# Revisiting Social Welfare in Bandits: UCB Is (Nearly) All You Need

**Dhruv Sarkar**
Indian Institute of Technology Kharagpur
dhruvsarkar@kgpian.iitkgp.ac.in

**Nishant Pandey**
Indian Institute of Technology Kanpur
nishantp22@iitk.ac.in

**Sayak Ray Chowdhury**
Indian Institute of Technology Kanpur
sayakrc@iitk.ac.in

## Abstract

Regret in stochastic multi-armed bandits traditionally measures the difference between the highest reward and either the arithmetic mean of accumulated rewards or the final reward. These conventional metrics often fail to address fairness among agents receiving rewards, particularly in settings where rewards are distributed across a population, such as patients in clinical trials. To address this, a recent body of work has introduced Nash regret, which evaluates performance using the geometric mean of accumulated rewards, aligning with the Nash social welfare function, which satisfies fairness axioms.

To minimize Nash regret, existing approaches require specialized algorithm designs and strong assumptions, such as multiplicative concentration inequalities and bounded, non-negative rewards, making them unsuitable for even Gaussian reward distributions. We demonstrate that an initial uniform exploration phase followed by a standard Upper Confidence Bound (UCB) algorithm achieves near-optimal Nash regret while relying only on additive Hoeffding bounds and naturally extends to sub-Gaussian rewards. Furthermore, we generalize the algorithm to a broad class of fairness metrics called the $p$-mean regret, proving (nearly) optimal regret bounds uniformly across all $p$ values. This is in contrast to prior work, which made extremely restrictive assumptions about the bandit instances and, even then, achieved suboptimal regret bounds. Numerical simulations validate our method's practical efficacy, broadening the accessibility of fairness in bandit algorithms.

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. 

## 1 INTRODUCTION

The multi-armed bandit (MAB) problem is a foundational framework for sequential decision-making under uncertainty, with applications that span healthcare, advertising, education, and beyond. In this setup, a decision-maker sequentially selects from a set of arms $[k] := \{1, 2, ...k\}$ – each having reward distribution $\rho_i$ with unknown mean $\mu_i \in \mathbb{R}$ – across a time horizon $T$, aiming to minimize regret, which quantifies the loss incurred by not always choosing the arm with the highest mean reward $\mu^* = \max_{i\in[k]} \mu_i$. Traditionally, regret is measured as the difference between $\mu^*$ and the arithmetic mean of accumulated rewards $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\mu_{I_t}]$, where $I_t$ is the arm selected at $t$-th round depending on past (random) observations. This definition of (average) regret often overlooks fairness considerations, particularly in settings where rewards correspond to values accruing to a population of users, such as patients in clinical trials. While average regret maximizes social welfare, as defined by the average reward, it may still permit significant disparities across users.

To address these fairness limitations, Barman et al. (2023) introduce Nash regret—a strengthened notion based on the Nash Social Welfare (NSW) function – defined using the geometric mean of expected rewards $\left(\prod_{t=1}^{T} \mathbb{E}[\mu_{I_t}]\right)^{\frac{1}{T}}$ – a function well-known in economics for satisfying key fairness axioms (Moulin, 2004). By

minimizing Nash regret, bandit algorithms promote equitable outcomes that balance efficiency and fairness, ensuring that users across all rounds benefit fairly from the decision-making process while still enabling exploration to identify optimal arms.

While Upper Confidence Bound (UCB, Bubeck et al. (2012)) is a classical index-based strategy to minimize average regret, Barman et al. (2023) argues it doesn't suffice to minimize Nash regret. They propose a new index, Nash Confidence Bound (NCB) $\widehat{\mu}_i + \sqrt{\frac{2\hat{\mu}_i \log T}{n_i}}$, where the confidence width depends on $\widehat{\mu}_i$–an unbiased estimate of $\mu_i$ from $n_i$ independent samples (e.g. sample mean). To show that NCBs are optimistic estimates of unknown true means, they resort to multiplicative Hoeffding/Chernoff bounds, which put a rather restrictive assumption on reward distributions– each $p_i$ to have support on a non-negative interval in $\mathbb{R}$. It also implicitly puts a cap on the value of $\mu^*$ and requires their algorithm to know it. Instead, it is desirable to have algorithms that work under generic reward distributions (e.g., Gaussian) and don't require any upper bound on $\mu^*$.

Krishna et al. (2025) study a more general class of fairness metric, the $p$-mean welfare – defined using the generalized power mean of expected rewards $\left(\frac{1}{T}\sum_{t=1}^{T}(\mathbb{E}[\mu_{I_t}])^p\right)^{\frac{1}{p}}$ – a function with roots in social choice theory (Moulin, 2004). By changing the value of the parameter $p \in (-\infty, 1]$, this single function can be made to behave like a utilitarian function (focused on maximizing total utility), a Rawlsian function (focused on the worst-off individual), or an intermediate function like Nash social welfare. This generalized metric thus helps us study all the social objectives under a single umbrella, based on the value of $p$.

To minimize $p$-mean regret (which includes Nash regret as a special case), Krishna et al. (2025) employs the UCB index. Their results depend on a restrictive assumption that every arm has an expected reward of at least the order $\sqrt{k}/T^{\frac{1}{8}}$. This is in stark contrast to a counter-example presented in Barman et al. (2023) that demonstrates the failure of UCB in minimizing Nash regret. It considers two Bernoulli arms with means $\mu_1 = (2e)^{-T}$ and $\mu_2 = 1$. The assumption of the minimum expected reward excludes this example, significantly limiting the applicability of their algorithm to bandit instances like this. Moreover, they assume bounded and non-negative rewards, as in Barman et al. (2023), further narrowing the scope.

We address these shortcomings in prior work by introducing a systematic framework to minimize both Nash and $p$-mean regrets. Our study demonstrates that the UCB index, combined with a data-adaptive initial exploration step, is sufficient to achieve non-trivial, nearly optimal regret bounds without any restrictive assumptions on the reward distributions and their (unknown) means. More specifically, we make the following contributions:

1. We introduce a reduction framework that enables us to minimize Nash regret via a short adaptive uniform exploration phase, followed by the execution of a standard bandit algorithm, such as UCB. Our data-adaptive stopping rule for the exploration phase is key to facilitating this reduction, thereby demonstrating the UCB index's versatility in minimizing Nash regret. Moreover, the reduction seamlessly adapts to all possible values of the fairness parameter $p$ (for Nash regret, $p = 0$). We utilize this insight to design Welfarist-UCB, a novel bandit algorithm that minimizes both Nash and $p$-mean regrets.

2. To bound Nash/$p$-mean regret of Welfarist-UCB, we work with the (additive) Hoeffding inequality instead of the multiplicative one (which is used in Barman et al. (2023)). This helps us sidestep restrictive (e.g., bounded, non-negative) assumptions on the rewards that multiplicative bounds often require. Notably, multiplicative bounds tend to be inapplicable or significantly looser in broader settings, such as sub-Gaussian distributions. By relying on additive Hoeffding bounds, our algorithm naturally works with sub-Gaussian rewards and doesn't require an upper bound on the optimal reward $\mu^*$.

3. (a) We prove that Welfarist-UCB attains $\widetilde{O}\left(\sigma\sqrt{\frac{k}{T}}\right)$ upper bound on Nash regret for $\sigma$-sub-Gaussian rewards. This bound is order-optimal and includes the result of Barman et al. (2023) for bounded, non-negative rewards as a special case.

   (b) For $p \in [0, 1]$, we obtain an order-optimal $p$-mean regret of $\widetilde{O}\left(\sigma\sqrt{\frac{k}{T}}\right)$ for Welfarist-UCB, which not only includes the bound of Krishna et al. (2025) for bounded, non-negative rewards as a sub-case but also gets rid of their unrealistic assumption of each $\mu_i \geq \widetilde{\Omega}\left(\sqrt{k}T^{-1/8}\right)$.

   (c) For any $p < 0$, we prove that $p$-mean regret of Welfarist-UCB is $\widetilde{O}\left(\frac{\sigma k^{\frac{|p|+1}{2}}}{\sqrt{T}} \cdot \max\{1, |p|\}\right)$. When $p \geq -1$, the regret can be further bounded by $\widetilde{O}\left(k/\sqrt{T}\right)$, which is tighter than the $\widetilde{O}\left(k^{3/4}T^{-\frac{1}{4}}\right)$ bound of Krishna et al. (2025) since $T > k$.

   (d) As $p$ becomes more negative (e.g., $p < -1$), our regret bound grows exponentially w.r.t. $|p|$ due to stricter fairness requirements while keeping the asymptotic scale w.r.t. time horizon $T$

| Reference | $p \in [0,1]$ | $p \in [-1,0)$ | $p < -1$ | Assumptions |
|---|---|---|---|---|
| Barman et al. (2023) | $\widetilde{O}\left(\sqrt{\frac{k}{T}}\right)$ | - | - | Rewards bounded in $[0,1]$ |
| Krishna et al. (2025) | $\widetilde{O}\left(\sqrt{\frac{k}{T}}\right)$ | $\widetilde{O}\left(k^{\frac{3}{4}}T^{\frac{-1}{4}}\right)$ | $\widetilde{O}\left(k^{\frac{1}{2}}T^{\frac{-1}{4\lvert p\rvert}}\right)$ | Rewards bounded in $[0,1]$<br>$\mu_i \geq 32\sqrt{\frac{k\log T\sqrt{\log k}}{T^{1/4}}}\ \forall\ i \in [k]$ |
| Ours | $\widetilde{O}\left(\sqrt{\frac{k}{T}}\right)$ | $\widetilde{O}\left(\frac{k^{\frac{\lvert p\rvert+1}{2}}}{\sqrt{T}}\right)$ | $\widetilde{O}\left(\frac{\lvert p\rvert k^{\frac{\lvert p\rvert+1}{2}}}{\sqrt{T}}\right)$ | Sub-Gaussian rewards<br>$\mu_i \geq 0\ \forall\ i \in [k]$ |

*Table 1: Summary of regret bounds for different $p \in (-\infty, 1]$. Our results not only do away with restrictive assumptions of prior work but also significantly improve the regret bounds to achieve order-optimality.*

same. In the extreme case when $p \to -\infty$ (i.e., when the regret is Rawlsian), the bound becomes vacuous unless $T > O\big(p^2 k^{|p|}\big)$, essentially highlighting a "no-free-lunch" principle in $p$-mean regret minimization. Krishna et al. (2025) avoids the exponential scaling w.r.t. $p$ at the expense of a worse $T^{-\frac{1}{4|p|}}$ scaling w.r.t. $T$ and resorting to a restrictive assumption of $\mu_i \geq \widetilde{\Omega}\big(\frac{\sqrt{k}}{T^{1/8}}\big)$ for all $i$.

(e) We validate the practical effectiveness of our approach through numerical simulations, comparing our algorithm with prior work across different values of the fairness parameter $p$ and demonstrating its utility in fairness-aware sequential decision-making. The theoretical results and comparisons are summarized in Table 1.

A comprehensive survey on other related works is presented in Section A of the Appendix.

## 2 PRELIMINARIES

In the stochastic multi-armed bandit setting, the learner (algorithm) has access to $k$ probability distributions $\rho_1, \ldots, \rho_k$ (referred to as arms). Upon pulling an arm $i \in [k] := \{1, \ldots, k\}$, the learner observes a (random) reward $R_i$ sampled independently from $\rho_i$. We assume that each $R_i$ is sub-gaussian with mean $\mu_i$ and variance-proxy $\sigma^2$, i.e.,

$$\mathbb{E}\left[R_i\right] = \mu_i\ ,\ \mathbb{E}\left[\exp\left(\lambda(R_i - \mu_i)\right)\right] \leq \exp\left(\frac{\sigma^2\lambda^2}{2}\right)\ .$$

We assume that the mean rewards are non-negative, i.e., $\mu_i \geq 0$ for all arms $i \in [k]$.

**Remark 1.** *This assumption is standard in literature (Barman et al., 2023; Sawarni et al., 2024; Krishna et al., 2025), motivated by social welfare applications such as clinical trials or resource allocation, where candidate arms are typically pre-screened to be non-harmful on average. The goal is thus to identify the most beneficial option among safe alternatives. Moreover, fairness metrics like Nash Social Welfare, based on the geometric mean, can only be defined for non-negative means. However, this does not preclude observing negative individual rewards: for example, a treatment may be beneficial on average ($\mu_i \geq 0$) yet can cause negative side effects in some instances. Our use of additive concentration bounds allows the model to handle these individual negative outcomes gracefully, even while the assumption of non-negative means remains essential for the coherence of the welfare functions and the subsequent regret analysis.*

The learning process unfolds over $T \geq 1$ rounds, where each round corresponds to a (distinct) user. At each round $t \in [T]$, the algorithm pulls an arm $I_t \in [k]$ and observes a reward $R_t \sim \rho_{I_t}$, where $I_t$ depends on the arm pulls and (random) observed rewards till round $t-1$. The sequence of expected rewards $\mathbb{E}[\mu_{I_t}], t \in [T]$, can be mapped to a social welfare value using a function $f : \mathbb{R}^T \to \mathbb{R}$ and then subtracted from the maximum welfare $\mu^*$ to quantify the algorithm's performance (regret) across $T$ users. When $f$ returns the arithmetic mean of expected rewards, we get the average regret $\mathrm{AR}_T := \mu^* - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mu_{I_t}]$. When $f$ returns the geometric mean of expected rewards, we get the Nash regret

$$\mathrm{NR}_T := \mu^* - \left(\prod\nolimits_{t=1}^{T}\mathbb{E}[\mu_{I_t}]\right)^{\frac{1}{T}}\ .$$

By AM-GM inequality, $\mathrm{NR}_T \geq \mathrm{AR}_T$, and thus minimizing Nash regret is harder compared to average regret. When $f$ returns the generalized power mean of expected rewards, we get, for $p \in \mathbb{R}$, the $p$-mean regret

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

Unlike average and Nash regrets, the $p$-mean regret captures the complete spectrum of interplay between fairness and utility.

For $p > 1$, the function $f$ (generalized power mean) puts more emphasis on rounds with high rewards, re-

flecting a more utilitarian perspective. In the extreme case when $p \to \infty$, $f$ returns the maximum of expected rewards $\max_{t=1}^{T} \mathbb{E}[\mu_{I_t}]$, which focuses on the best-off individual only and does not provide any fairness guarantee. For $p = 1$, $p$-mean regret coincides with the average regret, focusing on the average utility across $T$ individuals. For $p < 1$, the focus shifts to fairness. The function $f$ satisfies the Pigou-Dalton transfer principle, which ensures that transferring a small amount of reward from a well-off individual to another one with lower utility increases the overall welfare (Moulin, 2004). For $p = 0$, $p$-mean regret coincides with the Nash regret, focusing on average welfare across $T$ individuals. As $p$ decreases further, the function puts more emphasis on rounds with low rewards. In the extreme case when $p \to -\infty$, $f$ returns the minimum of expected rewards $\min_{t=1}^{T} \mathbb{E}[\mu_{I_t}]$, which focuses on the worst-off individual only and does not provide any utility guarantee. In summary, $p$ can be viewed as a parameter trading off fairness and utility, with the region of interest $p \leq 1$, where a smaller value of $p$ ensures more fairness and vice versa.

## 3 ALGORITHM AND RESULTS

In this section, we introduce our algorithm (Welfarist-UCB) that decomposes Nash/$p-$mean regret minimization into two phases: (I) Uniform exploration over a data-adaptive horizon, and (II) Explore-exploit optimization using the UCB index.

**Phase I (Uniform Exploration):** In this phase, at the beginning of each block of $k$ steps, we draw a uniform random permutation $\pi \in \Pi_k$ of the arms, where $\Pi_k$ denotes the set of all $k!$ permutations of $\{1, 2, \ldots, k\}$. We then select the arms sequentially in the order $\pi(1), \pi(2), \ldots, \pi(k)$. After $k$ steps, the permutation is exhausted, and a new independent uniform permutation is drawn for the next block.

Phase I continues until the accumulated reward of some arm $i$ exceeds an adaptive threshold. Formally, termination occurs at the first time $t$ such that for some arm $i \in [k]$, $\widehat{\mu}_i > 2\sqrt{\frac{2\sigma^2 \log T}{n_i}}$ and

$$n_i > \frac{192 p^2 \sigma^2 \log T}{\left(\widehat{\mu}_i - 2\sqrt{\frac{2\sigma^2 \log T}{n_i}}\right)^2}.$$

Here $n_i$ denotes the number of pulls of arm $i$ up to time $t$, and $\widehat{\mu}_i$ denotes the empirical mean of its observed rewards. $\sigma^2$ is the variance-proxy of sub-Gaussian rewards, and $p$ is the fairness parameter provided as input. We normalize it by setting $p = 1$ whenever $p \geq -1$, since for $p \geq -1$ case, the analysis for Phase I termination condition does not depend on $p$ (the careful choice of numeric inequalities in our analysis for $p \in [-1, 0)$ does away with the involvement of $p$, whereas for $p > 0$, the generalised mean inequality trivially helps us avoid $p$). The second terminating condition is critical because it ensures, with high probability, that Phase I ends after $\Theta\left(\frac{1}{(\mu^*)^2}\right)$ rounds, which, in turn, enables us to work with the UCB index in Phase II (see Remark 6). The additional constraint $\widehat{\mu}_i > 2\sqrt{\frac{2\sigma^2 \log T}{n_i}}$ ensures the condition is meaningful, ruling out cases where the denominator in the threshold expression is non-positive.

**Algorithm 1** Welfarist UCB

**Input:** Number of arms $k$, time horizon $T$, fairness measure $p$ and reward variance-proxy $\sigma^2$.

Initialize empirical means $\widehat{\mu}_i = 0$ and counts $n_i = 0$ for all $i \in [k]$, round index $t = 1$.
**if** $p \geq -1$ **then**
  Set $p \leftarrow 1$.
**end if**
`Phase I`
**while** for all $i \in [k]$, $\widehat{\mu}_i \leq 2\sqrt{\frac{2\sigma^2 \log T}{n_i}}$ **or** $n_i \leq 192 p^2 \sigma^2 \frac{\log T}{\left(\widehat{\mu}_i - 2\sqrt{\frac{2\sigma^2 \log T}{n_i}}\right)^2}$ **do**
  **if** $t \bmod k = 1$ **then**
    Draw a permutation $\pi$ uniformly at random from the set of all permutations of $[k]$.
  **end if**
  Pull arm $I_t = \pi\left(1 + (t-1) \bmod k\right)$.
  Observe reward $R_t \sim \rho_{I_t}$.
  Update $n_{I_t} \leftarrow n_{I_t} + 1$, $\widehat{\mu}_{I_t} \leftarrow \frac{(n_{I_t}-1)\widehat{\mu}_{I_t}}{n_{I_t}} + \frac{R_t}{n_{I_t}}$.
  Update $t \leftarrow t + 1$.
**end while**
`Phase II`
**while** $t \leq T$ **do**
  Pull arm $I_t = \arg\max_{i \in [k]} \left\{\widehat{\mu}_i + 2\sqrt{\frac{2\sigma^2 \log T}{n_i}}\right\}$.
  Observe reward $R_t \sim \rho_{I_t}$.
  Update $n_{I_t} \leftarrow n_{I_t} + 1$, $\widehat{\mu}_{I_t} \leftarrow \frac{(n_{I_t}-1)\widehat{\mu}_{I_t}}{n_{I_t}} + \frac{R_t}{n_{I_t}}$.
  Update $t \leftarrow t + 1$.
**end while**

**Phase II (Explore-exploit with UCB):** In this phase, we employ the UCB index $\widehat{\mu}_i + 2\sqrt{\frac{2\sigma^2 \log T}{n_i}}$ to pull arms for $\sigma$-sub-Gaussian rewards. For comparison, Barman et al. (2023) employs the Nash Confidence Bound (NCB) index $\hat{\mu}_i + 4\sqrt{\frac{\widehat{\mu}_i \log T}{n_i}}$ for bounded $[0, 1]$ rewards to minimize Nash regret. Our results show that the simpler UCB rule, along with the data-adaptive uniform exploration phase, suffices to achieve order-optimal bounds on Nash regret. Moreover, this strategy achieves the best known bounds for $p-$mean

regret across different regimes of the fairness parameter $p$. The pseudo-code of the strategy (Welfarist-UCB) is presented in Algorithm 1.

**Theorem 1.** (Nash Regret of Welfarist-UCB) *For any bandit instance with $k$ arms, each with $\sigma-$sub-gaussian rewards, and given any (moderately large) $T$, Welfarist-UCB achieves a Nash regret*

$$\mathrm{NR}_T = O\left(\sigma\sqrt{\frac{k\log T\log k}{T}}\right).$$

**Remark 2** (Comparison with NCB (Barman et al., 2023))**.** *Theorem 1 establishes a strict generalization of the regret bound achieved by the standard NCB algorithm. In contrast to NCB, which relies on restrictive assumptions requiring rewards to be bounded and non-negative, Welfarist-UCB applies to arbitrary sub-Gaussian rewards. Furthermore, it does not assume any prior upper bound on $\mu^*$. These improvements are enabled by the data-adaptive termination rule for Phase I and the UCB-based arm selection in Phase II, both of which play equally crucial roles (see Remark 6). A modified NCB algorithm achieves a tighter bound by removing the $\sqrt{\log k}$ factor, but it requires a much larger value of $T$, while $T \geq k$ suffices for us.*

**Theorem 2.** ($p-$mean regret of Welfarist-UCB) *Consider a $k$-armed bandit problem with $\sigma-$sub-gaussian rewards, time horizon $T$, and fairness parameter $p \in \mathbb{R}$. Then, the $p$-mean regret of Welfarist-UCB satisfies*

$$\mathrm{R}_T^p = \begin{cases} O\left(\frac{\sigma k^{\frac{|p|+1}{2}}\sqrt{\log T}}{\sqrt{T}}\cdot\max\{1,|p|\}\right), & p<0\ , \\ O\left(\frac{\sigma\sqrt{k\log T\log k}}{\sqrt{T}}\right), & p\geq 0\ . \end{cases}$$

**Remark 3** (Lower bound)**.** *Observe that for $p \leq 1$, $\left(\frac{\sum_{t=1}^T(\mathbb{E}[\mu_{I_t}])^p}{T}\right)^{\frac{1}{p}} \leq \frac{1}{T}\sum_{t=1}^T\mathbb{E}[\mu_{I_t}]$ since the generalized mean is strictly increasing in $p$. Hence, from the standard lower bound on average regret $\mathrm{AR}_T$ (see, e.g., Bubeck et al. (2012)), we can lower bound $p$-mean regret as $\mathrm{R}_T^p \geq \mathrm{AR}_T \geq \widetilde{\Omega}\big(\sqrt{k/T}\big)$. Thus, our upper bound is optimal up to poly-log factors for $p \in [0,1]$ and is only off by a factor at most $\sqrt{k}$ for $p \in [-1,0)$.*

**Remark 4** (Comparison with Explore-then-UCB (Krishna et al., 2025))**.** *Theorem 2 provides several improvements over Explore-then-UCB. First, for $p \geq 0$ (as in the case of NCB), it offers a strict generalization by extending the analysis from bounded, non-negative rewards to arbitrary sub-Gaussian rewards. Second, for $p \in [-1,0)$, it achieves a worst-case regret bound of $\widetilde{O}\Big(k/\sqrt{T}\Big)$, which is sharper than the $\widetilde{O}\big(k^{3/4}T^{-1/4}\big)$ bound attained by Explore-then-UCB. Most importantly, it eliminates the restrictive assumption that each unknown mean $\mu_i \geq \widetilde{\Omega}\Big(\sqrt{k}\,T^{-1/8}\Big)$.*[1]

*For $p < -1$, Krishna et al. (2025) obtain a bound $\widetilde{O}\big(k^{1/2}T^{-1/4|p|}\big)$. At first glance, their bound appears to scale better with $k$, while ours does with $T$. However, a more careful scaling analysis shows that our bound is strictly tighter. First, note that regret bounds are vacuous whenever they exceed $\mu^*$. Assuming $\mu^* \leq 1$ for parity with Krishna et al. (2025), we observe that our bound is non-trivial only when $T > k^{|p|+1}$, whereas their bound becomes non-trivial only when $T > k^{2|p|}$. Moreover, their bound is tighter than ours only in the regime $T \leq k^{\frac{2p^2}{2|p|-1}}$, but in this regime both bounds are vacuous (and the trivial bound $\mu^* \leq 1$ dominates), since $\frac{2p^2}{2|p|-1} < |p|+1$ for all $p < -1$. Hence, for all practically relevant values of $T$, our bound is tighter than that of Krishna et al. (2025).*

**Remark 5** (No free fairness)**.** *For $p < -1$, our upper bound grows exponentially with $|p|$ unless $T > p^2k^{\frac{|p|}{2}}$, suggesting that very strong fairness may be unattainable in the short term. In the extreme case when $p \to -\infty$, the bound becomes vacuous, aligning with the intuition that achieving vanishing Rawlsian regret is impossible. This highlights a 'no-free-lunch' phenomenon: although the dependence on $T$ remains asymptotically the same, the intrinsic hardness of the learning problem increases as the notion of fairness becomes increasingly stringent. A natural direction for future work is to formalize this by proving a matching lower bound in the region $p < -1$.*

## 4 PROOF TECHNIQUES

We start with a common analytical framework that underlies the analysis of both Nash and $p$-mean regrets (Theorems 1 and 2, respectively).

Our analysis will be based on conditioning on a "good" event $\mathcal{E}$, under which at every round $t$, each $\mu_i$ lies in the interval $\left[\widehat{\mu}_i - 2\sqrt{\frac{2\sigma^2\log T}{n_i}}, \widehat{\mu}_i + 2\sqrt{\frac{2\sigma^2\log T}{n_i}}\right]$ *with high probability*, where $n_i$ and $\widehat{\mu}_i$ are running updates at round $t$. The proof follows from a standard application of the (additive) Hoeffding bound.

We first bound the total length of Phase I.

**Lemma 3** (Rounds of uniform exploration)**.** *Let $\tau$ be the (random) number of rounds after which Phase I ends. Then, under the event $\mathcal{E}$, we have*

$$32\ kS \leq \tau \leq 128\ kS\ ,$$

[1] The justification that this assumption is equivalent to $\mu_i \geq 0$ as $T \to \infty$ is unsatisfactory: accepting it would amount to claiming that a regret of $O(T^{-1/4})$ is equivalent to the optimal $O(T^{-1/2})$ rate, since both vanish as $T$ grows.

*almost surely, where* $S := \frac{4p_a^2\sigma^2\log T}{(\mu^*)^2}$, *with* $p_a = 1$ *if* $p \geq -1$ *and* $p_a = p$ *if* $p < -1$.

The lemma stems from our carefully constructed terminating condition. One can show that if $\tau \leq 32kS$, then each arm $i$ satisfy $n_i < 192p^2\sigma^2\frac{\log T}{\left(\widehat{\mu}_i - 2\sqrt{\frac{2\sigma^2\log T}{n_i}}\right)^2}$, whereas for $\tau \geq 128kS$, there exists atleast one arm (specifically the optimal arm) which violates the above. Based on these, one can conclude that Phase I terminates between these time steps. See Lemma 7 and 8 in the appendix for details.

Recall that in Phase I, we sample a uniform random permutation of the arms every $k$ rounds, and then pull them sequentially. By Lemma 3, this ensures that each arm is pulled $\Theta\left(\frac{1}{(\mu^*)^2}\right)$ times. This procedure is equivalent to sampling arms uniformly at random with $\Pr[I_t = i] = 1/k$ for any arm $i \in [k]$ so that the ex-ante expected reward satisfies $\mathbb{E}[\mu_{I_t}] \geq \mu^*/k$. In other words, permutation-based sampling is a *static coupling* of sequential uniform sampling: it fixes the entire sampling order in advance, whereas sequential sampling decides the order dynamically (see Lemma 9).

We next ensure that if an arm is pulled in Phase II, then its mean must be close to $\mu^*$, so its contribution to Nash/$p$-mean regret is low.

**Lemma 4** (Near optimality of Phase II arms)**.** *Under the event* $\mathcal{E}$, $\mu_i \geq \mu^* - 4\sqrt{\frac{2\sigma^2\log T}{T_i - 1}}$ *for any arm* $i$ *that is pulled at least once in Phase II, where* $T_i$ *denotes the total number of pulls of arm* $i$.

*Proof.* Fix any arm $i$ that is pulled at least once in Phase II. When arm $i$ was pulled the $T_i$-th time during Phase II, it must have had the highest UCB. In particular, at that round $\mathrm{UCB}(i) \geq \mathrm{UCB}(i^*) \geq \mu^*$; the last inequality follows from the Hoeffding bound that the UCB index is an overestimate of the mean with high probability. Therefore, we have $\widehat{\mu}_i + 2\sqrt{\frac{2\sigma^2\log T}{T_i - 1}} \geq \mu^*$, where $\widehat{\mu}_i$ denotes the sample mean of arm $i$ after $T_i - 1$ pulls. Also, $\mu_i \geq \widehat{\mu}_i - 2\sqrt{\frac{2\sigma^2\log T}{T_i - 1}}$ under $\mathcal{E}$. Combining these two together, we have $\mu_i \geq \mu^* - 4\sqrt{\frac{2\sigma^2\log T}{T_i - 1}}$. □

### 4.1 Proof Sketch for Theorem 1

We split the analysis of Nash regret into two regimes depending on the magnitude of $\mu^*$. When $\mu^* \leq \frac{40\sigma\sqrt{2k\log T\log k}}{\sqrt{T}}$, the regret bound holds trivially. Thus in the sequel we only consider $\mu^* \geq \frac{40\sigma\sqrt{2k\log T\log k}}{\sqrt{T}}$. We first compute Nash social welfare in Phase I using the fact that $\mathbb{E}[\mu_{I_t}] \geq \mu^*/k$ for each round in Phase I.

**Lemma 5** (NSW in Phase I)**.** *Suppose Phase I runs for* $\overline{T}$ *rounds. If* $\mu^* \geq \frac{40\sigma\sqrt{2k\log T\log k}}{\sqrt{T}}$, *then*

$$\left(\prod_{t=1}^{\overline{T}}\mathbb{E}[\mu_{I_t}]\right)^{\frac{1}{\overline{T}}} \geq (\mu^*)^{\frac{\overline{T}}{T}}\left(1 - \frac{\overline{T}\log k}{T}\right).$$

Next, we lower bound $\left(\prod_{t=\overline{T}+1}^{T}\mathbb{E}[\mu_{I_t}]\right)^{\frac{1}{T}}$, the Nash social welfare in Phase II, which is further controlled by $\mathbb{E}\left[\left(\prod_{t=\overline{T}+1}^{T}\mu_{I_t}\right)^{\frac{1}{T}}\right]$ using multivariate Jensen's inequality. To bound this, consider the arms that are pulled at least once after the first $\overline{T}$ rounds. Let $\{1, 2, \ldots, \ell\}$ be the set of all those arms and $m_i \geq 1$ be the number of times arm $i \in [\ell]$ is pulled after the first $\overline{T}$ rounds. Then $\mathbb{E}\left[\left(\prod_{t=\overline{T}+1}^{T}\mu_{I_t}\right)^{\frac{1}{T}}\right] = \mathbb{E}\left[\prod_{i=1}^{\ell}\mu_i^{\frac{m_i}{T}}\right]$. Moreover, since we are conditioning on the good event $\mathcal{E}$, Lemma 4 applies to each arm $i \in [\ell]$. Hence, we get $\mathbb{E}\left[\prod_{i=1}^{\ell}\mu_i^{\frac{m_i}{T}}\right] \geq \mathbb{E}\left[\prod_{i=1}^{\ell}\left(\mu^* - 2\sqrt{\frac{2\sigma^2\log T}{T_i - 1}}\right)^{\frac{m_i}{T}}\right]$. Using $\sum_{i=1}^{\ell} m_i = T - \overline{T}$, this is further bounded by $(\mu^*)^{1-\frac{\overline{T}}{T}}\mathbb{E}\left[\prod_{i=1}^{\ell}\left(1 - \frac{2\sqrt{2\sigma^2\log T}}{\mu^*\sqrt{T_i - 1}}\right)^{\frac{m_i}{T}}\right]$.

Let $\xi_i := \frac{2\sqrt{2\sigma^2\log T}}{\mu^*\sqrt{T_i - 1}}$. Now, our sampling strategy in Phase I, along with Lemma 3, together imply that each arm is pulled at least $\frac{128\sigma^2\log T}{(\mu^*)^2}$ times during the first $\overline{T}$ rounds. Hence $T_i - 1 \geq \frac{512\sigma^2\log T}{(\mu^*)^2}$ for each arm $i \in [\ell]$, which yields $\xi_i \leq 1/4$. Thus, applying the inequality $(1-x)^a \geq 1 - 2ax$ for $x \in \left[0, \frac{1}{2}\right], a \geq 0$, we get

$$\begin{aligned}\mathbb{E}\left[\prod_{i=1}^{\ell}\mu_i^{\frac{m_i}{T}}\right] &\geq (\mu^*)^{1-\frac{\overline{T}}{T}}\mathbb{E}\left[\prod_{i=1}^{\ell}\left(1 - \frac{4m_i}{T\mu^*}\sqrt{\frac{2\sigma^2\log T}{T_i - 1}}\right)\right] \\ &\geq (\mu^*)^{1-\frac{\overline{T}}{T}}\mathbb{E}\left[\prod_{i=1}^{\ell}\left(1 - \frac{4}{T\mu^*}\sqrt{2m_i\sigma^2\log T}\right)\right],\end{aligned}$$

where we have used $T_i \geq m_i + 1$. Now, further simplification with a Cauchy-Schwarz inequality together with $\sum_{i=1}^{l} m_i \leq T$ and $\ell \leq k$ yields a bound on NSW.

**Lemma 6** (NSW in Phase II)**.** *Suppose Phase I runs for* $\overline{T}$ *rounds. If* $\mu^* \geq \frac{40\sigma\sqrt{2k\log T\log k}}{\sqrt{T}}$, *then*

$$\left(\prod_{t=\overline{T}+1}^{T}\mathbb{E}[\mu_{I_t}]\right)^{\frac{1}{T}} \geq (\mu^*)^{1-\frac{\overline{T}}{T}}\left(1 - \frac{4}{\mu^*}\sqrt{\frac{2k\sigma^2\log T}{T}}\right).$$

Finally, Lemma 5 and 6 together with the high probability bound on the good event $\mathcal{E}$ complete the proof of Theorem 1. Refer to the appendix for details.

**Remark 6** (UCB versus NCB)**.** *Observe that our terminating condition and permutation-based sampling strategy together ensure that each arm is pulled at least* $n_i \approx \widetilde{\Omega}\left(\frac{1}{(\mu^*)^2}\right)$ *times in phase I. This helps us bound*

*$\xi_i \approx \widetilde{O}\big(\frac{1}{\mu^*\sqrt{n_i}}\big)$ for any arm $i$ that has been pulled at least once in Phase II, with a constant less than $1/2$, which is crucial to control the Nash regret.*

*In contrast, algorithms in Barman et al. (2023) pull each arm at least either $\widetilde{\Omega}(\sqrt{T})$ or $\widetilde{\Omega}\big(\frac{1}{\mu^*}\big)$ times in phase I. In both of these cases, $\xi_i \approx \widetilde{O}\big(\frac{1}{\mu^*\sqrt{n_i}}\big)$ can be bounded by a constant only if one assumes $\mu^* \approx \widetilde{\Omega}\big(\frac{1}{T^{1/4}}\big)$ or $\widetilde{\Omega}(1)$, respectively, both of which render the regret under complementary cases sub-optimal. Instead, they resort to NCB-based arm selection in phase II, which, roughly, requires them to bound $\widetilde{O}\big(\frac{1}{\sqrt{\mu^* n_i}}\big)$ by $1/2$ for controlling Nash regret. This is achieved by assuming $\mu^* \approx \widetilde{\Omega}\big(\frac{1}{\sqrt{T}}\big)$ since the optimal regret is trivially attained when $\mu^* \approx \widetilde{O}\big(\frac{1}{\sqrt{T}}\big)$.*[2]

*The use of NCB instead of UCB as arm selection index compels them to condition their analysis on the event $\Big\{\forall i \in [k], \mu_i \in \Big[\widehat{\mu}_i - 4\sqrt{\frac{\widehat{\mu}_i \log T}{n_i}}, \widehat{\mu}_i + 4\sqrt{\frac{\widehat{\mu}_i \log T}{n_i}}\Big]\Big\}$ and employ the multiplicative Hoeffding inequality (instead of the additive one) to ensure that it is a "good" event. However, it restricts their approach to bounded and non-negative rewards only, whereas our approach works for sub-Gaussian rewards because of the use of UCB index and additive Hoeffding bound.*

### 4.2 Proof Sketch for Theorem 2

We begin by splitting the analysis into two cases for $p$, namely $p \geq 0$ and $p < 0$. We assume that the good event $\mathcal{E}$ holds with high probability.

**Case I ($p \geq 0$):** By monotonicity of power means, $\big(\prod_{t=1}^{T} \mathbb{E}[\mu_{I_t}]\big)^{1/T} \leq \big(\frac{1}{T}\sum_{t=1}^{T}(\mathbb{E}[\mu_{I_t}])^p\big)^{1/p}$ for any $p \geq 0$ (see Lemma 12). This implies that the $p$-mean regret is at most as large as the Nash regret, and hence, from Theorem 1, we get $\mathrm{R}_T^p = O\Big(\sigma\sqrt{\frac{k \log T \log k}{T}}\Big)$.

**Case II ($p < 0$):** The analysis further splits into two regimes depending on the magnitude of $\mu^*$. When $\mu^* \leq O\Big(\frac{\sigma|p|k^{(|p|+1)/2}\sqrt{\log T}}{\sqrt{T}}\Big)$, the regret bound holds trivially. Thus, we focus on the case when $\mu^* \geq \Omega\Big(\frac{\sigma|p|k^{(|p|+1)/2}\sqrt{\log T}}{\sqrt{T}}\Big)$. First, we set $q = -p > 0$ to convert the $p$-mean regret into the $q$-regret

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

Next, we split the sum into Phase I and II, and define

$$x = \frac{T}{\sum_{t=1}^{\overline{T}}(\mathbb{E}[\mu_{I_t}])^{-q}}, \quad y = \frac{T}{\sum_{t=\overline{T}+1}^{T}(\mathbb{E}[\mu_{I_t}])^{-q}},$$

to obtain $\mathrm{R}_T^q = \mu^* - \big(\frac{1}{1/x+1/y}\big)^{1/q}$. Thus, our goal would be to get bounds on $1/x$ and $1/y$. To bound $1/x$, note that in Phase I, uniform exploration ensures $\mathbb{E}[\mu_{I_t}] \geq \mu^*/k$, yielding $\frac{1}{x} \leq \frac{\overline{T}k^q}{(\mu^*)^q T}$.

To bound $1/y$, first note that $[\ell]$ denotes the set of arms that are pulled at least once in Phase II and $m_i$ denotes the number of times arm $i \in [\ell]$ is pulled in Phase II. Then an application of Jensen's inequality gives $\sum_{t=\overline{T}+1}^{T}(\mathbb{E}[\mu_{I_t}])^{-q} = \mathbb{E}\Big[\sum_{i=1}^{\ell} m_i \mu_i^{-q}\Big]$. Further applying Lemma 4 to each arm $i \in [\ell]$, we get $\frac{1}{y} \leq \frac{\mathbb{E}\big[\sum_{i=1}^{\ell} m_i(\mu^* - u_i)^{-q}\big]}{T}$, where $u_i = 4\sqrt{\frac{2\sigma^2 \log T}{T_i - 1}}$.

Next, we control the terms $(\mu^* - u_i)^{-q}$. Note that for sufficiently small $u_i/\mu^*$, one can linearize $(1 - u_i/\mu^*)^{-q}$ as $1 + \frac{2qu_i}{\mu^*}$. With the choice of $\mu^*$ and the lower bound $T_i - 1 \geq \frac{128 p_a^2 \sigma^2 \log T}{(\mu^*)^2}$ from Lemma 3, we argue that $u_i/\mu^*$ is sufficiently small to apply the above linearization. Thus, the inverse terms of the form $(\mu^* - u_i)^{-q}$ in the $q$-regret can be bounded linearly in $qu_i/\mu^*$, allowing us to control the contributions of suboptimal arms $i \in [l]$ in the $p$-mean regret. We use two slightly different variations of the above linearization for the two regimes $p \leq -1$ and $-1 < p < 0$ (see Claims 13 and 14 in the Appendix); however, the final result remains the same for both cases.

The reason behind linearizing $(\mu^* - u_i)^{-q}$ is that after some simplification, we can bound each of the terms $m_i(\mu^* - u_i)^{-q}$ by $(\mu^*)^{-q}\Big(m_i + \frac{4q\sqrt{m_i}}{\mu^*}\sqrt{2\sigma^2 \log T}\Big)$. The $\sqrt{m_i}$ terms help us apply Cauchy-Schwarz inequality to handle $\sum_{i\in[\ell]}\sqrt{m_i}$, since $\sum_{i\in[\ell]} m_i \leq T - \overline{T} \leq T$, which helps us deal only with $T$. Using these steps and making further simplifications, we get

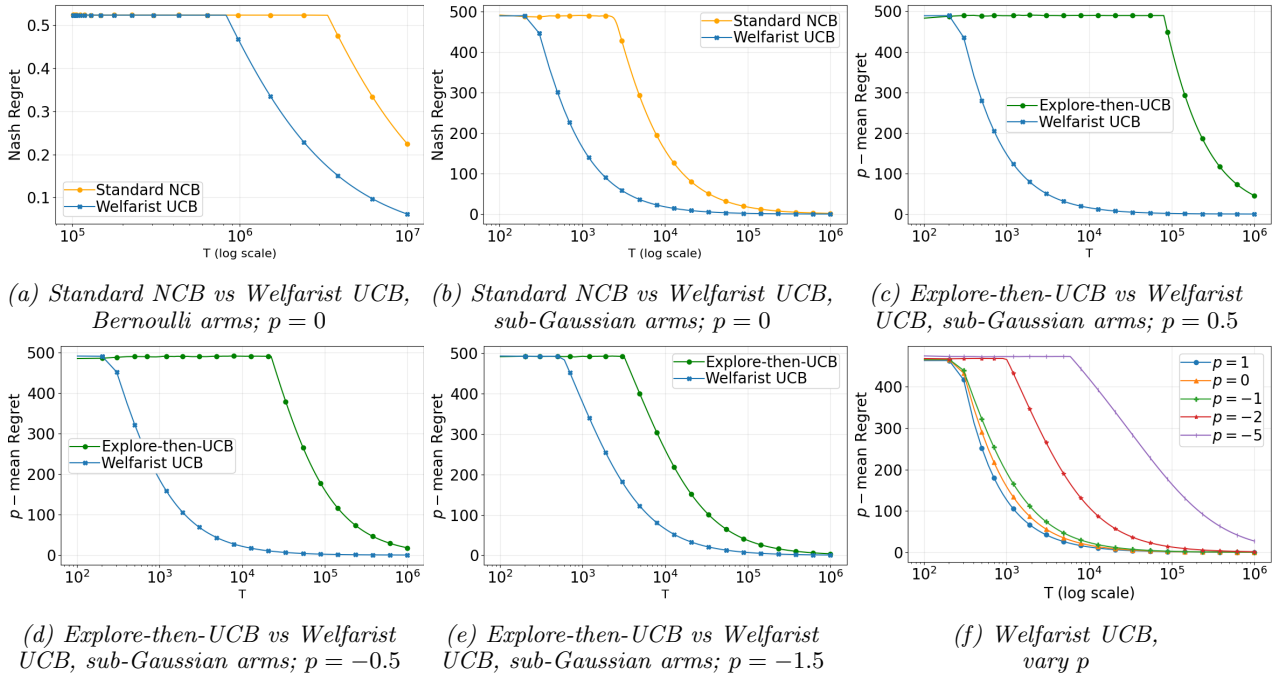
$$\frac{1}{y} \leq \frac{1}{(\mu^*)^q\Big(1 - \frac{4q\sqrt{2k\sigma^2 \log T}}{\mu^*\sqrt{T}}\Big)}.$$

Substituting the bounds on $1/x$ and $1/y$ in the denominator for the $q-$regret, and simplifying this further using the standard numeric inequality $(1+a)^{-1} \geq 1 - a$ for $a \in [0, 1)$, we get

$$\begin{aligned}\frac{1}{\frac{1}{x}+\frac{1}{y}} &\geq (\mu^*)^q\left(1 - \frac{4q\sqrt{2k\sigma^2 \log T}}{\mu^*\sqrt{T}} - \frac{\overline{T}k^q}{T}\right)\\ &\geq (\mu^*)^q\left(1 - \frac{4q\sqrt{2k\sigma^2 \log T}}{\mu^*\sqrt{T}} - \frac{512q^2k^{q+1}\sigma^2 \log T}{(\mu^*)^2 T}\right),\end{aligned}$$

where the last inequality follows from Lemma 3 since $\overline{T} \leq 128kS = \frac{512kq^2\sigma^2 \log T}{(\mu^*)^2}$. Defining $v$ as the sum of the two negative terms inside the parentheses, we can show that $v \leq 1/2$ in the working regime of $\mu^*$. Then, exponentiating both sides by $\frac{1}{q}$ and applying

[2] Krishna et al. (2025) could work with the UCB index as they sidestep this issue by enforcing an unrealistic assumption of $\mu_i \approx \widetilde{\Omega}\big(\frac{1}{T^{1/4}}\big)$ for each arm $i$.




*(a) Standard NCB vs Welfarist UCB, Bernoulli arms; $p = 0$*

*(b) Standard NCB vs Welfarist UCB, sub-Gaussian arms; $p = 0$*

*(c) Explore-then-UCB vs Welfarist UCB, sub-Gaussian arms; $p = 0.5$*

*(d) Explore-then-UCB vs Welfarist UCB, sub-Gaussian arms; $p = -0.5$*

*(e) Explore-then-UCB vs Welfarist UCB, sub-Gaussian arms; $p = -1.5$*

*(f) Welfarist UCB, vary $p$*

*Figure 1: Numerical results for Welfarist UCB. (a) illustrates the comparision between the Nash regret ($p = 0$) from Welfarist UCB and Standard NCB for Bernoulli rewards. (b) shows that Welfarist UCB achieves Standard NCB under sub-gaussian rewards. (c), (d) and (e) show that Welfarist UCB outperforms Explore-Then-UCB for the general $p-$mean regret in all three regimes. (f) shows the ablation on $p$; as $p$ decreases, we get more fairness, but the regret also increases.*

the binomial approximation $(1-v)^{1/q} \geq 1 - \frac{2v}{q}$ for $v \in [0, 1/2]$, we obtain

$$\left(\frac{1}{\frac{1}{x}+\frac{1}{y}}\right)^{\frac{1}{q}} \geq \mu^* - \frac{8\sqrt{2k\sigma^2 \log T}}{\sqrt{T}} - \frac{1024qk^{q+1}\sigma^2 \log T}{\mu^* T} .$$

Substituting the assumed regime for $\mu^*$, we get the desired terms of the form $\frac{k^{\frac{q+1}{2}}}{\sqrt{T}}$. The final step is to use this find upper bound the regret as $\mathrm{R}_T^q = \mu^* - \left(\frac{1}{\frac{1}{x}+\frac{1}{y}}\right)^{\frac{1}{q}}$, which gives us $\mathrm{R}_T^q \leq O\left(\frac{\sigma k^{(|p|+1)/2}\sqrt{\log T}}{\sqrt{T}}\right)$. Combining the two results from the two cases on $\mu^*$, we get the stated bound on the $p$-mean regret. Please refer to Appendix B.3 for the complete proof.

## 5 EXPERIMENTS

In this section, we present the results from our numerical simulations. All experiments report the average reward over 50 runs to estimate $\mathbb{E}[\mu_{I_t}]$. We compare our algorithm (Welfarist-UCB) with NCB (Barman et al., 2023) and Explore-then-UCB (Krishna et al., 2025).

We first consider $k = 50$ Bernoulli arms with means sampled uniformly at random from the interval $[0.005, 1]$, and compare the Nash regret ($p = 0$) of NCB and Welfarist-UCB. As shown in Figure (a), the Nash regret of Welfarist-UCB decreases much faster than that of NCB as the horizon $T$ increases.

Next, we consider $k = 50$ Gaussian arms with means sampled uniformly at random from the interval $[10, 1000]$, and a fixed standard deviation of $\sigma = 20$. As shown in Figure (b), Welfarist-UCB minimizes the Nash regret significantly faster than NCB.

Then, we compare the $p$-mean regret of Welfarist-UCB and Explore-then-UCB under the Gaussian bandit setting. We evaluate the algorithms in three regimes by selecting $p = 0.5$ ($0 < p \leq 1$), $p = -0.5$ ($-1 < p < 0$), and $p = -1.5$ ($p \leq -1$). Figures (c), (d), and (e) present the results for these three cases, respectively, and illustrate that Welfarist-UCB consistently minimizes the $p$-mean regret faster than Explore-then-UCB across all regimes.

Finally, we conduct an ablation study on the fairness parameter $p$, evaluating the $p$-mean regret of our algorithm across different regimes. As shown in Figure (f), the $p$-mean regret consistently increases as $p$ decreases, indicating that stronger fairness requirements come at the expense of higher regret. This trend corroborates the "no-free-lunch" hypothesis stated in Remark 5.

## 6 CONCLUSION

We show that the classic UCB algorithm, when combined with a data-adaptive exploration phase, yields a near-optimal solution for fairness-aware regret in

stochastic bandits. Our Welfarist-UCB algorithm extends existing guarantees from bounded, non-negative rewards to general sub-Gaussian reward distributions, achieving near-optimal regret under both Nash and $p$-mean welfare metrics. This makes fair sequential decision-making more practical and widely applicable. Moreover, our results highlight the versatility of UCB for optimizing social welfare objectives, while also uncovering a "no-free-lunch" principle: enforcing stricter fairness criteria fundamentally increases the difficulty of the learning problem, thereby requiring more samples to achieve low regret. An important direction for future work is to formalize this principle by establishing matching lower bounds.

## Acknowledgements

SRC would like to thank the Anusandhan National Research Foundation (ANRF), India, for an early-career research grant.

## References

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, pages 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, pages 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.

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*, pages 160–169.

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, pages 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.

Krishna, A., John, P. G., Barik, A., and Tan, V. Y. (2025). p-mean regret for stochastic bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17966–17973.

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.

Sarkar, D., Pandey, N., and Chowdhury, S. R. (2025). Dp-ncb: Privacy preserving fair bandits. *arXiv preprint arXiv:2508.03836*.

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

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*.

## Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes, Sections 3 and 4]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes, Section 4]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes, Sections 3 and 4]

(b) Complete proofs of all theoretical results. [Yes, Sections 3 and 4]

(c) Clear explanations of any assumptions. [Yes, Sections 3 and 4]

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes, Section 5]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. [Not Applicable]

(b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

# Supplementary Material

## A 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** In the study of Multi-Armed Bandits (MABs), fairness has been conceptualized in several distinct ways. One prominent approach, explored by Joseph et al. (2016), Celis et al. (2019), and Patil et al. (2021), centers on guaranteeing equitable treatment for the arms themselves. Another direction, particularly in multi-agent contexts, has focused on the fair allocation of rewards among agents per arm pull, as investigated by Hossain et al. (2021) and Jones et al. (2023). Our research diverges from these perspectives by addressing fairness across time, treating each sequential round as a distinct agent that requires fair consideration.

A foundational work by Barman et al. (2023) established the notion of Nash regret and proposed the Nash Confidence Bound algorithm to minimize it in stochastic multi-armed bandit environments. Their algorithm achieves tight regret guarantees that hold for both known and unknown ($T$-oblivious) time horizons. However, this approach requires a specialized algorithm and relies on strong assumptions, limiting its applicability. Specifically, its analysis is based on multiplicative concentration inequalities, which restrict rewards to be **bounded and non-negative**, making the method unsuitable for general distributions like Gaussian rewards. Additionally, their algorithm implicitly requires prior knowledge of an upper bound on the optimal mean reward, $\mu^*$.

Krishna et al. (2025) later extended this by generalizing the objective to $p$-mean regret, which is derived from the $p$-mean welfare function in social choice theory. They proposed using a standard Explore-then-UCB algorithm. However, to circumvent the known failure of UCB for Nash regret shown by Barman et al. (2023), their analysis relies on a restrictive assumption that all arms possess a minimum expected reward of at least $\tilde{\Omega}(\sqrt{k}T^{-1/8})$. This assumption significantly limits the applicability of their method, as it excludes common bandit instances. The justification for this assumption has also been found unsatisfactory, as it conflates different convergence rates. Furthermore, their approach is confined to bounded, non-negative rewards and achieves sub-optimal regret bounds in certain fairness regimes, such as $\tilde{O}(k^{3/4}T^{-1/4})$ for $p \in [-1, 0)$

**p-Mean Welfare and Fair Division** The concept of $p$-mean welfare is well-established within the field of fair division, which integrates principles from both mathematical economics and computer science. Drawing from social choice theory Moulin (2004), this welfare function serves as a tunable framework for navigating the trade-off between equity and efficiency, a topic explored in various works Barman et al. (2020); Garg et al. (2021); Barman et al. (2022); Eckart et al. (2024). It is axiomatically defined by five fundamental properties—anonymity, scale invariance, continuity, monotonicity, and symmetry—which together ensure its alignment with core principles of fair allocation. Furthermore, it adheres to the Pigou-Dalton principle, which states that welfare increases when a resource is transferred from a more advantaged individual to a less advantaged one; this principle constrains the parameter $p$ to values no greater than 1. By building on this robust theoretical grounding, our work avoids the need to formulate arbitrary or ad hoc fairness rules.

**Other Related Work** The pursuit of fairness is expanding into related domains, underscoring a broader trend of integrating such considerations into machine learning algorithms. For instance, Sawarni et al. (2024) investigated Nash regret within the context of stochastic linear bandits, for which they established tight upper bounds under the condition of sub-Poisson rewards.

The research by Zhang et al. (2024) on online Nash social welfare (NSW) maximization provides another relevant perspective. While their approach differs from ours—focusing on decisions that simultaneously affect multiple agents rather than our sequential, round-by-round fairness model—it similarly emphasizes the critical need to embed fairness into online decision-making frameworks.

Further illustrating this trend, Mandal and Gan (2022) apply an axiomatic framework to multi-agent reinforcement learning, showing that Nash Social Welfare uniquely satisfies specific fairness criteria and deriving regret bounds for policies that optimize for fair outcomes. Collectively, such studies demonstrate the increasing integration of fairness metrics like NSW into a wide array of learning algorithms, spanning from bandit problems to more complex Markov decision process environments.

Recent work has begun to address the intersection of privacy and fairness in sequential decision-making. Sarkar et al. (2025) bridges the gap between privacy-preserving and fairness-aware bandits by introducing the Differentially Private Nash Confidence Bound (DP-NCB) framework. Their work provides a unified approach to simultaneously guarantee $\epsilon$-differential privacy while minimizing Nash regret. The proposed algorithms operate under both global (GDP) and local (LDP) privacy models and are designed to be anytime, requiring no knowledge of the time horizon. For the GDP setting, they achieve an order-optimal Nash regret of $\tilde{O}(\sqrt{\frac{k}{T}} + \frac{k}{\epsilon T})$, and in the more restrictive LDP setting, they achieve a regret of $\tilde{O}(\sqrt{\frac{k}{T}} + \frac{1}{\epsilon}\sqrt{\frac{k}{T}})$, with both bounds matching known lower bounds up to logarithmic factors.

# B MISSING PROOFS

## B.1 Proof for Bound on Probability of Good Event $\mathcal{E}$

To prove this result, we will instead bound $\mathcal{E}^c$. Invoking Lemma 16 with $\epsilon = 2\sqrt{\frac{2\sigma^2 \log T}{n_i}}$, we have, for every arm $i$,

$$\Pr\left\{|\hat{\mu}_i - \mu_i| \geq 2\sqrt{\frac{2\sigma^2 \log T}{n_i}}\right\} \leq 2\exp\left(-\frac{8n_i\sigma^2 \log T}{2n_i\sigma^2}\right) = \frac{2}{T^4}$$

Thus, by union bound, we get

$$\Pr\{\mathcal{E}_2^c\} = \frac{2}{T^4} \cdot kT \leq \frac{2}{T} \tag{1}$$

Finally, we have

$$\Pr\{\mathcal{E}\} = 1 - \Pr\{\mathcal{E}^c\} \geq 1 - \frac{2}{T}\ .$$

## B.2 Proof of Supporting Lemmas

First, we state the following two lemmas, which help us derive the terminating condition and prove Lemma 3.

**Lemma 7.** *Under the event $\mathcal{E}$, for any arm $i$ and its sample count $n_i \leq 32S$, we have $n_i \leq 128p_a^2\sigma^2 \frac{\log T}{\left(\widehat{\mu}_{i,n} - 2\sqrt{\frac{2\sigma^2 \log T}{n_i}}\right)^2}$.*

*Proof.* Write $N \coloneqq 32S$. Note that, for any arm $i$, the quantity $n\,\widehat{\mu}_{i,n}$ is equal to the sum of the rewards observed for arm $i$ in the first $n$ samples. Therefore, for all $n \leq N$, we have

$$\begin{aligned}
n\,\widehat{\mu}_{i,n} &\leq n\,\left(\mu_i + 2\sqrt{\frac{2\sigma^2 \log T}{n}}\right) && \text{(via event } \mathcal{E}\text{)} \\
&= n\mu_i + 2\sqrt{2n\sigma^2 \log T} \leq 128p_a^2\sigma^2 \frac{\mu_i \log T}{(\mu^*)^2} + 2\sqrt{2n\sigma^2 \log T} && (n \leq 32S) \\
&\leq 128p_a^2\sigma^2 \frac{\log T}{\mu^*} + 2\sqrt{2n\sigma^2 \log T} && (\mu_i \leq \mu^*) \\
&\leq 128p_a^2\sigma^2 \frac{\log T}{\widehat{\mu}_{i,n} - 2\sqrt{\frac{2\sigma^2 \log T}{n}}} + 2\sqrt{2n\sigma^2 \log T}\ ,
\end{aligned}$$

$$\implies n\left(\widehat{\mu}_{i,n} - 2\sqrt{\frac{2\sigma^2\log T}{n}}\right) \le 128p_a^2\sigma^2\frac{\log T}{\left(\widehat{\mu}_{i,n} - 2\sqrt{\frac{2\sigma^2\log T}{n}}\right)}$$

$$\implies n \le 128p_a^2\sigma^2\frac{\log T}{\left(\widehat{\mu}_{i,n} - 2\sqrt{\frac{2\sigma^2\log T}{n}}\right)^2} \tag{2}$$

which completes the proof. □

**Lemma 8.** *Under the event $\mathcal{E}$, for the optimal arm $i^*$ and its sample count $n_{i^*} \ge 128S$, we have $n_{i^*} \ge 256p_a^2\sigma^2\frac{\log T}{\left(\widehat{\mu}_{i^*,n} - 2\sqrt{\frac{2\sigma^2\log T}{n_{i^*}}}\right)^2}$.*

*Proof.* Write $M \coloneqq 128S$ and note that, for all $n \ge M$, we have $n\ \widehat{\mu}_{i,n} \ge M\ \widehat{\mu}_{i,M}$. Thus, proving this lemma for $M = 128S$ is sufficient. Observe that

$$\begin{aligned}
\widehat{\mu}_{i^*,M} &\ge \mu^* - 2\sqrt{\frac{2\sigma^2\log T}{M}} && \text{(via event } \mathcal{E}\text{)}\\
&= \mu^* - \frac{\mu^*}{\sqrt{64p_a^2}} && \text{(since } M = 128S = \tfrac{512p_a^2\sigma^2\log T}{(\mu^*)^2}\text{)}\\
&= \mu^*\left(1 - \frac{1}{8|p_a|}\right) \ge \frac{7}{8}\mu^* && (|p_a| \ge 1)
\end{aligned}$$

Thus, the total observed reward satisfies:

$$M\ \widehat{\mu}^*_{i,M} \ge \frac{7}{8}\mu^*\ 128S = \frac{448p_a^2\sigma^2\log T}{\mu^*} \tag{3}$$

Now, consider the following terms

$$256p_a^2\sigma^2\frac{\log T}{\widehat{\mu}_{i^*,M} - 2\sqrt{\frac{2\sigma^2\log T}{M}}} \le 256p_a^2\sigma^2\frac{\log T}{\frac{6}{8}\mu^*} = 384p_a^2\sigma^2\frac{\log T}{\mu^*} \tag{4}$$

and,

$$2\sqrt{2\sigma^2 M\log T} = \frac{64|p_a|\sigma^2\log T}{\mu^*} \le \frac{64p_a^2\sigma^2\log T}{\mu^*} \qquad (|p_a| \ge 1)$$

Adding inequality (4) and the above inequality, we have

$$256p_a^2\frac{\sigma^2\log T}{\widehat{\mu}_{i^*,M} - 2\sqrt{\frac{2\sigma^2\log T}{M}}} + 2\sqrt{2\sigma^2 M\log T} \le \frac{448p_a^2\sigma^2\log T}{\mu^*} \le M\ \widehat{\mu}^*_{i,M} \qquad \text{(via inequality (3))}$$

$$\implies M \ge 256p_a^2\sigma^2\frac{\log T}{\left(\widehat{\mu}_{i^*,M} - 2\sqrt{\frac{2\sigma^2\log T}{M}}\right)^2} \tag{5}$$

This completes the proof of the lemma. □

Next, we restate and prove Lemma 3.

**Lemma 3** (Rounds of uniform exploration)**.** *Let $\tau$ be the (random) number of rounds after which Phase I ends. Then, under the event $\mathcal{E}$, we have*

$$32\ kS \le \tau \le 128\ kS\ ,$$

*almost surely, where $S \coloneqq \frac{4p_a^2\sigma^2\log T}{(\mu^*)^2}$, with $p_a = 1$ if $p \ge -1$ and $p_a = p$ if $p < -1$.*

*Proof.* Suppose Phase I terminates after $t = 32kS$ rounds. Then each arm has been pulled $32S$ times. Thus, via Lemma 7, for every arm $i$, the number of pulls $n_i$ is less than $128p_a^2\sigma^2\frac{\log T}{\left(\widehat{\mu}_{i,n}-2\sqrt{\frac{2\sigma^2\log T}{n_i}}\right)^2} \leq 192p_a^2\sigma^2\frac{\log T}{\left(\widehat{\mu}_{i,n}-2\sqrt{\frac{2\sigma^2\log T}{n_i}}\right)^2}$. Therefore, $\tau \geq 32kS$.

Similarly, suppose Phase I terminates after $t = 128kS$ rounds. Then each arm has been pulled $128S$ times. Therefore, Lemma 8 implies that, by round $t_2$ and for the optimal arm $i^*$, the number of pulls $n_{i^*}$ is at least $256p_a^2\sigma^2\frac{\log T}{\left(\widehat{\mu}_{i^*,n}-2\sqrt{\frac{2\sigma^2\log T}{n_{i^*}}}\right)^2} \geq 192p_a^2\sigma^2\frac{\log T}{\left(\widehat{\mu}_{i^*,n}-2\sqrt{\frac{2\sigma^2\log T}{n_{i^*}}}\right)^2}$. Hence, $\tau \leq 128kS$ □

**Lemma 9** (Permutation sampling is a static coupling of sequential uniform sampling)**.** *Let $\mathcal{B}$ be a multiset containing $N_i$ items of type $i$ for $i = 1, \ldots, K$, with $\sum_{i=1}^K N_i = N$. Consider two sampling schemes:*

*(A) (Permutation) Draw a uniform random permutation of the $N$ (labelled) items and set $I_t$ to be the type observed at position $t$ (sampling* without *replacement).*

*(B) (Sequential with replacement) Draw each $I_t$ independently* with *replacement, choosing type $i$ with probability $N_i/N$ at every draw.*

*Then for every $t \in \{1, \ldots, N\}$ and every type $i$, the one-step marginals coincide, i.e.,*

$$\Pr_A(I_t = i) = \Pr_B(I_t = i) = \frac{N_i}{N}\ .$$

*Proof.* Scheme (B) is immediate by definition: each draw chooses type $i$ with probability $N_i/N$, so $\Pr_B(I_t = i) = N_i/N$ for every $t$.

For scheme (A) consider the uniform permutation of the $N$ labelled items (items of the same type can be thought of as distinct labels that share the same type). By symmetry, every labelled item is equally likely to occupy position $t$ in the permutation. Since there are $N_i$ labelled items of type $i$ out of $N$ total labelled items, the probability that position $t$ holds an item of type $i$ is

$$\Pr_A(I_t = i) = \frac{N_i}{N}.$$

Thus $\Pr_A(I_t = i) = \Pr_B(I_t = i) = N_i/N$, as claimed. □

**Lemma 10.** *Consider a bandit instance with optimal mean $\mu^* \geq \frac{40\sigma\sqrt{2k\log T\log k}}{\sqrt{T}}$. Then, we have*

$$\left(\prod_{t=1}^T \mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} \geq \mu^* - 256\sigma\sqrt{\frac{k\log T\log k}{T}} - \frac{4\mu^* k\log T\log k}{T} - \frac{8\mu^*}{T}.$$

*Proof.* Firstly, for Nash regret, the parameter $p$ is set to 0. Consequently, we will take $p_a = 1$ wherever necessary in the analysis. Next, we assume that Phase 1 runs for at most $\overline{T} \leq 128k\mathcal{S}$ rounds. The existence of this upper bound on $\overline{T}$ is guaranteed by Lemma 3, which ensures that Algorithm 1 will complete Phase 1 by the $\overline{T}$-th round; specifically, the termination condition of the first while-loop (Line 5) in the algorithm must be satisfied by the $\overline{T}$-th round. Also, note that, the following inequality holds based on the assumption on $\mu^*$ stated in the lemma.

$$\frac{\overline{T}\ \log(k)}{T} = \frac{128\ kS\ \log(k)}{T} = \frac{512\ k\sigma^2\log T\ \log(k)}{(\mu^*)^2 T} \leq \frac{512}{3200} \leq 1 \tag{6}$$

Next, we split the Nash social welfare into the following two terms, for the two phases respectively:

$$\left(\prod_{t=1}^T \mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} = \left(\prod_{t=1}^{\overline{T}} \mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} \left(\prod_{t=\overline{T}+1}^{T} \mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} \tag{7}$$

We shall separately lower-bound these two products.

The term corresponding to Phase I, i.e. first term in the RHS of equation (7) can be bounded as:

$$\begin{aligned}\left(\prod_{t=1}^{\overline{T}}\mathbb{E}[\mu_{I_t}]\right)^{\frac{1}{T}} &\geq \left(1-\frac{2}{T}\right)^{\frac{\overline{T}}{T}}\left(\frac{\mu^*}{k}\right)^{\frac{\overline{T}}{T}} \geq \left(1-\frac{2}{T}\right)(\mu^*)^{\frac{\overline{T}}{T}}\left(\frac{1}{k}\right)^{\frac{\overline{T}}{T}} \\ &= \left(1-\frac{2}{T}\right)(\mu^*)^{\frac{\overline{T}}{T}}\left(\frac{1}{2}\right)^{\frac{\overline{T}\log(k)}{T}} = \left(1-\frac{2}{T}\right)(\mu^*)^{\frac{\overline{T}}{T}}\left(1-\frac{1}{2}\right)^{\frac{\overline{T}\log(k)}{T}} \\ &\geq (\mu^*)^{\frac{\overline{T}}{T}}\left(1-\frac{\overline{T}\log(k)}{T}\right)\left(1-\frac{2}{T}\right) \end{aligned} \tag{8}$$

To prove the last inequality, observe that the exponent $\frac{\overline{T}\log(k)}{T} \leq 1$ (see inequality (6)). Thus, we can apply Claim 11.

Now, for the Phase II term, i.e. the second term in the RHS of equation (7), we have the following

$$\begin{aligned}\left(\prod_{t=\overline{T}+1}^{T}\mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} &\geq \mathbb{E}\left[\left(\prod_{t=\overline{T}+1}^{T}\mu_{I_t}\right)^{\frac{1}{T}}\right] && \text{(Multivariate Jensen's inequality)} \\ &\geq \mathbb{E}\left[\left(\prod_{t=\overline{T}+1}^{T}\mu_{I_t}\right)^{\frac{1}{T}}\middle|\,\mathcal{E}\right]\Pr\{\mathcal{E}\} \end{aligned} \tag{9}$$

Since, as mentioned, Lemma 3 guarantees that Algorithm 1 completes Phase 1 by the $\overline{T}$-th round, any round $t > \overline{T}$ necessarily falls under Phase 2.

To bound the expected value on the right-hand side of Inequality (9), we consider only the arms that are pulled at least once after the first $\overline{T}$ rounds (i.e., in Phase 2). With re-indexing, let $\{1, 2, \ldots, \ell\}$ denote this set of arms. Let $m_i \geq 1$ be the number of times arm $i \in [\ell]$ is pulled in Phase 2, such that $\sum_{i=1}^{\ell} m_i = T - \overline{T}$.

Furthermore, let $T_i$ denote the total number of times arm $i$ is pulled throughout the algorithm. Note that $(T_i - m_i)$ is the number of times arm $i \in [\ell]$ is pulled during the first $\overline{T}$ rounds (Phase 1).

Using this notation, the expected value can be rewritten as:

$$\mathbb{E}\left[\left(\prod_{t=\overline{T}+1}^{T}\mu_{I_t}\right)^{\frac{1}{T}}\middle|\,E\right] = \mathbb{E}\left[\left(\prod_{i=1}^{\ell}\mu_i^{\frac{m_i}{T}}\right)\middle|\,E\right]$$

Moreover, since our analysis is conditioned on the good event $E$, we can apply Lemma 4 to each arm $i \in [\ell]$. Hence,

$$\begin{aligned}\mathbb{E}\left[\left(\prod_{t=\overline{T}+1}^{T}\mu_{I_t}\right)^{\frac{1}{T}}\middle|\,\mathcal{E}\right] = \mathbb{E}\left[\left(\prod_{i=1}^{\ell}\mu_i^{\frac{m_i}{T}}\right)\middle|\,\mathcal{E}\right] &\geq \mathbb{E}\left[\prod_{i=1}^{\ell}\left(\mu^* - 2\sqrt{\frac{2\sigma^2\log T}{T_i - 1}}\right)^{\frac{m_i}{T}}\middle|\,\mathcal{E}\right] && \text{(Lemma 4)} \\ &= (\mu^*)^{\frac{T-\overline{T}}{T}}\mathbb{E}\left[\prod_{i=1}^{\ell}\left(1 - 2\sqrt{\frac{2\sigma^2\log T}{(\mu^*)^2(T_i-1)}}\right)^{\frac{m_i}{T}}\middle|\,\mathcal{E}\right] \end{aligned} \tag{10}$$

For the last equality, we use $\sum_{i=1}^{\ell} m_i = T - \overline{T}$. We now state a numeric inequality, which helps us simplify the analysis. The proof is deferred to the Auxillary Lemmas section.

**Claim 11.** *For all reals $x \in \left[0, \frac{1}{2}\right]$ and all $a \geq 0$, we have $(1-x)^a \geq 1 - 2ax$.*

Recall that each arm is pulled at least $32S$ times during the first $\overline{T}$ rounds. Hence, $T_i > 32S$, for each arm $i \in [\ell]$. Since $S = \frac{4\sigma^2 \log T}{(\mu^*)^2}$, we have:

$$2\sqrt{\frac{2\sigma^2 \log T}{(\mu^*)^2(T_i-1)}} \leq 2\sqrt{\frac{2\sigma^2 \log T}{(\mu^*)^2(32S)}} = 2\sqrt{\frac{2\sigma^2 \log T}{(\mu^*)^2 \cdot 32 \cdot \frac{4\sigma^2 \log T}{(\mu^*)^2}}} = 2\sqrt{\frac{2\sigma^2 \log T}{128\sigma^2 \log T}} \leq \frac{1}{2} \quad \text{for each } i \in [\ell].$$

Therefore, we can apply Claim 11 to reduce the expected value in inequality (10) as follows

$$\begin{aligned}
&\mathbb{E}\left[\prod_{i=1}^{\ell}\left(1-\frac{2}{\mu^*}\sqrt{\frac{2\sigma^2\log T}{(T_i-1)}}\right)^{\frac{m_i}{T}} \;\middle|\; \mathcal{E}\right] \geq \mathbb{E}\left[\prod_{i=1}^{\ell}\left(1-\frac{4}{T}\frac{m_i}{\mu^*}\sqrt{\frac{2\sigma^2\log T}{(T_i-1)}}\right) \;\middle|\; \mathcal{E}\right] \\
&\geq \mathbb{E}\left[\prod_{i=1}^{\ell}\left(1-\frac{4}{T\mu^*}\sqrt{2m_i\sigma^2\log T}\right) \;\middle|\; \mathcal{E}\right] && \text{(since } T_i \geq m_i+1\text{)}
\end{aligned}$$

We can further simplify the above inequality by noting that $(1-x)(1-y) \geq 1-x-y$ for all $x, y \geq 0$.

$$\begin{aligned}
&\mathbb{E}\left[\prod_{i=1}^{\ell}\left(1-\frac{4}{T\mu^*}\sqrt{2m_i\sigma^2\log T}\right) \;\middle|\; \mathcal{E}\right] \geq \mathbb{E}\left[1-\sum_{i=1}^{\ell}\left(\frac{4}{T\mu^*}\sqrt{2m_i\sigma^2\log T}\right) \;\middle|\; \mathcal{E}\right] \\
&= 1-\left(\frac{4}{T\mu^*}\sqrt{2\sigma^2\log T}\right)\mathbb{E}\left[\sum_{i=1}^{\ell}\sqrt{m_i} \;\middle|\; \mathcal{E}\right] \geq 1-\left(\frac{4}{T\mu^*}\sqrt{2\sigma^2\log T}\right)\mathbb{E}\left[\sqrt{\ell}\,\sqrt{\sum_{i=1}^{\ell} m_i} \;\middle|\; \mathcal{E}\right] \\
& && \text{(Cauchy-Schwarz inequality)} \\
&\geq 1-\left(\frac{4}{T\mu^*}\sqrt{2\sigma^2\log T}\right)\mathbb{E}\left[\sqrt{\ell\,T} \;\middle|\; \mathcal{E}\right] = 1-\left(\frac{4}{\mu^*}\sqrt{\frac{2\sigma^2\log T}{T}}\right)\mathbb{E}\left[\sqrt{\ell} \;\middle|\; \mathcal{E}\right] && \text{(since } \textstyle\sum_i m_i \leq T\text{)} \\
&\geq 1-\left(\frac{4}{\mu^*}\sqrt{\frac{2k\sigma^2\log T}{T}}\right) && \text{(since } \ell \leq k\text{)}
\end{aligned}$$

Using this bound, along with inequalities (9), and (10), we obtain

$$\left(\prod_{t=\overline{T}+1}^{T}\mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} \geq (\mu^*)^{\frac{T-\overline{T}}{T}}\left(1-\frac{4}{\mu^*}\sqrt{\frac{2k\sigma^2\log T}{T}}\right)\Pr\{\mathcal{E}\} \tag{11}$$

Thus, inequalities (11) and (8) to obtain relevant bounds on the NSW from Phase I and II respectively. Hence, for the combined Nash social welfare of the algorithm, we have

$$\begin{aligned}
\left(\prod_{t=1}^{T}\mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} &\geq \mu^*\left(1-\frac{\overline{T}\cdot\log(k)}{T}-\frac{2}{T}\right)\left(1-\frac{4}{\mu^*}\sqrt{\frac{2k\sigma^2\log T}{T}}\right)\Pr\{\mathcal{E}\} \\
&\geq \mu^*\left(1-\frac{\overline{T}\cdot\log(k)}{T}-\frac{2}{T}\right)\left(1-\frac{4}{\mu^*}\sqrt{\frac{2k\sigma^2\log T}{T}}\right)\left(1-\frac{2}{T}\right) && \text{(via good event } \mathcal{E}\text{)}
\end{aligned}$$

Using the inequality $(1-x)(1-y) \geq 1-x-y \;\forall\; x, y \geq 0$ we have,

$$\begin{aligned}
\left(\prod_{t=1}^{T}\mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} &\geq \mu^*\left(1-\frac{\overline{T}\cdot\log(k)}{T}-\frac{4}{\mu^*}\sqrt{\frac{2k\sigma^2\log T}{T}}-\frac{4}{T}\right) \\
&\geq \mu^*\left(1-\frac{\log(k)}{T}\frac{512\;k\sigma^2\log T}{(\mu^*)^2}-\frac{4}{\mu^*}\sqrt{\frac{2k\sigma^2\log T}{T}}-\frac{4}{T}\right) \\
&\geq \mu^*\left(1-\frac{512\;k\sigma^2\log T\log(k)}{(\mu^*)^2 T}-\frac{4}{\mu^*}\sqrt{\frac{2k\sigma^2\log T}{T}}-\frac{4}{T}\right)
\end{aligned}$$

$$= \mu^* \left(1 - \frac{256\ \sqrt{k\sigma^2 \log T \log k}}{\mu^* \sqrt{T}} \cdot \frac{2\sqrt{k\sigma^2 \log T \log k}}{\mu^* \sqrt{T}} - \frac{4k \log T \log k}{T} - \frac{4}{\mu^*}\sqrt{\frac{2k\sigma^2 \log T}{T}} - \frac{4}{T}\right)$$

$$\geq \mu^* \left(1 - \frac{256\sqrt{k\sigma^2 \log T \log k}}{\mu^* \sqrt{T}} - \frac{4}{\mu^*}\sqrt{\frac{2k\sigma^2 \log T}{T}} - \frac{4}{T}\right) \qquad ()$$

Note that the last inequality holds since $\frac{2\sqrt{k\sigma^2 \log T \log k}}{\mu^* \sqrt{T}} \leq 1$ for large enought $T$ as $\mu^* \geq \frac{40\sigma\sqrt{2k \log T \log k}}{\sqrt{T}}$. Thus, we get

$$\left(\prod_{t=1}^{T} \mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} \geq \mu^* - 256\sigma\sqrt{\frac{k \log T \log k}{T}} - \frac{4\mu^* k \log T \log k}{T} - \frac{4\mu^*}{T}.$$

The lemma stands proved. □

### B.3 Proofs for p-means

We first state a fundamental inequality, known as the *power inequality* or the *generalised mean inequality*, in the following lemma.

**Lemma 12.** *Let $x_1, \ldots, x_n \geq 0$ and for $r \in \mathbb{R}$ define*

$$M_r = \begin{cases} \left(\frac{1}{n}\sum_{i=1}^{n} x_i^r\right)^{1/r}, & r \neq 0, \\ \left(\prod_{i=1}^{n} x_i\right)^{1/n}, & r = 0. \end{cases}$$

*Then $M_r$ is strictly increasing in $r$; in particular, if $a < b$ then $M_a < M_b$.*

**Theorem 2.** ($p-$mean regret of Welfarist-UCB) *Consider a $k$-armed bandit problem with $\sigma-$sub-gaussian rewards, time horizon $T$, and fairness parameter $p \in \mathbb{R}$. Then, the p-mean regret of Welfarist-UCB satisfies*

$$\mathrm{R}_T^p = \begin{cases} O\left(\frac{\sigma k^{\frac{|p|+1}{2}}\sqrt{\log T}}{\sqrt{T}} \cdot \max\{1, |p|\}\right), & p < 0\ , \\ O\left(\frac{\sigma\sqrt{k \log T \log k}}{\sqrt{T}}\right), & p \geq 0\ . \end{cases}$$

*Proof.* We split our analysis into three cases, where $p > 0$, and $p \leq 0$ respectively

**When** $p \geq 0$

Invoking the generalised mean inequality (Lemma 12 for $a = 0$ and $b = p > 0$, we have

$$\left(\prod_{t=1}^{T} \mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} \leq \left(\frac{\sum_{t=1}^{T}(\mathbb{E}_{I_t}[\mu_{I_t}])^p}{T}\right)^{\frac{1}{p}}$$

Thus, we get the following bound on the p-mean regret using Lemma 10

$$\begin{aligned} \mathrm{R}_T^p \triangleq \mu^* - \left(\frac{\sum_{t=1}^{T}(\mathbb{E}_{I_t}[\mu_{I_t}])^p}{T}\right)^{\frac{1}{p}} &\leq \mu^* - \left(\prod_{t=1}^{T} \mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} \\ &\leq \mu^* - \left(\mu^* - 256\sigma\sqrt{\frac{k \log T \log k}{T}} - \frac{4\mu^* k \log T \log k}{T} - \frac{8\mu^*}{T}\right) \\ &\leq O\left(\sigma\sqrt{\frac{k \log T \log k}{T}}\right) \end{aligned}$$

**When $p < 0$**

Notice that when $\mu^* \leq \frac{40\sigma k^{\frac{|p|+1}{2}}\sqrt{\log T}}{\sqrt{T}}$, the p-mean regret satisfies

$$\mathrm{R}_T^q = \mu^* - \left(\frac{\sum_{t=1}^T (\mathbb{E}_{I_t}[\mu_{I_t}])^p}{T}\right)^{\frac{1}{p}} \leq \mu^* \leq O\left(\frac{\sigma|p|k^{\frac{|p|+1}{2}}\sqrt{\log T}}{\sqrt{T}}\right) \tag{12}$$

Next, we consider the case when $\mu^* \geq \frac{40|p|\sigma k^{\frac{|p+1|}{2}}\sqrt{\log T}}{\sqrt{T}}$. Firstly, we set $q = -p$ for notational convenience. Hence, we have $|p_a| = |p| = q$ wherever required. Thus, our objective is converted to minimising the following quantity

$$\mathrm{R}_T^q \triangleq \mu^* - \left(\frac{T}{\sum_{t=1}^T \frac{1}{(\mathbb{E}_{I_t}[\mu_{I_t}])^q}}\right)^{\frac{1}{q}} \tag{13}$$

We will refer to the above quantity as $q-$regret. Towards analyzing $\mathrm{R}_T^q$, we first set the threshold for Phase I runs as $\overline{T} = 128kS$ (i.e., the maximum length of Phase I). Then define

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

so that we have

$$\mathrm{R}_T^q = \mu^* - \left(\frac{1}{\frac{1}{x} + \frac{1}{y}}\right)^{1/q}. \tag{14}$$

To obtain an upper bound for $\mathrm{R}_T^q$, 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

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

Hence,

$$\frac{1}{x} \leq \frac{\overline{T}k^q}{(\mu^*)^q T}. \tag{15}$$

Next, we will focus on $\frac{1}{y}$. Note that when we condition the expectation on the good event $\mathcal{E}$, the following inequality holds trivially

$$\mathbb{E}_{I_t}[\mu_{I_t}] \geq \Pr\{\mathcal{E}\}\mathbb{E}_{I_t}[\mu_{I_t}|\mathcal{E}]$$

Hence,

$$y \geq \frac{T}{\sum_{t=\overline{T}+1}^{T} \frac{1}{(\mathbb{E}_{I_t}[\mu_{I_t}|\mathcal{E}]\Pr\{\mathcal{E}\})^q}} = \frac{T(\Pr\{\mathcal{E}\})^q}{\sum_{t=\overline{T}+1}^{T} \frac{1}{(\mathbb{E}_{I_t}[\mu_{I_t}|\mathcal{E}])^q}} \tag{16}$$

$$\tag{17}$$

Now, we know that by Jensen's inequality, $f(z) = z^{-\frac{1}{q}}$ is convex on $\mathbb{R}_{>0}$, for $q > 0$. Utilising this result and the linearity of expectation, we get

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

For simplicity, we will drop the subscripts. 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

$$\frac{\mathbb{E}\left[\sum_{t=\overline{T}+1}^{T} \frac{1}{(\mu_{I_t})^q}\middle|\mathcal{E}\right]}{T(\Pr\{\mathcal{E}\})^q} = \frac{\mathbb{E}\left[\sum_{i=1}^{\ell} m_i(\mu_i)^{-q}\middle|\mathcal{E}\right]}{T(\Pr\{\mathcal{E}\})^q} \leq \frac{\mathbb{E}\left[\sum_{i=1}^{\ell} m_i\left(\mu^* - 4\sqrt{\frac{2\sigma^2\log T}{n_i}}\right)^{-q}\middle|\mathcal{E}\right]}{T(\Pr\{\mathcal{E}\})^q} \qquad \text{(via Lemma 4)}$$

Now, we split our analysis into two cases: when $p < -1$ and $-1 \leq p < 0$ respectively. We will use two separate claims for these cases

*Case 1: $p < -1$:* In this case, $q = |p_a| = |p|$. The following claim holds for this case (see Appendix C for proof).

**Claim 13.** *For all $q \geq 1$ and reals $x \in \left[0, \frac{1}{2q}\right]$, we have $(1-x)^{-q} \leq 1 + 2qx$.*

Now, let $u = 4\sqrt{\frac{2\sigma^2\log T}{n_i}}$. Then we have

$$\begin{aligned} u &= 4\sqrt{\frac{2\sigma^2\log T}{n_i}} \leq 4\sqrt{\frac{4k\sigma^2\log T}{\overline{T}}} && (n_i \geq \tfrac{\overline{T}}{2k}) \\ &= 4\sqrt{\frac{4k\sigma^2(\mu^*)^2\log T}{512kp^2\sigma^2\log T}} = \frac{4\mu^*}{\sqrt{128p^2}} && (\overline{T} = \tfrac{512k\sigma^2p^2\log T}{(\mu^*)^2}) \end{aligned}$$

Thus, the quantity $x = \frac{u}{\mu^*} = \frac{4}{\sqrt{128}|p|} \leq \frac{1}{2q}$ (as $|p| = q$).

*Case 2: $-1 \leq p < 0$:* For this case, $p_a = 1$; we modify the above claim slightly to get the following result (see Appendix C for proof).

**Claim 14.** *For all $0 < q \leq 1$ and reals $x \in \left[0, \frac{1}{2}\right]$, we have $(1-x)^{-q} \leq 1 + 2qx$.*

Again, let $u = 4\sqrt{\frac{2\sigma^2\log T}{n_i}}$. Then we have

$$\begin{aligned} u &= 4\sqrt{\frac{2\sigma^2\log T}{n_i}} \leq 4\sqrt{\frac{4k\sigma^2\log T}{\overline{T}}} && (n_i \geq \tfrac{\overline{T}}{2k}) \\ &= 4\sqrt{\frac{4k\sigma^2(\mu^*)^2\log T}{512k\sigma^2\log T}} = \frac{4\mu^*}{\sqrt{128}} && (\overline{T} = \tfrac{512k\sigma^2\log T}{(\mu^*)^2}) \end{aligned}$$

which implies, the quantity $x = \frac{u}{\mu^*} = \frac{4}{\sqrt{128}} \leq \frac{1}{2}$. Thus, we have, by Claim 13 and 14 with $x = \frac{u}{\mu^*}$, $\forall\ p < 0$,

$$\begin{aligned} \frac{1}{y} &\leq \frac{\mathbb{E}\left[\sum_{i=1}^{\ell} m_i(\mu^*)^{-q}\left(1 + \frac{4q}{\mu^*}\sqrt{\frac{2\sigma^2\log T}{n_i}}\right)\middle|\mathcal{E}\right]}{T(\Pr\{\mathcal{E}\})^q} \\ &\leq \frac{\mathbb{E}\left[\sum_{i=1}^{\ell} m_i(\mu^*)^{-q}\left(1 + \frac{4q}{\mu^*}\sqrt{\frac{2\sigma^2\log T}{m_i}}\right)\middle|\mathcal{E}\right]}{T(\Pr\{\mathcal{E}\})^q} && (\text{since } n_i \geq m_i) \\ &\leq \frac{\mathbb{E}\left[\sum_{i=1}^{\ell} (\mu^*)^{-q}\left(m_i + \frac{4q\sqrt{m_i}}{\mu^*}\sqrt{2\sigma^2\log T}\right)\middle|\mathcal{E}\right]}{T(\Pr\{\mathcal{E}\})^q} \end{aligned}$$

Now, invoking Cauchy-Schwarz inequality on $\sum_{i=1}^{\ell}\sqrt{m_i}$ and using $m_i \leq T - \overline{T}$, we have

$$\frac{1}{y} \leq \frac{\mathbb{E}\left[(\mu^*)^{-q}\left(T - \overline{T} + \frac{4q\sqrt{\ell(T-\overline{T})}}{\mu^*}\sqrt{2\sigma^2\log T}\right)\middle|\mathcal{E}\right]}{T(\Pr\{\mathcal{E}\})^q}$$

$$
\begin{aligned}
&\leq \frac{\mathbb{E}\left[(\mu^*)^{-q}\left(T+\frac{4q\sqrt{kT}}{\mu^*}\sqrt{2\sigma^2\log T}\right)\middle|\mathcal{E}\right]}{T(\Pr\{\mathcal{E}\})^q} && (\ell \leq k \text{ and } T-\overline{T}\leq T)\\
&= \frac{(\mu^*)^{-q}\left(1+\frac{4q\sqrt{2k\sigma^2\log T}}{\sqrt{T}\mu^*}\right)T}{T(\Pr\{\mathcal{E}\})^q}\\
&\leq \frac{1}{\left((\mu^*)^q-\frac{4q\sqrt{2k\sigma^2\log T}(\mu^*)^{q-1}}{\sqrt{T}}\right)(\Pr\{\mathcal{E}\})^q}, && (18)
\end{aligned}
$$

where inequality (18) holds using the standard inequality $(1+x)\leq\frac{1}{1-x}\ \forall\ 0\leq x\leq 1$. Using the above inequality and inequality (15), we get

$$
\begin{aligned}
\frac{1}{\frac{1}{x}+\frac{1}{y}} &\geq \frac{1}{\frac{\overline{T}k^q}{(\mu^*)^qT}+\frac{1}{\left((\mu^*)^q-\frac{4q\sqrt{2k\sigma^2\log T}(\mu^*)^{q-1}}{\sqrt{T}}\right)(\Pr\{\mathcal{E}\})^q}}\\
&= \frac{\left((\mu^*)^q-\frac{4q\sqrt{2k\sigma^2\log T}(\mu^*)^{q-1}}{\sqrt{T}}\right)(\Pr\{\mathcal{E}\})^q}{1+\frac{\overline{T}k^q}{T}\left(1-\frac{4q\sqrt{2k\sigma^2\log T}}{\mu^*\sqrt{T}}\right)(\Pr\{\mathcal{E}\})^q}
\end{aligned}
$$

Multiplying the numerator and denominator by $1-\frac{\overline{T}k^q}{T}\left(1-\frac{4q\sqrt{2k\sigma^2\log T}}{\mu^*\sqrt{T}}\right)(\Pr\{\mathcal{E}\})^q$, we have

$$
\begin{aligned}
\frac{1}{\frac{1}{x}+\frac{1}{y}} &\geq \frac{\left((\mu^*)^q-\frac{4q\sqrt{2k\sigma^2\log T}(\mu^*)^{q-1}}{\sqrt{T}}\right)(\Pr\{\mathcal{E}\})^q\left(1-\frac{\overline{T}k^q}{T}\left(1-\frac{4q\sqrt{2k\sigma^2\log T}}{\mu^*\sqrt{T}}\right)(\Pr\{\mathcal{E}\})^q\right)}{1-\left(\frac{\overline{T}k^q}{T}\left(1-\frac{4q\sqrt{2k\sigma^2\log T}}{\mu^*\sqrt{T}}\right)(\Pr\{\mathcal{E}\})^q\right)^2}\\
&\geq \left((\mu^*)^q-\frac{4q\sqrt{2k\sigma^2\log T}(\mu^*)^{q-1}}{\sqrt{T}}\right)(\Pr\{\mathcal{E}\})^q\left(1-\frac{\overline{T}k^q}{T}\left(1-\frac{4q\sqrt{2k\sigma^2\log T}}{\mu^*\sqrt{T}}\right)(\Pr\{\mathcal{E}\})^q\right) && (19)
\end{aligned}
$$

where the last inequality holds because of the denominator being less than 1.

We can expand Inequality (19) to get

$$
\begin{aligned}
\frac{1}{\frac{1}{x}+\frac{1}{y}} &\geq (\mu^*)^q\left(1-\frac{4q\sqrt{2k\sigma^2\log T}}{\mu^*\sqrt{T}}\right)(\Pr\{\mathcal{E}\})^q\left(1-\frac{\overline{T}k^q}{T}\left(1-\frac{4q\sqrt{2k\sigma^2\log T}}{\mu^*\sqrt{T}}\right)(\Pr\{\mathcal{E}\})^q\right)\\
&\geq (\mu^*)^q\left(1-\frac{4q\sqrt{2k\sigma^2\log T}}{\mu^*\sqrt{T}}\right)(\Pr\{\mathcal{E}\})^q\left(1-\frac{\overline{T}k^q}{T}\right) && (\text{since } \left(1-\tfrac{4q\sqrt{2k\sigma^2\log T}}{\mu^*\sqrt{T}}\right)(\Pr\{\mathcal{E}\})^q\leq 1)\\
&\geq (\mu^*)^q(\Pr\{\mathcal{E}\})^q\left(1-\frac{4q\sqrt{2k\sigma^2\log T}}{\mu^*\sqrt{T}}-\frac{\overline{T}k^q}{T}\right) && (\text{using } (1-x)(1-y)\geq(1-x-y)\ \forall\ x,y>0)\\
&= (\mu^*)^q(\Pr\{\mathcal{E}\})^q\left(1-\frac{4q\sqrt{2k\sigma^2\log T}}{\mu^*\sqrt{T}}-\frac{128Sk^{q+1}}{T}\right)\\
&= (\mu^*)^q(\Pr\{\mathcal{E}\})^q\left(1-\frac{4q\sqrt{2k\sigma^2\log T}}{\mu^*\sqrt{T}}-\frac{512q^2k^{q+1}\sigma^2\log T}{(\mu^*)^2T}\right) && (q=|p|)
\end{aligned}
$$

Exponentiating the last inequality by $\frac{1}{q}$, we have

$$
\left(\frac{1}{\frac{1}{x}+\frac{1}{y}}\right)^{\frac{1}{q}}\geq(\mu^*)(\Pr\{\mathcal{E}\})\left(1-\frac{4q\sqrt{2k\sigma^2\log T}}{\mu^*\sqrt{T}}-\frac{512q^2k^{q+1}\sigma^2\log T}{(\mu^*)^2T}\right)^{\frac{1}{q}}
$$

Now, consider the following term

$$\begin{aligned} v &= \frac{4q\sqrt{2k\sigma^2 \log T}}{\mu^* \sqrt{T}} + \frac{512 q^2 k^{q+1} \sigma^2 \log T}{(\mu^*)^2 T} \\ &\leq \frac{4}{40k^{\frac{q}{2}}} + \frac{512}{1600} \leq \frac{1}{2} && (\mu^* \geq \tfrac{40|p|\sigma k^{\frac{|p|+1}{2}} \sqrt{\log T}}{\sqrt{T}}) \end{aligned}$$

Thus, we can apply Claim 11 on $(1-v)^{\frac{1}{q}}$ to get

$$\left(\frac{1}{\frac{1}{x}+\frac{1}{y}}\right)^{\frac{1}{q}} \geq (\mu^*)(\Pr\{\mathcal{E}\}) \left(1 - \frac{8\sqrt{2k\sigma^2 \log T}}{\mu^* \sqrt{T}} - \frac{1024 q^2 k^{q+1} \sigma^2 \log T}{q(\mu^*)^2 T}\right)$$

Further, substituting $\Pr\{\mathcal{E}\} \geq 1 - \frac{2}{T}$, we have

$$\begin{aligned} \left(\frac{1}{\frac{1}{x}+\frac{1}{y}}\right)^{\frac{1}{q}} &\geq (\mu^*) \left(1 - \frac{2}{T}\right) \left(1 - \frac{8\sqrt{2k\sigma^2 \log T}}{\mu^* \sqrt{T}} - \frac{1024 q^2 k^{q+1} \sigma^2 \log T}{q(\mu^*)^2 T}\right) \\ &\geq (\mu^*) \left(1 - \frac{8\sqrt{2k\sigma^2 \log T}}{\mu^* \sqrt{T}} - \frac{1024 q^2 k^{q+1} \sigma^2 \log T}{q(\mu^*)^2 T} - \frac{2}{T}\right) \\ & && (\text{using } (1-x)(1-y) \geq (1-x-y) \ \forall \ x, y > 0) \end{aligned}$$

Thus, the $q-$regret satisfies

$$\begin{aligned} \mathrm{R}_T^q &\leq \mu^* - (\mu^*) \left(1 - \frac{8\sqrt{2k\sigma^2 \log T}}{\mu^* \sqrt{T}} - \frac{1024 q^2 k^{q+1} \sigma^2 \log T}{q(\mu^*)^2 T} - \frac{2}{T}\right) \\ &\leq \frac{8\sqrt{2k\sigma^2 \log T}}{\sqrt{T}} + \frac{1024 q^2 k^{q+1} \sigma^2 \log T}{q(\mu^*) T} + \frac{2\mu^*}{T} \\ &\leq \frac{8\sqrt{2k\sigma^2 \log T}}{\sqrt{T}} + \frac{256 k^{\frac{q+1}{2}} \sigma \sqrt{\log T}}{10\sqrt{T}} + \frac{2\mu^*}{T} && (\mu^* \geq \tfrac{40|p|\sigma k^{\frac{|p|+1}{2}} \sqrt{\log T}}{\sqrt{T}}) \end{aligned}$$

and as a result, the $p-$mean regret satisfies

$$\mathrm{R}_T^p \leq O\left(\frac{\sigma k^{\frac{|p|+1}{2}} \sqrt{\log T}}{\sqrt{T}}\right) \tag{20}$$

Thus, from inequalities (12) and (20), we get the final regret bound as

$$\mathrm{R}_T^p \leq O\left(\frac{\sigma k^{\frac{|p|+1}{2}} \sqrt{\log T}}{\sqrt{T}} \cdot \max(1, |p|)\right)$$

which completes the proof of the theorem. □

# C AUXILLARY LEMMAS

**Lemma 15** (Chernoff Bound)**.** *Let $Z_1, \ldots, Z_n$ be independent Bernoulli random variables. Consider the sum $S = \sum_{r=1}^n Z_r$ and let $\nu = \mathbb{E}[S]$ be its expected value. Then, for any $\varepsilon \in [0,1]$, we have*

$$Pr\{S \leq (1-\varepsilon)\nu\} \leq \exp\left(-\frac{\nu\varepsilon^2}{2}\right), \text{ and}$$
$$Pr\{S \geq (1+\varepsilon)\nu\} \leq \exp\left(-\frac{\nu\varepsilon^2}{3}\right).$$

**Lemma 16** (Hoeffding Inequality)**.** *Let $Z_1, \dots, Z_n$ be independent random variables, with mean $\mu$ and subgaussianity parameter $\sigma$. Consider the empirical mean $\hat{\mu} = \frac{1}{n}\sum_{r=1}^{n} Z_r$. Then, we have*

$$Pr\{|\hat{\mu} - \mu| \geq \epsilon\} \leq 2\exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right)$$

**Claim 11.** *For all reals $x \in \left[0, \frac{1}{2}\right]$ and all $a \geq 0$, we have $(1-x)^a \geq 1 - 2ax$.*

*Proof.* For $x \in [0, \frac{1}{2}]$ we have

$$-\ln(1-x) = \sum_{n\geq 1} \frac{x^n}{n} \leq \sum_{n\geq 1} x^n = \frac{x}{1-x} \leq 2x,$$

so $\ln(1-x) \geq -2x$. Hence for $a \geq 0$,

$$(1-x)^a = \exp\big(a\ln(1-x)\big) \geq \exp(-2ax).$$

Finally, using the standard inequality $e^{-y} \geq 1 - y$ for all $y \in \mathbb{R}$, with $y = 2ax \geq 0$ we get $\exp(-2ax) \geq 1 - 2ax$. Combining these results completes the proof. □

**Claim 13.** *For all $q \geq 1$ and reals $x \in \left[0, \frac{1}{2q}\right]$, we have $(1-x)^{-q} \leq 1 + 2qx$.*

*Proof.* Define $\Phi(x) := \ln(1+2qx) + q\ln(1-x)$. The inequality is equivalent to $\Phi(x) \geq 0$, since

$$(1-x)^{-q} \leq 1 + 2qx \iff -q\ln(1-x) \leq \ln(1+2qx) \iff \Phi(x) \geq 0.$$

Clearly, $\Phi(0) = 0$.

Differentiating with respect to $x$, we have

$$\Phi'(x) = \frac{2q}{1+2qx} - \frac{q}{1-x} = \frac{q\big(1 - 2x(1+q)\big)}{(1+2qx)(1-x)}.$$

Since the denominator is positive on $\left[0, \frac{1}{2q}\right]$, the sign of $\Phi'(x)$ is determined by $1 - 2x(1+q)$, we have

$$\begin{cases} \Phi'(x) \geq 0 & \text{for } x \leq \frac{1}{2(1+q)}, \\ \Phi'(x) \leq 0 & \text{for } x \geq \frac{1}{2(1+q)}. \end{cases}$$

Thus $\Phi$ increases on $\left[0, \frac{1}{2(1+q)}\right]$ and decreases on $\left[\frac{1}{2(1+q)}, \frac{1}{2q}\right]$. Hence the minimum of $\Phi$ on $\left[0, \frac{1}{2q}\right]$ is attained at an endpoint, so

$$\Phi(x) \geq \min\{\Phi(0), \Phi(1/(2q))\} = \min\{0, \Phi(1/(2q))\}.$$

Now evaluating at $x = \frac{1}{2q}$ we get

$$\Phi\left(\frac{1}{2q}\right) = \ln 2 + q\ln\left(1 - \frac{1}{2q}\right).$$

So it suffices to show

$$q\ln\left(1 - \frac{1}{2q}\right) \geq -\ln 2, \quad \text{i.e.} \quad \left(1 - \frac{1}{2q}\right)^q \geq \frac{1}{2}.$$

Define $\psi(q) := q\ln\left(1 - \frac{1}{2q}\right)$. Then

$$\psi'(q) = \ln\left(1 - \frac{1}{2q}\right) + \frac{1}{2q\left(1 - \frac{1}{2q}\right)}.$$

Using the bound $\ln(1-t) \geq -\frac{t}{1-t}$ for $0 \leq t < 1$ with $t = \frac{1}{2q}$, we obtain $\psi'(q) \geq 0$. Thus $\psi$ is increasing for $q \geq 1$. At $q = 1$,

$$\psi(1) = \ln\left(\frac{1}{2}\right) = -\ln 2.$$

So $\psi(q) \geq -\ln 2$ for all $q \geq 1$. Therefore,

$$\Phi\left(\frac{1}{2q}\right) = \ln 2 + \psi(q) \geq 0.$$

Combining the above, we conclude $\Phi(x) \geq 0$ for all $x \in [0, \frac{1}{2q}]$. Exponentiation yields

$$(1-x)^{-q} \leq 1 + 2qx.$$

Hence, the claim is proved.

□

**Claim 14.** *For all $0 < q \leq 1$ and reals $x \in \left[0, \frac{1}{2}\right]$, we have $(1-x)^{-q} \leq 1 + 2qx$.*

*Proof.* Fix $x \in \left[0, \frac{1}{2}\right]$ and consider the function

$$f(q) := (1-x)^{-q}, \qquad q \in [0, 1].$$

A straightforward differentiation gives

$$f'(q) = -\ln(1-x)\,(1-x)^{-q}, \qquad f''(q) = (\ln(1-x))^2(1-x)^{-q} \geq 0,$$

so $f$ is convex on $[0, 1]$. By the convexity (the chord inequality) we have, for every $q \in [0, 1]$,

$$f(q) \leq (1-q)f(0) + qf(1) = (1-q) \cdot 1 + q \cdot (1-x)^{-1} = 1 + q\big((1-x)^{-1} - 1\big).$$

Since $(1-x)^{-1} - 1 = \dfrac{x}{1-x}$ and $1/(1-x) \leq 2$ for $x \in [0, \frac{1}{2}]$, we obtain

$$(1-x)^{-q} \leq 1 + q\frac{x}{1-x} \leq 1 + 2qx,$$

which is the desired inequality. □