 1 + 6\ln T \log N + \sum_{j\in[N]} \sum_{a\in[K]} \pi_{a,t} w_{j,a,t}.$$

Lemma 4.4 bounds the error incurred by the $\frac{1}{N_{a,t}}$ term of Equation 1. We carefully analyze the martingale sequence to obtain a tighter bound than that of black-box approaches, i.e. in Lemma 4.4 we bound the martingale sequence to be of $O(K\ln T/K)$ while Azuma-Hoeffding inequality would give us $O(\sqrt{T})$.

Lemma 4.4. *With probability $1-\delta/2$,*

$$\sum_{t\in[T]} \sum_{a\in[K]} \pi_{a,t}/N_{a,t} \le 2K\left(\ln\frac{T}{K} + 1\right) + \ln(2/\delta).$$

With Lemma 4.3 and Lemma 4.4, we are ready to bound $\sum_{t\in[T]} \mathrm{NSW}(\pi_t, U_t) - \mathrm{NSW}(\pi_t, \mu^\star)$ by a Lipschitz-continuity analysis of the Nash social welfare function.

Theorem 4.5. *Suppose $\forall j, a, t$, $r_{j,t,a} \in [0,1]$ and $w_{j,a,t} = \sqrt{\frac{12(1-\hat{\mu}_{j,a,t})\ln(4NKT/\delta)}{N_{a,t}}} + \frac{12\ln(4NKT/\delta)}{N_{a,t}}$, for any $\delta \in (0,1)$, the regret of the Fair multi-agent UCB algorithm (Algorithm 1) is $R_T = \widetilde{O}\left(\sqrt{NKT} + NK\right)$ with probability at least $1-\delta$.*

Proof. Let $I' \subseteq [T]$ denote the set of all rounds t where there exists $p \ge 0$ such that $|S(t,p)| \ge 2^p 3\ln T$. Let $I = [K] \cup I'$. We have

$$\begin{aligned}
&\sum_{t\in I} \mathrm{NSW}(\pi_t, U_t) - \mathrm{NSW}(\pi_t, \mu^\star) \\
&\le \sum_{t\in I'} \mathrm{NSW}(\pi_t, U_t) + K \\
&\le \sum_{t\in I'} \prod_{j\in S(t,p)} (1 - g_{j,t}) + K \\
&\le T\left(1 - 2^{-p}\right)^{2^p \cdot 3\ln T} + K \\
&\le \frac{1}{T^2} + K.
\end{aligned} \tag{2}$$

The last inequality is due to the fact that $(1-\frac{1}{x})^x \le \frac{1}{e}\ \forall x \ge 1$. We bound the regret of rounds not in I. For any $\delta \in (0,1)$, the events in Lemma 4.2 and Lemma 4.4 hold with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{t\notin I} \mathrm{NSW}(\pi_t, U_t) - \mathrm{NSW}(\pi_t, \mu^\star) \\
&\le \sum_{t\notin I} \sum_{j\in[N]} \sum_{a\in[K]} \pi_{a,t} \left|U_{j,a,t} - \mu^\star_{j,a,t}\right| \\
&= \sum_{t\notin I} \sum_{j\in[N]} \sum_{a\in[K]} \pi_{a,t} \left|U_{j,a,t} - \hat{\mu}_{j,a,t} + \hat{\mu}_{j,a,t} - \mu^\star_{j,a,t}\right| \\
&\le 2\sum_{t\notin I} \sum_{j\in[N]} \sum_{a\in[K]} \pi_{a,t} \sqrt{\frac{12(1-\hat{\mu}_{j,a,t})\ln(4NKT/\delta)}{N_{a,t}}} \\
&\quad + \sum_{t\notin I} \sum_{j\in[N]} \sum_{a\in[K]} \pi_{a,t} \frac{24\ln(4NKT/\delta)}{N_{a,t}}.
\end{aligned} \tag{3}$$

The first inequality follows from Lemma 4.1. The last inequality follows from Lemma 4.2 and the definition of $U_{j,a,t}$.

By Lemma 4.4, we can bound the error incurred by the linear term of the confidence interval in Equation 3,

$$\begin{aligned}
&\sum_{t\notin I} \sum_{j\in[N]} \sum_{a\in[K]} \pi_{a,t} \frac{24\ln(4NKT/\delta)}{N_{a,t}} \\
&\le \sum_{t\in[T]} \sum_{j\in[N]} \sum_{a\in[K]} \pi_{a,t} \frac{24\ln(4NKT/\delta)}{N_{a,t}}
\end{aligned}$$

$$= \sum_{j\in[N]} \sum_{t\in[T]} \sum_{a\in[K]} \pi_{a,t} \frac{24\ln(4NKT/\delta)}{N_{a,t}}$$
$$\leq 24\ln(4NKT/\delta)N\left(2K\left(\ln\frac{T}{K}+1\right)+\ln(2/\delta)\right). \tag{4}$$

We are done after bounding the remaining term of Equation 3. For brevity, we bound it without the log term,

$$\sum_{t\notin I} \sum_{j\in[N]} \sum_{a\in[K]} \pi_{a,t}\sqrt{\frac{1-\hat{\mu}_{j,a,t}}{N_{a,t}}}$$
$$\leq \sum_{t\notin I} \sum_{j\in[N]} \sum_{a\in[K]} \pi_{a,t}\left(\frac{q(1-\hat{\mu}_{j,a,t})}{2}+\frac{1}{2q\cdot N_{a,t}}\right). \tag{5}$$

The inequality follows from Young's inequality for $q \geq 0$. Applying Lemma 4.3 to the first term and Lemma 4.4 to the second term of Equation 5, we have:

$$\sum_{t\notin I} \sum_{j\in[N]} \sum_{a\in[K]} \pi_{a,t}\sqrt{\frac{1-\hat{\mu}_{j,a,t}}{N_{a,t}}}$$
$$\leq \frac{q}{2}\sum_{t\notin I}\left(1+6\ln T+\sum_{j\in[N]}\sum_{a\in[K]}\pi_{a,t}w_{j,a,t}\right)$$
$$+\frac{N}{2q}\left(2K\left(\ln\frac{T}{K}+1\right)+\ln(2/\delta)\right)$$
$$\leq \frac{q}{2}\sum_{t\notin I}\sum_{j\in[N]}\sum_{a\in[K]}\sqrt{\frac{12(1-\hat{\mu}_{j,a,t})\ln(4NKT/\delta)}{N_{a,t}}}$$
$$+\frac{q}{2}\sum_{t\notin I}\sum_{j\in[N]}\sum_{a\in[K]}\frac{12\ln(4NKT/\delta)}{N_{a,t}}+6q\cdot T\cdot\ln T$$
$$+\frac{N}{2q}\left(2K\left(\ln\frac{T}{K}+1\right)+\ln(2/\delta)\right),$$

where we use the fact that $6\ln T \geq 6\ln 2 \geq 1$. Suppose $q \in (0,1]$, applying Lemma 4.4 to the linear term of the confidence interval, we have

$$\sum_{t\notin I} \sum_{j\in[N]} \sum_{a\in[K]} \pi_{a,t}\sqrt{\frac{1-\hat{\mu}_{j,a,t}}{N_{a,t}}}$$
$$\leq \frac{q}{2}\sum_{t\notin I}\sum_{j\in[N]}\sum_{a\in[K]}\sqrt{\frac{12(1-\hat{\mu}_{j,a,t})\ln(4NKT/\delta)}{N_{a,t}}}$$
$$+\frac{N\ln(4NKT/\delta)}{q}\left(2K\left(\ln\frac{T}{K}+1\right)+\ln(2/\delta)\right)$$
$$+6q\cdot T\cdot\ln T.$$

Setting $q = \frac{\sqrt{KN}}{(\sqrt{KN}+\sqrt{T})\sqrt{12\ln(4NKT/\delta)}} \leq 1$ and re-arranging the terms, we have

$$\frac{1}{2}\sum_{t\notin I} \sum_{j\in[N]} \sum_{a\in[K]} \pi_{a,t}\sqrt{\frac{1-\hat{\mu}_{j,a,t}}{N_{a,t}}}$$
$$\leq (\ln(4NKT/\delta))^{3/2}\left(1+\frac{\sqrt{T}}{\sqrt{KN}}\right)2NK\left(\ln\frac{T}{K}+1\right)$$
$$+(\ln(4NKT/\delta))^{3/2}\left(1+\frac{\sqrt{T}}{\sqrt{KN}}\right)N\ln(2/\delta)$$
$$+\frac{6T\sqrt{KN}}{(\sqrt{KN}+\sqrt{T})}. \tag{6}$$

From Equations 2, 3, 4, and 6, we have

$$\sum_{t\in[T]} \mathrm{NSW}(\pi_t, U_t) - \mathrm{NSW}(\pi_t, \mu^\star)$$
$$= O\left((\sqrt{NKT}+NK)\cdot\mathrm{polylog}(NKT/\delta)\right).$$

By monotonicity of the Nash-social welfare function and the optimization step in the algorithm, we have $\mathrm{NSW}(\pi_t, U_t) \geq \mathrm{NSW}(\pi^\star, U_t) \geq \mathrm{NSW}(\pi^\star, \mu^\star)$. Thus,

$$\sum_{t\in[T]} \mathrm{NSW}(\pi^\star, \mu^\star) - \sum_{t\in[T]} \mathrm{NSW}(\pi_t, \mu^\star)$$
$$= O\left((\sqrt{NKT}+NK)\cdot\mathrm{polylog}(NKT/\delta)\right)$$

■

An Inefficient Algorithm with Improved Regret

Algorithm 2: Fair multi-agent UCB algorithm with high start-up cost

input: K, N, T, δ
for $t = 1$ to T **do**
  if $t \leq 180N^2K\ln(6NTK/\delta)\ln T$ **then**
    $\pi_t \leftarrow$ policy that puts probability 1 on arm $t \bmod (K+1)$
  else
    $\forall j, a, \hat{\mu}_{j,a,t} = \frac{1}{N_{a,t}}\sum_{\tau=1}^{t-1} r_{j,a_\tau,\tau}\mathbf{1}\{a_\tau = a\}$
    $S_\pi = \{\pi \in \Delta^K : \sum_{a\in[K],j\in[N]} \pi_a(1-\hat{\mu}_{j,a,t}) \leq 1+2\ln T\}$
    if $S_\pi \neq \emptyset$ **then**
      $\pi_t \leftarrow \mathrm{argmax}_{\pi\in S_\pi}\left(\mathrm{NSW}(\pi, \hat{\mu}_t) + \pi\cdot\eta_t\right)$
    else
      $\pi_t \leftarrow$ policy that puts probability 1 on a random arm $a \in [K]$
    end if
  end if
  Sample a_t from π_t
  Observe rewards $\{r_{j,a_t,t}\}_{j\in N}$
  $N_{a_t,t+1} \leftarrow N_{a_t,t} + 1$
end for

Algorithm 2 is able to obtain a tighter regret bound in terms of T by first pulling each arm $\widetilde{O}(N^2)$ times, then selecting π_t to optimize the Nash Social Welfare on $\hat{\mu}_t$ plus an additive term $\pi\cdot\eta_t$, where $\eta_t = \{\eta_{a,t}\}_{a\in[K]}$ is a vector. The additive term gives an upper bound on $|\mathrm{NSW}(\pi, \mu^*) -$

$\text{NSW}(\pi, \hat{\mu}_t)|$, and by the optimization step we obtain that for all $t > \tilde{O}(N^2K)$,

$$\begin{aligned}\text{NSW}(\pi^*, \mu^*) &\leq \text{NSW}(\pi^*, \hat{\mu}_t) + \sum_{a\in[K]} \pi^*_a \cdot \eta_{a,t} \\ &\leq \text{NSW}(\pi_t, \hat{\mu}_t) + \sum_{a\in[K]} \pi_{a,t} \cdot \eta_{a,t} \\ &\leq \text{NSW}(\pi_t, \mu^*) + 2\sum_{a\in[K]} \pi_{a,t} \cdot \eta_{a,t}.\end{aligned}$$

Thus, by bounding $\sum_{t\in[\tilde{O}(N^2K),T]} \sum_{a\in[K]} \pi_{a,t} \cdot \eta_{a,t}$, we obtain a bound on the total regret:

Theorem 4.6. *Suppose* $\forall j, a, t,\ r_{j,t,a} \in [0,1]$, $w_{j,a,t} = \sqrt{\frac{12(1-\hat{\mu}_{j,a,t})\ln(6NKT/\delta)}{N_{a,t}}} + \frac{12\ln(6NKT/\delta)}{N_{a,t}}$, *and*

$$\begin{aligned}\eta_{a,t} =& \tilde{O}(\sqrt{K/T}) \sum_{j\in[N]} (1 - \hat{\mu}_{j,a,t} \\ &+ \tilde{O}(\sqrt{T/K} + \sqrt{N})\frac{1}{N_{a,t}} \\ &+ O(1/\sqrt{N}) \sum_{j\in[N]} w_{j,a,t}\end{aligned}$$

for any $\delta \in (0,1)$, *the regret of Algorithm 2 is* $R_T = \widetilde{O}\left(\sqrt{KT} + N^2K\right)$ *with probability at least* $1-\delta$.

We defer the proof of this theorem and the details of the constants and log terms in the O and $\tilde{O}$ of η to the appendix. At a high level, this bound comes from using a tighter bound in place of Lemma 4.2, where we bound $\left|\sum_{j\in[N]} \mu^*_{j,a} - \hat{\mu}_{j,a,t}\right|$ instead of $\left|\mu^*_{j,a} - \hat{\mu}_{j,a,t}\right|$. We also analyze the regret at each time step t using

$$\text{NSW}(\pi, \hat{\mu}_t) = \prod_{j\in[N]} \left(\mathbb{E}_{a\sim\pi}\mu^*_{j,a} + \mathbb{E}_{a\sim\pi}(\hat{\mu}_{j,a,t} - \mu^*_{j,a})\right).$$

Since $\text{NSW}(\pi, \mu^*) = \prod_{j\in[N]} \mathbb{E}_{a\sim\pi}\mu^*_{j,a}$, we can bound $\text{NSW}(\pi, \mu^*) - \text{NSW}(\pi, \hat{\mu}_t)$ by

$$\sum_{m=1}^{N} \sum_{\{B\subseteq[N]:|B|=m\}} \prod_{j\in B} \frac{\mathbb{E}_{a\sim\pi}(\hat{\mu}_{j,a,t} - \mu^*_{j,a})}{\mathbb{E}_{a\sim\pi}\mu^*_{j,a}},$$

which drops the leading factor $\prod_j \mathbb{E}_{a\sim\pi}\mu^*_{j,a} \leq 1$ from the bound. Note that here the regret is bounded by $\mathbb{E}_{a\sim\pi}(\hat{\mu}_{j,a,t} - \mu^*_{j,a})$ as opposed to $\mathbb{E}_{a\sim\pi}|\hat{\mu}_{j,a,t} - \mu^*_{j,a}|$ in Lemma 4.1. Analyzing the terms with $m = 1$ and $m \geq 2$ separately allows us to derive the bound $\eta_{a,t}$ in our algorithm.

The algorithm obtains a bound in T which matches the known lower bound $O(\sqrt{KT})$ up to logarithmic terms, at the tradeoff of a high initial cost for pulling each arm $\tilde{O}(N^2)$ times. Additionally, similar to the UCB algorithm of (Hossain, Micha, and Shah 2021) the algorithm uses an additive regularization term in its optimization step and therefore does not have a known efficient implementation.

5 Experiments

We test Algorithm 1 and the UCB algorithm from (Hossain, Micha, and Shah 2021) with both $\alpha_t = N$ and $\alpha_t = \sqrt{12NK\log(NKt)}$, although we exclude the second α_t from the results because it was outperformed by the other algorithms in every test. We test three pairs of (N, K): a small size $(4, 2)$, a medium size $(20, 4)$, and a large size $(80, 8)$. For each pair (N, K) we test the algorithms on 5 values of $\mu^\star$ chosen randomly from 1 minus an exponential distribution with mean 0.04, and rounded up to 0.1 in extreme cases. When an action $a \in [K]$ is taken, we draw the rewards for each agent $j \in [N]$ from a Bernoulli distribution with $p = \mu^\star_{j,a}$. For the optimization step, we round the empirical mean up to 10^{-3} since this is a divisor in gradient computations. We also optimized the additive terms in both algorithms using a constant factor found through empirical binary search: the additive term $w_{j,a,t}$ in Algorithm 1 is scaled by 0.5, and the additive term in the optimization step of (Hossain, Micha, and Shah 2021) is scaled by 0.8. [1].

In each algorithm we compute π_t using the projected gradient ascent. For Algorithm 1, we take advantage of the log-concavity of the Nash Social Welfare function and the monotonicity of the logarithm and optimize the log of the objective in the gradient ascent.. The UCB algorithm from (Hossain, Micha, and Shah 2021) computes the policy π_t as $\arg\max_\pi \text{NSW}(\pi, \hat{\mu}_t) + \alpha_t \sum_{a\in[K]} \left(\pi_a \cdot \sqrt{\frac{\log(NKt)}{N_{a,t}}}\right)$, which is no longer a log-concave objective due to the linear terms (Hossain, Micha, and Shah 2021). We can address this issue to some degree by allowing the gradient ascent to run substantially longer: for Algorithm 1 the gradient ascent terminates after the objective changes by less than $2 \cdot 10^{-4}$ after 20 iterations, and for (Hossain, Micha, and Shah 2021) the ascent terminates after changing less than 10^{-6} after 30 iterations. By tightening the termination conditions we make it more difficult for the ascent to hang on a suboptimal position in (Hossain, Micha, and Shah 2021) at the tradeoff of longer runtime. Even at this point, we still see instability in the regret graph over time with the non-concave optimization. We include the algorithm none-the-less as it is the only other existing algorithm for fair multi-agent multi-armed bandits with regret on the order of $O(\sqrt{T})$.

Table 2 shows the average regret of the two algorithms after 200,000 iterations and 500,000 iterations over the 5 independent instances for each size, which are the same set of instances for both algorithms. Figure 1 shows cumulative regret graphs of the algorithms' instances for each setting of N and K.

In all cases, Algorithm 1 outperforms the previous best algorithm as T becomes large. Once the $\sqrt{T}$ terms in the regret bounds become dominant, the $\sqrt{K} \cdot \sqrt{\min N, K}$ factor saved in Algorithm 1 over the previous algorithm becomes apparent. This is especially true as N and K are larger. For the small case where $(N, K) = (4, 2)$, both regret curves take on the shape $\sqrt{T}$, with a growing separation between the two

[1] The repository is hosted at github.com/MetricJones/Fair-MultiAgent-MultiArmed-Bandits.

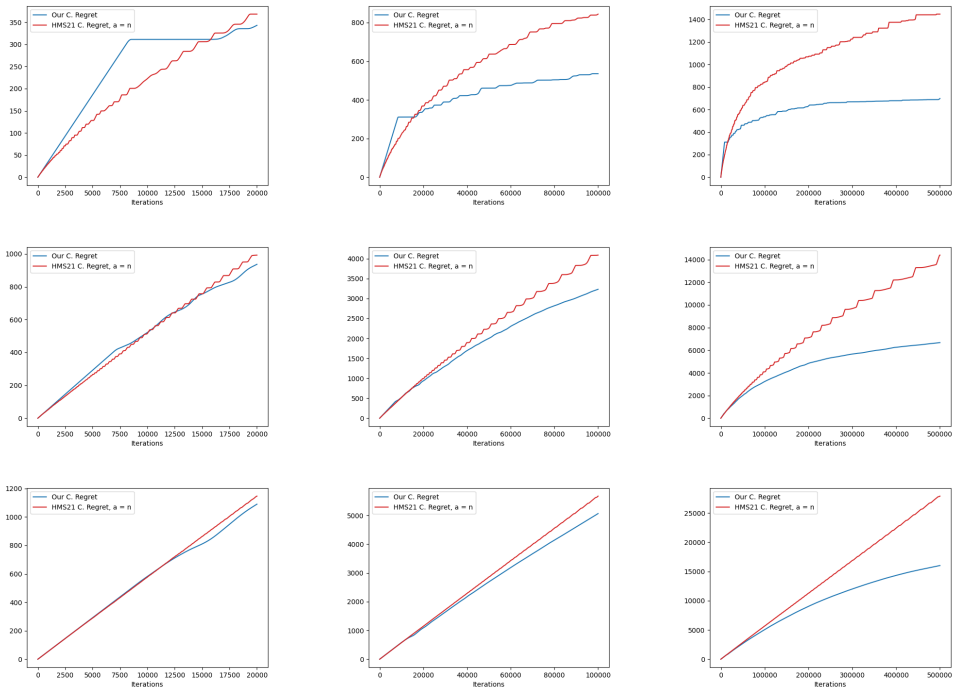

Figure 1: Sample Cumulative Regret Graphs for each setting. (N, K) is (4,2) in the top row, (20,4) in the center row, and (80,8) on the bottom row. The regret graphs show up to 20,000 iterations in the first column, 100,000 iterations in the second column, and 500,000 iterations in the third.

Table 2: Average cumulative regrets over 10 values of μ^*, with $T = 5 \cdot 10^5$

N	K	Algorithm 1		(Hossain, Micha, and Shah 2021)		NSW$(\pi^\star, \mu^\star)$
		$t = 2 \cdot 10^5$	$t = 5 \cdot 10^5$	$t = 2 \cdot 10^5$	$t = 5 \cdot 10^5$	
4	2	1222	1806	1226	1832	0.9428 ± 0.0370
20	4	8313	15322	10801	21655	0.6353 ± 0.0358
80	8	4521	9966	5230	12874	0.0808 ± 0.0154

as T increases. However, it is worth noting that (Hossain, Micha, and Shah 2021) does outperform at early iterations. This is due to the difference in the two UCB approaches. Specifically, (Hossain, Micha, and Shah 2021) does not make any changes to $\hat{\mu}$ in the optimization step, so their algorithm is able to begin improving immediately. Algorithm 1 adds the confidence bound to $\hat{\mu}$ and then caps all elements in $\hat{\mu}$ at 1, so until one of the terms in the upper confidence bound drops below 1 the algorithm will choose a uniform π, which accounts for the large linear cumulative regret in the first 7500 iterations before the upper confidence bound is non-trivial and Algorithm 1 begins to outperform (Hossain, Micha, and Shah 2021). For the small size, it seems that Algorithm 1 performs better as long as the number of iterations is at least 25,000.

In the medium case, the linear section in the regret curve of Algorithm 1 still does not outperform (Hossain, Micha, and Shah 2021), but (Hossain, Micha, and Shah 2021) sees a steeper regret curve which narrows the gap for small T and creates a larger regret gap for large T. In the largest case, where $(N, K) = (80, 8)$, we see that (Hossain, Micha, and Shah 2021) barely ever outperforms Algorithm 1 even at small T when Algorithm 1's upper confidence bound is a matrix of 1s. At high values of T we still see that Algorithm 1 outperform and we observe that the other factors in the $\sqrt{T}$ terms begin to play a significant role. Algorithm 1's regret curve still takes on a $\sqrt{T}$ shape although it is much gentler than the smaller cases, which can be at least partially attributed to the fact that the instantaneous regret at each round is bounded by the optimal value of the Nash social welfare, which is the product of expected rewards. The algorithm from (Hossain, Micha, and Shah 2021) still appears almost linear even at 500,000 iterations, as substantially larger values of T are required to overcome the factors of N and K and see the $\sqrt{T}$ shape. There is still a substantial difference between the cumulative regrets of the two algorithms as T increases.

Our experiments support the theoretical gains of our results. We see that at small values of T, under 10,000, our algorithm may be outperformed by (Hossain, Micha, and Shah 2021) for small sizes of N and K. This effect becomes significantly

weaker as N and K increase. At sufficiently large T, on the order of 25,000, Algorithm 1 outperforms the previous state-of-the-art with increasing significance and consistency as K, N, and T increase. Additionally, as T increases the regret curves of Algorithm 1 are significantly smoother than those of the previous algorithm due to the efficiency of the optimization step.

6 Acknowledgments

The authors are supported in part by NSF grants 1750716 and 1909314.

References

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 2002. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1): 48–77.

Bargiacchi, E.; Verstraeten, T.; Roijers, D.; Nowé, A.; and Hasselt, H. 2018. Learning to coordinate with coordination graphs in repeated single-stage multi-agent decision problems. In *International conference on machine learning*, 482–490. PMLR.

Bastani, H.; and Bayati, M. 2020. Online decision making with high-dimensional covariates. *Operations Research*, 68(1): 276–294.

Bouneffouf, D.; Bouzeghoub, A.; and Gançarski, A. L. 2012. A contextual-bandit algorithm for mobile context-aware recommender system. In *International conference on neural information processing*, 324–331. Springer.

Bouneffouf, D.; Bouzeghoub, A.; and Gançarski, A. L. 2013. Contextual bandits for context-based information retrieval. In *International Conference on Neural Information Processing*, 35–42. Springer.

Bubeck, S.; and Eldan, R. 2016. Multi-scale exploration of convex functions and bandit convex optimization. In *Conference on Learning Theory*, 583–589. PMLR.

Bubeck, S.; Lee, Y. T.; and Eldan, R. 2017. Kernel-based methods for bandit convex optimization. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, 72–85.

Chakraborty, M.; Chua, K. Y. P.; Das, S.; and Juba, B. 2017. Coordinated Versus Decentralized Exploration In Multi-Agent Multi-Armed Bandits. In *IJCAI*, 164–170.

Chen, L.; Zhang, M.; and Karbasi, A. 2019. Projection-free bandit convex optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2047–2056. PMLR.

Chung, F.; and Lu, L. 2006. Concentration inequalities and martingale inequalities: a survey. *Internet mathematics*, 3(1): 79–127.

Durand, A.; Achilleos, C.; Iacovides, D.; Strati, K.; Mitsis, G. D.; and Pineau, J. 2018. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Machine learning for healthcare conference*, 67–82. PMLR.

Flaxman, A. D.; Kalai, A. T.; and McMahan, H. B. 2004. Online convex optimization in the bandit setting: gradient descent without a gradient. *arXiv preprint cs/0408007*.

Gillen, S.; Jung, C.; Kearns, M.; and Roth, A. 2018. Online learning with an unknown fairness metric. *Advances in neural information processing systems*, 31.

Hazan, E.; and Levy, K. 2014. Bandit convex optimization: Towards tight bounds. *Advances in Neural Information Processing Systems*, 27.

Hazan, E.; and Megiddo, N. 2007. Online learning with prior knowledge. In *International Conference on Computational Learning Theory*, 499–513. Springer.

Hoffman, M.; Brochu, E.; de Freitas, N.; et al. 2011. Portfolio Allocation for Bayesian Optimization. In *UAI*, 327–336. Citeseer.

Hossain, S.; Micha, E.; and Shah, N. 2021. Fair Algorithms for Multi-Agent Multi-Armed Bandits. *Advances in Neural Information Processing Systems*, 34.

Huo, X.; and Fu, F. 2017. Risk-aware multi-armed bandit problem with application to portfolio selection. *Royal Society open science*, 4(11): 171377.

Joseph, M.; Kearns, M.; Morgenstern, J.; Neel, S.; and Roth, A. 2016. Fair algorithms for infinite and contextual bandits. *arXiv preprint arXiv:1610.09559*.

Kaneko, M.; and Nakamura, K. 1979. The Nash Social Welfare Function. *Econometrica*, 47(2): 423–435.

Kleinberg, R. 2004. Nearly tight bounds for the continuum-armed bandit problem. *Advances in Neural Information Processing Systems*, 17.

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

Kleinberg, R.; Slivkins, A.; and Upfal, E. 2019. Bandits and experts in metric spaces. *Journal of the ACM (JACM)*, 66(4): 1–77.

Landgren, P.; Srivastava, V.; and Leonard, N. E. 2016. Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, 167–172. IEEE.

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

Nash, J. F. 1950. The Bargaining Problem. *Econometrica*, 18(2): 155–162.

Patil, V.; Ghalme, G.; Nair, V.; and Narahari, Y. 2020. Achieving Fairness in the Stochastic Multi-Armed Bandit Problem. In *AAAI*, 5379–5386.

Shen, W.; Wang, J.; Jiang, Y.-G.; and Zha, H. 2015. Portfolio choices with orthogonal bandit learning. In *Twenty-fourth international joint conference on artificial intelligence*.

Slivkins, A. 2019. Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272*.

Wang, L.; and Joachims, T. 2020. Fairness and Diversity for Rankings in Two-Sided Markets. *arXiv preprint arXiv:2010.01470*.

Yue, Y.; Broder, J.; Kleinberg, R.; and Joachims, T. 2012. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5): 1538–1556.

Yue, Y.; and Joachims, T. 2009. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 1201–1208.

Zhou, Q.; Zhang, X.; Xu, J.; and Liang, B. 2017. Large-scale bandit approaches for recommender systems. In *International Conference on Neural Information Processing*, 811–821. Springer.

Zhu, H.; and Walker, A. 2018. Pension system reform in China: Who gets what pensions? *Social Policy & Administration*, 52(7): 1410–1424.

A Notations

N	Number of agents
K	Number of arms
T	Time horizon
π_t	Arm selection policy at round t
a_t	Arm selected at round t
$N_{a,t}$	Number of times that arm a has been selected up to round t
$\mu^\star_{j,a}$	Mean reward of arm a from agent j
$\hat{\mu}_{j,a,t}$	Empirical mean reward of arm a from agent j up to round t
$w_{j,a,t}$	Confidence interval of mean reward of arm a from agent j up to round t
$U_{j,a,t}$	Upper confidence bound of mean reward of arm a from agent j up to round t

Table 3: Table of notations.

B Missing Proofs

We first analyze the Lipschitz-continuity of $\mathrm{NSW}(\pi, \mu)$ when the policy π is fixed.

Lemma B.1 (Lemma 4.1)**.** *Given a policy $\pi \in \Delta^k$ and reward matrices $\mu^1, \mu^2 \in [0,1]^{N\times K}$, we have*

$$\left|\mathrm{NSW}(\pi, \mu^1) - \mathrm{NSW}(\pi, \mu^2)\right| \le \sum_{j\in[N]} \sum_{a\in[K]} \pi_a \left|\mu^1_{j,a} - \mu^2_{j,a}\right|$$

Proof. We express the difference as the telescoping sum,

$$\begin{aligned}
\mathrm{NSW}(\pi, \mu^1) - \mathrm{NSW}(\pi, \mu^2) &= \prod_{j\in[N]} \sum_{a\in[K]} \pi_a \mu^1_{j,a} - \prod_{j\in[N]} \sum_{a\in[K]} \pi_a \mu^2_{j,a} \\
&= \sum_{j\in[N]} \sum_{a\in[K]} \pi_a (\mu^1_{j,a} - \mu^2_{j,a}) \prod_{j'=1}^{j-1} \sum_{a'\in[K]} \pi_{a'} \mu^1_{j',a} \prod_{j''=j+1}^{N} \sum_{a''\in[K]} \pi_a \mu^2_{j'',a} \\
&\le \sum_{j\in[N]} \sum_{a\in[K]} \pi_a \left|\mu^1_{j,a} - \mu^2_{j,a}\right|
\end{aligned}$$

The inequality is due to the fact that $\sum_{a\in[K]} \pi_a \mu_{j,a} \le 1$. Similarly, we have,

$$\mathrm{NSW}(\pi, \mu^2) - \mathrm{NSW}(\pi, \mu^1) \le \sum_{j\in[N]} \sum_{a\in[K]} \pi_a \left|\mu^1_{j,a} - \mu^2_{j,a}\right|$$

■

Next, we derive a variant of the Chernoff bound.

Lemma B.2. *Suppose $X_1 \dots X_m$ are independent random variable taking values in $[0,1]$. Let $X = \sum_{i=1}^m X_i, \mu = \mathbb{E}[X]$, with probability $1-\delta$,*

$$|X - \mu| \le \sqrt{3\mu \ln(2/\delta)} + 3\ln(2/\delta)$$

Proof. Recall the multiplicative Chernoff bound,

$$\Pr[|X - \mu| \ge \alpha\mu] \le 2\exp(-\min(\alpha^2, \alpha)\mu/3)$$

If $\mu \le 3\ln(2/\delta)$, set $\alpha = 3\ln(2/\delta)/\mu$, we have:

$$\Pr[|X - \mu| \ge 3\ln(2/\delta)] \le \exp(-\alpha\mu/3) = \delta$$

Else, set $\alpha = \sqrt{3\ln(2/\delta)/\mu}$,

$$\Pr[|X - \mu| \ge \sqrt{3\mu\ln(2/\delta)}] \le \exp(-\alpha^2\mu/3) = \delta$$

■

Lemma B.3 (Lemma 4.2)**.** *For any $\delta \in (0,1)$, with probability at least $1 - \delta/2$, $\forall\, t > K, a \in [K], j \in [N]$, $\left|\mu^\star_{j,a} - \hat{\mu}_{j,a,t}\right| \le \sqrt{\frac{12(1-\hat{\mu}_{j,a,t})\ln(4NKT/\delta)}{N_{a,t}}} + \frac{12\ln(4NKT/\delta)}{N_{a,t}} = w_{j,a,t}$.*

Proof. By lemma B.2, we have with probability at least $1-\delta/(2NKT)$, :

$$\left|(1-\hat{\mu}_{j,a,t})-(1-\mu^\star_{j,a})\right| \le \sqrt{\frac{3(1-\mu^\star_{j,a})\ln(4NKT/\delta)}{N_{a,t}}}+\frac{3\ln(4NKT/\delta)}{N_{a,t}}.$$

We also have:

$$(1-\mu^\star_{j,a})-(1-\hat{\mu}_{j,a,t}) \le \sqrt{\frac{3(1-\mu^\star_{j,a})\ln(4NKT/\delta)}{N_{a,t}}}+\frac{3\ln(4NKT/\delta)}{N_{a,t}}$$

If $(1-\mu^\star_{j,a})-\sqrt{\frac{3(1-\mu^\star_{j,a})\ln(4NKT/\delta)}{N_{a,t}}} \ge 0$:

$$\sqrt{1-\mu^\star_{j,a}-\sqrt{\frac{3(1-\mu^\star_{j,a})\ln(4NKT/\delta)}{N_{a,t}}}} \ge \sqrt{1-\mu^\star_{j,a}}-\sqrt{\frac{3\ln(4NKT/\delta)}{N_{a,t}}}$$

In either cases, we have:

$$\begin{aligned}
&\sqrt{1-\hat{\mu}_{j,a,t}+\frac{3\ln(4NKT/\delta)}{N_{a,t}}} \ge \sqrt{1-\mu^\star_{j,a}}-\sqrt{\frac{3\ln(4NKT/\delta)}{N_{a,t}}} \\
\implies &\sqrt{1-\hat{\mu}_{j,a,t}+\frac{3\ln(4NKT/\delta)}{N_{a,t}}}+\sqrt{\frac{3\ln(4NKT/\delta)}{N_{a,t}}} \ge \sqrt{1-\mu^\star_{j,a}} \\
\implies &2\sqrt{1-\hat{\mu}_{j,a,t}+\frac{3\ln(4NKT/\delta)}{N_{a,t}}}\cdot\sqrt{\frac{3\ln(4NKT/\delta)}{N_{a,t}}}+\frac{6\ln(4NKT/\delta)}{N_{a,t}} \ge \hat{\mu}_{j,a,t}-\mu^\star_{j,a} \\
\implies &\sqrt{\frac{12(1-\hat{\mu}_{j,a,t})\ln(4NKT/\delta)}{N_{a,t}}}+\frac{12\ln(4NKT/\delta)}{N_{a,t}} \ge \hat{\mu}_{j,a,t}-\mu^\star_{j,a}
\end{aligned}$$

Similarly for the other direction, we have:

$$\left|\hat{\mu}_{j,a,t}-\mu^\star_{j,a}\right| \le \sqrt{\frac{12(1-\hat{\mu}_{j,a,t})\ln(4NKT/\delta)}{N_{a,t}}}+\frac{12\ln(4NKT/\delta)}{N_{a,t}}$$

The lemma follows by applying the union bound. ∎

Lemma B.4 (Lemma 4.3)**.** *Define* $g_{j,t}=\sum_{a\in[K]}\pi_{a,t}\,(1-U_{j,a,t})$*, and* $S(t,p)=\{j \text{ for } j\in[N] : g_{j,t}\ge 2^{-p}\}$ *. If* $|S(t,p)|<2^p\cdot 3\ln T$ *for all* $p\ge 0$*, then*

$$\begin{aligned}
&\sum_{j\in[N]}\sum_{a\in[K]}\pi_{a,t}(1-\hat{\mu}_{j,a,t}) \\
&\le 1+6\ln T\log N+\sum_{j\in[N]}\sum_{a\in[K]}\pi_{a,t}w_{j,a,t}.
\end{aligned}$$

Proof. Define $S'(t,p)=S(t,p+1)\setminus S(t,p)$. Note that $\sum_{j\in S'(t,p)}g_{j,t}\le|S(t,p+1)|\cdot 2^{-p}<2^{p+1}\cdot 3\ln T\cdot 2^{-p}$. We have:

$$\begin{aligned}
\sum_{j\in[N]}g_{j,t} &\le \sum_{j\notin S(t,\log N)}g_{j,t}+\sum_{j'\in S(t,\log N)}g_{j',t} \\
&\le \sum_{j\notin S(t,\log N)}g_{j,t}+\sum_{j'\in S(t,\lfloor\log N\rfloor)}g_{j',t} \\
&\le N\cdot\frac{1}{N}+\sum_{p=0}^{\lfloor\log N\rfloor}\sum_{j'\in S'(t,p)}g_{j',t}
\end{aligned}$$

$$\leq N \cdot \frac{1}{N} + \sum_{p=0}^{\lfloor \log N \rfloor} 2^{p+1} \cdot 3 \ln T \cdot 2^{-p}$$
$$\leq 1 + 6 \ln T \log N. \tag{7}$$

Thus,

$$\sum_{j\in[N]} \sum_{a\in[K]} \pi_{a,t}(1-\hat{\mu}_{j,a,t}) - 1 - 6\ln T \log N$$
$$\leq \sum_{j\in[N]} \sum_{a\in[K]} \pi_{a,t}(1-\hat{\mu}_{j,a,t}) - \sum_{j\in[N]} g_{j,t}$$
$$= \sum_{j\in[N]} \sum_{a\in[K]} \pi_{a,t}(1-\hat{\mu}_{j,a,t}) - \pi_{a,t}\left(1 - U_{j,a,t}\right)$$
$$= \sum_{j\in[N]} \sum_{a\in[K]} \pi_{a,t}(U_{j,a,t} - \hat{\mu}_{j,a,t})$$
$$\leq \sum_{j\in[N]} \sum_{a\in[K]} \pi_{a,t} w_{j,a,t}.$$

The first inequality is due to Equation 7. The last inequality is due to the definition of $U_{j,a,t}$. ■

Lemma B.5 (Lemma 4.4)**.** *With probability* $1-\delta/2, \sum_{t\in[T]} \sum_{a\in[K]} \pi_{a,t}/N_{a,t} \leq 2K\left(\ln \frac{T}{K} + 1\right) + \ln(2/\delta)$.

Proof. Let $v_t = \sum_{a\in[K]} \pi_{a,t}/N_{a,t}$. Consider the σ-algebra $\mathcal{F}_t = (a_1, a_2 \ldots a_t)$. Note that $v_{a,t}$ is $\mathcal{F}_t$-measurable. Define $g(y) = 2\sum_{i=2}^{\infty} \frac{y^{i-2}}{i!} = 2\frac{e^y - 1 - y}{y^2}$. We have

$$\mathbb{E}_{a\sim\pi_t}[\exp(zv_t)|\mathcal{F}_{t-1}] = \mathbb{E}_{a\sim\pi_t}\left[\sum_{i=0}^{\infty} \frac{z^i v_t^i}{i!}|\mathcal{F}_{t-1}\right]$$
$$= 1 + \mathbb{E}_{a\sim\pi_t}\left[zv_t|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[\sum_{i=2}^{\infty} \frac{(zv_t)^i}{i!}|\mathcal{F}_{t-1}\right]$$
$$= 1 + \mathbb{E}_{a\sim\pi_t}\left[zv_t|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[\frac{z^2 v_t^2}{2} g(zv_t)|\mathcal{F}_{t-1}\right]$$
$$\leq 1 + \mathbb{E}_{a\sim\pi_t}\left[zv_t|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[\frac{z^2 v_t}{2} g(z)|\mathcal{F}_{t-1}\right]$$
$$= 1 + \mathbb{E}_{a\sim\pi_t}\left[\left(z + \frac{z^2 g(z)}{2}\right) v_t|\mathcal{F}_{t-1}\right]$$
$$= 1 + \mathbb{E}_{a\sim\pi_t}\left[\left(z + \frac{z^2 g(z)}{2}\right) \frac{1}{N_{a,t}}|\mathcal{F}_{t-1}\right]$$
$$\leq \exp\left(\mathbb{E}_{a\sim\pi_t}\left[\left(z + \frac{z^2 g(z)}{2}\right) \frac{1}{N_{a,t}}|\mathcal{F}_{t-1}\right]\right)$$
$$\leq \mathbb{E}_{a\sim\pi_t}\left[\exp\left(\left(z + \frac{z^2 g(z)}{2}\right) \frac{1}{N_{a,t}}\right)|\mathcal{F}_{t-1}\right]$$

The first inequality is due to $v_t \leq 1$ and $g(y)$ is monotonically increasing for $y \geq 0$. The last two inequalities follow from the fact that $1 + x \leq e^x$ and $e^{\mathbb{E}[x]} \leq \mathbb{E}[e^x]\ \forall x$. Thus by induction over t,

$$\mathbb{E}_{a_1,\ldots,a_T}\left[\exp\left(z\sum_{t\in[T]} v_t\right)\right] \leq \mathbb{E}_{a_1,\ldots,a_T}\left[\exp\left(\left(z + \frac{z^2 g(z)}{2}\right)\sum_t \frac{1}{N_{a_t,t}}\right)\right]$$
$$\leq \exp\left(\left(z + \frac{z^2 g(z)}{2}\right) K\left(\ln\frac{T}{K} + 1\right)\right)$$

By Markov's inequality,

$$\Pr\left[\sum_{t\in[T]} v_t \geq \frac{\ln\alpha}{z}\right] = \Pr\left[\exp\left(z\sum_{t\in[T]} v_t\right) \geq \alpha\right]$$
$$\leq \exp\left(\left(z + \frac{z^2 g(z)}{2}\right) K\left(\ln\frac{T}{K}+1\right)\right)/\alpha$$

The lemma follows by setting $z = 1, \alpha = \exp\left(\left(z + \frac{z^2 g(z)}{2}\right) K\left(\ln\frac{T}{K}+1\right) + \ln(2/\delta)\right)$. ■

The Inefficient Algorithm with Improved Regret

We begin by reviewing the algorithm. We first pull each arm a total of $180N^2\ln(6NTK/\delta)$ times to establish $\hat{\mu}$ with high accuracy. Then, we select the policy at each time step in order to optimize the Nash social welfare plus the dot product with an error vector η. Note that the vector η defined in Theorem B.6 is slightly different than in Theorem 4.6. This error vector matches the proof below, and yields a simpler proof. Additionally, notice that we constrain π_t to the region S_π in Algorithm 3, which also eases the proof and is a linear constraint, and therefore a convex set, at each iteration.

Algorithm 3: Fair multi-agent UCB algorithm with high start-up cost

input: K, N, T, δ
for $t = 1$ to T **do**
  if $t \leq 180N^2K\ln(6NTK/\delta)\ln T$ **then**
    $\pi_t \leftarrow$ policy that puts probability 1 on arm $t \bmod (K+1)$
  else
    $\forall j, a, \hat{\mu}_{j,a,t} = \frac{1}{N_{a,t}}\sum_{\tau=1}^{t-1} r_{j,a_\tau,\tau}\mathbf{1}\{a_\tau = a\}$
    Let $S_\pi = \Delta^K \cap \{\pi : \sum_{a\in[K]}\pi_a\left(\sum_{j\in[N]}(1-\hat{\mu}_{j,a,t})\right) \leq 1 + 2\ln T\}$
    if $S_\pi \neq \emptyset$ **then**
      $\pi_t \leftarrow \operatorname{argmax}_{\pi\in S_\pi}\left(\mathrm{NSW}(\pi, \hat{\mu}_t) + \pi\cdot\eta_t\right)$
    else
      $\pi_t \leftarrow$ policy that puts probability 1 on a random arm $a \in [K]$
    end if
  end if
  Sample a_t from π_t
  Observe rewards $\{r_{j,a_t,t}\}_{j\in N}$
  $N_{a_t,t+1} \leftarrow N_{a_t,t} + 1$
end for

Theorem B.6 (Theorem 4.6)**.** *Suppose* $\forall j, a, t, r_{j,t,a} \in [0,1]$, $w_{j,a,t} = \sqrt{\frac{12(1-\hat{\mu}_{j,a,t})\ln(6NKT/\delta)}{N_{a,t}}} + \frac{12\ln(6NKT/\delta)}{N_{a,t}}$, *and*

$$\eta_{a,t} = \left(4\sqrt{\ln(6KT/\delta)} + 6\sqrt{2}\ln(6NKT/\delta)\sqrt{2+2\ln T}\right)\sqrt{\frac{K}{T}}\sum_{j\in[N]}(1-\hat{\mu}_{j,a,t})$$
$$+\Big(\left(4\sqrt{\ln(6KT/\delta)} + \sqrt{1+\ln T}\right)\sqrt{\frac{T}{K}}$$
$$+ 12\sqrt{2}\sqrt{N}\ln(6NKT/\delta)\sqrt{2+2\ln T}\Big)\frac{1}{N_{a,t}}$$
$$+\frac{1}{20\sqrt{N}/19 - 1}\sum_{j\in[N]} w_{j,a,t}$$

for any $\delta \in (0,1)$*, the regret of Algorithm 3 is* $R_T = \widetilde{O}\left(\sqrt{KT} + N^2K\right)$ *with probability at least* $1-\delta$*.*

We will use several lemmas from earlier to prove this, but we will need to condition one more event in lemma B.7. Therefore, we will need to adapt Lemmas 4.2 and 4.4 such that they have failure probability $\delta/3$ instead of $\delta/2$.

In order to prove this theorem we introduce the following notation:

$$a_j(\pi) = \sum_{a\in[K]} \pi_a \mu^*_{j,a} = 1 - \sum_{a\in[K]} \pi_a(1-\mu^*_{j,a})$$
$$b_{j,t}(\pi) = \sum_{a\in[K]} \pi_a(\hat{\mu}_{j,a,t} - \mu^*_{j,a})$$

where $a_j(\pi)$ is the true expected reward of agent j under policy π and $b_{j,t}(\pi)$ is the expected error $\hat{\mu}_t - \mu^*$ for agent j under policy π. Note that $\text{NSW}(\pi,\mu^*) = \text{NSW}'(\pi,\mu^*) = \prod_{j=1}^N a_j(\pi)$ and $\text{NSW}(\pi,\hat{\mu}_t) = \text{NSW}'(\pi,\hat{\mu}_t) = \prod_{j=1}^N (a_j(\pi)+b_{j,t}(\pi))$. We can expand the product $\text{NSW}'(\pi,\hat{\mu})$ to obtain

$$\begin{aligned}
\text{NSW}'(\pi,\hat{\mu}_t) &= \prod_{j=1}^N (a_j(\pi)+b_{j,t}(\pi)) \\
&= \left(\prod_{j=1}^N a_j(\pi)\right)\left(1+\sum_{m=1}^N \sum_{B\in\{S\subseteq[N]:|S|=m\}} \prod_{i\in B} \frac{b_{j,t}(\pi)}{a_j(\pi)}\right) \\
\implies \text{NSW}(\pi,\hat{\mu}_t) - \text{NSW}(\pi,\mu^*) &= \text{NSW'}(\pi,\hat{\mu}_t) - \text{NSW}'(\pi,\mu^*) \\
&= \left(\prod_{j=1}^N a_j(\pi)\right)\left(\sum_{m=1}^N \sum_{B\in\{S\subseteq[N]:|S|=m\}} \prod_{j\in B} \frac{b_{j,t}(\pi)}{a_j(\pi)}\right).
\end{aligned}$$

We will omit the t from the subscript of b and we will omit the arguments π for conciseness in sections of this proof.

The proof structure will be ordered as follows. First, we prove a Chernoff bound in Lemma B.7 and bounds on $\sum_{a\in[K]} \pi_a \left(\sum_{j\in[N]}(1-\mu_{j,a})\right)$ for useful pairs of π and μ in Lemmas B.9, B.12, and B.13 to bound terms of the form $\sum_{j\in[N]} b_{j,t}(\pi)$. Then, we bound the terms with $m \geq 2$ using Lemma B.14 and we bound the terms with $m = 1$ in Lemma B.15 and Lemma B.16 combined in Lemma B.17. With all the terms bounded, we can obtain a bound on the entire error summed over $t \in [T]$ in order to prove Theorem B.6.

Lemma B.7. *With probability at least $1-\delta/3$, for all $a \in [K]$ and all $t \in (180N^2K\ln(6NTK/\delta)\ln(T), T]$ we have*

$$\left|\sum_{j\in[N]} \left(\hat{\mu}_{j,a,t} - \mu^*_{j,a}\right)\right| \leq \sqrt{\frac{4\ln(6KT/\delta)\sum_{j\in[N]}(1-\mu^*_{j,a})}{N_{a,t}}}.$$

Proof. Fix t to be a value between 1 and T and $a \in [K]$. Let $v_{j,a,t}$ be the value we obtain for agent j when we pull arm a for the t-th time. Let $x_{j,a,t} = v_{j,a,t} - \mu^*_{j,a}$. Observe that $x_{j,a,t}$ is in the range $[-1,1]$ with mean 0 and variance $\mu^*_{j,a}(1-\mu^*_{j,a}) \leq 1-\mu^*_{j,a}$. Thus, $X_{t_0} = \sum_{j\in[N]} x_{j,a,t_0}$ has mean 0 and variance at most $\sigma^2 = \sum_{j\in[N]}(1-\mu^*_{j,a})$. By the Chernoff bound (see e.g. Theorem 3.1 in (Chung and Lu 2006)),

$$\Pr[|\sum_{t_0=1}^t X_{t_0}| \geq k\sigma] \leq 2\exp(-k^2/(4t))$$

By choosing $k = \sqrt{4t\ln(6KT/\delta)}$, we have that with probability $1-\delta/(3KT)$,

$$\left|\sum_{t_0=1}^t X_{t_0}\right| \leq \sqrt{4t\ln(6KT/\delta)\sum_{j\in[N]}(1-\mu^*_{j,a})}$$
$$\left|\frac{1}{t}\sum_{t_0=1}^t X_{t_0}\right| \leq \sqrt{\frac{4\ln(6KT/\delta)\sum_{j\in[N]}(1-\mu^*_{j,a})}{t}}$$

By union bound over all choices of $t \in [T]$ and $a \in [K]$, the above inequality holds for all t and a simultaneously with probability $1-\delta/3$. Note that for all time horizon $t \in [T]$, because $N_{a,t} \in [T]$ and $\hat{\mu}_{j,a,t}$ is the average of the $N_{a,t}$ values we obtained for agent j and arm a, it follows that with probability $1-\delta/3$,

$$\left|\sum_{j\in[N]} \left(\hat{\mu}_{j,a,t} - \mu^*_{j,a}\right)\right| \leq \sqrt{\frac{4\ln(6KT/\delta)\sum_{j\in[N]}(1-\mu^*_{j,a})}{N_{a,t}}}\ \forall a,t.$$

■

Before we begin to bound the terms in our expansion of the regret, we introduce a few lemmas in the vein of Lemma 4.3 in order to bound $\sum_{a\in[K]} \pi_a \left(\sum_{j\in[N]}(1-\mu_{j,a})\right)$ for useful pairs of π and μ. We show that these hold when conditioning on an event whose failure gives us a trivial regret bound in Lemma B.10.

Lemma B.8. *Given N values $0 \le a_1, ..., a_N \le 1$, if $\prod_{j\in[N]} a_j \ge T^{-2}$ then $\sum_{j\in[N]}(1-a_j) \le 2\ln T$ for all $T \ge 1$.*

Proof. By the inequality $1 + x \le e^x$, we have

$$\begin{aligned} 1 + (a_j - 1) &\le e^{a_j - 1} \\ \prod_{j\in[N]} a_j &\le e^{\sum_j (a_j - 1)} \\ \implies \ln\left(\prod_{j\in[N]} a_j\right) &\le \sum_j (a_j - 1) \\ \implies -2\ln(T) &\le \sum_{j\in[N]} (a_j - 1) \\ \implies 2\ln(T) &\ge \sum_{j\in[N]} (1 - a_j) \end{aligned}$$

■

For any π, μ in our problem, then, we have that

$$\mathrm{NSW}(\pi, \mu) = \prod_{j\in[N]} \left(\sum_{a\in[K]} \pi_a \mu_{j,a}\right) \le T^{-2}$$

or

$$\sum_{j\in[N]} \left(1 - \sum_{a\in[K]} \pi_a \mu_{j,a}\right) = \sum_{j\in[N]} \left(\sum_{a\in[K]} \pi_a (1 - \mu_{j,a})\right) \le 2\ln T$$

which yields the immediate corollary:

Lemma B.9. *If $\mathrm{NSW}(\pi^*, \mu^*) \ge T^{-2}$ then $\sum_{j\in[N]} \left(\sum_{a\in[K]} \pi^*_a (1 - \mu^*_{j,a})\right) \le 2\ln T$.*

Additionally, since we have that $\mathrm{NSW}(\pi_t, \mu^*) \ge 0$, we have the following trivial regret bound:

Lemma B.10. *If $\sum_{j\in[N]} \left(\sum_{a\in[K]} \pi^*_a (1 - \mu^*_{j,a})\right) \ge 2\ln T$, then we obtain the total regret bound*

$$R_T = \sum_{t\in[T]} (\mathrm{NSW}(\pi^*, \mu^*) - \mathrm{NSW}(\pi_t, \mu^*)) \le \sum_{t\in[T]} \mathrm{NSW}(\pi^*, \mu^*) \le \frac{1}{T}.$$

Since we achieve a trivial regret bound in the other case, we condition on the event that $\sum_{j\in[N]} \left(\sum_{a\in[K]} \pi^*_a (1 - \mu^*_{j,a})\right) \le 2\ln T$ going forward.

Next, we want to achieve a similar bound with $\hat{\mu}_t$ in place of μ^*:

Lemma B.11. *If $\sum_{j\in[N]} \left(\sum_{a\in[K]} \pi^*_a (1 - \mu^*_{j,a})\right) \le 2\ln T$, then*

$$\sum_{j\in[N]} \sum_{a\in[K]} \pi^*_a (1 - \hat{\mu}_{j,a,t}) \le 1 + 2\ln T$$

for all $t \ge 180N^2 \ln(6NKT/\delta) \ln T$.

Proof. We can achieve this bound using Lemma B.9 and a bound on the term $\sum_{j\in[N]} \sum_{a\in[K]} \pi^*_a (\mu^*_{j,a} - \hat{\mu}_{j,a,t})$:

$$\sum_{j\in[N]} \sum_{a\in[K]} \pi^*_a (1 - \hat{\mu}_{j,a,t}) = \sum_{j\in[N]} \sum_{a\in[K]} \pi^*_a (1 - \mu^*_{j,a}) + \sum_{j\in[N]} \sum_{a\in[K]} \pi^*_a (\mu^*_{j,a} - \hat{\mu}_{j,a,t})$$

$$\leq 2\ln T + \sum_{a\in[K]} \pi^*_a \left(\sum_{j\in[N]} \left(\mu^*_{j,a} - \hat{\mu}_{j,a,t}\right) \right)$$
$$\leq 2\ln T + \sum_{a\in[K]} \pi^*_a \left| \sum_{j\in[N]} \left(\mu^*_{j,a} - \hat{\mu}_{j,a,t}\right) \right|$$
$$\leq 2\ln T + \sum_{a\in[K]} \pi^*_a \sqrt{\frac{4\ln(6KT/\delta)\sum_{j\in[N]}(1-\mu^*_{j,a})}{N_{a,t}}}$$

where the first inequality follows from Lemma B.9, the last inequality follows from conditioning on Lemma B.7. With $N_{a,t} \geq 180N^2\ln(6NKT/\delta)\ln(T)$ for $t \geq 180N^2K\ln(6NKT/\delta)\ln(T)$, we have

$$\sqrt{\frac{4\ln(6KT/\delta)\sum_{j\in[N]}(1-\mu^*_{j,a})}{N_{a,t}}} \leq \sqrt{\frac{4\ln(6KT/\delta)\cdot N}{180N^2\ln(6NKT/\delta)\ln T}} \leq \frac{1}{5\sqrt{N}}$$

for all $a \in [K]$. The bound holds since $\sum_{a\in[K]} \pi^*_a = 1$. ■

We then restrict π_t to the linear constraint $\sum_{a\in[K]} \pi_{a,t}\left(\sum_{j\in[N]}(1-\hat{\mu}_{j,a,t})\right) \leq 1 + 2\ln T$, as given by the set S_π in Algorithm 3. By contraposition, if no feasible solution π_t exists then we know our conditioning has failed, and any policy π will give us the trivial regret bound in Lemma B.10. Otherwise, a feasible solution exists and therefore we can guarantee that we have $\sum_{a\in[K]} \pi_{a,t}\left(\sum_{j\in[N]}(1-\hat{\mu}_{j,a,t})\right) \leq 1 + 2\ln T$:

Lemma B.12.

$$\sum_{a\in[K]} \pi_{a,t} \left(\sum_{j\in[N]} (1-\hat{\mu}_{j,a,t}) \right) \leq 1 + 2\ln T$$

for all $t \geq 180N^2K\ln(6NKT/\delta)\ln T$ conditioned on the events in Lemma B.9 and B.7.

Additionally, we repeat the analysis from B.11 to obtain

Lemma B.13.

$$\sum_{a\in[K]} \pi_{a,t} \left(\sum_{j\in[N]} (1-\mu^*_{j,a}) \right) \leq 2 + 2\ln T.$$

At this point we have all of the tools necessary to continue.

Lemma B.14. *If $N_{a,t} \geq 180N^2(1+\log N)\ln(6NTK/\delta)\ln T$ for all a, then*

$$\left| \sum_{m=2}^{N} \sum_{B\in\{S:S\subset[N]\wedge|S|=m\}} \left(\prod_{j\in B} b_{j,t}(\pi) \prod_{j'\notin B} a_{j'}(\pi) \right) \right| \leq \frac{1}{20\sqrt{N}/19 - 1} \left(\sum_{j\in[N]} \sum_{a\in[K]} \pi_a w_{j,a,t} \right).$$

conditioned on the event in adapted Lemma 4.2.

Proof. For higher-order terms ($2 \leq |S| \leq N$), we have

$$\left| \sum_{m=2}^{N} \sum_{B\in\{S:S\subset[N]\wedge|S|=m\}} \prod_{j\in B} b_j(\pi) \prod_{j'\notin B} a_{j'}(\pi) \right| \leq \sum_{m=2}^{N} \sum_{B\in\{S:S\subset[N]\wedge|S|=m\}} \prod_{j\in B} |b_j(\pi)|$$
$$\leq \sum_{m=2}^{N} \left(\sum_{j\in[N]} |b_j(\pi)| \right)^m$$
$$= \left(\sum_{j\in[N]} |b_j(\pi)| \right) \sum_{m=2}^{N} \left(\sum_{j\in[N]} |b_j(\pi)| \right)^{m-1}$$

We start by bounding $\sum |b_j|$:

$$\begin{aligned}\sum_{j\in[N]} |b_j(\pi)| &= \sum_{j\in[N]} \left| \sum_{a\in[K]} \pi_a(\hat{\mu}_{j,a,t} - \mu^*_{j,a}) \right| \\ &\leq \sum_{j\in[N]} \sum_{a\in[K]} \pi_a w_{j,a,t} \\ &= \sum_{j\in[N]} \sum_{a\in[K]} \pi_a \left(\sqrt{\frac{12(1-\hat{\mu}_{j,a,t})\ln(6NKT/\delta)}{N_{a,t}}} + \frac{12\ln(6NKT/\delta)}{N_{a,t}} \right)\end{aligned}$$

We analyze the two summands on the innermost term separately. First, the right summand yields

$$\begin{aligned}\sum_{j\in[N]} \sum_{a\in[K]} \pi_a \left(\frac{12\ln(6NKT/\delta)}{N_{a,t}} \right) &= N \sum_{a\in[K]} \pi_a \left(\frac{12\ln(6NKT/\delta)}{N_{a,t}} \right) \\ &\leq N \left(\frac{1}{2N^2 \ln T} \right) \\ &\leq 1/(2N)\end{aligned}$$

which also gives us the bound $1/(2\sqrt{N})$. The left summand gives

$$\begin{aligned}\sum_{j\in[N]} \sum_{a\in[K]} \pi_a \left(\sqrt{\frac{12(1-\hat{\mu}_{j,a,t})\ln(6NKT/\delta)}{N_{a,t}}} \right) &= \sum_{a\in[K]} \sum_{j\in[N]} \pi_a \cdot \sqrt{12\ln(6NKT/\delta)} \cdot \sqrt{\frac{(1-\hat{\mu}_{j,a,t})}{N_{a,t}}} \\ &\leq \sum_{a\in[K]} \pi_a \sum_{j\in[N]} \left(\frac{1}{12\sqrt{N}\ln T}(1-\hat{\mu}_{j,a,t}) \right) \\ &\quad + \sum_{a\in[K]} \pi_a \sum_{j\in[N]} \left(36 \cdot \ln(6NKT/\delta)\sqrt{N}\ln T \frac{1}{N_{a,t}} \right) \\ &\leq \frac{\sum_{a\in[K]} \sum_{j\in[N]} \pi_a(1-\hat{\mu}_{j,a,t})}{12\sqrt{N}\ln T} + \frac{36\sqrt{N}\ln T \ln(6NKT/\delta) * N}{180\ln(6NKT/\delta)N^2 \ln T} \\ &\leq \frac{1+2\ln T}{12\sqrt{N}\ln T} + \frac{1}{5\sqrt{N}} \\ &= \frac{1}{12\sqrt{N}\ln T} + \frac{1}{6\sqrt{N}} + \frac{1}{5\sqrt{N}}\end{aligned}$$

where the first inequality is an application of Young's inequality with $z = 6\sqrt{12N\ln(6NKT/\delta)}\ln T$ and the third inequality holds for $\pi = \pi^*$ and $\pi = \pi_t$ because of Lemmas B.11 and B.12 respectively. This is bounded by $\frac{9}{20\sqrt{N}}$, which combined with the other bound, $1/2N$, we have the bound $19/(20\sqrt{N})$. Since $N \geq 1$, this bound is at most 19/20 < 1. We use the earlier bound $\sum_j |b_j(\pi)| \leq \sum_j \sum_{a\in[K]} \pi_a w_{j,a,t}$ and include the sum over the order of the terms m:

$$\begin{aligned}\left| \sum_{m=2}^{N} \sum_{B\in\{S:S\subset[N]\wedge|S|=m\}} \left(\sum_{j\in B} b_{j,t}(\pi) \prod_{j'\notin B} a_{j'}(\pi) \right) \right| &\leq \sum_{m=2}^{N} \left(\sum_{j\in[N]} |b_{j,t}(\pi)| \right)^m \\ &\leq \sum_{m=2}^{N} \left(\left(19N^{-1/2}/20\right)^{m-1} \sum_{j\in[N]} \sum_{a\in[K]} |b_{j,t}(\pi)| \right) \\ &\leq \sum_{m=2}^{N} \left(\left(19N^{-1/2}/20\right)^{m-1} \sum_{j\in[N]} \sum_{a\in[K]} \pi_a w_{j,a,t} \right) \\ &\leq \frac{(19N^{-1/2}/20)}{1-(19N^{-1/2}/20)} \sum_{j\in[N]} \sum_{a\in[K]} \pi_a w_{j,a,t}\end{aligned}$$

$$= \frac{1}{20\sqrt{N}/19 - 1} \sum_{j\in[N]} \sum_{a\in[K]} \pi_a w_{j,a,t}.$$

■

This gives us the bound on the higher order terms, so we continue with the lower order terms.

Lemma B.15.

$$\left|\sum_{j\in[N]} b_{j,t}(\pi)\right| \leq 4\sqrt{\ln(6KT/\delta)} \left(\sqrt{\frac{K}{T}} \sum_{a\in[K]} \pi_a \sum_{j\in[N]} (1-\hat{\mu}_{j,a,t}) + \sqrt{\frac{T}{K}} \sum_{a\in[K]} \pi_a/N_{a,t}\right)$$

conditioned on the events of Lemma B.7.

Proof.

$$\begin{aligned}
\left|\sum_{j\in[N]} b_{j,t}(\pi)\right| &= \left|\sum_{j\in[N]} \left(\sum_{a\in[K]} \pi_a(\hat{\mu}_{j,a,t} - \mu^*_{j,a})\right)\right| \\
&= \left|\sum_{a\in[K]} \pi_a \left(\sum_{j\in[N]} (\hat{\mu}_{j,a,t} - \mu^*_{j,a})\right)\right| \\
&\leq \sum_{a\in[K]} \pi_a \left|\sum_{j\in[N]} (\hat{\mu}_{j,a,t} - \mu^*_{j,a})\right| \\
&\leq \sum_{a\in[K]} \pi_a \sqrt{\frac{4\ln(6KT/\delta)\cdot\sum_{j\in[N]}(1-\mu^*_{j,a})}{N_{a,t}}} \\
&\leq 2\sqrt{\ln(6KT/\delta)} \left(\frac{1}{z}\sum_{a\in[K]} \pi_a \sum_{j\in[N]} (1-\mu^*_{j,a}) + z\sum_{a\in[K]} \pi_a/N_{a,t}\right)
\end{aligned}$$

where the second-to-last inequality follows from Lemma B.7 and last inequality is an application of Young's Inequality which holds for $z > 0$. It is important that we remove the algorithmically unknown term $\mu^*_{j,a}$ from our bound, so we split that term.

$$\begin{aligned}
\left|\sum_{j\in[N]} b_{j,t}(\pi)\right| &\leq 2\sqrt{\ln(6KT/\delta)} \left(\frac{1}{z}\sum_{a\in[K]} \pi_a\Big(\sum_{j\in[N]} (1-\hat{\mu}_{j,a,t}) + \sum_{j\in[N]} (\hat{\mu}_{j,a,t} - \mu^*_{j,a})\Big) + z\sum_{a\in[K]} \pi_a/N_{a,t}\right) \\
&\leq 2\sqrt{\ln(6KT/\delta)} \left(\frac{1}{z}\sum_{a\in[K]} \pi_a \left(\sum_{j\in[N]} (1-\hat{\mu}_{j,a,t}) + \left|\sum_{j\in[N]} (\hat{\mu}_{j,a,t} - \mu^*_{j,a})\right|\right) + z\sum_{a\in[K]} \pi_a/N_{a,t}\right)
\end{aligned}$$

Note that the second term is an earlier step in the bound with a factor $\frac{2\sqrt{\ln(6KT/\delta)}}{z}$, specifically from the first inequality in the proof. Therefore,

$$\begin{aligned}
\left|\sum_{j\in[N]} b_{j,t}(\pi)\right| \leq \sum_{a\in[K]} \pi_a \left|\sum_{j\in[N]} (\hat{\mu}_{j,a,t} - \mu^*_{j,a})\right| &\leq 2\sqrt{\ln(6KT/\delta)}\Big(\frac{1}{z}\sum_{a\in[K]} \pi_a \sum_{j\in[N]} (1-\hat{\mu}_{j,a,t}) \\
&\quad + \frac{1}{z}\sum_{a\in[K]} \pi_{a,t}\Big|\sum_{j\in[N]} (\hat{\mu}_{j,a,t} - \mu^*_{j,a})\Big| + z\sum_{a\in[K]} \pi_a/N_{a,t}\Big) \\
\implies \left(1 - \frac{2\sqrt{\ln(6KT/\delta)}}{z}\right)\sum_{a\in[K]} \pi_a \left|\sum_{j\in[N]} (\hat{\mu}_{j,a,t} - \mu^*_{j,a})\right| &\leq 2\sqrt{\ln(6KT/\delta)} \left(\frac{1}{z}\sum_{a\in[K]} \pi_a \sum_{j\in[N]} (1-\hat{\mu}_{j,a,t}) + z\sum_{a\in[K]} \pi_a/N_{a,t}\right) \\
\implies \left|\sum_{j\in[N]} b_{j,t}(\pi)\right| \leq \sum_{a\in[K]} \pi_a \left|\sum_{j\in[N]} (\hat{\mu}_{j,a,t} - \mu^*_{j,a})\right| &\leq \frac{2z\sqrt{\ln(6KT/\delta)}}{z - 2\sqrt{\ln(6KT/\delta)}} \left(\frac{1}{z}\sum_{a\in[K]} \pi_a \sum_{j\in[N]} (1-\hat{\mu}_{j,a,t}) + z\sum_{a\in[K]} \pi_a/N_{a,t}\right)
\end{aligned}$$

$$\leq 4\sqrt{\ln(6KT/\delta)}\left(\sqrt{\frac{K}{T}}\sum_{a\in[K]}\pi_a\sum_{j\in[N]}(1-\hat{\mu}_{j,a,t})+\sqrt{\frac{T}{K}}\sum_{a\in[K]}\pi_a/N_{a,t}\right)$$

where we use $z=\sqrt{T/K}$ and $T\geq 16K\sqrt{\ln(6KT/\delta)}$. ■

Lemma B.16. *If $N_{a,t}\geq 180N^2\ln(6NTK/\delta)\ln T$ for all a, then*

$$\sqrt{\sum_{j\in[N]}b_{j,t}^2(\pi)}\leq\left(\frac{\sqrt{T}}{\sqrt{2}\sqrt{K}}+12\sqrt{2}\sqrt{N}\ln(6NKT/\delta)\right)\sum_{a\in[K]}\frac{\pi_a}{N_{a,t}}+\frac{6\sqrt{2}\sqrt{K}\ln(6NKT/\delta)}{\sqrt{T}}\sum_{a\in[K]}\sum_{j\in[N]}\pi_a(1-\hat{\mu}_{j,a,t})$$

conditioned on the event in the adapted Lemma 4.2.

Proof.

$$\begin{aligned}
\sqrt{\sum_{j\in[N]}b_j^2(\pi)}&=\sqrt{\sum_{j\in[N]}\Big(\sum_{a\in[K]}\pi_a(\hat{\mu}_{j,a,t}-\mu_{j,a}^*)\Big)^2}\\
&=\sqrt{\sum_{j\in[N]}\sum_{a_1,a_2\in[K]}\pi_{a_1}\pi_{a_2}(\hat{\mu}_{j,a_1,t}-\mu_{j,a_1}^*)(\hat{\mu}_{j,a_2,t}-\mu_{j,a_2}^*)}\\
&=\sqrt{\sum_{a_1,a_2\in[K]}\pi_{a_1}\pi_{a_2}\sum_{j\in[N]}(\hat{\mu}_{j,a_1,t}-\mu_{j,a_1}^*)(\hat{\mu}_{j,a_2,t}-\mu_{j,a_2}^*)}\\
&\leq\sqrt{\sum_{a_1,a_2\in[K]}\pi_{a_1}\pi_{a_2}\sqrt{\left(\sum_{j\in[N]}(\hat{\mu}_{j,a_1,t}-\mu_{j,a_1}^*)^2\right)\left(\sum_{j\in[N]}(\hat{\mu}_{j,a_2,t}-\mu_{j,a_2}^*)^2\right)}}\\
&=\sqrt{\left(\sum_{a\in[K]}\pi_a\sqrt{\sum_{j\in[N]}(\hat{\mu}_{j,a.t}-\mu_{j,a}^*)^2}\right)^2}\\
&=\sum_{a\in[K]}\pi_a\sqrt{\sum_{j\in[N]}(\hat{\mu}_{j,a.t}-\mu_{j,a}^*)^2}\\
&\leq\sum_{a\in[K]}\pi_a\sqrt{\sum_{j\in[N]}w_{j,a,t}^2}\\
&=\sum_{a\in[K]}\pi_a\sqrt{\sum_{j\in[N]}\left(\sqrt{\frac{12(1-\hat{\mu}_{j,a,t})\ln(6NKT/\delta)}{N_{a,t}}}+\frac{12\ln(6NKT/\delta)}{N_{a,t}}\right)^2}\\
&\leq\sum_{a\in[K]}\pi_a\sqrt{2\sum_{j\in[N]}\left(\frac{12(1-\hat{\mu}_{j,a,t})\ln(6NKT/\delta)}{N_{a,t}}+\left(\frac{12\ln(6NKT/\delta)}{N_{a,t}}\right)^2\right)}\\
&\leq\sqrt{2}\sum_{a\in[K]}\pi_a\left(\sqrt{\sum_{j\in[N]}\frac{12(1-\hat{\mu}_{j,a,t})\ln(6NKT/\delta)}{N_{a,t}}}+\sqrt{\sum_{j\in[N]}\left(\frac{12\ln(6NKT/\delta)}{N_{a,t}}\right)^2}\right)
\end{aligned}$$

The first inequality holds by the Cauchy-Schwarz inequality, the second holds by the adapted Lemma 4.2, the third holds because $2(x^2+y^2)\geq(x+y)^2$, and the last holds because $\sqrt{x+y}\leq\sqrt{x}+\sqrt{y}$ for all $x,y\geq 0$. We bound these sums separately:

$$\sum_{a\in[K]}\pi_a\sqrt{\sum_{j\in[N]}\frac{12(1-\hat{\mu}_{j,a,t})\ln(6NKT/\delta)}{N_{a,t}}}\leq\frac{z}{2}\sum_{a\in[K]}\frac{\pi_a}{N_{a,t}}+\frac{1}{2z}\sum_{a\in[K]}\pi_a\sum_{j\in[N]}12(1-\hat{\mu}_{j,a,t})\ln(6NKT/\delta)$$

$$= \frac{\sqrt{T}}{2\sqrt{K}} \sum_{a\in[K]} \frac{\pi_a}{N_{a,t}} + \frac{6\sqrt{K}\ln(6NKT/\delta)}{\sqrt{T}} \sum_{a\in[K]} \sum_{j\in[N]} \pi_a(1-\hat{\mu}_{j,a,t})$$

$$\sum_{a\in[K]} \pi_a \sqrt{\sum_{j\in[N]} \left(\frac{12\ln(6NKT/\delta}{N_{a,t}}\right)^2} = 12\sqrt{N}\ln(6NKT/\delta) \sum_{a\in[K]} \frac{\pi_a}{N_{a,t}}$$

where we use Young's inequality in the first sum. Combining these bounds gives the lemma. ■

Lemma B.17.

$$\left(\prod_{j=1}^{N} a_j(\pi)\right)\left(\sum_{j\in[N]} \frac{b_j(\pi)}{a_j(\pi)}\right) \le \left(4\sqrt{\ln(6KT/\delta)} + 6\sqrt{2}\ln(6NKT/\delta)\sqrt{2+2\ln T}\right)\sqrt{\frac{K}{T}} \sum_{a\in[K]} \pi_a \sum_{j\in[N]} (1-\hat{\mu}_{j,a,t})$$
$$+ \Bigg(\left(4\sqrt{\ln(6KT/\delta)} + \sqrt{1+\ln T}\right)\sqrt{\frac{T}{K}}$$
$$+ 12\sqrt{2}\sqrt{N}\ln(6NKT/\delta)\sqrt{2+2\ln T}\Bigg) \sum_{a\in[K]} \pi_a/N_{a,t}$$

Proof. We begin our bound on first-order ($|S| = 1$) terms by separating into two sums.

$$\left(\prod_{j=1}^{N} a_j(\pi)\right)\left(\sum_{j\in[N]} \frac{b_j(\pi)}{a_j(\pi)}\right) \le \left(\prod_{j=1}^{N} a_j(\pi)\right)\left(\left|\sum_{j\in[N]} b_j(\pi)\right| + \sum_{j\in[N]} |b_j(\pi)||1/a_j(\pi) - 1|\right)$$
$$\le \left(\prod_{j=1}^{N} a_j(\pi)\right)\left|\sum_{j\in[N]} b_j(\pi)\right| + \sum_{j\in[N]} |b_j(\pi)||1-a_j(\pi)|$$
$$\le \left|\sum_{j\in[N]} b_j(\pi)\right| + \sqrt{\left(\sum_{j\in[N]} b_j^2(\pi)\right)\left(\sum_{j\in[N]} (1-a_j(\pi))^2\right)}$$
$$\le \left|\sum_{j\in[N]} b_j(\pi)\right| + \sqrt{\left(\sum_{j\in[N]} b_j^2(\pi)\right)\sum_{j\in[N]} 1-a_j(\pi)}$$
$$\le \left|\sum_{j\in[N]} b_j(\pi)\right| + \sqrt{2+2\ln T}\cdot\sqrt{\sum_{j\in[N]} b_j^2(\pi)}$$

where the second inequality holds because $|1/a_j(\pi) - 1| = |1 - a_j(\pi)|/a_j(\pi)$ and the third inequality is an application of Cauchy-Schwarz. We then combine with the bounds from Lemmas B.15 and B.16 to obtain our bound. Note that our $\sqrt{2+2\ln T}$ term comes from B.9 for π^* and B.13 for π_t. ■

We are now ready to prove Theorem B.6:

Proof.

$$|\text{NSW}(\pi, \hat{\mu}_t) - \text{NSW}(\pi, \mu^*)| = \left|\prod_{j=1}^{N}(a_j(\pi) + b_{j,t}(\pi)) - \prod_{j=1}^{N} a_j(\pi)\right|$$
$$\le \left|\sum_{j\in[N]} b_{j,t}(\pi)\right| + \sqrt{2+2\ln T}\sqrt{\sum_{j\in[N]} b_{j,t}^2(\pi)} + \sum_{m=2}^{N}\left(\sum_{j\in[N]} |b_{j,t}(\pi)|\right)^m$$

$$
\begin{aligned}
&\leq \left(4\sqrt{\ln(6KT/\delta)} + 6\sqrt{2}\ln(6NKT/\delta)\sqrt{2+2\ln T}\right)\sqrt{\frac{K}{T}}\sum_{a\in[K]}\pi_a\sum_{j\in[N]}(1-\hat{\mu}_{j,a,t}) \\
&\quad + \left(\left(4\sqrt{\ln(6KT/\delta)} + \sqrt{1+\ln T}\right)\sqrt{\frac{T}{K}}\right. \\
&\quad \left. + 12\sqrt{2}\sqrt{N}\ln(6NKT/\delta)\sqrt{2+2\ln T}\right)\sum_{a\in[K]}\pi_a/N_{a,t} \\
&\quad + \frac{1}{20\sqrt{N}/19-1}\sum_{j\in[N]}\sum_{a\in[K]}\pi_a w_{j,a,t} \\
&= f_t(\pi,\hat{\mu}_t)
\end{aligned}
$$

gives us a bound on the error of the Nash social welfare for a policy between the true and estimated rewards. Note that all of the values in this bound, N, K, T, $N_{a,t}$, $\hat{\mu}$, and w, are available to the algorithm at iteration t. Additionally, we see that all of these terms can be rewritten as the dot product of a vector with π. Therefore, we are able to optimize $f_t(\pi,\hat{\mu}_t)$ by computing the vector η_t as in Algorithm 3. We define π_t at each time step $t > 180N^2K\ln(6NTK/\delta)\ln T$ as

$$
\pi_t = \arg\max_{\pi\in\Delta_K} \text{NSW}(\pi,\hat{\mu}_t) + f_t(\pi,\hat{\mu}_t)
$$

Then we have at each time step $t > 180N^2K\ln(6NTK/\delta)\ln T$:

$$
\begin{aligned}
\text{NSW}(\pi^*,\mu^*) &\leq \text{NSW}(\pi^*,\hat{\mu}_t) + f_t(\pi^*,\hat{\mu}_t) \\
&\leq \text{NSW}(\pi_t,\hat{\mu}_t) + f_t(\pi_t,\hat{\mu}_t) \\
&\leq \text{NSW}(\pi_t,\mu^*) + 2f_t(\pi_t,\hat{\mu}_t),
\end{aligned}
$$

where the second bound holds because we condition on $\pi^* \in S_\pi$, which shows that our instantaneous regret at time t is bounded by $\text{NSW}(\pi^*,\mu^*) - \text{NSW}(\pi_t,\mu^*) \leq 2f_t(\pi_t,\hat{\mu}_t)$. At this point we recall that we condition on the event in Lemma B.9, otherwise we have the trivial regret bound T^{-1} by Lemma B.10.

We first pull each arm $\tilde{O}(N^2)$ times, each with an instantaneous regret of at most 1, for a $\tilde{O}(N^2K)$ term in the regret bound. The remainder of the regret bound comes from bounding the sum $2\sum_{t\in[T_0,T]} f_t(\pi_t,\hat{\mu}_t)$ over all other rounds:

$$
\begin{aligned}
\sum_t f_t(\pi_t,\hat{\mu}_t) &= \left(4\sqrt{\ln(6KT/\delta)} + 6\sqrt{2}\ln(6NKT/\delta)\sqrt{2+2\ln T}\right)\sqrt{\frac{K}{T}}\sum_t\sum_{a\in[K]}\pi_a\sum_{j\in[N]}(1-\hat{\mu}_{j,a,t}) \\
&\quad + \left(\left(4\sqrt{\ln(6KT/\delta)} + \sqrt{1+\ln T}\right)\sqrt{\frac{T}{K}}\right. \\
&\quad \left. + 12\sqrt{2}\sqrt{N}\ln(6NKT/\delta)\sqrt{2+2\ln T}\right)\sum_{a\in[K]}\sum_t\pi_a/N_{a,t} \\
&\quad + \frac{1}{20\sqrt{N}/19-1}\sum_t\sum_{j\in[N]}\sum_{a\in[K]}\pi_a w_{j,a,t} \\
&\leq \tilde{O}(\sqrt{KT}) + \tilde{O}(\sqrt{KT}+K\sqrt{N}) + \tilde{O}\left(\frac{\sqrt{NKT}}{\sqrt{N}}\right)
\end{aligned}
$$

where we bound $\sum_{a\in[K]}\sum_{j\in[N]}\pi_{a,t}(1-\hat{\mu}_{j,a,t}) \leq 1 + 2\ln T$ using Lemma B.12, we bound $\sum_t\sum_{a\in[K]}\pi_{a,t}/N_{a,t}$ using Lemma 4.4, and we bound $\sum\sum\sum\pi_{a,t}w_{j,a,t}$ using the analysis in Theorem 4.5 (Equations 3, 4, 6) with Lemma B.12 in the place of 4.3. Putting this together with the trivial regret bound for the first $T_0 = 180N^2K\ln(6NTK/\delta)\ln T$ rounds gives us the entire bound $\tilde{O}(\sqrt{KT}+N^2K)$. We obtain our failure probability from the events we condition on in Lemma B.7 and the adapted Lemmas 4.2 and 4.4, which each have failure probability $\delta/3$. Hence, with probability at least $(1-\delta)$ we have regret at most $\tilde{O}(\sqrt{KT}+N^2K)$.

■

C Additional Experimental Data

Our error bounds tend to be quite large, since changing the value of μ^* significantly affects the resulting regret. We include them here for completeness, reported as mean plus/minus standard deviation. Additionally, we implement the Epsilon-Greedy and

Explore-First algorithms from (Hossain, Micha, and Shah 2021) for completeness. Algorithm 1 consistently outperforms the two at all sizes for a sufficiently large time horizon. Epsilon-Greedy tends to yield a very smooth regret curve which is always worse than Algorithm 1 once the algorithms move from exploration to exploitation. Explore-First performs very well after it is finished exploring, showing nearly optimal behavior. However, the startup cost is prohibitively high, and is truncated on the $(80, 8)$ instances even at 500,000 iterations.

Table 4: Average cumulative regrets for $T^{1/2}$-regret algorithms over 10 values of μ^*, with $T = 5 \cdot 10^5$

N	K	Algorithm 1		(Hossain, Micha, and Shah 2021)		$\mathrm{NSW}(\pi^\star, \mu^\star)$
		$t = 2 \cdot 10^5$	$t = 5 \cdot 10^5$	$t = 2 \cdot 10^5$	$t = 5 \cdot 10^5$	
4	2	1222 ± 127	1806 ± 501	1226 ± 201	1832 ± 549	0.9428 ± 0.0370
20	4	8313 ± 2211	15322 ± 4000	10801 ± 2975	21655 ± 5690	0.6353 ± 0.0358
80	8	4521 ± 1500	9966 ± 3049	5230 ± 1800	12874 ± 4398	0.0808 ± 0.0154

Table 5: Average running times over 10 values of μ^*, with $T = 5 \cdot 10^5$

N	K	Algorithm 1	(Hossain, Micha, and Shah 2021)
4	2	365 ± 12	802 ± 179
20	4	675 ± 49	2324 ± 496
80	6	1862 ± 65	14128 ± 629

We also include time measurements for the algorithms here, reported in seconds. Recall that the running time is tuneable by controlling the stopping conditions of the gradient ascent, these are the times reported for the same settings as the results in the Experiments section.

The experiments were run without the GPU, on a system with an Intel i5-11600K CPU and a Radeon RX 580 GPU. We omit experiments for Algorithm 2 because it performs poorly in practice due to its large constants and prohibitive start-up cost in terms of N, although it is of theoretical interest.

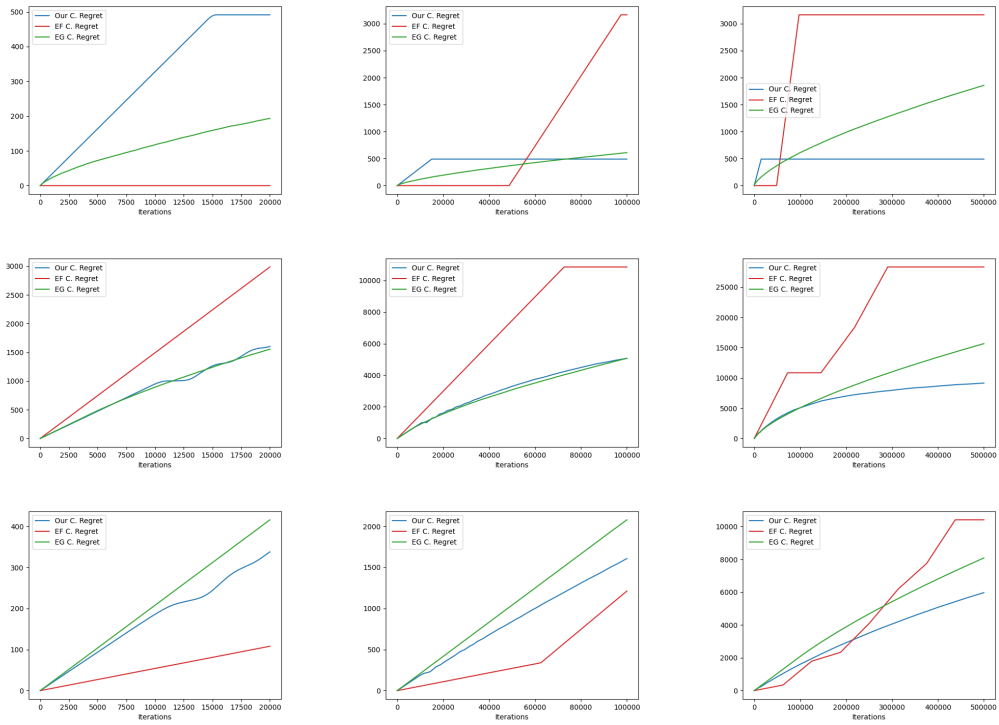

Figure 2: Sample Cumulative Regret Graphs for each setting. (N, K) is (4,2) in the top row, (20,4) in the center row, and (80,8) on the bottom row. The regret graphs show up to 20,000 iterations in the first column, 100,000 iterations in the second column, and 500,000 iterations in the third.