Voting-Based Multi-Agent Bandits

Alex Kozik, Eric Yachbes, Michael Ngo[*]

Abstract

Our paper examines multi-agent multi-armed bandit settings with a voting mechanism (aggregator) that considers agents' votes and pulls arms accordingly. Our setting is motivated by many real-world examples in which decisions (arms) are made through votes of a group of people (agents), and these decisions affect all people in this group differently. We show that with full knowledge of how these arms affect agents, agents can act optimally to achieve zero individual regret. We further show, by having the aggregator employ statistical learning algorithms, that agents can achieve sublinear individual regret. Finally, we examine games in which having all agents play optimally to minimize their individual regret does not lead to good social welfare outcomes, showing a price of anarchy of $O(N)$ and a social welfare regret of $O(NT)$.

[*] Authors listed in alphabetical order.

Contents

1 Introduction

There are many real-world examples where one decision affects a group of individuals, or one action affects the reward across multiple agents. A hedge fund that decides which stock to buy or sell will affect each of its investors. A company that amends its policy to increase parental leave will affect its employees in different ways. A dictator's decisions will affect the lives of the country's individuals. A dictator need not factor in their people's preferences, and make decisions that primarily benefit themselves. *But how can decisions be made that are fair to all individuals?*

We will consider a more, but not fully, democratic setting where the decision is made via a *voting system.* For example, the President of the United States is elected via the Electoral College. A CEO is chosen by a board of directors, and a board of directors is, in turn chosen by a company's shareholders. A recommendation algorithm makes decisions about what to recommend by incorporating user feedback and advertiser feedback. The voting system need not be that each individual has an equal vote, but rather that the ultimate decision is made after every individual's preferences are considered.

Lastly, note that it is natural to model individual preference via *bandit-style feedback.* For example, a user might not know what kind of content they like until they are shown it. A citizen might not know the true effects of electing their new mayor until decisions are made and implemented.

This work. The question we ask is: *under what strategies and voting systems, can fair decisions be made?* Specifically, we model this problem as an online, multi-agent, multi-armed bandits setting called Voting Bandits, where every round, one shared arm is pulled and decided via a voting algorithm called the Aggregator. We study Voting Bandits and aim to characterize how fairness can be achieved under different assumptions. Our results are as follows.

First, we define the novel bandit setting of Multi-Agent *Voting Bandits.* We define appropriate notions of individual and group regret, where group regret measures regret with respect to group fairness. We define T time steps, N agents, and K arms.

On the positive side, we show that

1. Assuming agents know their reward distributions, also called the known-reward model, we show that zero individual regret with respect to the aggregator is possible if the aggregator is a random sampler. Agents do as well as they can, proportional to the influence they have.

2. Assuming rewards across agents are perfectly, positively correlated, we show that total reward, called social welfare, is equivalent to a group fairness notation called Nash social welfare. If the agents are sufficiently similar, then being greedy can be fair.

3. We extend and analyze an explore–first-style strategy for Voting Bandits and show it incurs $\widetilde{O}(\sqrt{KT}/N)$ individual regret with respect to the aggregator. Thus, in some way, we have shown that voting bandits with regret w.r.t. aggregator has analogous results to the single-agent MAB problem.

4. We analyze an ε-greedy-style strategy for Voting Bandits which while it also has $O(T)$ individual regret w.r.t. self, under the assumption that agents are sufficiently similar enough to share the same best arm, we show it obtains $O(\log(T))$ individual regret w.r.t. self. When a

best decision exists for everyone, it is possible to have a strategy that is optimally greedy for all agents simultaneously.

On the negative side, when we get rid of these assumptions, we show impossibility of maximizing Nash social welfare, let alone regular social welfare under simple voting systems.

1. We show that our explore-first voting bandits strategy is tightly linear in individual regret w.r.t T, and we show experimentally the same holds for our ε-greedy voting strategy.
2. If the aggregator is a random sampler and all agents are rational, we show an example where neither social welfare nor Nash social welfare are maximized.
3. We generalize the result above to show that if the aggregator is a random sampler and all agents are rational, then the Price-of-Anarchy (PoA), the ratio between maximum social welfare and actual social welfare is $O(N)$.

1.1 Related Literature

[Foster and Rakhlin, 2023] describes the long studied problem of stochastic multi-armed bandits, formalizing regret minimization and the exploration-exploitation tradeoff. It analyzes algorithms such as ϵ-greedy and Upper Confidence Bound (UCB), which our work builds on and extends to the multi-agent voting setting.

Recently, fairness with respect to arms has been studied in the bandit setting. [Joseph et al., 2016] formalizes a fairness notion requiring that a fair strategy assigns a higher selection probability to an arm over another only when there is sufficient evidence that it has a higher expected reward. [Liu et al., 2017] propose a fairness framework for stochastic bandits based on smooth and calibrated fairness, requiring that similar arms be treated similarly and that arms be sampled in proportion to their probability of being optimal. These works showed a tradeoff between fairness and regret. Similarly, our work highlights the tradeoffs between fairness and regret, and continues to propose new fairness notions in the multi-agent voting bandit setting.

[Hossain et al., 2021] and [Harada et al., 2025] study fairness objectives in multi-agent bandit settings, proposing Nash Social Welfare and max-min fairness, respectively. Both works analyze centralized bandit algorithms, such as explore-first, ε-greedy, UCB, and multiplicative weights, that learn arm distributions optimizing these fairness metrics and obtain regret bounds. In contrast, our work does not impose a centralized fairness objective, but instead studies decentralized, strategic agents who vote over actions via an aggregator. This shows that individual no-regret behavior need not lead to optimal social or fairness outcomes.

Fairness in voting and public decision-making has been extensively studied in computational social choice. [Aziz et al., 2019] study fair mixing for public outcomes under dichotomous preferences (like/dislike), and introduce a hierarchy of fairness notions, such as Individual Fair Share, Unanimous Fair Share, and Average Fair Share. [Conitzer et al., 2017] generalize fair allocation to public decision-making, focusing on proportionality and its relaxations, and analyzing which properties these notions satisfy. These works informed our project by motivating the study of tradeoffs between individual and social incentives in collective decision-making settings. They also motivate our use of Nash social welfare as a computationally efficient, group fairness metric.

2 Preliminaries

2.1 Voting Bandits

Consider the multi-agent multi-armed bandits setting as follows. We say there are K arms, N agents, and T rounds. We say arms are indexed $a \in [K]$, agents are indexed $i \in [N]$, and rounds are indexed $t \in [T]$. Each agent-arm pair has a reward distribution $\mathcal{D}_{i,j}$ on support $[0, 1]$, where the reward for player i is drawn from this distribution when arm j is pulled. For each agent i and arm j, we denote $\mu_{ij} := \mathbb{E}_{D_{ij}}[r] := \mathbb{E}_{r \sim D_{ij}}[r]$ as the mean reward for each agent-arm pair.

For each round $t \in [T]$, each agent i selects a voting policy $p_i^t \in \Delta[K]$ without knowledge of other agents' selected policies. Then, each agent i draws an arm $a_i^t \sim p_i^t$ and votes for this arm. After that, the aggregator $\mathcal{A} : [K]^n \to \Delta([K])$ (which may be stateful) maps from all agents' selected arms to a distribution over arms and picks an arm $a^t \sim \mathcal{A}(a_1^t, \dots, a_n^t)$ from this distribution. Each agent i then receives independent reward $r_i^t \sim \mathcal{D}_{i,a^t}$. We call a strategy for the voting bandits setting the set of all K agents' and the aggregator's strategies.

Regarding the information each agent sees, we assume in this base setting of bandit-style feedback that the agents do not know the reward distributions $\mathcal{D}_{i,j}$ and that at the end of each round, each agent only sees their realized reward r_i^t.

We have a few ways to vary the above base setting and design goals for the agents and for the aggregator. These are described below.

2.1.1 Full vs Bandit Information

Our paper will analyze settings beyond the standard bandit settings. Specifically, we will refer to the information model where agents have access to full information of their own reward distributions. That is, each agent i has a complete description of $\mathcal{D}_{i,j}$ for all arms $j \in [K]$. In attempting to maximize expected regret, it suffices in many settings for agents to even just know their expected reward for their distributions, or $\mathbb{E}[\mathcal{D}_{i,j}]$ for all j, which is implied by the full information model.

2.1.2 Reward Correlation Models

In some cases, we apply additional assumptions to correlate the rewards or reward distributions over the agents. The strongest is the *Full Correlation Setting*, where all agents have identical reward distributions over arms, and when an arm is pulled, a single reward is realized and shared by all agents.

Definition 2.1 (Full Correlation Setting). *$\mathcal{D}_{i,j} = \mathcal{D}_j$ for all agents i and arms j. And at round t, if arm a^t is selected, then $r^t \sim \mathcal{D}_{a^t}$ is the reward all agents receive.*

Then, we consider a relaxation called the *I.I.D. Setting*, in which agents share the same reward distribution for each arm, but receive independent reward realizations. Therefore, rewards agree in distribution, but not in stochasticity.

Definition 2.2 (I.I.D. Setting). *$\mathcal{D}_{i,j} = \mathcal{D}_j$ for all agents i and arms j. And at round t, if arm a^t is selected, then $r_i^t \sim \mathcal{D}_{a^t}$ is i.i.d. sampled as agent i's reward.*

Finally, weakening this assumption yields ε-*Similarity*, which captures the notion that agents have approximately similar preferences: for each arm, all agents' expected rewards differ by at most ϵ.

Definition 2.3 (Similarity)**.** *For a fixed $\varepsilon \in [0,1]$, the agents are ε-similar when $|\mu_{i,j} - \mu_{i',j}| \leq \varepsilon$ for all agents i, i' and arms j.*

2.1.3 Individual Regret

While each agent would ideally want to maximize their cumulative reward, this is difficult because agents do not directly control which arm is selected. So, we define two different notions of regret.

The first is called regret with respect to self, which compares an agent's realized reward to the reward they would have obtained if their optimal arm had been selected on every round. This regret measures the agent's loss due to the fact that they don't have full control over the decision. Formally:

Definition 2.4 (Individual Regret w.r.t. Self)**.**

$$\mathrm{Reg}_i^{\mathrm{self}} = \mathbb{E}_G\left[\max_{j\in[K]}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{D}_{i,j}}[r_i^t] - \sum_{t=1}^{T}\mathbb{E}_{a^t\sim\mathcal{A}(a_1^t,\ldots,a_n^t)}[\mathbb{E}_{\mathcal{D}_{i,a^t}}[r_i^t]]\right]$$

The second is called regret with respect to the aggregator, which compares an agent's realized reward to the best reward they could have achieved by voting optimally, while holding the aggregator's and all other agents' behavior fixed. This notion of regret captures whether an agent is acting optimally given their partial influence on the outcome. Formally:

Definition 2.5 (Individual Regret w.r.t. Aggregator)**.**

$$\mathrm{Reg}_i^{\mathrm{agg}} = \mathbb{E}_G\left[\max_{j\in[K]}\sum_{t=1}^{T}\mathbb{E}_{a^t\sim\mathcal{A}(a_1^t,\ldots,a_i^t=j,\ldots,a_n^t)}[\mathbb{E}_{\mathcal{D}_{i,a^t}}[r_i^t]] - \sum_{t=1}^{T}\mathbb{E}_{a^t\sim\mathcal{A}(a_1^t,\ldots,a_n^t)}[\mathbb{E}_{\mathcal{D}_{i,a^t}}[r_i^t]]\right]$$

In both settings, the regret is computed in expectation over all possible games G. Notice that in $\mathrm{Reg}_i^{\mathrm{self}}$, the left term in the difference is under the assumption that the voted arm is always the optimal arm for player i, whereas in the $\mathrm{Reg}_i^{\mathrm{agg}}$, the voted arm is still dependent on other players' votes, but player i now acts to vote optimally.

2.1.4 Social Welfare Regret

We also define a notion of social welfare and social welfare regret, which we hope for our aggregator $\mathcal{A}$ to maximize and minimize respectively. Formally, we define:

Definition 2.6 (Social Welfare (SW))**.**

$$\mathrm{SW}(\mathcal{A}(a_1,\ldots,a_n)) = \mathbb{E}_G\left[\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}_{a^t\sim\mathcal{A}(a_1^t,\ldots,a_n^t)}[\mathbb{E}_{\mathcal{D}_{i,a^t}}[r_i^t]]\right]$$

Definition 2.7 (Social Welfare Regret)**.**

$$Reg^{\mathrm{SW}}_{\mathcal{A}(a_1,\ldots,a_n)} = \max_{\mathcal{A}^*} \mathrm{SW}(\mathcal{A}^*) - \mathrm{SW}(\mathcal{A}(a_1,\ldots,a_n))$$

Notice that for social welfare, the left side of the difference is equivalent to the social welfare achieved by always picking the arm j with highest expected sum of reward over all agents.

Similarly, we may define Nash social welfare and its regret by swapping the sum over agents with a product. Thus, maximizing this metric means to try to maximize the reward of each agent, invariant to scale. This is known as a fairness metric that acts a medium between social welfare and having total equality (egalitarianism), where all agents have equal reward.

Definition 2.8 (Nash Social Welfare (NSW))**.**

$$\mathrm{NSW}(\mathcal{A}(a_1,\ldots,a_n)) = \mathbb{E}_G\left[\prod_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}_{a^t\sim\mathcal{A}(a_1^t,\ldots,a_n^t)}[\mathbb{E}_{\mathcal{D}_{i,a^t}}[r_i^t]]\right]$$

Definition 2.9 (Nash Social Welfare Regret)**.**

$$Reg^{\mathrm{NSW}}_{\mathcal{A}(a_1,\ldots,a_n)} = \max_{\mathcal{A}^*} \mathrm{NSW}(\mathcal{A}^*) - \mathrm{NSW}(\mathcal{A}(a_1,\ldots,a_n))$$

2.1.5 Price of Anarchy

The price of anarchy is a measure of the ratio between the social optimal social welfare and the social welfare achieved by some other algorithm:

Definition 2.10 (Price of Anarchy)**.**

$$PoA_{\mathcal{A}(a_1,\ldots,a_n)} = \frac{\max_{\mathcal{A}^*} SW(\mathcal{A}^*)}{SW(\mathcal{A}(a_1,\ldots,a_n))}$$

3 Zero Regret w.r.t. Aggregator under Random Sampling

In this setting, we analyze the extent to which agents can minimize their individual regret when they have full information about their reward distributions. We find that in this setting, agents can play their highest expected value arm to achieve zero regret. To prove this, we split our regret term into two parts: one when the agent's vote is chosen by the aggregator, and the other when it is not. The first part contributes no regret, since the best choice in this setting is to play the highest-EV arm, and the second part also contributes zero regret, since the agent's vote does not affect their reward. This proof generally informs our strategy of each agent picking their empirically optimal arm in bandit information settings during "exploitation" rounds, which we explain later.

Theorem 3.1. *If each agent $i \in [n]$ knows its reward distributions $\mathcal{D}_{i,1}, \ldots, \mathcal{D}_{i,K}$ and the aggregator $\mathcal{A}$ chooses a^t uniformly random from $\{a_1^t, \ldots, a_n^t\}$, then if each agent always selects its own optimal arm*

$$j^* = \arg\max_{j \in [K]} \mathbb{E}_{\mathcal{D}_{i,j}}[r_i^t],$$

each agent i achieves zero aggregator regret

$$\mathrm{Reg}_i^{\mathrm{agg}} = 0.$$

Proof. In general, for a random sampler aggregator, we can write the expected reward for agent i^* as

$$\sum_{t=1}^{T} \mathbb{E}_{a^t \sim \mathcal{A}(a_1^t, \ldots, a_n^t)}[\mathbb{E}_{\mathcal{D}_{i^*,a^t}}[r_i^t]] = \sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} [\mathbb{E}_{\mathcal{D}_{i^*,a_i^t}}[r_i^t]] = \sum_{t=1}^{T} \frac{1}{n} \mathbb{E}_{\mathcal{D}_{i^*,a_{i^*}^t}}[r_{i^*}^t] + C$$

where C is independent of $a_{i^*}^t$. These C's therefore cancel out in the two terms of $\mathrm{Reg}_i^{\mathrm{agg}}$, and what remains is

$$\left(\max_{j \in [K]} \sum_{t=1}^{T} \frac{1}{n} \mathbb{E}_{\mathcal{D}_{i^*,j}}[r_{i^*}^t] \right) - \left(\sum_{t=1}^{T} \frac{1}{n} \mathbb{E}_{\mathcal{D}_{i^*,j^*}}[r_{i^*}^t] \right) = 0$$

by definition of j^*. Thus, we have zero regret in this case for every agent.

□

4 "Greedy is Fair" under High Reward Correlation

Intuitively, if agents act greedily to maximize their own reward, there are obvious cases where the greedy choice is unfair. For example, if arm i represents allocating a limited beneficial resource to agent i, then pulling arm i necessarily favors agent i over all other agents. In some problems, though, some decisions should benefit an entire group of people.

Thus in this section, we study the relationship between optimal-greedy and fair solutions in the high reward correlation settings of full-correlation and i.i.d. reward. First, we show that in the full-correlation setting, the multi-agent voting bandits are equivalent to single-agent bandits. This is because the agents all have identical information, and so the best strategy is equivalent to the best strategy for the single-agent case. Then, to go beyond the equivalence, we relax fully correlated rewards to i.i.d. rewards and show that SW is equivalent to the more fair group metric NSW.

Theorem 4.1. *In the full correlation setting, the best possible voting bandit strategy with a random sample aggregator $\mathcal{A}$ incurs regret $\mathrm{Reg}_n^{\mathrm{SW}} = K\mathrm{Reg}_n^{Bandit}$ where Reg_n^{Bandit} is the lowest possible regret of a single-agent bandit.*

A random sample aggregator randomly selects one of the agents' votes.

Proof. First, we show that $\mathrm{Reg}_n^{\mathrm{SW}} \geq \mathrm{Reg}_n^{\mathrm{Bandit}}$. Let $\mathcal{C}$ be the optimal K-agent voting bandits strategy in the full correlation model that achieve SW regret $\mathrm{Reg}_n^{\mathrm{SW}}$. We devise a single-agent bandit strategy as follows. We can simulate $\mathcal{C}$ since all agents incur identical reward. Specifically, for each round t, we simulate the vote of each arm, then simulate the aggregator. If arm a_t is picked, and arm reward r_t is sampled, then we send all simulated K-agents the reward r_t. This is identical to the full-correlation setting, but total acrues at a $1/K$ rate since we compare single to K agents. So this single-agent bandit strategy incurs regret Reg_n/K. Therefore, $\mathrm{Reg}_n^{\mathrm{SW}} \geq K\mathrm{Reg}_n^{\mathrm{Bandit}}$.

Now we show $\mathrm{Reg}_n^{\mathrm{SW}} \leq K\mathrm{Reg}_n^{\mathrm{Bandit}}$. Let $\mathcal{B}$ be the optimal single-agent bandits strategy the achieves regret $\mathrm{Reg}_n^{\mathrm{Bandit}}$. We construct a K-agent voting bandits strategy where all K agents take on strategy $\mathcal{B}$, and the aggregator copies the first agent's vote. Then notice the individual regret of agent 1 (w.r.t. self and w.r.t. the aggregator) is equal to $\mathrm{Reg}_n^{\mathrm{Bandit}}$. Since all rewards and reward distributions across agents are identical, and since we are technically talking about expected regret, all other agent's regret is $\mathrm{Reg}_n^{\mathrm{Bandit}}$. Therefore, the SW regret is $K\mathrm{Reg}_n^{\mathrm{Bandit}}$. Then, $\mathrm{Reg}_n^{\mathrm{SW}} \leq K\mathrm{Reg}_n^{\mathrm{Bandit}}$, and the claim is shown. □

Now we relax to the i.i.d. setting, where all agents have the same reward distributions, but the rewards are sampled i.i.d. for each agent. In this way, when agents' preferences are sufficiently similar, then we hope that we can find solutions that both maximize the total welfare while being fair. We go further and show that SW and NSW are equivalent in minima and maxima as they are directly correlated.

Theorem 4.2. *In the i.i.d. setting,* $\mathrm{SW}(p)$ *and* $\mathrm{NSW}(p)$ *are directly correlated. Therefore, maximizing* SW *maximizes* NSW*, and vice versa.*

Proof. Fix p and define $\mu(p) := \mathbb{E}_{a^t \sim p, r \sim D_{a^t}}[r]$. Since all agents have identical reward distributions and use the same policy p, we have $\mathbb{E}[r_i^t] = \mu(p)$ for each i, t. So, we have

$$\mathrm{SW}(p) = \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{E}[r_i^t] = NT\mu(p)$$

and

$$\mathrm{NSW}(p) = \prod_{i=1}^{N} \sum_{t=1}^{T} \mathbb{E}[r_i^t] = (T\mu(p))^N = T^N(\mu(p))^N$$

Solving for $\mu(p)$, we obtain $\mu(p) = \frac{1}{T}\left(\mathrm{NSW}(p)\right)^{\frac{1}{N}}$, which when substituted into $\mathrm{SW}(p)$, we get

$$\mathrm{SW}(p) = NT\mu(p) = NT\left(\frac{1}{T}\left(\mathrm{NSW}(p)\right)^{\frac{1}{N}}\right) = N\left(\mathrm{NSW}(p)\right)^{\frac{1}{N}}$$

Thus, $\mathrm{SW}(p) = N(\mathrm{NSW}(p))^{\frac{1}{N}}$. Multiplication by N and raising to the power of $\frac{1}{N}$ are monotonically increasing functions on $\mathbb{R}_{\geq 0}$, and both SW and NSW are non-negative since reward is always non-negative. Therefore SW and NSW are directly correlated, and p maximizes SW $\iff$ p maximizes NSW.

□

5 Explore-First: A Voting Bandits Strategy

Let us return to the bandit setting, without making any assumptions about the correlation between agents. What are some basic voting bandit strategies that will maximize SW or NSW? We first turn to the standard approaches for the single-agent setting: explore-then-exploit. This strategy has two phases: exploration and exploitation. During exploration, the aggregator tries every arm a uniform number of times. Each agent computes which arm empirically maximizes its reward. Then, during the exploitation phase, each agent votes for its maximum empirical mean arm, and the aggregator randomly samples a vote for the final decision.

Strategy: Explore-First.
Explore phase.

- For the first cK time steps, $\mathcal{A}$ picks each of the K arms c times. That is, for rounds $(j-1)c+1$ through jc, $\mathcal{A}$ pulls arm j.
- Each agent i keeps track of the empirical mean $\hat{\mu}_{ij} := \frac{1}{c}\sum_{i=(j-1)c+1}^{jc} r_i^t$ where r_i^t is drawn i.i.d. from $\mathcal{D}_{i,j}$ for all $j \in [K]$.
- Each agent i computes the optimal arm (breaking ties randomly), $\hat{a}_i = \text{argmax}_{j\in[K]}\, \hat{\mu}_{ij}$.

Exploit phase.

- For all the remaining rounds, $t = cK+1$ through T, each agent i votes for arm $\hat{a}_i$.
- The aggregator $\mathcal{A}$ randomly selects $i_t \sim [N]$ and pulls arm $\hat{a}_{i_t}$.

Goal. We analyze how our algorithm performs for the individual regret of the agents with respect to both self and the aggregator. First, we show that the agents have good estimates of the means of each of their arms at the end of the exploration phase using Hoeffding's inequality. Second, we show that the agent's voted arm during the exploitation phases is close to optimal via triangle inequality. Third, we convert our high probability expressions into expectations, so we can analyze expected regret. Finally, we choose parameters carefully and show that the agents' regret w.r.t. aggregator is sublinear in K and T. On the negative side, we show a example that shows regret w.r.t. self is linear in T.

Notation. Let a_i^* be the optimal arm for agent i, and μ_i^* be its corresponding mean reward. Recall that $\hat{a}_i$ is the arm that agent i actually votes for, and $\hat{\mu}_{ij}$ is the empirical mean reward of agent i for arm j. Recall for arm j and agent i, we denote the true mean reward as μ_{ij}.

Lemma 5.1 (Concentration of Means). *For any $d > 0$ and $\delta > 0$, if $\frac{2\log(2NK/\delta)}{d^2} \leq c \leq \frac{T}{K}$, then with probability at least $1-\delta$, for all $i \in [N]$ and $j \in [K]$, it holds that $|\hat{\mu}_{ij} - \mu_{ij}| \leq \frac{d}{2}$.*

Proof. Fix i and j. $\hat{\mu}_{i,j}$ is the average of c independent draws from $\mathcal{D}_{i,j}$, and each draw is bounded between 0 and 1. Then, Hoeffding's Inequality implies

$$\mathbb{P}\left(|\hat{\mu}_{i,j} - \mu_{i,j}| > \frac{d}{2}\right) \leq 2\exp\left(-\frac{d^2 c}{2}\right) \leq \frac{\delta}{NK},$$

and the latter inequality follows by choice of c. By union bound over all $i \in [N]$ and $j \in [K]$, we have

$$\mathbb{P}\left(\exists i \in [N], j \in [K].\ |\hat{\mu}_{ij} - \mu_{ij}| > \frac{d}{2}\right) \leq \delta.$$

Finally, we take the complement and loosen the inequality from $>$ to $\geq$

$$\mathbb{P}\left(\forall i \in [N], j \in [K].\ |\hat{\mu}_{ij} - \mu_{ij}| \leq \frac{d}{2}\right) \geq 1 - \delta.$$

□

Now, using that the means are sufficiently concentrated, we can show that the regret incurred from agent i by choosing (and playing) arm $\hat{a}_i$ over the optimal arm a_i^* is sufficiently small.

Lemma 5.2 (Approximate Optimality of Votes). *For any $d > 0$ and $\delta > 0$, if $\frac{2\log(2NK/\delta)}{d^2} \leq c \leq \frac{T}{K}$, then with probability at least $1 - \delta$, it holds that for all $i \in [N]$, $\mu_i^* - \mu_{i,\hat{a}_i} \leq d$.*

Proof. Notice that for all $i \in [N]$:

$$\mu_i^* - \mu_{i,\hat{a}_i} = (\mu_{i,a_i^*} - \hat{\mu}_{i,a_i^*}) + (\hat{\mu}_{i,a_i^*} - \hat{\mu}_{i,\hat{a}_i}) + (\hat{\mu}_{i,\hat{a}_i} - \mu_{i,\hat{a}_i})$$

Since $\hat{a}_i$ is chosen to achieve the maximum mean reward, $\hat{\mu}_{i,a_i^*} - \hat{\mu}_{i,\hat{a}_i} \leq 0$. And by choice of c, we can apply Lemma 5.1 with probability $1 - \delta$ to get $\mu_{i,a_i^*} - \hat{\mu}_{i,a_i^*} \leq \frac{d}{2}$ and $\hat{\mu}_{i,\hat{a}_i} - \mu_{i,\hat{a}_i} \leq \frac{d}{2}$ for all $i \in [N]$. Thus. by triangle inequality, $\mu_i^* - \mu_{i,\hat{a}_i} \leq \frac{d}{2} + 0 + \frac{d}{2} = d$. □

Now, we turn $1 - \delta$ probability into expectation by using the assumption that reward distributions are positive and bounded by 1.

Lemma 5.3 (Expected Optimality of Votes). *For any $d > 0$ and $\delta > 0$, if $\frac{2\log(2NK/\delta)}{d^2} \leq c \leq \frac{T}{K}$, then for all $i \in [N]$, $\mathbb{E}[\mu_i^* - \mu_{i,\hat{a}_i}] \leq d + \delta$.*

Proof. Let $\mathcal{E}$ be the event where for all $i \in [N]$, $\mu_i^* - \mu_{i,\hat{a}_i} \leq d$. We know by Lemma 5.2 this holds with probability $1 - \delta$. For the other δ probability, we note that $\mu_i^* - \mu_{i,\hat{a}_i} \leq 1$ since rewards are bounded in $[0, 1]$. Thus,

$$\begin{aligned}
\mathbb{E}[\mu_i^* - \mu_{i,\hat{a}_i}] &= \mathbb{E}[\mu_i^* - \mu_{i,\hat{a}_i} \mid \mathcal{E}]\,\mathbb{P}(\mathcal{E}) + \mathbb{E}[\mu_i^* - \mu_{i,\hat{a}_i} \mid \mathcal{E}^c]\,\mathbb{P}(\mathcal{E}^c) && \text{(law of total expectation)} \\
&\leq \mathbb{E}[\Delta_{i,\hat{a}_i} \mid \mathcal{E}]\,\mathbb{P}(\mathcal{E}) + \mathbb{P}(\mathcal{E}^c) && (\mu_i^* - \mu_{i,\hat{a}_i} \leq 1) \\
&\leq d \cdot \mathbb{P}(\mathcal{E}) + \mathbb{P}(\mathcal{E}^c) && \text{(definition of } \mathcal{E}) \\
&\leq d + \mathbb{P}(\mathcal{E}^c) && (\mathbb{P}(\mathcal{E}) \leq 1) \\
&\leq d + \delta && \text{(Lemma 5.2)}
\end{aligned}$$

□

Now we analyze the regret w.r.t. aggregator and show it achieves sublinear regret w.r.t. T and K. Note when c is sufficiently large, regret w.r.t aggregator is equal to 0. However, this is a degenerate case since the agents have no autonomy as the exploitation phase vanishes.

Theorem 5.4. *By choosing* $c = \frac{T}{K}$, $\mathrm{Reg}_i^{\mathrm{agg}} = 0$. *Otherwise when* $\alpha\frac{T}{K} < c < \frac{T}{K}$ *for any constant* $\alpha(0,1)$, $\mathrm{Reg}_i^{\mathrm{agg}} = \widetilde{O}(\frac{\sqrt{KT}}{N})$.

Proof. During the exploration phase, since $\mathcal{A}$ disregards the vote of each agent during this phase, the regret w.r.t. aggregator is 0. When $c = \frac{T}{K}$, then all T rounds are exploration, implying $\mathrm{Reg}_i^{\mathrm{agg}} = 0$. Our analysis proceeds assuming $c < \frac{T}{K}$.

For the exploitation phase, agent i only affects $\mathcal{A}$ with $\frac{1}{N}$ probability, in which case, $\mathcal{A}$ chooses agent i's vote of $\hat{a}_i$. We have:

$$\mathrm{Reg}_i^{\mathrm{agg}} = \mathbb{E}_G\left[\frac{T-cK}{N}\mu_i^* - \frac{T-cK}{N}\mu_{i,\hat{a}_i}\right] = \mathbb{E}_G\left[\frac{T-cK}{N}(\mu_i^* - \mu_{i,\hat{a}_i})\right]$$

Recall that if $c > \frac{2\log(2NK/\delta)}{d^2}$ for some choice of $\delta > 0$ and $d > 0$, then Lemma 5.3 implies $\mathbb{E}_G[\mu_i^* - \mu_{i,\hat{a}_i}] \leq d + \delta$. So

$$\mathrm{Reg}_i^{\mathrm{agg}} \leq \frac{T-cK}{N}(d+\delta)$$

By choosing $\delta < \frac{N}{T}$, we have that for $c > \frac{2\log(2KT)}{d^2}$:

$$\mathrm{Reg}_i^{\mathrm{agg}} < \frac{T-cK}{N}d + \frac{T-cK}{T} < \frac{T-cK}{N}d + 1$$

Taking any constant $\alpha \in (0,1)$, if $\alpha\frac{T}{K} < c < \frac{T}{K}$, then the smallest value of d where $c > \frac{2\log(2KT)}{d^2}$ holds is $d = \sqrt{\frac{2K\log(2KT)}{\alpha T}}$. So we have

$$\begin{aligned}\mathrm{Reg}_i^{\mathrm{agg}} - 1 &< \frac{T-cK}{N}d \\ &= \frac{\sqrt{2KT\log(2KT)}}{\sqrt{\alpha}\cdot N} - \frac{cK\sqrt{2K\log(2KT)}}{\sqrt{\alpha T}\cdot N} \\ &= \left(\frac{1}{\sqrt{\alpha}} - \sqrt{\alpha}\right)\frac{\sqrt{2KT\log(2KT)}}{N} && (\alpha\tfrac{T}{K} \leq c)\end{aligned}$$

Thus, $\mathrm{Reg}_i^{\mathrm{agg}} \leq \widetilde{O}(\frac{\sqrt{KT}}{N})$, as desired. □

Now we analyze regret w.r.t self, and show through counterexample that it is linear in the worst case. We present the worst case where a large portion of agents disagree on the best arms.

Theorem 5.5. *If* $N \geq 2$ *and* $K \geq 2$, $\mathrm{Reg}_i^{\mathrm{self}} = \Theta(T)$.

Proof. The upper bound $\mathrm{Reg}_i^{\mathrm{self}} = O(T)$ is trivial as rewards are bounded in $[0,1]$, so regret accumulates at most $O(T)$ for T rounds.

For the lower bound, we provide a tight example. Suppose W.L.O.G. $N \geq 2$ is even. Let half the agents have constant reward distributions where arm 1 gives reward 1, and all other arms give reward 0. Let the other half of the agents have constant reward distributions where arm 2 gives reward 1, and all other arms give reward 0.

Take any agent $i \in [N]$. We analyze $\mathrm{Reg}_i^{\mathrm{self}}$. During the cK rounds of exploration, each of the K arms is explored c times. Since agent i gets reward 1 from a single arm, and 0 from all others, it accumulates regret $c(K-1)$ during exploration. Meanwhile during exploitation, half the agents vote for arm 1 while the other half vote for arm 2. Therefore in expectation, agent i incurs regret $1/2$ per round of exploitation, in which there are $T - cK$ total rounds. Therefore, we have

$$\begin{aligned} \mathrm{Reg}_i^{\mathrm{self}} &= c(K-1) + \frac{T - cK}{2} \\ &\geq \frac{cK}{2} + \frac{T - cK}{2} = \frac{T}{2} && (K - 1 \geq K/2) \end{aligned}$$

Therefore, $\mathrm{Reg}_i^{\mathrm{self}} \geq \Omega(T)$. So, $\mathrm{Reg}_i^{\mathrm{self}} = \Theta(T)$. □

6 Extending Epsilon-Greedy Under Similarity

6.1 Notation and Definitions

Setting. We consider the voting bandits setting with $K = 2$ arms, $N = 2$ agents, and T rounds. Recall at each round $t \in [T]$, each agent i submits a vote a_i^t, which is the arm it wants to select. Then, an aggregator algorithm $\mathcal{A} : [T]\times[K]^2 \to \Delta([K])$ chooses an arm A_t based on the agents votes, and both agents pull A_t. When arm A_t is selected, agent i receives a stochastic reward $r_i^t \sim \mathcal{D}^i_{A_t}$, where $\mathcal{D}^i_{A_t} \in \Delta([0, 1])$ is agent i's reward distribution for arm j. Rewards are independent across agents and i.i.d. over rounds for a fixed arm.

We assume we are in the ϵ-similarity reward correlation setting, where the means of the agents' reward distributions are ϵ-close, meaning $|\mu_j^1 - \mu_j^2| \le \epsilon$ for each arm j.

Strategy: Epsilon-Greedy. We extend the Epsilon-Greedy algorithm as follows. Roughly, the agents act greedily, voting for the arm that empirically maximizes their reward so far. Then the aggregator injects a bit of random noise by either randomly sampling a vote or, with small probability, uniformly selecting an arm $j \in \{1, 2\}$.

The agents' policy is as follows: each agent maintains empirical estimates of its expected reward for each arm using the observed history, and at each round, votes for the arm with the highest empirical mean. Ties are broken arbitrarily. Formally, each agent i maintains empirical means based on the public history $\hat{\mu}^i_{j,t} = \frac{1}{N_j(t)} \sum_{s:A_s=j,s<t} r_s^{i*}$ where $N_j(t) = \sum_{s=1}^{t-1} \mathbf{1}(A_s = j)$ is the number of times arm j has been pulled before time t. When $N_j(t) = 0$, we set $\hat{\mu}^i_{j,t} = 1$ Then, agent i votes greedily for the arm with highest empirical mean, which is $a_t^i \in \arg\max_{j\in\{1,2\}} \hat{\mu}^i_{j,t}$.

The aggregator $\mathcal{A}$ works as follows: at time t, selects $A_t \sim \mathrm{Unif}(\{1, 2\})$ with probability $\epsilon_{\text{greedy}}(t) = \min\{1, 4c/t\}$, and exploits $A_t \sim \mathrm{Unif}(\{a_t^1, a_t^2\})$ otherwise. Importantly, note that the aggregator has an exploration parameter $c > 0$, which controls the exploration probability.

Goal. We analyze the individual regret of the agents w.r.t. self in this ϵ-similarity setting, and show it is possible to achieve logarithmic regret. Our analysis starts by decomposing regret into exploration and exploitation phases. We first bound the expected number of exploration rounds, which gives a logarithmic contribution to regret. Next, we define a high-probability concentration event under which all empirical means remain close to their true values. We condition on this event and show that an agent can vote for a suboptimal arm only when its reward gap is below a threshold. We then bound the number of times the suboptimal arm is selected during exploitation. Finally, we convert the high-probability bounds into expectations, yielding a logarithmic upper bound on regret for all agents.

6.2 Analysis

In the proofs, we define the **gap for agent** i to be $\Delta^i = \mu^i_{j_i^*} - \mu^i_{j_{-i}}$, where j_{-i} denotes the suboptimal arm. Assuming $\Delta^1 = \mu_1^1 - \mu_2^1$, we call $\epsilon < \frac{\Delta^1}{2}$ the **aligned case** and $\epsilon \ge \frac{\Delta^1}{2}$ the **misaligned case**.

We are now ready to prove results about this setting.

Lemma 6.1. *If $\epsilon < \frac{\Delta^1}{2}$ where $\Delta^1 = \mu_1^1 - \mu_2^1 > 0$, Then both agents have the same optimal arm, namely arm 1.*

Proof sketch

We know that $\Delta^1 = \mu_1^1 - \mu_2^1 > 0$. Therefore, agent 1 prefers arm 1. By the similarity constraint, for each arm $j \in \{1, 2\}$, we have

$$|\mu_j^1 - \mu_j^2| \leq \epsilon$$

which implies that

$$\mu_1^2 \geq \mu_1^1 - \epsilon, \mu_2^2 \leq \mu_2^1 + \epsilon$$

From this, it follows that

$$\Delta^2 = \mu_1^2 - \mu_2^2 \geq (\mu_1^2 - \epsilon) - (\mu_2^1 + \epsilon) = \Delta^1 - 2\epsilon > 0$$

using the assumption that $\epsilon < \frac{\Delta^1}{2}$.

Therefore, agent 2 also prefers arm 1, so both agents share the same optimal arm.

Full proof in Appendix B

Lemma 6.2. *Define the event $\mathcal{E}$ as follows*

$$\mathcal{E} = \left\{ \forall i \in \{1, 2\},\ \forall j \in \{1, 2\},\ \forall t \in [T] : \left|\hat{\mu}_{j,t}^i - \mu_j^i\right| \leq \sqrt{\frac{2\log(4T)}{N_j(t)}} \right\}.$$

Then $\mathbb{P}(\mathcal{E}) \geq 1 - \frac{2}{T}$

This lemma essentially says that, with high probability, all empirical mean estimates remain close to their true values throughout the entire horizon. Specifically, the estimation error for each arm-agent pair is bounded by $\sqrt{\frac{2\log(4T)}{N_j(t)}}$, which decreases as arm j is pulled more frequently.

This result serves as the foundation for analyzing the agents' voting behavior and will ultimately allow us to derive a logarithmic regret bound.

Proof sketch

For any fixed agent i, arm j, and time t, Hoeffding's inequality tells us that the empirical mean $\hat{\mu}_{j,t}^i$ deviates from its true mean μ_j^i by more than

$$\sqrt{\frac{2\log(4T)}{N_j(t)}}$$

with probability bounded above by $O(T^{-2})$. Taking a union bound over all agents, arms, and times, we obtain that with probability at least $1 - \frac{2}{T}$, none of these deviations occur. On this event, all empirical means remain close to their true values.

Full proof in Appendix B.1

Lemma 6.3 (Condition for voting for suboptimal arm)**.** *On event $\mathcal{E}$, if agent i votes for arm* 2 *(the suboptmal arm) at time t, then, we have*

$$\Delta^i \leq \sqrt{\frac{2\log(4T)}{N_2(t)}} + \sqrt{\frac{2\log(4T)}{N_1(t)}}$$

This lemma captures a key insight: if an agent is voting for arm 2 (the suboptimal arm), it must be because the empirical estimates are ”fooling” the agent into thinking arm 2 is the better one. However, event $\mathcal{E}$ bounds how much the estimates can deviate from the truth. Therefore, if the agent votes for arm 2, the gap Δ^i cannot be too large; otherwise, even with estimation errors, the empirical mean of arm 1 would dominate the noise.

As we collect more samples and $N_1(t), N_2(t) \to \infty$, the R.H.S. of the inequality shrinks to 0. This suggests that once we have pulled each arm sufficiently many times, the Δ^i will exceed the estimation error bound $|\hat{\mu}^i_{j,t} - \mu^i_j| \leq ....$, making it impossible for agents to vote for arm 2. This is the foundation for the next lemma, which will formalize when agents stop voting for the suboptimal arm.

Proof sketch

On the high-probability event $\mathcal{E}$, all empirical means are close to their true means. If agent i votes for arm 2 at time t, then by greedy voting we must have

$$\hat{\mu^{i}}_{2,t} \geq \hat{\mu}^i_{1,t}$$

Using the concentration bounds from $\mathcal{E}$ to upper bound $\hat{\mu}^i_{2,t}$ and lower bound $\hat{\mu}^i_{1,t}$, this inequality implies that the true gap Δ^i is at most the sum of the confidence radii for arms 1 and 2.

Thus, voting for the suboptimal arm can only occur when the gap is smaller than the current estimation error.

Full proof in Appendix B.3

Lemma 6.4 (Agents stop voting for suboptimal arm after sufficient samples)**.** *On event $\mathcal{E}$, once $N_1(t), N_2(t) \geq m$, where*

$$m = \left\lceil \frac{8\log(4T)}{(\Delta^1 - 2\epsilon)^2} \right\rceil$$

Both agents vote exclusively for arm 1. WLOG assume that $\frac{8\log(4T)}{(\Delta^1-2\epsilon)^2} \notin \mathbb{Z}$ which implies that $m > \frac{8\log(4T)}{(\Delta^1-2\epsilon)^2}$. This is a practical assumption to make because, in practice, $\frac{8\log(4T)}{(\Delta^1-2\epsilon)^2}$ won't be an integer.

Proof sketch

We condition on $\mathcal{E}$. If an agent i ever votes for suboptimal arm 2 at time t by Lemma 6.3, we must have

$$\Delta^i \leq \sqrt{\frac{2\log(4T)}{N_2(t)}} + \sqrt{\frac{2\log(4T)}{N_1(t)}}$$

If additionally $N_1(t), N_2(t) \geq m$, the right hand side is at most

$$2\sqrt{\frac{2\log(4T)}{m}}$$

By the similarity constraint and Lemma B.4, both agents' gaps satisfy

$$\Delta^i \geq \Delta^1 - 2\epsilon$$

Choosing

$$m = \left\lceil \frac{8\log(4T)}{(\Delta^1 - 2\epsilon)^2} \right\rceil$$

makes it such that

$$2\sqrt{\frac{2\log(4T)}{m}} < \Delta^1 - 2\epsilon \leq \Delta^i$$

Therefore, the voting condition cannot hold. This means once each arm has been sampled at least m times and $\mathcal{E}$ holds, no agent will vote for arm 2, so both vote exclusively for arm 1 from that point forward.

Full proof in Appendix B.4

Theorem 6.5. *Suppose that $\epsilon < \frac{\Delta^1}{2}$, where $\Delta^1 = \mu_1^1 - \mu_2^1 > 0$. Further suppose that $c > 0$ is a constant. Then for both agents $i \in \{1, 2\}$, we have the regret bound*

$$\mathbb{E}[\mathcal{R}_i(T)] \;\leq\; O\left(c\log T + \frac{\log T}{(\Delta^1 - 2\epsilon)^2}\right).$$

Proof sketch

We decompose the regret into exploration and exploitation. Exploration occurs with probability $\min(1, \frac{4c}{t})$. We use an integral upper bound of the form $\int_a^b \frac{dx}{x}$ to show that the expected number of exploration rounds is $O(c\log T)$. Since per-round regret is at most 1, this yields

$$\mathbb{E}\left[\mathcal{R}_i^{\text{explore}}\right] = O(c\log T)$$

For exploitation, we use union bound and Hoeffding's inequality to show that

$$\mathbb{P}(\mathcal{E}) \geq 1 - \frac{2}{T}$$

On $\mathcal{E}$, an agent can vote for the suboptimal arm only when Δ^i is smaller than the sum of confidence radii. Choosing

$$m = \left\lceil \frac{8\log(4T)}{(\Delta^1 - 2\epsilon)^2} \right\rceil$$

ensures that once both arms have been pulled at least m times, this condition can no longer hold, so both agents vote only for the optimal arm after that.

Therefore, conditioned on $\mathcal{E}$, the suboptimal arm can be chosen during exploitation at most m times, giving

$$\mathbb{E}[\mathcal{R}_i^{\text{exploit}}(T)|\mathcal{E}] = O\left(\frac{\log T}{(\Delta^1 - 2\epsilon)^2}\right)$$

Finally, since rewards are bounded in $[0,1]$, the regret incurred on the failure event $\mathcal{E}^C$ is at most T. This is because it can be shown that

$$\mathbb{P}\left(\mathcal{E}^C\right) = O\left(\frac{1}{T}\right)$$

therefore, its contribution to the expected regret is $O(1)$, which doesn't change the asymptotic bound.

Adding the two regrets, we have

$$\mathbb{E}[\mathcal{R}_i(T)] \;\le\; O\Bigg(c\log T + \frac{\log T}{(\Delta^1 - 2\epsilon)^2}\Bigg).$$

for both agents $i \in \{1,2\}$.

Full proof in Appendix B.5

6.3 Experimental Results

Below we display empirical results for our epsilon greedy algorithm in the cases where both agents have the same optimal arm and where both agents have differing optimal arms:

Experiment 1:

In this experiment, we consider the following reward structure: for agent 1, arm 1 is Unif$(0.7, 0.8)$ and arm 2 is Unif$(0, 0.2)$. For agent 2, arm 1 is Unif$(0.5, 0.7)$ and arm 2 is Unif$(0, 0.1)$.

The mean rewards are

$$\mu_1^1 = 0.75,\ \ \mu_2^1 = 0.1,\ \ \mu_1^2 = 0.6,\ \ \mu_2^2 = 0.05$$

Thus, we have

$$\Delta^1 = \mu_1^1 - \mu_2^1 = 0.75 - 0.1 = 0.65 > 0$$

so, arm 1 is optimal for agent 1. Also, arm 1 is optimal for agent 2 because $\mu_1^2 > \mu_2^2$.

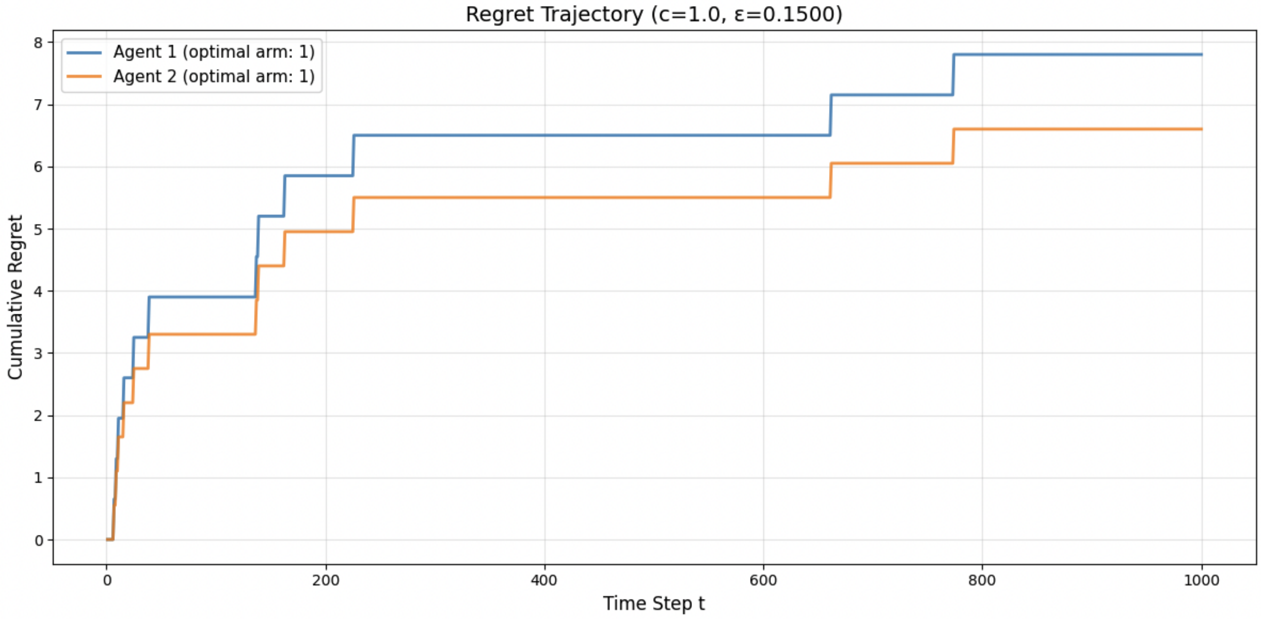
The similarity parameter is

$$\epsilon = \max\{|\mu_1^1 - \mu_1^2|, |\mu_2^1 - \mu_2^2|\} = \max\{0.15, 0.05\} = 0.15$$

Thus, we have

$$\epsilon < \frac{\Delta^1}{2}$$

so the preconditions of Theorem 6.5 are satisfied. Thus, we observe logarithmic cumulative regret in the plot below.



Experiment 2:

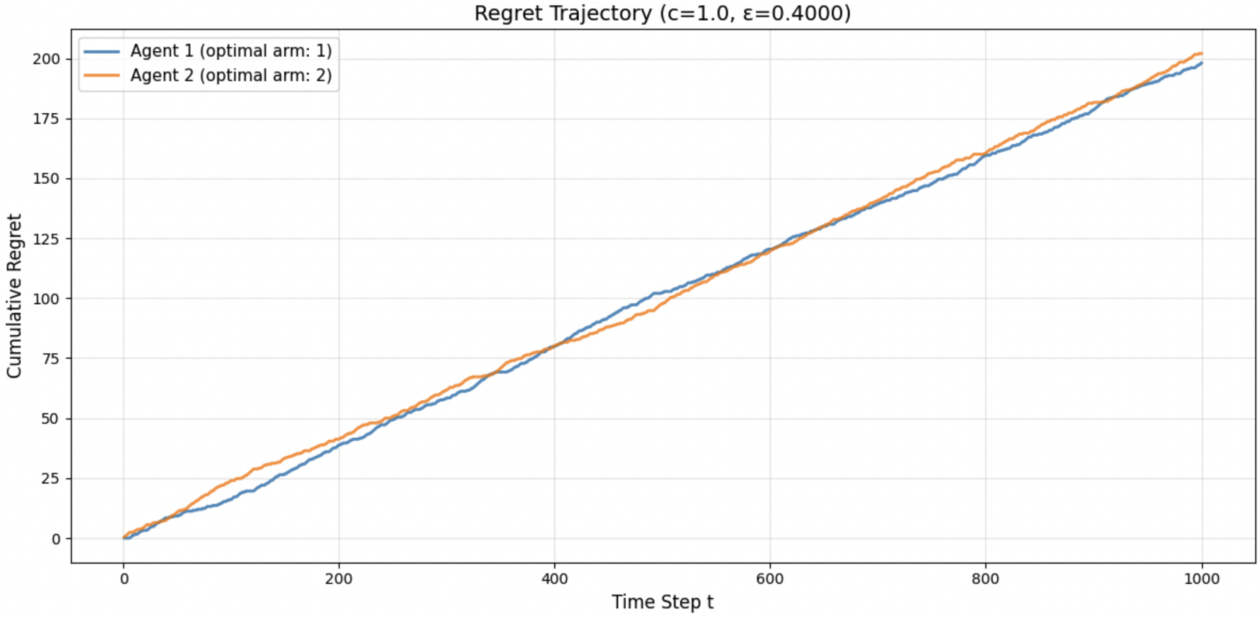
In this experiment, we consider the following reward structure: for agent 1, arm 1 is Unif$(0.6, 0.8)$ and arm 2 is Unif$(0.2, 0.4)$. For agent 2, arm 1 is Unif$(0.2, 0.4)$ and arm 2 is Unif$(0.6, 0.8)$. Here, the agents 1's optimal arm is 2 while agent 2's optimal arm is 1.

$$\Delta^1 = \mu_1^1 - \mu_2^1 = 0.7 - 0.3 = 0.4 > 0$$

$$\implies \frac{\Delta^1}{2} = 0.2$$

$$\epsilon = \max\{|\mu_1^1 - \mu_1^2|, |\mu_2^1 - \mu_2^2|\} = \max\{|0.7 - 0.3|, |0.3 - 0.7|\} = 0.4$$

Therefore, $\epsilon < \frac{\Delta^1}{2}$ does not hold. Thus, we observe linear cumulative regret in the plot below.



7 Random Sampling Doesn't Maximize Nash Social Welfare

Now we present our impossibility results. One would hope that if each agent votes for their selfishly optimal arm, then there exists an aggregator that can maximize SW, however this is inherently unfair because maximizing SW can lead to the *tyranny of the majority* [Moulin, 2004].

Example. Consider $N = 2$ agents with $K = 2$ arms, when agent 1 and agent 2 receive constant rewards $(0.9, 0)$ and $(0, 0.5)$ from the arms, respectively. Then no matter what the votes are, the best aggregator will select arm 1 because the SW is maximized at $0.9T$. However, agent 2 receives no reward. In fact, with respect to NSW, this strategy achieves the worst possible score of 0.

As a compromise, we might hope that the most fair aggregators are uniformly random sampling from the votes it gets. Then each agent will achieve equal influence compared to every other agent. In fact, this aggregator maximizes NSW in the example above.

We will show that each agent minimizing individual regret and random sampling aggregation is not enough to optimize NSW. The idea is that minimizing individual regret approaches a Nash optimum, which can differ than the max NSW.

Theorem 7.1. *There exists a voting bandits instance with $N = 2$ agents and $K = 3$ arms such that, under random-sampling aggregation, agents achieve zero regret w.r.t. the aggregator, but NSW is not maximized.*

Proof. Let agent 1 have constant reward distribution $(1, \frac{2}{3}, 0)$ and agent 2 have constant reward distribution $(0, \frac{2}{3}, 1)$. To achieve 0 regret w.r.t. aggregator, each agent will eventually, mostly vote for their optimal arm. W.L.O.G. let, agent 1 always voting for arm 1 and agent 2 always vote for arm 3. Under random sampling, $\mathcal{A}$ chooses arms according to the distribution $(\frac{1}{2}, 0, \frac{1}{2})$. This achieves NSW $\frac{1}{4}T^2$.

However, to maximize NSW, both agents should always vote for arm 2 so that $\mathcal{A}$ chooses the policy $(0, 1, 0)$, thus achieving NSW $\frac{4}{9}T^2$. □

8 Price of Anarchy is Linear in Worst-Case

We have seen that through statistical learning algorithms for our aggregator where each agent votes for their empirically best arm, we can generally achieve sublinear individual regret w.r.t. aggregator. In this section, we analyze the potential drawbacks of this greedy behavior by observing that, in the worst case, social welfare can decrease by a factor of n when all agents are playing to minimize their own regret. Specifically, for simplicity, we analyze the setting in which the aggregator randomly samples over agents' voted arms and show that if each agent is to achieve sublinear asymptotic regret, there exists a "prisoners dilemma-esque" case in which the social welfare is n times worse than the social optimal. This ultimately shows that minimizes individual regret does not necessarily lead to net good outcomes for all players.

Theorem 8.1 (Price of Anarchy is $O(N)$)**.** *There exists a multi-agent multi-armed bandits game with N agents and $K = N + 1$ arms, for which the price of anarchy is $O(N)$ between a social optimal aggregator and an instance in which all agents achieve sublinear individual regret w.r.t. aggregator. Further, the social welfare regret in these instances is $O(NT)$.*

Proof. For example, consider a case with n agents and $K = n + 1$ arms. For each agent i,

$$\mathcal{D}_{i,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \text{ and } j \in [n] \\ 0.99 & \text{if } j = n + 1 \end{cases}$$

For simplification, we assume there is no variance on these distributions, so they are all point masses. We note that if any agent i picks any arm, that is not arm i, $O(T)$ times, they will achieve at least $(1 - 0.99)O(T) = O(T)$ regret, which is not sublinear. Therefore, every agent i must pick arms that are not arm i a sublinear amount of times $o(T)$ by assumption.

From this we can derive that arm $n+1$ will be picked $o(T)$ times and, since the social welfare from all other arms is 1, the social welfare of this instance will be $O(T)$ and $o(NT)$.

In contrast, the optimal social welfare is to have the aggregator pick arm $n+1$ every time resulting in a social welfare of $0.99NT$. From this, we get that the social welfare regret is $O(NT)$ and the price of anarchy is $O(N)$, as needed.

□

9 Conclusion

9.1 Summary of Results

We studied a multi-agent stochastic bandit setting in which agents vote on arms, and a central aggregator decides which arms are pulled in each round—a novel setting we call voting bandits. We characterize voting bandits across different informational models, construct generalizations of standard bandit strategies, and provide counterexamples showing the impossibility of maximizing NSW or SW under rationality assumptions.

Of note, when agents know their reward distributions, a random-sampling aggregator allows each agent to achieve zero regret w.r.t. the aggregator by always voting for its optimal arm. More generally, regret with respect to the aggregator reduces to single-agent bandit regret. When rewards are fully correlated, voting bandits reduce to single-agent bandits. And even when rewards are just i.i.d. correlated, our fairness notion NSW is equivalent to SW.

We analyzed Explore-First and ε-Greedy-style strategies under bandit-style feedback of the voting aggregation system, showing that agents achieve individual sublinear regret w.r.t. the aggregator. Where the worst-case regret w.r.t. self is linear, assuming the agent's reward means are sufficiently similar, we show we can even achieve logarithmic regret w.r.t. self.

Finally, we showed that voting-based aggregation can lead to inefficiencies in SW. We constructed instances in which *rational* no-regret agents interacting through a random-sampling aggregator fail to maximize SW, let alone NSW, yielding a price of anarchy that is linear in the number of agents.

9.2 Future Directions

In the short term, there are several things to complete our analysis. For example, does Explore-First under ε-similarity also have logarithmic regret w.r.t. self? Can we generalize our ε-Greedy analysis for N agents and K arms? What happens if we generalize the UCB algorithm to our setting? Also pertinently, although under rationality assumptions we cannot maximize SW or NSW, how much SW or NSW can we achieve with our strategies?

Is ε-similarity the right reward correlation assumption? Our analysis holds under the assumption that ε is sufficiently small, since this implies that all agents share the same optimal arm. However, this is unreasonable in cases of dichotomous preferences.

Can we lower the price of anarchy by increasing the aggregator's influence or complexity? We would like to achieve zero-regret w.r.t. SW or NSW. The fact that random samplers and rational agents cannot achieve zero regret indicates an informational bottleneck in our voting model. Since our existing counterexamples arise from a discrepancy between minimizing individual and social regret, one way we hope to achieve better price-of-anarchy bounds is by imposing a penalty on selfish agent actions. By including a penalty when an agent picks an arm that differs from the arm the aggregator picked, we generally incentivize choosing an arm that maximizes social welfare, and analyzing how such penalties affect individual and social regret would be a promising direction for future work.

References

[Aziz et al., 2019] Aziz, H., Bogomolnaia, A., and Moulin, H. (2019). Fair mixing: the case of dichotomous preferences. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 753–781.

[Conitzer et al., 2017] Conitzer, V., Freeman, R., and Shah, N. (2017). Fair public decision making. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 629–646.

[Foster and Rakhlin, 2023] Foster, D. J. and Rakhlin, A. (2023). Foundations of reinforcement learning and interactive decision making. *arXiv preprint arXiv:2312.16730*.

[Harada et al., 2025] Harada, T., Ito, S., and Sumita, H. (2025). Bandit max-min fair allocation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 380–397. Springer.

[Hossain et al., 2021] Hossain, S., Micha, E., and Shah, N. (2021). Fair algorithms for multi-agent multi-armed bandits. *Advances in Neural Information Processing Systems*, 34:24005–24017.

[Joseph et al., 2016] Joseph, M., Kearns, M., Morgenstern, J. H., and Roth, A. (2016). Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems*, 29.

[Liu et al., 2017] Liu, Y., Radanovic, G., Dimitrakakis, C., Mandal, D., and Parkes, D. C. (2017). Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875*.

[Moulin, 2004] Moulin, H. (2004). *Fair division and collective welfare*. MIT press.

A Literature Survey

A.1 Foster & Rakhlin - Foundations of Reinforcement Learning

Since our project focuses on multi-agent bandits, we read the multi-armed bandits chapter of [Foster and Rakhlin, 2023].

The bandits chapter presents the classical stochastic multi-armed bandit framework, in which an agent repeatedly selects an arm $a_t \in [K]$ and receives a reward from it. The goal is to minimize pseudo-regret across time.

$$\text{Reg}(T) = T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} r_t\right]$$

The core problem here is balancing exploration and exploitation. Exploration means trying each arm enough times such that we have a good estimate of what its mean reward is, while exploitation means picking the most optimal arm a sufficient number of times. This chapter analyzes standard algorithms, including ϵ-greedy and upper confidence bound (UCB).

The ϵ-greedy algorithm balances exploration and exploitation by selecting the empirically best arm most of the time, while exploring uniformly at random with a small probability.

$$a_t \sim \begin{cases} \arg\max_{a \in [K]} \hat{\mu}_a(t); \text{with probability } 1 - \epsilon \\ \text{Unif}([K]); \text{with probability } \epsilon \end{cases}$$

where $\hat{\mu}_a(t)$ is the empirical mean reward of arm a up to round t and $\epsilon \in (0, 1)$ controls the rate of exploration.

The UCB algorithm is defined as follows: for each arm $a \in [A]$ and round t, define

$$\hat{f}_t(a) = \text{empirical mean reward of arm } a \text{ at time } t$$

$$n_t(a) = \text{the number of times arm } a \text{ has been selected before round } t$$

Then, we define

$$\overline{f_t}(a) = \hat{f}_t(a) + \sqrt{\frac{2\log\left(2T^2\frac{A}{\delta}\right)}{n_t(a)}}$$

$$\underline{f_t}(a) = \hat{f}_t(a) - \sqrt{\frac{2\log\left(2T^2\frac{A}{\delta}\right)}{n_t(a)}}$$

It is shown that, in the stochastic multi-armed bandit setting, it holds with probability at least $1 - \delta$ that

$$\text{Reg}(T) \lesssim \sqrt{AT\log\left(\frac{AT}{\delta}\right)}$$

And, the UCB algorithm selects the arm with the highest optimistic estimate.

$$a_t = \arg\max_{a \in A} \overline{f_t}(a)$$

where A is the number of arms. Their derivation uses confidence intervals based on Hoeffding-style concentration. This result captures the core exploration/exploitation trade-off and provides the theoretical foundation for the single-agent regret notions on which our Voting Bandit model is based.

A recurring proof technique in this chapter is to define a high-probability event under which the empirical means of all arms are close to their true values. Using Hoeffding's inequality and a union bound across arms and time, it is shown that with probability at least $1-\delta$, it holds that for all a, t,

$$|\hat{\mu}_a(t) - \mu_a| \leq \sqrt{\frac{2\log(2T^2\frac{A}{\delta})}{n_t(a)}}$$

Then, conditioning on this event, it is shown that suboptimal arms are pulled only when their confidence intervals are still large, and such events only happen at $O(\log(T))$ times. Outside this event, the regret is at most T, and the probability of this happening is at most δ. This decomposition into a high-probability event plus a low-probability failure is a core proof technique that we use in our own analysis.

A.2 Joseph et al. - Classic and Contextual Bandits

[Joseph et al., 2016] introduces a notion of individual fairness in bandit learning and examines how fairness constraints affect the achievable regret.

Each arm $j \in [k]$ has an unknown mean reward μ_j, and the learner selects arm i_t each round and observes a stochastic reward.

An algorithm is defined as **fair** when it assigns a higher probability to arm j than to arm j' only if $\mu_j > \mu_{j'}$. Formally, a bandit algorithm A is δ-fair when with probability at least $1-\delta$ for all histories h, round t, and arms j, j', it holds that

$$\pi_{t,j|h} > \pi_{t,j'|h} \implies \mu_j > \mu_{j'}$$

where h is the entire sequence of past actions and rewards up to time $t-1$ and $\pi_{t,j|h} \in [0,1]$ is the probability that the algorithm chooses arm j at round t, given that history. More formally, we have

$$\pi_{t,j|h} = \mathbb{P}(i_t = j|h)$$

This fairness constraint prevents the algorithm from favoring an arm unless it has sufficient evidence that the arm is actually better than others.

The authors introduce FAIRBANDITS, a δ-fair algorithm that, after $T = \Omega(k^3)$ rounds, achieves regret

$$\text{Reg}(T) = \tilde{O}\left(\sqrt{k^3 T}\right).$$

The authors also prove a matching lower bound, which states that every δ-fair algorithm must incur constant per-round regret for at least $\Omega\left(k^3 \log\left(\frac{1}{\delta}\right)\right)$.

Overall, this paper formalizes a strong fairness constraint on decision-making and shows that such fairness can dramatically slow the learning process. In contrast, our Voting Bandits framework studies multi-agent bandits with a different notion of fairness and welfare. Instead of requiring monotonicity in selection probabilities, we analyze how aggregated individual votes affect individual and social regret under partial control over the selected action.

A.3 Liu et al. Calibrated Fairness in Bandits

[Liu et al., 2017] introduce a fairness framework for stochastic multi-armed bandits built around two principles: smooth fairness (treating similar arms similarly) and calibrated fairness (sampling arms in proportion to their probability of being the best).

Smooth fairness requires that if two arms have similar reward distributions, then their selection probabilities must be similar. Formally, we say that for some divergence function D, a bandit algorithm is $(\epsilon_1, \epsilon_2, \delta)$-fair when with probability at least $1-\delta$, for all rounds t and all arms i, j, it holds that

$$D(\pi_t(i)||\pi_t(j)) \leq \epsilon_1 D(r_i||r_j) + \epsilon_2$$

where $\pi_t(i)$ is the probability of selecting arm i at time t, and $D(r_i||r_j)$ measures divergence between the reward distributions of arms i and j. Smooth fairness prevents an algorithm from heavily favoring one arm over another when they have nearly identical distributions.

Calibrated fairness strengthens this notion by requiring that an algorithm select each arm with probability equal to the probability that the arm's realized reward is the highest.

$$\pi_t(a) = P^*(a)$$

where

$$P^*(a) = \mathbb{P}\left(a = \arg\max_j r_j\right)$$

This notion of fairness ensures that empirically weaker arms still receive opportunities in proportion to how often they could still be the best.

The authors introduce Fair SD-TS, a Thompson sampling-based algorithm with an exploration phase. They show that this algorithm is $(2, 2\epsilon_2, \delta)$-fair and achieves fairness regret

$$R_{f,T} \leq \tilde{O}((kT)^{\frac{2}{3}})$$

Overall, this work provides a calibrated, distribution-aware approach to fairness in bandits and demonstrates that fairness constraints impose performance costs. Compared to our voting bandits framework, their definition of fairness centers on similarity across arms, whereas we study how aggregated multi-agent votes influence individual and social regret under partial control of the chosen action.

A.4 Hossain, Micha, and Shah - Fair Algorithms for Multi-Agent Multi-Armed Bandits

[Hossain et al., 2021] propose a multi-agent variant of the stochastic multi-armed bandit problem, where pulling an arm yields a possibly different stochastic reward for each of the N agents. Unlike the single-agent setting, there is no universally best arm in the general case, since each agent may prefer a different arm.

The authors define the Nash Social Welfare (NSW) as the objective

$$\mathrm{NSW}(p, \mu^*) = \prod_{i=1}^{N} \left(\sum_{j=1}^{K} p_j \mu^*_{i,j} \right)$$

For a policy $p \in \Delta_K$ that assigns probability p_j to arm j, the expected utility of agent i is

$$\sum_j p_j \mu^*_{i,j}$$

The goal in this problem is to learn a distribution over arms that maximizes the NSW. The authors define cumulative regret as

$$R_T = \sum_{t=1}^{T} (\mathrm{NSW}(p^*, \mu^*) - \mathrm{NSW}(p_t, \mu^*))$$

where p^* is a policy that maximizes NSW.

The authors generalize the Explore-First, ϵ-Greedy, and UCB algorithms by choosing distributions over arms instead of individual arms. The Explore-First algorithm first pulls each arm L times, then computes the empirical best NSW-maximizing distribution, and plays it for the remaining iterations. For this algorithm, the authors show regret bounds.

$$\mathbb{E}[R_T] = \tilde{O}\left(N^{\frac{2}{3}} K^{\frac{1}{3}} T^{\frac{2}{3}} \right)$$

and

$$\mathbb{E}[R_T] = \tilde{O}\left(N^{\frac{1}{3}} K^{\frac{2}{3}} T^{\frac{2}{3}} \right)$$

depending on the choice of L.

The most sophisticated algorithm in this paper is a multi-agent UCB variant, which augments the NSW estimate with optimism terms of the form $\alpha_t \sum_{j \in [K]} p_j r^t_j$, where

$$r^t_j = \sqrt{\frac{\log(NKT)}{n^t_j}}$$

is a confidence radius for arm j. They prove that UCB achieves $\tilde{O}(\sqrt{T})$ regret.

$$\mathbb{E}[R_T] = \tilde{O}(NKT^{\frac{1}{2}})$$
$$\mathbb{E}[R_T] = \tilde{O}(N^{\frac{1}{2}} K^{\frac{3}{2}} T^{\frac{1}{2}})$$

depending on the confidence parameter α_t.

The authors also prove a lower bound of

$$\mathbb{E}[R_T] = \Omega(\sqrt{KT})$$

Overall, this paper formalizes fairness through NSW in a multi-agent setting and establishes regret bounds for natural extensions of originally single-agent algorithms. Reading this paper was useful because it helped us identify objectives to optimize in our work.

A.5 Harada - Ito, and Sumita - Bandit Max-Min Fair Allocation

[Harada et al., 2025] introduce the Bandit Max-Min Fair Allocation problem, a new sequential decision-making framework that combines fairness in resource allocation with bandit-style uncertainty. There are n agents and m indivisible items. In each round t, the algorithm must allocate every item to exactly one agent, observing semi-bandit feedback $v_{i,e}^t$ for allocated pairs (i, e). We seek to maximize egalitarian social welfare, which is defined as the minimum cumulative utility across agents.

Each agent's value for item e at round t is sampled from an unknown distribution $D_{i,e} \subseteq [0, 1]$ with mean $\mu_{i,e}$. The cumulative utility of agent i is

$$X_i = \sum_{t=1}^{T} \sum_{e \in M} v_{i,e}^t a_{i,e}^t$$

and the algorithm's payoff is $\min_i X_i$.

Regret is defined as follows: let $x^1, ..., x^T$ be an optimal sequence of allocations computed using only the expectations $\mu_{i,e}$. The expected regret is defined as

$$R_T = \mathbb{E}[\mathrm{OPT} - \mathrm{ALG}]$$

where

$$\mathrm{OPT} = \min_i \sum_{t,e} v_{i,e}^t x_{i,e}^t$$

$$\mathrm{ALG} = \min_i \sum_{t,e} v_{i,e}^t a_{i,e}^t$$

The authors also define a surrogate regret using expected values.

$$R_T^{\mu} = \mathbb{E}[\mathrm{OPT}_{\mu} - \mathrm{ALG}_{\mu}]$$

and show that R_T^{μ} is not so far from R_T. Formally, they show that

$$|R_T - R_T^{\mu}| = O\left(m\sqrt{T \log(T)}\right)$$

The core algorithm introduced in the paper combines UCB for estimating means with a multiplicative-weight-style discount based on each agent's past cumulative utility. The algorithm essentially learns which agents have been historically "starved" and boosts them via discounting.

The main regret bound proved in the paper is

$$R_T = O\left(m\sqrt{T}\log(T)/n + m\sqrt{T \log(mnT)}\right)$$

The authors also prove a lower regret bound of

$$\Omega(m\sqrt{T}/n)$$

In summary, this paper introduces a bandit model with a max-min cumulative utility (egalitarian social welfare) fairness objective and proves regret bounds for an algorithm in that setting.

A.6 Aziz, Bogomolnaia, Moulin - Fair Mixing: the Case of Dichotomous Preferences

[Aziz et al., 2019] study fair mixing, which is selecting a probabilistic mixture of public outcomes when agents have dichotomous (like/dislike) preferences.

A set of n agents votes over outcomes in A. Each agent i has utility $u_i \in \{0,1\}^A$, indicating which outcomes they like. A mixture $z \in \Delta(A)$ assigns probabilities/shares to outcomes, and agent i's utility is $U_i = u_i \cdot z$. The goal is to select fair mixtures balancing majority preferences with minority protection.

The key fairness properties are

- **Individual Fair Share (IFS):** Each agent gets utility at least $\frac{1}{n}$. The intuition here is that each agent owns a $\frac{1}{n}$-th share of decision power; therefore, they should be able to ensure an outcome they like at least $\frac{1}{n}$ of the time.
- **Unanimous Fair Share (UFS):** This is a stronger notion than IFS. It says that if s agents have the same preferences, then they should collectively get utility at least $\frac{s}{n}$. This captures the idea that "numbers matter".
- **Average Fair Share (AFS):** For any group of agents sharing a common liked outcome, their average utility is at least their proportional size.
- **Core Fair Share (CFS):** No group of agents can block the outcome by pooling their decision power to enforce a better mixture.

The chain of implications is as follows:

$$\text{CFS} \implies \text{AFS} \implies \text{UFS} \implies \text{IFS}$$

The paper defines a notion called strategyproofness, which essentially means an agent cannot benefit by misreporting their preferences, regardless of what false preferences u_i'. In other words, the true utility u_i evaluated at the honest outcome is never worse than the true utility evaluated at any outcome from misreporting a preference.

The paper analyzes three voting rules.

- **Conditional Utilitarian (CUT):** Each agent gets $\frac{1}{n}$ of the total decision power and uses it on their preferred outcomes that others also like most. This algorithm is strategyproof, which means there is no incentive for agents to lie. It also guarantees fair shares to groups. However, it can be inefficient.
- **Egalitarian (EGAL):** Maximizes the leximin welfare ordering. It is efficient and satisfies excludable strategyproofness, which is where agents can be excluded from outcomes they claimed to dislike.
- **Nash Max Product (NMP):** Maximizes $\sum_i \ln(U_i)$. Efficient and gives the strongest fairness guarantees, but fails excludable strategyproofness.

The main result in this paper is that no rule can be simultaneously efficient, strategyproof, and satisfy IFS. This result informed our project by motivating us to study tradeoffs between individual and social welfare metrics.

A.7 Conitzer, Freeman, and Shah - Fair Public Decision Making

[Conitzer et al., 2017] generalizes the problem of fairly allocating private goods to a public decision-making setting in which decisions are made on multiple issues simultaneously, and a single decision can benefit several players.

There are m issues and n players, and each issue has a set of alternatives. Each player $i \in N$ has a utility function for each issue, and the total utility is additive across issues. The goal is to choose an outcome (one alternative per issue) that is both fair and efficient.

The authors focus on proportionality, meaning that each player should receive at least $\frac{1}{n}$ of the utility they would get if their preferred alternative were chosen for every issue. Since proportionality cannot always be guaranteed, the paper introduces three relaxations of this notion.

- **Proportionality up to one issue (Prop1)**: A player can achieve their proportional share if allowed to change the outcome of a single issue in their favor.
- **Round Robin Share (RRS):** Each player receives at least what they would get from the round robin mechanism if they were last in the ordering.
- **Pessimistic Proportional Share (PPS):** Each player receives at least their maximum utility from an adversarially chosen set of $\lfloor \frac{m}{n} \rfloor$ issues.

The chain of implications is as follows:

$$\text{Prop} \implies \text{RRS} \implies \text{PPS}, \text{Prop} \implies \text{Prop1}$$

The paper defines Pareto optimality (PO) as the property where no alternative outcome can make any player strictly better off without making at least one player strictly worse off.

The authors show that the round robin mechanism satisfies RRS, PPS, and Prop1. However, it fails PO. They show the leximin mechanism satisfies RRS, PPS, and PO. Finally, they show the Maximum Nash Welfare (MNW) mechanism, analogous to Nash social welfare, satisfies Prop1 and PO.

Overall, the paper provides a useful framework connecting fair division and voting. This paper informed our project by motivating the analysis of welfare trade-offs in multi-agent decision-making.

B Missing Proofs

Proof of Lemma 6.1

Proof. We know that

$$\Delta^1 = \mu_1^1 - \mu_2^1 > 0$$

Therefore, agent 1's optimal arm is arm 1. We will now prove that agent 2's optimal arm is also arm 1.

By the similarity constraint, we know that

$$|\mu_j^1 - \mu_j^2| \leq \epsilon$$

for each arm $j \in \{1, 2\}$.

This gives us

$$\mu_1^2 \geq \mu_1^1 - \epsilon$$
$$\mu_2^2 \leq \mu_2^1 + \epsilon$$

Therefore, the gap for agent 2 is

$$\Delta^2 = \mu_1^2 - \mu_2^2 \geq (\mu_1^1 - \epsilon) - (\mu_2^1 + \epsilon) = \mu_1^1 - \mu_2^1 - 2\epsilon = \Delta^1 - 2\epsilon$$

$$\implies \Delta^2 \geq \Delta^1 - 2\epsilon$$

Since $\epsilon < \frac{\Delta^1}{2}$, we know that

$$\Delta^2 \geq \Delta^1 - 2\epsilon > \Delta^1 - 2 \cdot \frac{\Delta^1}{2} = 0$$

$$\implies \Delta^2 > 0$$

$$\implies \mu_1^2 - \mu_2^2 > 0$$

$$\implies \mu_1^2 > \mu_2^2$$

Therefore, agent 2's optimal arm is arm 1, so both agents have the same optimal arm. □

B.1 Proof of Lemma 6.2

Proof. By Hoeffding's inequality, for each arm j, agent i, and time t, we have

$$\mathbb{P}\left(|\hat{\mu}_{j,t}^i - \mu_j^i| > \sqrt{\frac{2\log(4T)}{N_j(t)}}\right) \leq \frac{1}{2T^2}$$

Define the bad event

$$E_{i,j,t} = \left\{|\hat{\mu}_{j,t}^i - \mu_j^i| > \sqrt{\frac{2\log(4T)}{N_j(t)}}\right\}$$

By taking the union bound over all 2 × 2 × T bad events, we have

$$\mathbb{P}(\mathcal{E}^c) = \mathbb{P}\left(\bigcup_{i,j,t} E_{i,j,t}\right) \leq \sum_{i,j,t} \mathbb{P}(E_{i,j,t}) = 4T \cdot \frac{1}{2T^2} = \frac{2}{T}$$

$$\implies \mathbb{P}(\mathcal{E}^c) \leq \frac{2}{T}$$

$$\implies \mathbb{P}(\mathcal{E}) \geq 1 - \frac{2}{T}$$

□

B.2 Proof of Lemma B.4

Proof. By the similarity constraint, we know that

$$|\mu_j^1 - \mu_j^2| \leq \epsilon$$

for all arms $j \in \{1, 2\}$. This gives us

$$\mu_1^2 \geq \mu_1^1 - \epsilon$$
$$\mu_2^2 \leq \mu_2^1 + \epsilon$$

Therefore, the gap for agent 2 is

$$\Delta^2 = \mu_1^2 - \mu_2^2 \geq (\mu_1^1 - \epsilon) - (\mu_2^1 + \epsilon) = \mu_1^1 - \mu_2^1 - 2\epsilon = \Delta^1 - 2\epsilon$$

Since $\epsilon < \frac{\Delta^1}{2}$, we know that

$$\Delta^2 \geq \Delta^1 - 2\epsilon > \Delta^1 - 2 \cdot \frac{\Delta^1}{2} = 0$$

Therefore,

$$\Delta^2 > 0$$

□

B.3 Proof of Lemma 6.3

Proof. Condition on event $\mathcal{E}$. Suppose that agent i votes for arm 2 at time t. Then, by the greedy voting policy, we know that

$$\hat{\mu}_{2,t}^i \geq \hat{\mu}_{1,t}^i$$

By conditioning on $\mathcal{E}$, we know that for each arm i and each agent j, it holds that

$$\hat{\mu}_{j,i}^i \in \left[\mu_j^i - \sqrt{\frac{2\log(4T)}{N_j(T)}}, \mu_j^i + \sqrt{\frac{2\log(4T)}{N_j(t)}}\right]$$

Therefore, we obtain the following upper bound for $\hat{\mu}^i_{2,t}$

$$\hat{\mu}^i_{2,t} \leq \mu^i_2 + \sqrt{\frac{2\log(4T)}{N_2(t)}}$$

and the following lower bound for $\hat{\mu}^i_{1,t}$

$$\hat{\mu}^i_{1,t} \geq \mu^i_1 - \sqrt{\frac{2\log(4T)}{N_1(t)}}$$

Combining these with $\hat{\mu}^i_{2,t} \geq \hat{\mu}^i_{1,t}$, we obtain

$$\mu^i_2 + \sqrt{\frac{2\log(4T)}{N_2(t)}} \geq \mu^i_1 - \sqrt{\frac{2\log(4T)}{N_1(t)}}$$

Rearranging the terms, we obtain

$$\mu^i_1 - \mu^i_2 \leq \sqrt{\frac{2\log(4T)}{N_2(t)}} + \sqrt{\frac{2\log(4T)}{N_1(t)}}$$

Since $\Delta^i = \mu^i_1 - \mu^i_2$ we have

$$\Delta^i \leq \sqrt{\frac{2\log(4T)}{N_2(t)}} + \sqrt{\frac{2\log(4T)}{N_1(t)}}$$

□

B.4 Proof of Lemma 6.4

Proof. Suppose, for contradiction, that agent $i \in \{1,2\}$ votes for arm 2 at some time t, when $N_1(t), N_2(t) \geq m$, and event $\mathcal{E}$. We will now derive a contradiction.

We know, from the previous lemma, that

$$\Delta^i \leq \sqrt{\frac{2\log(4T)}{N_2(t)}} + \sqrt{\frac{2\log(4T)}{N_1(t)}}$$

Since $N_1(t), N_2(t) \geq m$, we know that

$$\Delta^i \leq \sqrt{\frac{2\log(4T)}{N_2(t)}} + \sqrt{\frac{2\log(4T)}{N_1(t)}} \leq \sqrt{\frac{2\log(4T)}{m}} + \sqrt{\frac{2\log(4T)}{m}} = 2\sqrt{\frac{2\log(4T)}{m}}$$

$$\implies \Delta^i \leq 2\sqrt{\frac{2\log(4T)}{m}}$$

By , we know that

$$\Delta^i \geq \Delta^1 - 2\epsilon$$

Combining these inequalities, we obtain

$$\Delta^1 - 2\epsilon \le \Delta^i \le 2\sqrt{\frac{2\log(4T)}{m}}$$

which implies

$$\Delta^1 - 2\epsilon \le 2\sqrt{\frac{2\log(4T)}{m}}$$
$$\implies (\Delta^1 - 2\epsilon)^2 \le \frac{8\log(4T)}{m}$$

Rearranging, we get

$$m \le \frac{8\log(4T)}{(\Delta^1 - 2\epsilon)^2}$$

However, by the assumption we made in the lemma, we have

$$m > \frac{8\log(4T)}{(\Delta^1 - 2\epsilon)^2}$$

which is a contradiction.

Therefore, agent i cannot vote for arm 2 when $N_1(t), N_2(t) \ge m$, and $\mathcal{E}$ holds. Since this argument applies to both agents, both agents only vote for arm 1.

□

B.5 Proof of Lemma 6.5

Proof. We decompose the regret into exploration and exploitation phases:

$$\mathbb{E}[\mathcal{R}_i(T)] = \mathbb{E}[\mathcal{R}_i^{\text{explore}}(T)] + \mathbb{E}[\mathcal{R}_i^{\text{exploit}}(T)]$$

Now, we will find an upper bound on the regret incurred during both phases.

Step 1. Exploration Regret Recall that the aggregator explores with probability $p_{\text{explore}}(t) = \min\left\{1, \frac{4c}{t}\right\}$

Therefore, the expected number of exploration rounds is

$$\mathbb{E}[\#\text{ explore rounds}] = \sum_{t=1}^{T} \min\{1, 4c/t\} = 4c + \sum_{t=4c+1}^{T} \frac{4c}{t}$$

We can use an integral to obtain an upper bound on the sum

$$\le 4c + 4c\int_{4c}^{T} \frac{1}{x}dx = 4c\left(1 + \log\left(\frac{T}{4c}\right)\right) = O(c\log(T))$$

Note that the integral $\int_{4c}^{T} \frac{1}{x}dx$ converges because $c > 0$, so there are no vertical asymptotes in $[c, T]$.

So, we have

$$\mathbb{E}[\#\text{ explore rounds}] = O(c\log(T))$$

During exploration, each arm is pulled uniformly. The regret per exploration round is at most 1.

Therefore, we obtain

$$\boxed{\mathbb{E}[\mathcal{R}_i^{\text{explore}}(T)] = O(c\log T)}$$

We now switch our attention to finding an asymptotic upper bound on the regret incurred during exploitation.

Step 2. Concentration of mean estimates By Hoeffding's inequality, we know that for each arm j, each agent i, and each time t, the following tail bound holds

$$\mathbb{P}\left(|\hat{\mu}_{j,t}^i - \mu_j^i| > \sqrt{\frac{2\log(4T)}{N_j(t)}}\right) \le \frac{2}{(4T)^4} \le \frac{1}{2T^2}$$

Define the event $E_{i,j,t}$ to be the event in which agent i's approximation of the mean of the reward distribution of arm j at time t is "far" from the true mean of the reward distribution.

$$E_{i,j,t} = \left(|\hat{\mu}_{j,t}^i - \mu_j^i| > \sqrt{\frac{2\log(4T)}{N_j(t)}}\right)$$

We now define $\mathcal{E}$ to be the event that *none* of these bad deviations occur for any agent, arm, or time up to T:

$$\mathcal{E} = \left\{\forall i \in \{1,2\},\ \forall j \in \{1,2\},\ \forall t \in [T] : \left|\hat{\mu}_{j,t}^i - \mu_j^i\right| \le \sqrt{\frac{2\log(4T)}{N_j(t)}}\right\}$$

We now take the union bound over all $E_{i,j,t}$, of which there are $2 \times 2 \times T$.

Note that

$$\mathbb{P}(E_{i,j,t}) \le \frac{1}{2T^2}$$

Now, applying the union bound, we have

$$\mathbb{P}(\mathcal{E}^c) = \mathbb{P}\left(\bigcup_{i,j,t} E_{i,j,t}\right) \le \sum_{i,j,t} \mathbb{P}(E_{i,j,t}) = 4T \cdot \frac{1}{2T^2} = \frac{2}{T}$$

$$\implies \mathbb{P}(\mathcal{E}^c) \le \frac{2}{T}$$

$$\implies \mathbb{P}(\mathcal{E}) \ge 1 - \frac{2}{T}$$

From now on, we condition on $\mathcal{E}$.

Step 3. Analyze Agent 2's Gap We now examine the gap for agent 2. Since $|\mu_j^1 - \mu_j^2| \leq \epsilon$ for both arms, we know that $\mu_1^2 \geq \mu_1^1 - \epsilon$ and $\mu_2^2 \leq \mu_2^1 + \epsilon$.

Therefore, the gap of agent 2 is

$$\Delta^2 = \mu_1^2 - \mu_2^2 \geq (\mu_1^1 - \epsilon) - (\mu_2^1 + \epsilon) = \Delta^1 - 2\epsilon > 0$$

We know that $\Delta^2 \geq 0$ because $\epsilon < \frac{\Delta^1}{2}$.

Similarly, $\Delta^1 \geq \Delta^1 - 2\epsilon$ (the gap of agent 1 is at least as large).

Step 4. Analyze when agents stop voting for arm 2 Note that arm 2 is the suboptimal arm for both agents, so, intuitively, both agent's will eventually stop voting for arm 2 as their empirical reward mean estimates converge to the true means.

On the clean event $\mathcal{E}$, we know that agent i votes for arm 2 only if

$$\hat{\mu}_{2,t}^i \geq \hat{\mu}_{1,t}^i$$

By the conditioning on $\mathcal{E}$, we know that

$$\left\{ \forall i \in \{1,2\},\ \forall j \in \{1,2\},\ \forall t \in [T] : \left| \hat{\mu}_{j,t}^i - \mu_j^i \right| \leq \sqrt{\frac{2\log(4T)}{N_j(t)}} \right\}.$$

Therefore, we know that for each arm j,

$$\hat{\mu}_{j,t}^i \in \left[\mu_j^i - \sqrt{\frac{2\log(4T)}{N_j(t)}},\ \mu_j^i + \sqrt{\frac{2\log(4T)}{N_j(t)}} \right].$$

From this, it follows that

$$\hat{\mu}_{2,t}^i \leq \mu_2^i + \sqrt{\frac{2\log(4T)}{N_2(t)}}$$

$$\hat{\mu}_{1,t}^i \geq \mu_1^i - \sqrt{\frac{2\log(4T)}{N_1(t)}}$$

Applying these upper and lower bounds to $\hat{\mu}_{2,t}^i \geq \hat{\mu}_{1,t}^i$, we obtain

$$\mu_2^i + \sqrt{\frac{2\log(4T)}{N_2(t)}} \geq \mu_1^i - \sqrt{\frac{2\log(4T)}{N_1(t)}}$$

$$\implies \mu_2^i - \mu_1^i \geq -\sqrt{\frac{2\log(4T)}{N_2(t)}} - \sqrt{\frac{2\log(4T)}{N_1(t)}}$$

$$\implies \mu_1^i - \mu_2^i \leq \sqrt{\frac{2\log(4T)}{N_2(t)}} + \sqrt{\frac{2\log(4T)}{N_1(t)}}$$

$$\Delta^i \leq \sqrt{\frac{2\log(4T)}{N_2(t)}} + \sqrt{\frac{2\log(4T)}{N_1(t)}}$$

Let

$$m = \left\lceil \frac{8\log(4T)}{(\Delta^1 - 2\epsilon)^2} \right\rceil$$

and suppose that $N_1(t), N_2(t) \geq m$

Now, the right hand side is at most

$$2\sqrt{\frac{2\log(4T)}{m}} = 2\sqrt{\frac{2\log(4T) \cdot (\Delta^1 - 2\epsilon)^2}{8\log(4T)}} = 2\sqrt{\frac{(\Delta^1 - 2\epsilon)^2}{4}} = \Delta^1 - 2\epsilon$$

Therefore, we have that

$$\Delta^i \leq \Delta^1 - 2\epsilon$$

From step 2.2, recall that $\Delta^1 - 2\epsilon \leq \Delta^i$. Therefore, we have

$$\Delta^i \leq \Delta^1 - 2\epsilon \leq \Delta^i$$

$$\implies \Delta^i = \Delta^1 - 2\epsilon$$

When $N_1(t), N_2(t) > m$, we know that

$$\Delta^i > \sqrt{\frac{2\log(4T)}{N_2(t)}} + \sqrt{\frac{2\log(4T)}{N_1(t)}}$$

This contradicts the requirement for voting for arm 2. Therefore, agent i cannot vote for arm 2.

From this we have the following lemma

Lemma B.1. *Once both arms have been pulled at least m times, and event $\mathcal{E}$ holds, then both agents will only vote for arm 1.*

Step 5 : Bounding exploitation regret The key to this step is we will use the law of total expectation conditioned on $\mathcal{E}$

$$\mathbb{E}\big[\mathcal{R}_i^{\text{exploit}}(T)\big] = \mathbb{E}\big[\mathcal{R}_i^{\text{exploit}}(T) \mid \mathcal{E}\big] \cdot \mathbb{P}(\mathcal{E}) + \mathbb{E}\big[\mathcal{R}_i^{\text{exploit}}(T) \mid \mathcal{E}^c\big] \cdot \mathbb{P}(\mathcal{E}^c).$$

Define τ to be the first time during which both arms have been pulled at least m times

$$\tau = \inf\{t : N_1(t) \geq m, N_2(t) \geq m\}$$

Before time τ, arm 2 can be pulled during exploration. However, there will be at most m pulls of arm 2 to reach the threshold.

$$N_2^{\text{exploit}}(\tau) \leq N_2(\tau) \leq max(N_1(\tau), N_2(\tau)) = m$$

After time τ, both agents will vote for arm 1. During exploration, the aggregator picks uniformly from $\{a_t^1, a_t^2\} = \{1, 1\}$, so it deterministically selects 1 for each iteration. Arm 2 is never pulled during exploration rounds after time τ.

Therefore, we know that

$$\mathbb{E}[N_2^{\text{exploit}}(T) \mid \mathcal{E}] \leq m = O\left(\frac{\log T}{(\Delta^1 - 2\epsilon)^2}\right)$$

From this, we derive that the exploitation regret is

$$\mathbb{E}[\mathcal{R}_i^{\text{exploit}}(T) \mid \mathcal{E}] = \Delta^i \cdot \mathbb{E}[N_2^{\text{exploit}}(T) \mid \mathcal{E}] \leq \Delta^i \cdot m$$

Since rewards are bounded in $[0, 1]$, we know that $\Delta^i \leq 1$. Therefore,

$$\mathbb{E}[\mathcal{R}_i^{\text{exploit}}(T) \mid \mathcal{E}]\mathbb{P}(\mathcal{E}) \leq \mathbb{E}[\mathcal{R}_i^{\text{exploit}}(T) \leq m = O\left(\frac{\log T}{(\Delta^1 - 2\epsilon)^2}\right)$$

Add the contribution from $\mathcal{E}$. Again, when $\mathcal{E}^c$ holds, regret can be arbitrarily bad, so, since rewards are bounded in $[0, 1]$, T is an upper bound for the expected exploitation regret term conditioned in $\mathcal{E}^c$

$$\mathbb{E}[\mathcal{R}_i^{\text{exploit}}(T) \mid \mathcal{E}^c] \cdot \mathbb{P}(\mathcal{E}^c) \leq T \cdot \frac{2}{T} = O(1)$$

Adding the upper bounds, we obtain

$$\mathbb{E}[\mathcal{R}_i^{\text{exploit}}(T)] = O\left(\frac{\log T}{(\Delta^1 - 2\epsilon)^2}\right)$$

Finally, adding the two regrets, we obtain

$$\boxed{\mathbb{E}\left[\mathcal{R}_i(T)\right] \leq O\left(c \log T + \frac{\log T}{(\Delta^1 - 2\epsilon)^2}\right)}$$

□