p-Mean Regret for Stochastic Bandits

Anand Krishna[1], **Philips George John**[2], **Adarsh Barik**[3], **Vincent Y. F. Tan**[4, 1]

[1] Department of Electrical and Computer Engineering, NUS
[2] CNRS-CREATE & Department of Computer Science, NUS
[3] Institute of Data Science, NUS
[4] Department of Mathematics, NUS
anandkrishna011095@gmail.com, philips.george.john@u.nus.edu, {abarik,vtan}@nus.edu.sg

Abstract

In this work, we extend the concept of the p-mean welfare objective from social choice theory to study p-mean regret in stochastic multi-armed bandit problems. The p-mean regret, defined as the difference between the optimal mean among the arms and the p-mean of the expected rewards, offers a flexible framework for evaluating bandit algorithms, enabling algorithm designers to balance fairness and efficiency by adjusting the parameter p. Our framework encompasses both average cumulative regret and Nash regret as special cases. We introduce a simple, unified UCB-based algorithm (EXPLORE-THEN-UCB) that achieves novel p-mean regret bounds. Our algorithm consists of two phases: a carefully calibrated uniform exploration phase to initialize sample means, followed by the UCB1 algorithm of Auer et al. Under mild assumptions, we prove that our algorithm achieves a p-mean regret bound of $\tilde{O}\left(\sqrt{\frac{k}{T^{\frac{1}{2|p|}}}}\right)$ for all $p \leq -1$, where k represents the number of arms and T the time horizon. When $-1 < p < 0$, we achieve a regret bound of $\tilde{O}\left(k^{\frac{3}{4}}T^{-\frac{1}{4}}\right)$. For the range $0 < p \leq 1$, we achieve a p-mean regret scaling as $\tilde{O}\left(\sqrt{\frac{k}{T}}\right)$, which matches the previously established lower bound up to logarithmic factors. This result stems from the fact that the p-mean regret of any algorithm is at least its average cumulative regret for $p \leq 1$. In the case of Nash regret (the limit as p approaches zero), our unified approach differs from prior work of Barman et al., which requires a new Nash Confidence Bound algorithm. Notably, we achieve the same regret bound up to constant factors using our more general method.

Code — https://github.com/philips-george/p-mean-regret-stochastic-bandits

Extended version — http://arxiv.org/abs/2412.10751

Introduction

The multi-armed bandit (MAB) problem has long served as a cornerstone in the study of sequential decision-making under uncertainty (Thompson 1933; Lai and Robbins 1985; Bubeck, Cesa-Bianchi et al. 2012). In the stochastic MAB framework, a decision-maker interacts with a set of k arms, each associated with an unknown probability distribution of rewards. Over T rounds, the algorithm selects an arm in each round and receives a reward drawn from that arm's distribution. Ideally, if the distributions were known, the optimal strategy would be to always choose the arm with the highest expected reward. However, because the statistical properties of the arms are unknown, the algorithm's selections may not always be optimal, leading to suboptimal rewards. This performance loss is quantified by the concept of regret, which measures the difference between the optimal mean and the algorithm's performance. Regret thus serves as a performance metric, reflecting the algorithm's efficiency in learning and adapting to the best choices over time. Traditional approaches to MAB problems have primarily focused on maximizing cumulative rewards, often overlooking crucial considerations of fairness and social welfare. As machine learning algorithms increasingly influence resource allocation and decision-making in societally impactful domains—such as healthcare (Villar, Bowden, and Wason 2015), education (Clement et al. 2013), and online advertising (Schwartz, Bradlow, and Fader 2017)—there is a growing imperative to incorporate principles of fairness and social welfare into our algorithmic frameworks.

The MAB literature has extensively studied two main regret formulations: average regret, which measures the difference between the optimal mean and the arithmetic mean of the rewards from the chosen arms over time (Lattimore and Szepesvári 2020), and simple regret, which quantifies the expected suboptimality of the final recommended arm, (Bubeck, Munos, and Stoltz 2009). However, these traditional formulations often fall short in addressing the nuanced fairness considerations that arise in many real-world applications. To appreciate the relevance of a more comprehensive perspective, consider settings where the algorithm's rewards correspond to values distributed across a population of T agents. One such scenario can be found in the context of a large-scale renewable energy initiative. For example, consider a government program testing different types of residential solar panel installations across a diverse region. In each round $t \in \{1, ..., T\}$, a new household is selected to receive a solar panel installation, and the program must choose one of k available solar panel technologies to install for the t-th household. The observed reward could be the energy output efficiency of the installed system over a fixed period. While there might be slight variations due to factors like roof

orientation or local weather patterns, there is likely one solar panel technology that performs best on average across the entire region. This scenario encapsulates the challenge of balancing between exploiting the currently best-known technology and exploring other options to ensure a potentially superior technology hasn't been overlooked, all while ensuring fair treatment for each household regardless of when they join the program.

In the context of such applications, the average regret equates the algorithm's performance to the social welfare across all T households. However, this utilitarian approach might not induce a fair outcome, as high social welfare can be achieved even if less efficient solar technologies are inconsiderately installed for an initial set of households. On the other hand, simple regret maps to the egalitarian (Rawlsian) welfare but only after excluding an initial set of households that served as test subjects for various technologies. These limitations highlight the need for a more flexible and comprehensive framework that can balance fairness and efficiency in MAB settings where the algorithm's rewards correspond to agents' values (e.g., energy output efficiency, household energy savings, or overall carbon footprint reduction). Such a framework should ensure that households joining the program at different times all benefit from increasingly effective technology choices while also allowing for necessary exploration to identify the best solar panel technology for the region.

To address these shortcomings, we study a novel approach that incorporates the concept of p-mean welfare (Moulin 2004) from social choice theory into the MAB framework. The p-mean welfare, also known as the generalized mean or power mean, lends itself as a flexible metric for aggregating individual utilities or rewards. It is defined as:

$$W_p = \left(\frac{1}{T} \sum_{t=1}^{T} u_i^p \right)^{\frac{1}{p}},$$

where p is a parameter that determines the sensitivity of the welfare measure to individual utilities $u_i \in \mathbb{R}_+$ of each agent $t \in [T]$.

This formulation allows for a spectrum of welfare considerations, ranging from the utilitarian approach (emphasizing total utility) to the Rawlsian approach (emphasizing the welfare of the worst-off individual). By adjusting the parameter p, one can tailor the welfare measure to reflect different societal preferences. This generalized framework is axiomatically well-supported. Notably, for $p \leq 1$, it adheres to the Pigou-Dalton principle, i.e., redistributing a small portion of reward from an agent with higher utility to one with lower utility tends to enhance overall welfare. The magnitude of this effect intensifies as p decreases, reflecting a stronger emphasis on equity. Concurrently, the p-mean welfare construct allows for a nuanced approach to allocative efficiency. As p increases, the function becomes more sensitive to aggregate reward maximization. By offering this continuous spectrum between utilitarian and egalitarian objectives, with Nash Social Welfare (Kaneko and Nakamura 1979) situated as an intermediate case, the p-mean welfare framework empowers researchers and decision-makers to fine-tune the balance

between efficiency and fairness, tailoring their approach to the specific ethical considerations and practical constraints of diverse application scenarios.

Motivated by the flexibility and comprehensiveness of p-mean welfare, we utilize this concept in bandit problems by defining p-mean regret. This metric is given by

$$R_p = \mu^* - \left(\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\mu_{I_t}]^p \right)^{\frac{1}{p}} \tag{1}$$

where I_t is a random variable representing the arm pulled in round t, μ_{I_t} is the expected reward of the arm I_t, and μ^* is the expected reward of the optimal arm. The p-mean regret measures the difference between the expected reward of the optimal arm and the p-mean of the expected rewards of the chosen arms.

The introduction of p-mean regret allows for a more nuanced evaluation of bandit algorithms. By varying p, we can capture different aspects of the regret distribution. For instance, when $p \to 1$, the p-mean regret approaches the traditional regret measure, focusing on the average performance. When $p \to -\infty$, it emphasizes the performance of the worst-off rounds, analogous to the Rawlsian welfare function. Conversely, when $p \to +\infty$, it highlights the best-performing rounds, reflecting a more utilitarian perspective without any fairness guarantees. For $p \to 0$, our formulation converges to the Nash regret that was previously studied by Barman et al. (2023), which is defined as the difference between the optimal mean μ^* and the geometric mean of the expected rewards. This special case aligns with the Nash social welfare (NSW), an axiomatically-supported welfare objective that satisfies fundamental principles including symmetry, independence of unconcerned agents, scale invariance, and the Pigou-Dalton transfer principle (Moulin 2004).

Incorporating p-mean regret into the analysis of bandit problems provides a richer understanding of algorithm performance, accounting for diverse criteria of fairness and efficiency. This approach not only aligns with the broader objectives of social choice theory but also offers a robust framework for designing and evaluating bandit algorithms in various practical scenarios.

p-range	Explore period $\tilde{T}$	Regret
$p \in (0, 1]$ (Theorem 5)	$16\sqrt{\frac{Tk^p \log T}{\log k}}$	$\tilde{O}\left(\sqrt{\frac{k}{T}}\right)$
$p \to 0$ (Nash) (Theorem 6)	$16\sqrt{\frac{Tk \log T}{\log k}}$	$\tilde{O}\left(\sqrt{\frac{k}{T}}\right)$
$p \in [-1, 0)$ (Theorem 4)	$16\sqrt{\frac{T \log T}{k^{|p|}}}$	$\tilde{O}\left(k^{\frac{3}{4}} T^{\frac{-1}{4}}\right)$
$p \in (-\infty, -1)$ (Theorem 4)	$16\sqrt{\frac{T \log T}{k^{|p|}}}$	$\tilde{O}\left(k^{\frac{1}{2}} T^{\frac{-1}{4|p|}}\right)$

Table 1: Summary of results

Our Contributions

We build upon the work of Barman et al. (2023), where they develop a new Nash Confidence Bound (NCB) algorithm specifically for Nash regret, i.e., the $p \to 0$ case. In their work, the authors present a stochastic MAB instance where the vanilla UCB1 algorithm incurs a Nash regret of $1 - \frac{1}{T}$. This counterexample illustrates that while UCB is effective for traditional regret metrics, it may not perform well when evaluated using alternative metrics that incorporate fairness or welfare considerations.

In this work, we develop an algorithm for any $p \leq 1$ by introducing a mild yet significant assumption that every arm has a minimum expected reward of $\tilde{\Omega}(\sqrt{k/T^{\frac{1}{4}}})$, which allows us to bypass the counter-example where UCB fails. This approach contrasts with the solution proposed in the aforementioned paper where they required a specialized algorithm for handling the case of Nash regret.

Our UCB-based algorithm achieves vanishing p-mean regret, i.e., $o(1)$ in the regime where $-\Omega(\log T) < p \leq 1$. As p decreases beyond $-T \log T$, the p-mean welfare for any bandit instance becomes a constant factor approximation of egalitarian welfare. Consequently, achieving vanishing p-mean regret is not possible with even two arms for $p \leq -T \log T$ (see Proposition 1 in (Barman et al. 2020)).

In summary, we explore the concept of p-mean regret, inspired by p-mean welfare, and demonstrate the effectiveness of a UCB-based algorithm under this new metric. By overcoming the limitations highlighted in previous work through a minimal assumption, we offer a comprehensive solution that balances fairness and efficiency in the evaluation of bandit algorithms.

We make the following technical contributions:

- We study p-mean regret in stochastic multi-armed bandit problems, providing a flexible framework for evaluating bandit algorithms that balances fairness and efficiency via the parameter p.

- We propose EXPLORE-THEN-UCB, a unified algorithm designed to achieve novel p-mean regret bounds. This algorithm consists of two distinct phases: an initial phase involving a calibrated uniform exploration followed by the UCB1 algorithm.

- Our algorithm achieves a p-mean regret bound of $\tilde{O}\left(\sqrt{\frac{k}{T^{1/(2|p|)}}}\right)$ for all $p \leq -1$, where k is the number of arms and T is the time horizon and for the range $-1 < p < 0$, we achieve a regret bound of $\tilde{O}\left(\sqrt{\frac{k^{3/2}}{T^{1/2}}}\right)$.

- For the special case of Nash regret ($p \to 0$), our approach achieves the same regret bound as prior work (Barman et al. 2023), but with a simpler algorithm.

- For $0 < p \leq 1$, we show that the p-mean regret scales as $\tilde{O}\left(\sqrt{k/T}\right)$, matching the known minimax lower bound that applies to this range via the generalized mean inequality.

Related Work

The incorporation of fairness considerations into MAB problems has garnered significant attention in recent years, driven by the increasing deployment of learning algorithms in domains with far-reaching social implications.

Fairness in Multi-Armed Bandits   Recent works have examined various notions of fairness in MAB contexts. Joseph et al. (2016), Celis et al. (2019), and Patil et al. (2021) primarily focused on ensuring fairness for the arms themselves. In a multi-agent setting, Hossain, Micha, and Shah (2021) and Jones, Nguyen, and Nguyen (2023) studied scenarios where each arm pull yields potentially different rewards for each agent, aiming to find a fair distribution over arms. While these approaches highlight the importance of fairness in MAB settings, our work diverges by ensuring fairness across time, treating each round as a distinct agent deserving of fair treatment.

Barman et al. (2023) introduced the concept of Nash regret and developed the Nash Confidence Bound algorithm to minimize it for stochastic MAB settings. This algorithm provides tight Nash regret guarantees for both known and unknown time horizons (including T-oblivious settings). Our research extends this work by studying the more general p-mean regret, which allows for a flexible balance between fairness and efficiency, with Nash regret as a special case when p approaches 0.

p-Mean Welfare and Fair Division   The p-mean welfare objective has been extensively studied in fair division research, an area at the interface of mathematical economics and computer science. As characterized in social choice theory (Moulin 2004), the p-mean welfare function provides a parameterized framework to balance equity and efficiency (Barman et al. 2020; Garg et al. 2021; Barman, Khan, and Maiti 2022; Eckart, Psomas, and Verma 2024). This family of welfare functions is defined by five natural axioms: anonymity, scale invariance, continuity, monotonicity, and symmetry, ensuring that it reflects various principles of fairness in resource allocation. Moreover, the Pigou-Dalton principle—where a transfer from a better-off individual to a worse-off one improves welfare—restricts p to be less than or equal to 1. Our work leverages this rich theoretical foundation, avoiding the need for ad-hoc fairness constraints.

Other related work   Several recent works have explored fairness concepts in domains adjacent to our research, highlighting the growing interest in incorporating fairness considerations into various learning algorithms. Sawarni, Pal, and Barman (2024) studied Nash regret for stochastic linear bandits, proving tight upper bounds for sub-Poisson rewards. Their work extends the concept of Nash regret to more complex bandit settings, complementing our generalization to p-mean regret in the MAB setup.

The work of Zhang, Vuong, and Luo (2024) investigated online multi-agent Nash social welfare (NSW) maximization. Their setting, where the learner's decision affects multiple agents simultaneously, differs from our round-by-round fairness approach but underscores the importance of considering fairness in online decision-making processes.

In a multi-agent reinforcement learning context, Mandal and Gan (2022) adopts an axiomatic approach, demonstrating that Nash Social Welfare (NSW) uniquely satisfies certain fairness axioms and provides regret bounds derived for policies maximizing various fair objectives. These approaches highlight the growing interest in incorporating fairness metrics like NSW into various learning algorithms, from bandits to complex Markov decision process environments.

Preliminaries

In the current work, we consider the stochastic MAB problem. Here, we are presented with k unknown probability distributions (arms) each with mean $\mu_i \geq \mu_{\min}$ each supported on the interval $[0, 1]$. Our goal is to design a learning algorithm that adaptively uses sample access (rewards) to these arms over a time horizon T to minimize the p-mean regret. This setup corresponds to an agent arriving in every round $t \in [T]$ of the algorithm, which receives a reward X_t based on the arm that the (decision maker) algorithm pulls.

In our approach to p-mean regret minimization, we employ a dual-phase strategy that combines uniform exploration with the Upper Confidence Bound (UCB1) algorithm. Here, we provide a brief overview of UCB1, introduced by Auer, Cesa-Bianchi, and Fischer (2002). It is renowned for its effectiveness in balancing exploration and exploitation in stochastic multi-armed bandit (MAB) problems.

The Upper Confidence Bound (UCB1) algorithm is a fundamental method for solving MAB problems, embodying the principle of 'optimism in the face of uncertainty'. UCB1 skillfully balances exploration and exploitation by constructing optimistic estimates of each action's expected payoff. At its core, UCB1 selects the action with the highest upper confidence bound, calculated as $\hat{\mu}_j + 4\sqrt{\frac{\log T}{n_j}}$, where $\hat{\mu}_j$ is the empirical mean payoff of arm j and n_j is the number of times arm j has been played.

By setting the confidence term to $4\sqrt{\frac{\log T}{n_j}}$, UCB1 ensures that the probability of the true mean exceeding this bound decreases as $O(T^{-4})$, rapidly converging to zero as the rounds progress. This approach addresses the exploration-exploitation dilemma: actions with high empirical means are favored (exploitation), while rarely-played actions maintain high upper bounds due to uncertainty (exploration).

The elegance of UCB1 lies in its simplicity and strong theoretical guarantees, achieving worst-case average regret bound of $O(\sqrt{\frac{k \log T}{T}})$ without prior knowledge of reward distributions. This makes it particularly suitable for our p-mean regret minimization context.

The EXPLORE-THEN-UCB Algorithm

Our algorithm begins with a uniform exploration phase, which serves to establish an empirical foundation by collecting initial reward estimates for each arm. This phase is crucial in the p-mean regret context, where fairness is guaranteed over the entire time horizon. This phase is technically motivated by our use of a stronger property for

---

Algorithm 1 The EXPLORE-THEN-UCB

Parameters: Time horizon T, number of arms k, exploration period $\tilde{T}$.

for $t \leftarrow 1, \ldots, \tilde{T}$ **do**
  Uniformly sample i_t from $[k]$.
  Pull arm i_t and observe the reward X_t.
  Increment $n_{i_t, t}$ by one and update $\hat{\mu}_{i_t, t}$

for $t \leftarrow \tilde{T} + 1, \ldots, T$ **do**
  Let $\mathsf{UCB}_{i, t-1} \triangleq \hat{\mu}_{i, t-1} + 4\sqrt{\frac{\log T}{n_{i, t-1}}}$.
  Select $i_t \in \arg\max_{i \in [k]} \mathsf{UCB}_{i, t-1}$.
  Pull arm i_t and observe the reward X_t.
  Update $n_{i_t, t}$ and $\hat{\mu}_{i_t, t}$.

---

the confidence-bound (as detailed in Lemmas 2 and 3) in Phase II, compared to the standard UCB1 analysis. However, striking a balance is crucial, as an overly lengthy exploration phase would negatively impact fairness for a large subset of agents. Consequently, for more negative values of p, we employ a shorter exploration phase to maintain fairness while still benefiting from the initial uniform exploration.

Following this calibrated exploration, the algorithm transitions to UCB1, capitalizing on the gathered data to drive the decision-making process. UCB1 dynamically balances exploration and exploitation based on the observed rewards, adapting effectively to the underlying stochastic environment while ensuring controlled worst-case regret. This dual-phase strategy achieves novel p-mean regret bounds by combining an initial phase of uniform exploration with the proven effectiveness of UCB1. The pseudocode of our algorithm is given in Algorithm 1.

In this section we show that the EXPLORE-THEN-UCB algorithm, when configured with suitable exploration periods $\tilde{T}$, achieves appropriate $o(1)$ regret bounds when $-\tilde{\Omega}(\log(T)) < p \leq 1$. To establish these bounds, we employ the following assumption on the expected rewards.

Assumption 1. *For all arms $i \in [k]$, we have that the expected reward $\mu_i \geq 32\sqrt{\frac{k \log T \sqrt{\log k}}{T^{1/4}}}$.*

In the context of our asymptotic analysis (as $T \to \infty$), this assumption essentially reduces to a positivity constraint on rewards, i.e., $\mu_i > 0$ for all $i \in [k]$ since we can consider T to be sufficiently large. In addition, this assumption is only required in the regret analysis for $p \leq 0$. We use the standard assumption that the exploration period $\tilde{T}$, and consequently the time horizon T, is sufficiently large compared to k.

Assumption 2. *For all bandit instances with k arms where we learn for T time steps, the choice of exploration period $\tilde{T}$ satisfies $\tilde{T} \geq 8k \log(Tk) + 16\sqrt{\frac{\sqrt{T}}{\log k}}$.*

Note that, as long as T is sufficiently large (say $T \geq 8k^2 \log(k)$), Assumption 2 is satisfied automatically:

(i) Our choices of $\tilde{T}$s for $p \in [-1, 1]$ (see Table 1) satisfy Assumption 2.

(ii) Our choice of $\tilde{T}$ for $p < -1$ satisfies Assumption 2 as long as $|p| \leq \frac{\log T}{2 \log k}$.

Remark 1. *A consequence of assumptions 1 and 2 is that* $\mu_i \geq 128\sqrt{\frac{k\log T}{\tilde{T}}}$ *for each* $i \in [k]$.

Regret Analysis

Analogous to the events used in the analysis of NCB in (Barman et al. 2023), we define 'good events' G_1, G_2 and set $G \triangleq G_1 \cap G_2$. Let $\hat{\mu}_{i,s}$ denote the empirical mean of arm i's rewards *after* seeing s samples from arm i, and let $n_{i,t}$ denote the number of times arm i is pulled in steps $\{1, \ldots, t\}$.

- G_1: Every arm $i \in [k]$ is sampled at least $\frac{\tilde{T}}{2k}$ times in Phase I, i.e. $n_{i,\tilde{T}} \geq \frac{\tilde{T}}{2k}$.
- G_2: For all arms $i \in [k]$ (under Assumption 1), and for all sample counts s such that $\frac{\tilde{T}}{2k} \leq s \leq T$, we have $|\mu_i - \hat{\mu}_{i,s}| \leq 2\sqrt{\frac{\log T}{s}}$.

Here, we represent all the events in the canonical bandit model (Lattimore and Szepesvári 2020). In the following lemma, we show that event G holds with high probability.

Lemma 1. *As long as Assumptions 1 and 2 are satisfied,* $P(G) \geq (1 - \frac{2}{T})$, *where* $G = G_1 \cap G_2$.

We can now prove the following lemma, which is analogous to Lemma 3 in (Barman et al. 2023), but with the UCB bound rather than NCB.

Lemma 2 (UCB correctness). *Let* $\text{UCB}_{i^*,t}$ *be the upper confidence bound of the optimal arm* i^* *after round* t. *Assume that the good event* G *holds. Then, for all rounds* $t > \tilde{T}$ *(i.e., for all rounds in Phase II), we have* $\text{UCB}_{i^*,t} \geq \mu^*$

The good event and the above UCB correctness lemma allow us to prove the following lemma, which is a linchpin of our various analyses and is analogous to Lemma 5 of (Barman et al. 2023), Lemma 8.2 in (Lattimore and Szepesvári 2020) (but significantly stronger) etc.

Lemma 3 (Only good arms in phase two). *Consider a bandit instance that satisfies Assumption 1 and assume that the good event* G *holds. Then, for any arm* i *that is pulled at least once in Phase II, we have*

$$\mu_i \geq \mu^* - 6\sqrt{\frac{\log T}{T_i - 1}},$$

where T_i *is the total number of times that arm* i *is pulled in the algorithm.*

We defer the proof of the above supporting lemmas to the full version of our paper (Krishna et al. 2024). In this paper, we establish upper bounds for p-mean regret across the comprehensive range of $p \in (-\infty, 1]$ in the asymptotic ($T \to \infty$) regime (see Theorems 4, 5, 6).

Regret Analysis of p-Mean Regret for $p < 0$

In this section, we establish regret upper bounds for p-mean regret when p is negative. We rewrite the p-mean regret definition, substituting $q = -p$. That is, we upper bound

$$\text{R}_T^q \triangleq \mu^* - \left(\frac{T}{\sum_{t=1}^{T} \frac{1}{(\mathbb{E}_{I_t}[\mu_{I_t}])^q}}\right)^{\frac{1}{q}}$$

for $q > 0$ and refer to this as q-negative-mean-regret. Taking $q > 0$ and considering q-negative-mean-regret rather than $(-p)$-mean-regret is for notational convenience.

The following theorem establishes an upper bound on the p-mean regret when $p < 0$ employing the EXPLORE-THEN-UCB algorithm. Here, we restate the theorem in terms of q.

Theorem 4. *Given a bandit instance with* k *arms, time horizon* T, *and regret parameter* $q > 0$ *where* $q = |p|$. *By setting the exploration period to* $\tilde{T} = 16\sqrt{\frac{T\log T}{k^q}}$, *the q-negative-mean regret of the Explore-Then-UCB algorithm satisfies*

$$\text{R}_T^q \leq \begin{cases} \tilde{O}\left(\sqrt{\frac{k^{3/2}}{T^{1/2}}}\right) & \text{when } q \in (0, 1] \\ \tilde{O}\left(\sqrt{\frac{k\log(T)}{T^{1/2q}}}\right) & \text{when } q > 1 \end{cases}$$

Proof. Towards analyzing R_T^q, we define

$$x \triangleq \frac{T}{\sum_{t=1}^{\tilde{T}} \frac{1}{\mathbb{E}_{I_t}[\mu_{I_t}]^q}} \text{ and } y \triangleq \frac{T}{\sum_{t=\tilde{T}+1}^{T} \frac{1}{\mathbb{E}_{I_t}[\mu_{I_t}]^q}},$$

so that we have

$$\text{R}_T^q = \mu^* - \left(\frac{1}{\frac{1}{x} + \frac{1}{y}}\right)^{1/q}. \tag{2}$$

Hence, to obtain an upper bound for R_T^q, we need to upper bound $\frac{1}{x}$ and $\frac{1}{y}$. Let us start by focusing on $\frac{1}{x}$. By uniform exploration in Phase I, we have that

$$\mathbb{E}_{I_t}[\mu_{I_t}] \geq \frac{\mu^*}{k} \Leftrightarrow \frac{1}{(\mathbb{E}_{I_t}[\mu_{I_t}])^q} \leq \left(\frac{k}{\mu^*}\right)^q.$$

Hence,

$$\frac{1}{x} \leq \frac{\tilde{T}k^q}{(\mu^*)^q T}. \tag{3}$$

By Jensen's inequality and linearity of expectation

$$\frac{1}{y} \leq \frac{\sum_{t=\tilde{T}+1}^{T} \mathbb{E}_{I_t}\left[\frac{1}{(\mu_{I_t})^q}\right]}{T} = \frac{\mathbb{E}_{I_1,\ldots,I_t}\left[\sum_{t=\tilde{T}+1}^{T} \frac{1}{(\mu_{I_t})^q}\right]}{T}$$

For simplicity, we drop the subscripts in the expectation. By reindexing the arms so that $\{1, 2, \ldots, \ell\}$ are the arms pulled at least once in Phase II, and letting m_i be the number of times (the reindexed) arm i is pulled in Phase II, we have

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=\tilde{T}+1}^{T} \frac{1}{(\mu_{I_t})^q}\right] = \frac{1}{T}\mathbb{E}\left[\sum_{i=1}^{\ell} \frac{m_i}{(\mu_i)^q}\right].$$

By conditioning on the good event G (see Lemma 1) and noting that $\sum_{i\in[\ell]} m_i = T - \tilde{T}$ and $\mu_i \geq \mu_{\min}$, we have

$$\frac{\mathbb{E}\left[\sum_{i=1}^{\ell}\frac{m_i}{(\mu_i)^q}\right]}{T} \leq \frac{\mathbb{E}\left[\sum_i \frac{m_i}{(\mu_i)^q} \mid G\right]\mathbb{P}\{G\} + \frac{T-\tilde{T}}{(\mu_{\min})^q}(1-\mathbb{P}\{G\})}{T}$$

We have $(1 - \mathbb{P}\{G\}) \leq \frac{2}{T}$ from Lemma 1. Hence, $\frac{T-\tilde{T}}{(\mu_{\min})^q}(1 - \mathbb{P}\{G\}) \leq \frac{2}{(\mu_{\min})^q}$. Thus,

$$\frac{1}{y} \leq \frac{\mathbb{E}\left[\sum_{i=1}^{\ell}\frac{m_i}{(\mu_i)^q} \mid G\right] \cdot \mathbb{P}\{G\}}{T} + \frac{2}{(\mu_{\min})^q T}. \tag{4}$$

Consider the first term in RHS. Conditioned on event G, by Lemma 3, we have $\mu_i \geq \mu^* - \beta_i$ for all arms $i \in [\ell]$ pulled in Phase II — where $\beta_i \triangleq 6\sqrt{\frac{\log T}{T_i - 1}}$. Note that, conditioned on the good event G (specifically G_1), we have

$$\beta_i = 6\sqrt{\frac{\log T}{T_i - 1}} \leq 6\sqrt{\frac{2k \log T}{\tilde{T}}} \triangleq \beta$$

for any arm $i \in [\ell]$. Hence, we have

$$\frac{\mathbb{E}[\sum_{i=1}^{\ell} \frac{m_i}{(\mu_i)^q} \mid G]\mathbb{P}\{G\}}{T} \leq \frac{\mathbb{E}[\sum_{i=1}^{\ell} \frac{m_i}{(\mu^* - \beta)^q}]}{T},$$

using $\mathbb{P}(G) \leq 1$ and $\mu_i \geq \mu^* - \beta$. Note that Assumption 1 implies that $\mu_i - \beta > 0$ for all $i \in [k]$, and in particular that $\mu^* - \beta > 0$. Also, from the assumptions we can see that $\frac{\beta}{\mu^*} \leq \frac{1}{2}$. Note that we have $(\mu^* - \beta)^q \geq (\mu^*)^q - q\beta(\mu^*)^{q-1}$ for $q > 0$. Combining this with (3) and (4) we get

$$\frac{1}{\frac{1}{x} + \frac{1}{y}} \geq \frac{1}{\frac{\tilde{T}k^q}{(\mu^*)^q T} + \frac{\mathbb{E}[\sum_{i=1}^{\ell} m_i]}{T((\mu^*)^q - q\beta(\mu^*)^{q-1})} + \frac{c}{T}}$$
$$\text{(taking } \frac{2}{(\mu_{\min})^q} \triangleq c\text{)}$$
$$\geq \frac{(\mu^* - \beta)T}{k^p\tilde{T} + (T - \tilde{T}) + c} \triangleq (*_1)$$
$$\text{(cancelling } \mu^* \text{ and using } c \geq c(\mu^*)^q\text{)}$$

Continuing, we can divide the numerator and denominator by T and rearrange the terms to get

$$(*_1) = \frac{((\mu^*)^q - q\beta(\mu^*)^{q-1})}{(k^q \frac{\tilde{T}}{T} + \frac{(T - \tilde{T})}{T} + \frac{c}{T})} = \frac{((\mu^*)^q - q\beta(\mu^*)^{q-1})}{\left(1 + \frac{(k^q - 1)\tilde{T} + c}{T}\right)}$$
$$= \frac{((\mu^*)^q - q\beta(\mu^*)^{q-1})\left(1 - \frac{(k^q - 1)\tilde{T} + c}{T}\right)}{\left(1 - \left(\frac{(k^q - 1)\tilde{T} + c}{T}\right)^2\right)} \triangleq (*_2)$$

In the last step, we multiply both numerator and denominator by $\left(1 - \frac{(k^q - 1)\tilde{T} + c}{T}\right) \in (0, 1)$. Thus, $1 - \left(\frac{(k^q - 1)\tilde{T} + c}{T}\right)^2 \in (0, 1)$ and we get

$$(*_2) \geq (\mu^*)^q \left(1 - \frac{q\beta}{\mu^*}\right)\left(1 - \frac{(k^q - 1)\tilde{T} + c}{T}\right) \triangleq (*_3)$$

When $q \geq 1$ We can expand $(*_3)$ to get

$$\frac{1}{\frac{1}{x} + \frac{1}{y}} \geq (\mu^*)^q \left(1 - \frac{(k^q - 1)\tilde{T} + c}{T} - \frac{q\beta}{\mu^*} + \frac{q\beta((k^q - 1)\tilde{T} + c)}{\mu^* T}\right)$$
$$\geq (\mu^*)^q \left(1 - \frac{(k^q - 1)\tilde{T} + c}{T} - \frac{q\beta}{\mu^*}\right)$$
$$= (\mu^*)^q \left(1 - \frac{(k^q - 1)\tilde{T} + c}{T} - \frac{6q\sqrt{2k \log T}}{\sqrt{\tilde{T}}\mu^*}\right)$$

Thus, we have (for $q > 1$), that $\left(\frac{1}{\frac{1}{x} + \frac{1}{y}}\right)^{\frac{1}{q}}$ is

$$\geq \left((\mu^*)^q - (\mu^*)^q \left(\frac{(k^q - 1)\tilde{T} + c}{T} + \frac{6q\sqrt{2k \log T}}{\sqrt{\tilde{T}}\mu^*}\right)\right)^{\frac{1}{q}}$$
$$\geq \mu^* - \mu^* \left(\frac{(k^q - 1)\tilde{T} + c}{T} + \frac{6q\sqrt{2k \log T}}{\sqrt{\tilde{T}}\mu^*}\right)^{1/q}$$

The last inequality uses the fact that $(a - b)^{1/q} \geq a^{1/q} - b^{1/q}$ for all $a \geq b \geq 0$ (from binomial expansion) as long as $q \geq 1$. Then, we have from (2) that

$$R_T^q \leq \mu^* \left(\frac{(k^q - 1)\tilde{T} + c}{T} + \frac{6q\sqrt{2k \log T}}{\sqrt{\tilde{T}}\mu^*}\right)^{1/q}.$$

Since $\tilde{T} = 16\sqrt{\frac{T \log T}{k^q}}$, we can substitute the value to get

$$R_T^q \leq \mu^* \left(\frac{16(k^q - 1)\sqrt{T \log T}}{\sqrt{k^q}T} + \frac{c}{T} + \frac{6q\sqrt{2k \log T}k^{q/4}}{4(T \log T)^{1/4}\mu^*}\right)^{\frac{1}{q}}$$
$$\leq \tilde{O}\left(\frac{k^{\frac{1}{2}}}{T^{\frac{1}{4q}}}\right). \qquad \text{(since } \frac{1}{4} + \frac{1}{2q} \leq \frac{1}{2} \text{ for } q \geq 1\text{)}$$

When $0 < q < 1$ From $(*_3)$, we get

$$\left(\frac{1}{\frac{1}{x} + \frac{1}{y}}\right)^{\frac{1}{q}} \geq (\mu^*)\left(1 - \frac{q\beta}{\mu^*}\right)^{\frac{1}{q}}\left(1 - \frac{(k^q - 1)\tilde{T} + c}{T}\right)^{\frac{1}{q}} \triangleq (*_4)$$

Since $0 < q \leq 1$, $\frac{q\beta}{\mu^*} \leq \frac{1}{2}$ and $\frac{(k^q - 1)\tilde{T} + c}{T} \leq \frac{1}{2}$, by Weierstrass inequality (Kozma 2021), we get

$$(*_4) \geq (\mu^*)\left(1 - \frac{\beta}{\mu^*} - \frac{(k^q - 1)\tilde{T} + c}{qT}\right)$$
$$\geq \mu^* - \mu^*\left(\frac{\beta}{\mu^*} + \frac{(k^q - 1)\tilde{T}}{qT} + \frac{c}{qT}\right)$$

Since $\tilde{T} = 16\sqrt{\frac{T \log T}{k^q}}$, we can substitute the value to get

$$R_T^q \leq \mu^*\left(\frac{\beta}{\mu^*} + \frac{(k^q - 1)\tilde{T} + c}{qT}\right)$$
$$\leq \frac{6\sqrt{2k \log T}}{\sqrt{\tilde{T}}} + \frac{(k^q - 1)\tilde{T}}{qT} + \frac{c}{qT}$$
$$\leq \frac{6\sqrt{2k \log T}k^{q/4}}{4(T \log T)^{1/4}} + \frac{(k^q - 1)16\sqrt{\log T}}{k^{\frac{q}{2}}q\sqrt{T}} + \frac{c}{qT}$$
$$\leq \tilde{O}\left(\frac{k^{\frac{3}{4}}}{T^{\frac{1}{4}}}\right) (\max\{\frac{1}{2} + \frac{q}{4}, \frac{q}{2}\} \leq \frac{3}{4}; q \geq 4/(kT)^{\frac{3}{4}})$$

$\square$

Remark 2. *Our bound for p-mean regret R_T^p when $p < -1$ is $\lesssim \frac{\sqrt{k \log T}}{T^{1/(4|p|)}}$. For any $\kappa > 0$, this bound will be at least κ when $|p| \geq \frac{\log T}{4(\log(\sqrt{k \log T}) + \log(\frac{1}{\kappa}))} \gtrsim \frac{\log T}{\log \log T + \log k + \log(1/\kappa)}$. This provides the transition point when the algorithm stops achieving vanishing regret.*

Regret Analysis for p-Mean Regret for $0 < p \leq 1$

In this section, we establish regret upper bounds for p-mean regret when p is between 0 and 1. That is, we upper bound

$$\mathrm{R}_T^p = \mu^* - \left(\frac{\sum_{t=1}^T (\mathbb{E}_{I_t}[\mu_{I_t}])^p}{T} \right)^{\frac{1}{p}}$$

for $p \in (0, 1]$. The following theorem establishes an $\tilde{O}\left(\sqrt{k/T}\right)$ upper bound on p-mean regret when using the EXPLORE-THEN-UCB algorithm.

Theorem 5. *Given a bandit instance with K arms, time horizon T, and regret parameter $p \in (0, 1]$, choosing the exploration period $\tilde{T} = 16\sqrt{\frac{Tk^p \log T}{\log k}}$, the p-mean regret of the* EXPLORE-THEN-UCB *algorithm satisfies*

$$\mathrm{R}_T^p \leq \tilde{O}\left(\sqrt{\frac{k}{T}} \right).$$

Proof Sketch. Towards analyzing R_T^p, we define

$$x \triangleq \frac{\sum_{t=1}^{\tilde{T}} \mathbb{E}_{I_t}[\mu_{I_t}]^p}{T} \text{ and } y \triangleq \frac{\sum_{t=\tilde{T}+1}^{T} \mathbb{E}_{I_t}[\mu_{I_t}]^p}{T},$$

so that we have

$$\mathrm{R}_T^p = \mu^* - (x + y)^{1/p}. \tag{5}$$

To obtain an upper bound for R_T^p, we need to lower bound x and y. We show that $x \geq \frac{\mu^{*p}\tilde{T}}{k^p T}$ since we use uniform exploration in Phase I. By reindexing the arms so that $\{1, 2, \ldots, \ell\}$ are the arms pulled at least once in Phase II, and letting m_i be the number of times (the reindexed) arm i is pulled in Phase II, we have $y \geq \frac{\mathbb{E}[\sum_{i=1}^{\ell} m_i \mu_i^p \mid G]\mathbb{P}\{G\}}{T}$. Using these facts and simplifying further, as in the case of $p < 0$ we get

$$(x + y)^{\frac{1}{p}} \geq \mu^* - \frac{(k^p - 1)\tilde{T}'}{p\sqrt{k}T} - \frac{6\sqrt{k \log T}}{\sqrt{T}} - \frac{4}{pT}$$

where $\tilde{T}' = \frac{\tilde{T}}{k^{p-0.5}}$. Then, substituting for $\tilde{T}$,

$$\begin{aligned}
\mathrm{R}_T^p &= \mu^* - (x + y)^{1/p} \\
&\leq \frac{(k^p - 1)\tilde{T}'}{p\sqrt{k}T} + \frac{6\sqrt{k \log T}}{\sqrt{T}} + \frac{4}{pT} \\
&= \frac{(k^p - 1)\tilde{T}}{pk^p T} + \frac{6\sqrt{k \log T}}{\sqrt{T}} + \frac{4}{pT} \\
&= \frac{16(k^p - 1)\sqrt{Tk^p \log(T)}}{pk^p T\sqrt{\log k}} + \frac{6\sqrt{k \log T}}{\sqrt{T}} + \frac{4}{pT} \\
&\leq \tilde{O}\left(\sqrt{\frac{k}{T}} \right) \qquad\qquad (\text{for } p \geq \tfrac{4}{\sqrt{kT}})
\end{aligned}$$

We also provide a bound for Nash regret NR_T, as studied in (Barman et al. 2023), achieving essentially the same regret bound as theirs but under our assumptions and using

the EXPLORE-THEN-UCB algorithm instead of their NCB algorithm, which employs a different confidence bound following the exploration phase. The proof is in the full version of our paper (Krishna et al. 2024).

Theorem 6. *Given a bandit instance with k arms and for time horizon T, choosing the exploration period $\tilde{T} = 16\sqrt{\frac{Tk \log T}{\log k}}$, the Nash regret of the* EXPLORE-THEN-UCB *algorithm satisfies*

$$\mathrm{NR}_T \leq \tilde{O}\left(\sqrt{\frac{k}{T}} \right).$$

Conclusion

Building on the p-mean welfare concept from social choice theory, this work examined p-mean regret—a flexible metric that allows the decision maker to effectively balance fairness and efficiency considerations. We proposed a unified EXPLORE-THEN-UCB algorithm that achieves p-mean regret bounds across a wide range of p values. Specifically, our analysis demonstrates that for $p \leq -1$, the p-mean regret of our algorithm scales as $\tilde{O}(k^{\frac{1}{2}} T^{-\frac{1}{4|p|}})$, while for $0 \leq p \leq 1$, the regret scales as $\tilde{O}(k^{\frac{1}{2}} T^{-\frac{1}{2}})$. This unified approach simplifies the design and analysis of bandit algorithms, particularly when compared to prior work that required specialized techniques for the Nash regret case ($p \to 0$). We have also performed some synthetic experiments to compare the p-mean-regrets achieved by EXPLORE-THEN-UCB versus NCB and UCB1 as baselines; these results can be found in the extended version (Krishna et al. 2024).

Several promising directions for future work emerge from our findings. Extending the analysis of p-mean regret to other bandit settings, such as linear and contextual bandits, could provide deeper insights into the interplay between fairness, efficiency, and bandit structure. Developing tight meta-algorithms that replace the UCB1 subroutine with other average-regret minimizing algorithms could be another interesting direction for future work. Extending the work for anytime guarantees and unknown time horizons would also improve the practical applicability of our work. By continuing to explore the connections between fairness, efficiency, and bandit learning, the study of p-mean regret contributes to the development of more socially responsible and widely applicable decision-making algorithms.

Acknowledgements

References

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47: 235–256.

Barman, S.; Bhaskar, U.; Krishna, A.; and Sundaram, R. G. 2020. Tight approximation algorithms for p-mean welfare under subadditive valuations. *arXiv preprint arXiv:2005.07370*.

Barman, S.; Khan, A.; and Maiti, A. 2022. Universal and tight online algorithms for generalized-mean welfare. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36(5), 4793–4800.

Barman, S.; Khan, A.; Maiti, A.; and Sawarni, A. 2023. Fairness and welfare quantification for regret in multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37(6), 6762–6769.

Bubeck, S.; Cesa-Bianchi, N.; et al. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1): 1–122.

Bubeck, S.; Munos, R.; and Stoltz, G. 2009. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory: 20th International Conference, ALT 2009, Porto, Portugal, October 3-5, 2009. Proceedings 20*, 23–37. Springer.

Celis, L. E.; Kapoor, S.; Salehi, F.; and Vishnoi, N. 2019. Controlling polarization in personalization: An algorithmic framework. In *Proceedings of the conference on fairness, accountability, and transparency*, 160–169.

Clement, B.; Roy, D.; Oudeyer, P.-Y.; and Lopes, M. 2013. Multi-armed bandits for intelligent tutoring systems. *arXiv preprint arXiv:1310.3174*.

Eckart, O.; Psomas, A.; and Verma, P. 2024. On the Fairness of Normalized p-Means for Allocating Goods and Chores. *arXiv preprint arXiv:2402.14996*.

Garg, J.; Husić, E.; Murhekar, A.; and Végh, L. 2021. Tractable fragments of the maximum nash welfare problem. *arXiv preprint arXiv:2112.10199*.

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.

Jones, M.; Nguyen, H.; and Nguyen, T. 2023. An efficient algorithm for fair multi-agent multi-armed bandit with low regret. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37(7), 8159–8167.

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.

Kaneko, M.; and Nakamura, K. 1979. The Nash social welfare function. *Econometrica: Journal of the Econometric Society*, 423–435.

Kozma, L. 2021. Useful Inequalities.

Krishna, A.; George John, P.; Barik, A.; and Tan, V. Y. F. 2024. p-Mean Regret for Stochastic Bandits. *arXiv preprint arxiv:2412.10751*.

Lai, T. L.; and Robbins, H. 1985. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1): 4–22.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit algorithms*. Cambridge University Press.

Mandal, D.; and Gan, J. 2022. Socially fair reinforcement learning. *arXiv preprint arXiv:2208.12584*.

Moulin, H. 2004. *Fair division and collective welfare*. MIT press.

Patil, V.; Ghalme, G.; Nair, V.; and Narahari, Y. 2021. Achieving fairness in the stochastic multi-armed bandit problem. *Journal of Machine Learning Research*, 22(174): 1–31.

Sawarni, A.; Pal, S.; and Barman, S. 2024. Nash regret guarantees for linear bandits. *Advances in Neural Information Processing Systems*, 36.

Schwartz, E. M.; Bradlow, E. T.; and Fader, P. S. 2017. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4): 500–522.

Thompson, W. R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4): 285–294.

Villar, S. S.; Bowden, J.; and Wason, J. 2015. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2): 199.

Zhang, M.; Vuong, R. D.-C.; and Luo, H. 2024. No-Regret Learning for Fair Multi-Agent Social Welfare Optimization. *arXiv preprint arXiv:2405.20678*.