Fairness and Welfare Quantification for Regret in Multi-Armed Bandits

Siddharth Barman[1], **Arindam Khan**[1], **Arnab Maiti**[2], **Ayush Sawarni**[1]

[1] Indian Institute of Science
[2] University of Washington
barman@iisc.ac.in, arindamkhan@iisc.ac.in, maitiarnab9@gmail.com, ayushsawarni@iisc.ac.in

Abstract

We extend the notion of regret with a welfarist perspective. Focussing on the classic multi-armed bandit (MAB) framework, the current work quantifies the performance of bandit algorithms by applying a fundamental welfare function, namely the Nash social welfare (NSW) function. This corresponds to equating algorithm's performance to the geometric mean of its expected rewards and leads us to the study of *Nash regret*, defined as the difference between the—a priori unknown—optimal mean (among the arms) and the algorithm's performance. Since NSW is known to satisfy fairness axioms, our approach complements the utilitarian considerations of average (cumulative) regret, wherein the algorithm is evaluated via the arithmetic mean of its expected rewards.

This work develops an algorithm that, given the horizon of play T, achieves a Nash regret of $O\left(\sqrt{\frac{k \log T}{T}}\right)$, here k denotes the number of arms in the MAB instance. Since, for any algorithm, the Nash regret is at least as much as its average regret (the AM-GM inequality), the known lower bound on average regret holds for Nash regret as well. Therefore, our Nash regret guarantee is essentially tight. In addition, we develop an anytime algorithm with a Nash regret guarantee of $O\left(\sqrt{\frac{k \log T}{T}} \log T\right)$.

1 Introduction

Regret minimization is a pre-eminent objective in the study of decision making under uncertainty. Indeed, regret is a central notion in multi-armed bandits (Lattimore and Szepesvári 2020), reinforcement learning (Sutton and Barto 2018), game theory (Young 2004), decision theory (Peterson 2017), and causal inference (Lattimore, Lattimore, and Reid 2016). The current work extends the formulation of regret with a welfarist perspective. In particular, we quantify the performance of a decision maker by applying a fundamental welfare function—namely the Nash social welfare—and study Nash regret, defined as the difference between the (a priori unknown) optimum and the decision maker's performance. We obtain essentially tight upper bounds for Nash regret in the classic multi-armed bandit (MAB) framework.

Recall that the MAB framework provides an encapsulating abstraction for settings that entail sequential decision making under uncertainty. In this framework, a decision maker (online algorithm) has sample access to k distributions (arms), which are a priori unknown. For T rounds, the online algorithm sequentially selects an arm and receives a reward drawn independently from the arm-specific distribution. Here, ex ante, the optimal solution would be to select, in every round, the arm with the maximum mean. However, since the statistical properties of the arms are unknown, the algorithm accrues—in the T rounds—expected rewards that are not necessarily the optimal. The construct of regret captures this sub-optimality and, hence, serves as a key performance metric for algorithms. A bit more formally, regret is defined as the difference between the optimal mean (among the arms) and the algorithm's performance.

Two standard forms of regret are average (cumulative) regret (Lattimore, Lattimore, and Reid 2016) and instantaneous (simple) regret (Slivkins 2019). Specifically, average regret considers the difference between the optimal mean and the average (arithmetic mean) of the expected rewards accumulated by the algorithm. Hence, in average regret the algorithm's performance is quantified as the arithmetic mean of the expected rewards it accumulates in the T rounds. Complementarily, in simple regret the algorithm's performance is equated to its expected reward precisely in the Tth round, i.e., the algorithm's performance is gauged only after $(T-1)$ rounds.

We build upon these two formulations with a welfarist perspective. To appreciate the relevance of this perspective in the MAB context, consider settings in which the algorithm's rewards correspond to values that are distributed across a population of T agents. In particular, one can revisit the classic motivating example of clinical trials (Thompson 1933): in each round $t \in \{1, \ldots, T\}$ the decision maker (online algorithm) administers one of the k drugs to the tth patient. The observed reward in round t is the selected drug's efficacy for patient t. Hence, in average regret one equates the algorithm's performance as the (average) social welfare across the T patients. It is pertinent to note that maximizing social welfare (equivalently, minimizing average regret) might not induce a fair outcome. The social welfare can in fact be high even if the drugs are inconsiderately tested on an initial set of patients. By contrast, in instantaneous regret, the algorithm's performance maps to the egalitarian (Rawlsian) welfare (which is a well-studied fairness criterion) but

only after excluding an initial set of test subjects.

The work aims to incorporate fairness and welfare considerations for such MAB settings, in which the algorithm's rewards correspond to agents' values (e.g., drug efficacy, users' experience, and advertisers' revenue). Towards this, we invoke a principled approach from mathematical economics: apply an axiomatically-justified welfare function on the values and thereby quantify the algorithm's performance. Specifically, we apply the Nash social welfare (NSW), which is defined as the geometric mean of agents' values (Moulin 2004); note that, in the current framework, for each round $t \in [T]$ of the algorithm, we have a distinct agent $t \in [T]$. Hence, we equate the algorithm's performance to the geometric mean of its T expected rewards. This leads us to the notion of Nash regret, defined as the difference between the optimal mean, μ^*, and the geometric mean of the expected rewards (see equation (1)). Note that here we are conforming to an ex-ante assessment–the benchmark (in particular, μ^*) is an expected value and the value associated with each agent is also an expectation; see Section 5 for discussions on variants of Nash regret.

NSW is an axiomatically-supported welfare objective (Moulin 2004). That is, in contrast to ad-hoc constraints or adjustments, NSW satisfies a collection of fundamental axioms, including symmetry, independence of unconcerned agents, scale invariance, and the Pigou-Dalton transfer principle (Moulin 2004). At a high level, the Pigou-Dalton principle ensures that NSW will increase under a policy change that transfers, say, δ reward from a well-off agent t to an agent $\widetilde{t}$ with lower current value.[1] At the same time, if the relative increase in $\widetilde{t}$'s value is much less than the drop in t's value, then NSW would not favor such a transfer (i.e., it also accommodates for allocative efficiency). The fact that NSW strikes a balance between fairness and economic efficiency is also supported by the observation that it sits between egalitarian and (average) social welfare: the geometric mean is at least as large as the minimum reward and it is also at most the arithmetic mean (the AM-GM inequality).

In summary, Nash social welfare induces an order among profiles of expected rewards (by considering their geometric means). Profiles with higher NSW are preferred. Our well-justified goal is to develop an algorithm that induces high NSW among the agents. Equivalently, we aim to develop an algorithm that minimizes Nash regret.

It is relevant here to note the conceptual correspondence with average regret, wherein profiles with higher social welfare are preferred. That is, while average regret is an appropriate primitive for utilitarian concerns, Nash regret is furthermore relevant when fairness is an additional desideratum.

At a conceptual level, Nash regret provides a novel quantification of the proverbial explore-vs-exploit tradeoff: Each round $t \in [T]$ corresponds to a distinct agent t. On the explore end of the tradeoff, one can pull near-optimal arms for later agents by first inconsiderately exploring in the initial rounds. Such an approach, however, can hamper the rewards of a significant fraction of the initial agents. Complementarily, towards the other end of the tradeoff, the algorithm can explore less and continue to follow its best estimate. Indeed, Nash welfare provides an axiomatically-justified quantification (as the geometric mean) towards balancing this fundamental tradeoff.

[1]Recall that NSW is defined as the geometric mean of rewards and, hence, a more balanced collection of rewards will have higher NSW.

1.1 Our Results and Techniques

We develop an algorithm that achieves Nash regret of $O\left(\sqrt{\frac{k \log T}{T}}\right)$; here, k denotes the number of arms in the bandit instance and T is the given horizon of play (Theorem 1 and Theorem 2). Note that, for any algorithm, the Nash regret is at least as much as its average regret.[2] Therefore, the known $\Omega\left(\sqrt{\frac{k}{T}}\right)$ lower bound on average regret (Auer et al. 2002) holds for Nash regret as well. This observation implies that, up to a log factor, our guarantee matches the best-possible bound, even for average regret.

We also show that the standard upper confidence bound (UCB) algorithm (Lattimore and Szepesvári 2020) does not achieve any meaningful guarantee for Nash regret (see the full version of the paper (Barman et al. 2022)). This barrier further highlights that Nash regret is a more challenging benchmark than average regret. In fact, it is not obvious if one can obtain any nontrivial guarantee for Nash regret by directly invoking upper bounds known for average (cumulative) regret. For instance, a reduction from Nash regret minimization to average regret minimization, by taking logs of the rewards (i.e., by converting the geometric mean to the arithmetic mean of logarithms), faces the following hurdles: (i) for rewards that are bounded between 0 and 1, the log values can be in an arbitrarily large range, and (ii) an additive bound for the logarithms translates back to only a multiplicative guarantee for the underlying rewards.

Our algorithm (Algorithm 1) builds upon the UCB template with interesting technical insights; see Section 3 for a detailed description. The two distinctive features of the algorithm are: (i) it performs uniform exploration for a judiciously chosen number of initial rounds and then (ii) it adds a novel (arm-specific) confidence width term to each arm's empirical mean and selects an arm for which this sum is maximum (see equation (2)). Notably, the confidence width includes the empirical mean as well. These modifications enable us to go beyond standard regret analysis.[3]

The above-mentioned algorithmic result focusses on settings in which the horizon of play (i.e., the number of rounds) T is given as input. Extending this result, we also establish a Nash regret guarantee for T-oblivious settings.

[2]This follows from the AM-GM inequality: The average regret is equal to the difference between the optimal mean, μ^*, and the arithmetic mean of expected rewards. The arithmetic mean is at least the geometric mean, which in turn is considered in Nash regret.

[3]Note that the regret decomposition lemma (Lattimore and Szepesvári 2020), a mainstay of regret analysis, is not directly applicable for Nash regret.

In particular, we develop an anytime algorithm with a Nash regret of $O\left(\sqrt{\frac{k \log T}{T}} \log T\right)$ (Theorem 3). This extension entails an interesting use of empirical estimates to identify an appropriate round at which the algorithm can switch out of uniform exploration.

1.2 Additional Related Work and Application

Given that learning algorithms are increasingly being used to guide socially sensitive decisions, there has been a surge of research aimed at achieving fairness in MAB contexts; see, e.g., (Joseph et al. 2016; Celis et al. 2019; Patil et al. 2020; Bistritz et al. 2020) and references therein. This thread of research has predominantly focused on achieving fairness for the arms. By contrast, the current work establishes fairness (and welfare) guarantees across time.

In addition, (Hossain, Micha, and Shah 2021) considers a multi-agent setting: each arm pull generates a (possibly distinct) reward among N agents. The goal in (Hossain, Micha, and Shah 2021) is to find a distribution (over the arms) that is fair for the N agents. This objective differs from identifying an arm with a high mean reward, since, for each arm, the rewards can vary across the agents. On the other hand, our work conforms to the classic MAB setup and considers fairness across rounds; each round $t \in [T]$ represents a distinct agent.

The significance of Nash social welfare and its axiomatic foundations (Kaneko and Nakamura 1979; Nash Jr 1950) in fair division settings are well established; see (Moulin 2004) for a textbook treatment. Specifically in the context of allocating divisible goods, NSW is known to uphold other important fairness and efficiency criteria (Varian 1974). In fact, NSW corresponds to the objective function considered in the well-studied convex program of Eisenberg and Gale (Eisenberg and Gale 1959). NSW is an object of active investigation in discrete fair division literature as well; see, e.g., (Caragiannis et al. 2019).

While the focus of the current paper is to develop provable algorithmic guarantees for Nash regret, we provide here an example to highlight the applicability of this fairness metric: Consider the use of MAB methods for displaying ad impressions (Schwartz, Bradlow, and Fader 2017). In this application, different ad configurations correspond to different arms, and the online users are the T agents. In round $t \in [T]$, the tth user visits the website and is shown an ad configuration (i.e., a chosen arm). The reward, for every user $t \in [T]$, is stochastic and based on the selected arm (i.e., the selected ad configuration). In this application, maximizing Nash welfare of rewards is a meaningful objective, since it qualitatively supports an online algorithm that is fair across the T agents. Indeed, Nash regret would dissuade sacrificing the experience of an initial set of users for overall utilitarian benefits.

2 Notation and Preliminaries

We study the classic (stochastic) multi-armed bandit problem. Here, an online algorithm (decision maker) has sample access to k (unknown) distributions, that are supported on $[0, 1]$. The distributions are referred to as arms $i \in \{1, \ldots, k\}$. The algorithm must iteratively select (pull) an arm per round and this process continues for $T \geq 1$ rounds overall. Successive pulls of an arm i yield i.i.d. rewards from the ith distribution. We will, throughout, write $\mu_i \in [0, 1]$ to denote the (a priori unknown) mean of the the ith arm and let μ^* be maximum mean, $\mu^* := \max_{i \in [k]} \mu_i$. Furthermore, given a bandit instance and an algorithm, the random variable $I_t \in [k]$ denotes the arm pulled in round $t \in \{1, \ldots, T\}$. Note that I_t depends on the draws observed before round t.

We address settings in which the rewards are distributed across a population of T agents. Specifically, for each agent $t \in \{1, \ldots, T\}$, the expected reward received is $\mathbb{E}[\mu_{I_t}]$ and, hence, the algorithm induces rewards $\{\mathbb{E}[\mu_{I_t}]\}_{t=1}^{T}$ across all the T agents. Notably, one can quantify the algorithm's performance by applying a welfare function on these induced rewards. Our focus is on Nash social welfare, which, in the current context, is equal to the geometric mean of the agents' expected rewards: $\left(\prod_{t=1}^{T} \mathbb{E}[\mu_{I_t}]\right)^{1/T}$. Here, the overarching aim of achieving a Nash social welfare as high as possible is quantitatively captured by considering *Nash regret*, NR_T; this metric is defined as

$$\mathrm{NR}_T := \mu^* - \left(\prod_{t=1}^{T} \mathbb{E}[\mu_{I_t}]\right)^{1/T} \tag{1}$$

Note that the optimal value of Nash social welfare is μ^*, and our objective is to minimize Nash regret.

Furthermore, the standard notion of average (cumulative) regret is obtained by assessing the algorithm's performance as the induced social welfare $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\mu_{I_t}]$. Specifically, we write R_T to denote the average regret, $\mathrm{R}_T := \mu^* - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\mu_{I_t}]$. The AM-GM inequality implies that Nash regret, NR_T, is a more challenging benchmark than R_T; indeed, for our algorithm, the Nash regret is $O\left(\sqrt{\frac{k \log T}{T}}\right)$ and the same guarantee holds for the algorithm's average regret as well.

3 The Nash Confidence Bound Algorithm

Our algorithm (Algorithm 1) consists of two phases. Phase I performs uniform exploration for $\widetilde{T} := 16\sqrt{\frac{kT \log T}{\log k}}$ rounds. In Phase II, each arm is assigned a value (see equation (2)) and the algorithm pulls the arm with the highest current value. Based on the observed reward, the values are updated and this phase continues for all the remaining rounds.

We refer to the arm-specific values as the *Nash confidence bounds*, NCB_i-s. For each arm $i \in [k]$, we obtain NCB_i by adding a 'confidence width' to the empirical mean of arm i; in particular, NCB_i depends on the number of times arm i has been sampled so far and rewards experienced for i. Formally, for any round t and arm $i \in [k]$, let $n_i \geq 1$ denote the number of times arm i has been pulled before this round.[4]

[4]Note that n_i is a random variable.

Also, for each $1 \le s \le n_i$, random variable $X_{i,s}$ be the observed reward when arm i is pulled the sth time. At this point, arm i has empirical mean $\widehat{\mu}_i \coloneqq \frac{1}{n_i}\sum_{s=1}^{n_i} X_{i,s}$ and we define the Nash confidence bound as

$$\mathrm{NCB}_i \coloneqq \widehat{\mu}_i + 4\sqrt{\frac{\widehat{\mu}_i \log T}{n_i}} \tag{2}$$

It is relevant to observe that, in contrast to standard UCB (see, e.g., (Lattimore and Szepesvári 2020)), here the confidence width includes the empirical mean (i.e., the additive term has $\widehat{\mu}_i$ under the square-root). This is an important modification that enables us to go beyond standard regret analysis. Furthermore, we note that the Nash regret guarantee of Algorithm 1 can be improved by a factor of $\sqrt{\log k}$ (see Theorem 1 and Theorem 2). The initial focus on Algorithm 1 enables us to highlight the key technical insights for Nash regret. The improved guarantee is detailed in the full version of the paper (Barman et al. 2022).

Algorithm 1: Nash Confidence Bound Algorithm

Input: Number of arms k and horizon of play T.

Initialize empirical means $\widehat{\mu}_i = 0$ and counts $n_i = 0$, for all arms $i \in [k]$. Also, set $\widetilde{T} \coloneqq 16\sqrt{\frac{kT\log T}{\log k}}$.
`Phase I`
for $t = 1$ to $\widetilde{T}$ **do**
  Select I_t uniformly at random from $\{1, 2, \ldots, k\}$.
  Pull arm I_t and observe reward X_t.
  For arm I_t, increment the count n_{I_t} (by one) and update the empirical mean $\widehat{\mu}_{I_t}$.
end for
`Phase II`
for $t = (\widetilde{T}+1)$ to T **do**
  Pull the arm I_t with the highest Nash confidence bound, i.e., $I_t = \arg\max_{i\in[k]} \mathrm{NCB}_i$.
  Observe reward X_t and update $\widehat{\mu}_{I_t}$.
  Update the Nash confidence bound for I_t (see equation (2)).
end for

The following theorem is the main result of this section and it establishes that Algorithm 1 achieves a tight—up to log factors—guarantee for Nash regret.

Theorem 1. *For any bandit instance with k arms and given any (moderately large) time horizon T, the Nash regret of Algorithm 1 satisfies*

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

3.1 Regret Analysis

We first define a "good" event G and show that it holds with high probability (Lemma 1); our Nash regret analysis is based on conditioning on G. In particular, we will first define three sub-events G_1, G_2, G_3 and set $G \coloneqq G_1 \cap G_2 \cap G_3$. For specifying these events, write $\widehat{\mu}_{i,s}$ to denote the empirical mean of arm i's rewards, based on the first s samples (of i).

G_1: Every arm $i \in [k]$ is sampled at least $\frac{\widetilde{T}}{2k}$ times in Phase I,[5] i.e., for each arm i we have $n_i \ge \frac{\widetilde{T}}{2k}$ at the end of the first phase in Algorithm 1.

G_2: For all arms $i \in [k]$, with $\mu_i > \frac{6\sqrt{k\log k\log T}}{\sqrt{T}}$, and all sample counts $\frac{\widetilde{T}}{2k} \le s \le T$ we have $|\mu_i - \widehat{\mu}_{i,s}| \le 3\sqrt{\frac{\mu_i \log T}{s}}$.

G_3: For all arms $j \in [k]$, with $\mu_j \le \frac{6\sqrt{k\log k\log T}}{\sqrt{T}}$, and all $\frac{\widetilde{T}}{2k} \le s \le T$, we have $\widehat{\mu}_{j,s} \le \frac{9\sqrt{k\log k\log T}}{\sqrt{T}}$.

Here,[6] all the events are expressed as in the canonical bandit model (see, e.g., (Lattimore and Szepesvári 2020, Chapter 4)). In particular, for events G_2 and G_3, one considers a $k \times T$ reward table that populates T independent samples for each arm $i \in [k]$. All the empirical means are obtained by considering the relevant entries from the table; see (Barman et al. 2022) for a more detailed description of the associated probability space. Also note that, conceptually, the algorithm gets to see the (i,s)th entry in the table only when it samples arm i the sth time.

The lemma below lower bounds the probability of event G; its proof appears in the full version of the paper (Barman et al. 2022).

Lemma 1. $\mathbb{P}\{G\} \ge \left(1 - \frac{4}{T}\right)$.

Next, we state a useful numeric inequality; for completeness, we provide its proof in the full version of the paper (Barman et al. 2022).

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

Now, we will show that the following key guarantees (events) hold under the good event G:

- Lemma 2: The Nash confidence bound of the optimal arm i^* is at least its true mean, μ^*, throughout Phase II.
- Lemma 3: Arms j with sufficiently small means (in particular, $\mu_j \le \frac{6\sqrt{k\log k\log T}}{\sqrt{T}}$) are never pulled in Phase II.
- Lemma 4: Arms i that are pulled many times in Phase II have means μ_i close to the optimal μ^*. Hence, such arms i do not significantly increase the Nash regret.

The proofs of these three lemmas are deferred to the full version of the paper (Barman et al. 2022). In these results, we will address bandit instances wherein the optimal mean $\mu^* \ge \frac{32\sqrt{k\log k\log T}}{\sqrt{T}}$. Note that in the complementary case (wherein $\mu^* < \frac{32\sqrt{k\log k\log T}}{\sqrt{T}}$) the Nash regret directly satisfies the bound stated in Theorem 1.

Lemma 2. *Let $\mathrm{NCB}_{i^*,t}$ be the Nash confidence bound of the optimal arm i^* at round t. Assume that the good event*

[5] Recall that $\widetilde{T} \coloneqq 16\sqrt{\frac{kT\log T}{\log k}}$.

[6] Note that if, for all arms $i \in [k]$, the means $\mu_i \le \frac{6\sqrt{k\log k\log T}}{\sqrt{T}}$, then, by convention, $\mathbb{P}\{G_2\} = 1$. Similarly, if all the means are sufficiently large, then $\mathbb{P}\{G_3\} = 1$.

G holds and also $\mu^* \geq \frac{32\sqrt{k \log k \log T}}{\sqrt{T}}$. Then, for all rounds $t > \widetilde{T}$ (i.e., for all rounds in Phase II*), we have* $\mathrm{NCB}_{i^*,t} \geq \mu^*$.

Lemma 3. *Consider a bandit instance with optimal mean $\mu^* \geq \frac{32\sqrt{k \log k \log T}}{\sqrt{T}}$ and assume that the good event G holds. Then, any arm j, with mean $\mu_j \leq \frac{6\sqrt{k \log k \log T}}{\sqrt{T}}$, is never pulled in all of Phase* II*.*

Lemma 4. *Consider a bandit instance with optimal mean $\mu^* \geq \frac{32\sqrt{k \log k \log T}}{\sqrt{T}}$ and assume that the good event G holds. Then, for any arm i that is pulled at least once in Phase* II *we have*

$$\mu_i \geq \mu^* - 8\sqrt{\frac{\mu^* \log T}{T_i - 1}},$$

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

3.2 Proof of Theorem 1

For bandit instances in which the optimal mean $\mu^* \leq \frac{32\sqrt{k \log k \log T}}{\sqrt{T}}$, the theorem holds directly; specifically, the Nash regret $\mathrm{NR}_T = \mu^* - \left(\prod_{t=1}^{T} \mathbb{E}[\mu_{I_t}]\right)^{1/T} \leq \mu^*$. Therefore, in the remainder of the proof we will address instances wherein $\mu^* \geq \frac{32\sqrt{k \log k \log T}}{\sqrt{T}}$.

The Nash social welfare of the algorithm satisfies[7]

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

In this product, the two terms account for the rewards accrued in the two phases, respectively. Next, we will lower bound these two terms.

Phase I: In each round of the first phase, the algorithm selects an arm uniformly at random. Hence, $\mathbb{E}[\mu_{I_t}] \geq \frac{\mu^*}{k}$, for each round $t \leq \widetilde{T}$. Therefore, for Phase I we have

$$\begin{aligned}
\left(\prod_{t=1}^{\widetilde{T}} \mathbb{E}[\mu_{I_t}]\right)^{\frac{1}{T}} &\geq \left(\frac{\mu^*}{k}\right)^{\frac{\widetilde{T}}{T}} = (\mu^*)^{\frac{\widetilde{T}}{T}} \left(\frac{1}{k}\right)^{\frac{16\sqrt{k \log k \log T}}{\sqrt{T \log k}}} \\
&= (\mu^*)^{\frac{\widetilde{T}}{T}} \left(\frac{1}{2}\right)^{\frac{16\sqrt{k \log k \log T} \log k}{\sqrt{T \log k}}} \\
&= (\mu^*)^{\frac{\widetilde{T}}{T}} \left(1 - \frac{1}{2}\right)^{\frac{16\sqrt{k \log k \log T}}{\sqrt{T}}} \\
&\geq (\mu^*)^{\frac{\widetilde{T}}{T}} \left(1 - \frac{16\sqrt{k \log k \log T}}{\sqrt{T}}\right) \quad (3)
\end{aligned}$$

To obtain the last inequality we note that the exponent $\frac{16\sqrt{k \log k \log T}}{\sqrt{T}} < 1$ (for appropriately large T) and, hence, the inequality follows from Claim 1.

[7]Recall that $\widetilde{T} \coloneqq 16\sqrt{\frac{kT \log T}{\log k}}$.

Phase II: For the second phase, the product of the expected rewards can be bounded as follows

$$\begin{aligned}
\left(\prod_{t=\widetilde{T}+1}^{T} \mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} &\geq \mathbb{E}\left[\left(\prod_{t=\widetilde{T}+1}^{T} \mu_{I_t}\right)^{\frac{1}{T}}\right] \\
&\geq \mathbb{E}\left[\left(\prod_{t=\widetilde{T}+1}^{T} \mu_{I_t}\right)^{\frac{1}{T}} \,\middle|\, G\right] \mathbb{P}\{G\} \quad (4)
\end{aligned}$$

Here, the first inequality follows from the multivariate Jensen's inequality and the second one is obtained by conditioning on the good event G. To bound the expected value in the right-hand-side of inequality (4), we consider the arms that are pulled at least once in Phase II. In particular, with reindexing, let $\{1, 2, \ldots, \ell\}$ denote the set of all arms that are pulled at least once in the second phase. Also, let $m_i \geq 1$ denote the number of times arm $i \in [\ell]$ is pulled in Phase II and note that $\sum_{i=1}^{\ell} m_i = T - \widetilde{T}$. Furthermore, let T_i denote the total number of times any arm i is pulled in the algorithm. Indeed, $(T_i - m_i)$ is the number of times arm $i \in [\ell]$ is pulled in Phase I. With this notation, the expected value in the right-hand-side of inequality (4) can be expressed as

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

Moreover, since we are conditioning on the good event G, Lemma 4 applies to each arm $i \in [\ell]$. Hence,

$$\begin{aligned}
\mathbb{E}\left[\left(\prod_{t=\widetilde{T}}^{T} \mu_{I_t}\right)^{\frac{1}{T}} \,\middle|\, G\right] &= \mathbb{E}\left[\left(\prod_{i=1}^{\ell} \mu_i^{\frac{m_i}{T}}\right) \,\middle|\, G\right] \\
&\geq \mathbb{E}\left[\prod_{i=1}^{\ell} \left(\mu^* - 8\sqrt{\frac{\mu^* \log T}{T_i - 1}}\right)^{\frac{m_i}{T}} \,\middle|\, G\right] \quad \text{(Lemma 4)} \\
&= (\mu^*)^{1-\frac{\widetilde{T}}{T}} \mathbb{E}\left[\prod_{i=1}^{\ell} \left(1 - 8\sqrt{\frac{\log T}{\mu^*(T_i - 1)}}\right)^{\frac{m_i}{T}} \,\middle|\, G\right] \quad (5)
\end{aligned}$$

For the last equality, we use $\sum_{i=1}^{\ell} m_i = T - \widetilde{T}$. Now, recall that, under event G, each arm is pulled at least $\frac{\widetilde{T}}{2k} = \frac{8}{k}\sqrt{\frac{kT \log T}{\log k}}$ times in Phase I. Hence, $T_i > \frac{\widetilde{T}}{2k}$ for each arm $i \in [\ell]$. Furthermore, since $\mu^* \geq \frac{32\sqrt{k \log k \log T}}{\sqrt{T}}$, we have $8\sqrt{\frac{\log T}{\mu^*(T_i - 1)}} \leq 8\sqrt{\frac{1}{256}} = \frac{1}{2}$ for each $i \in [\ell]$. Therefore, we can apply Claim 1 to reduce the expected value in inequality

(5) as follows

$$\mathbb{E}\left[\prod_{i=1}^{\ell}\left(1-8\sqrt{\frac{\log T}{\mu^*(T_i-1)}}\right)^{\frac{m_i}{T}}\,\middle|\, G\right] \geq \mathbb{E}\left[\prod_{i=1}^{\ell}\left(1-\frac{16m_i}{T}\sqrt{\frac{\log T}{\mu^*(T_i-1)}}\right)\,\middle|\, G\right] \geq \mathbb{E}\left[\prod_{i=1}^{\ell}\left(1-\frac{16}{T}\sqrt{\frac{m_i\log T}{\mu^*}}\right)\,\middle|\, G\right] \quad \text{(as } T_i \geq m_i+1\text{)}$$

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

$$\begin{aligned}
&\mathbb{E}\left[\prod_{i=1}^{\ell}\left(1-\frac{16}{T}\sqrt{\frac{m_i\log T}{\mu^*}}\right)\,\middle|\, G\right] \\
&\geq \mathbb{E}\left[1-\sum_{i=1}^{\ell}\left(\frac{16}{T}\sqrt{\frac{m_i\log T}{\mu^*}}\right)\,\middle|\, G\right] \\
&= 1-\left(\frac{16}{T}\sqrt{\frac{\log T}{\mu^*}}\right)\mathbb{E}\left[\sum_{i=1}^{\ell}\sqrt{m_i}\,\middle|\, G\right] \\
&\geq 1-\left(\frac{16}{T}\sqrt{\frac{\log T}{\mu^*}}\right)\mathbb{E}\left[\sqrt{\ell}\sqrt{\sum_{i=1}^{\ell}m_i}\,\middle|\, G\right] \quad \text{(Cauchy-Schwarz inequality)} \\
&\geq 1-\left(\frac{16}{T}\sqrt{\frac{\log T}{\mu^*}}\right)\mathbb{E}\left[\sqrt{\ell\,T}\,\middle|\, G\right] \quad \text{(as } \textstyle\sum_i m_i \leq T\text{)} \\
&= 1-\left(16\sqrt{\frac{\log T}{\mu^* T}}\right)\mathbb{E}\left[\sqrt{\ell}\,\middle|\, G\right] \\
&\geq 1-\left(16\sqrt{\frac{k\log T}{\mu^* T}}\right) \qquad (6)
\end{aligned}$$

Here, the final inequality holds since $\ell \leq k$. Using (6), along with inequalities (4), and (5), we obtain for Phase II:

$$\left(\prod_{t=\widetilde{T}+1}^{T}\mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} \geq (\mu^*)^{1-\frac{\widetilde{T}}{T}}\left(1-16\sqrt{\frac{k\log T}{\mu^* T}}\right)\mathbb{P}\{G\} \qquad (7)$$

Inequalities (7) and (3) provide relevant bounds for Phase II and Phase I, respectively. Hence, for the Nash social welfare of the algorithm we have

$$\begin{aligned}
&\left(\prod_{t=1}^{T}\mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} \\
&\geq \mu^*\left(1-\frac{16\sqrt{k\log k\log T}}{\sqrt{T}}\right)\left(1-16\sqrt{\frac{k\log T}{\mu^* T}}\right)\mathbb{P}\{G\} \\
&\geq \mu^*\left(1-\frac{16\sqrt{k\log k\log T}}{\sqrt{T}}\right)\left(1-16\sqrt{\frac{k\log T}{\mu^* T}}\right)\left(1-\frac{4}{T}\right) \quad \text{(via Lemma 1)} \\
&\geq \mu^*\left(1-\frac{32\sqrt{k\log k\log T}}{\sqrt{\mu^* T}}\right)\left(1-\frac{4}{T}\right) \\
&\geq \mu^*-\frac{32\sqrt{\mu^* k\log k\log T}}{\sqrt{T}}-\frac{4\mu^*}{T} \\
&\geq \mu^*-\frac{32\sqrt{k\log k\log T}}{\sqrt{T}}-\frac{4}{T} \quad \text{(since } \mu^* \leq 1\text{)}
\end{aligned}$$

Therefore, the Nash regret of the algorithm satisfies

$$\begin{aligned}
\mathrm{NR}_T &= \mu^*-\left(\prod_{t=1}^{T}\mathbb{E}\left[\mu_{I_t}\right]\right)^{\frac{1}{T}} \\
&\leq \frac{32\sqrt{k\log k\log T}}{\sqrt{T}}+\frac{4}{T}.
\end{aligned}$$

Overall, we get that $\mathrm{NR}_T = O\left(\sqrt{\frac{k\log k\log T}{T}}\right)$. The theorem stands proved.

Remark. Algorithm 1 is different from standard UCB, in terms of both design and analysis. For instance, here the empirical means appear in the confidence width and impact the concentration bounds utilized in the analysis.

3.3 Improved Nash Regret Guarantee

As mentioned previously, the Nash regret guarantee obtained in Theorem 1 can be improved by a factor of $\sqrt{\log k}$. To highlight the key technical insights, in Algorithm 1 we fixed the number of rounds in Phase I (to $\widetilde{T}$). However, with an adaptive approach, one can obtain a Nash regret of $O\left(\sqrt{\frac{k\log T}{T}}\right)$, as stated below in Theorem 2. A detailed description of the modified algorithm and the proof of Theorem 2 are deferred to the full version of the paper (Barman et al. 2022).

Theorem 2. *For any bandit instance with k arms and given any (moderately large) T, there exists an algorithm that achieves Nash regret*

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

4 Anytime Algorithm

The section provides a Nash regret guarantee for settings in which the horizon of play T is not known in advance. Here, in addition to the doubling trick, multiple new ideas are required. A key insight in the anytime algorithm[8] appears in the subroutine Modified NCB (see the full version of the paper (Barman et al. 2022)). This subroutine is different from Algorithm 1. Recall that in Algorithm 1 we uniformly sample the arms for essentially $\sqrt{T}$ initial rounds. To achieve similar sampling estimates in the anytime setting, we would need to uniformly sample about $\sqrt{T}$ times, whenever the window size W (guessed via doubling) is greater than $\sqrt{T}$. This, however, is no longer possible, since we do not know T to begin with, i.e., we do not know how the current window size W compares with T.

We overcome this barrier via a novel idea: we uniform sample until the sum of rewards for any arm i exceeds a judiciously chosen threshold. Specifically, let n_i denote the number of times an arm i has been pulled so far and $X_{i,s}$ denote the reward observed for arm i when it is pulled the sth time. The anytime exploration continues as long as $\max_i \sum_{s=1}^{n_i} X_{i,s} \leq 420c^2 \log W$; here c is an absolute constant.

Moreover, in the Nash context, one needs to modify the standard doubling trick as well. In the standard doubling-trick method, the algorithm calls a time-dependent subroutine with an initial guess for the time horizon, W, and subsequently doubles this guess if the number of rounds exceeds W. In the case of standard (cumulative) regret, this idea works because, even if an algorithm performs poorly for an initial guess of W, it could cover up in later rounds due to the additive nature of the net reward. We do not have this luxury in the case of Nash regret, and a direct implementation of the doubling trick would perform poorly. Hence, in addition to doubling the guess W, the algorithm also performs uniform exploration with probability $1 - \frac{1}{W^2}$.

We develop an algorithm in the full version of the paper (Barman et al. 2022) that builds on these ideas and leads to the following theorem.

Theorem 3. *There exists an anytime algorithm that, at any (moderately large) round T, achieves a Nash regret*

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

5 Ostensible Alternative

Recall that each round $t \in [T]$ corresponds to a distinct agent t and this work followed a standard ex-ante assessment; the value associated with each agent t is the expected reward in round t, i.e., $\mathbb{E}\left[\mu_{I_t}\right]$. With the hindsight optimal (in particular, μ^*) being an expected value as well, we obtained the construct of Nash regret, NR_T, by considering the difference between μ^* and the geometric mean of the agents' expected rewards: $\mathrm{NR}_T := \mu^* - \left(\prod_{t=1}^{T} \mathbb{E}[\mu_{I_t}]\right)^{1/T}$. As mentioned previously, NR_T is a more challenging benchmark than average regret (the AM-GM inequality) and, hence, our algorithm provably strengthens standard regret guarantees.

Now, from a technical point of view, one can also define the following variant of Nash regret

$$\mathrm{NR}_T^{(1)} := \mu^* - \mathbb{E}_{I_1,\ldots I_T}\left[\left(\prod_{t=1}^{T} \mu_{I_t}\right)^{\frac{1}{T}}\right].$$

While $\mathrm{NR}_T^{(1)}$ upper bounds Nash regret NR_T (see the full version of the paper (Barman et al. 2022)), it does not conform to a per-agent ex ante assessment. Instead, $\mathrm{NR}_T^{(1)}$ directly considers the expected value of the Nash welfare generated across the population.

Moreover, one *cannot* obtain bounds for $\mathrm{NR}_T^{(1)}$ that hold through all the rounds: consider a bandit instance in which all, expect one of the arms (i.e., all except the optimal arm), have zero rewards. As soon as, in the initial (say k) rounds one of these arms get pulled, $\mathrm{NR}_T^{(1)}$ becomes as high as μ^* and cannot be salvaged.

At the same time, we note the regret guarantee we obtain for Phase II of Algorithm 1 in fact holds for $\mathrm{NR}_T^{(1)}$; see inequality (4) and the following analysis. This observation implies that Algorithm 1 obtains a guarantee even in terms of $\mathrm{NR}_T^{(1)}$ for the last $\left(T - \widetilde{T}\right)$ agents.

6 Conclusion and Future Work

This work considers settings in which a bandit algorithm's expected rewards, $\{\mathbb{E}\left[\mu_{I_t}\right]\}_{t=1}^{T}$, correspond to values distributed among T agents. In this ex ante framework, we apply Nash social welfare (on the expected rewards) to evaluate the algorithm's performance and thereby formulate the notion of Nash regret. Notably, in cumulative regret, the algorithm is assessed by the social welfare it generates. That is, while cumulative regret captures a utilitarian objective, Nash regret provides an axiomatically-supported primitive for achieving both fairness and economic efficiency.

We establish an instance-independent (and essentially tight) upper bound for Nash regret. Obtaining a Nash regret bound that explicitly depends on the gap parameters, $\Delta_i := \mu^* - \mu$, is an interesting direction of future work. It would also be interesting to formulate regret under more general welfare functions. Specifically, one can consider the generalized-mean welfare (Moulin 2004) which—in the current context and for parameter $p \in (-\infty, 1]$—evaluates to $\left(\frac{1}{T}\sum_t \mathbb{E}\left[\mu_{I_t}\right]^p\right)^{1/p}$. Generalized-means encompass various welfare functions, such as social welfare ($p = 1$), egalitarian welfare ($p \to -\infty$), and Nash social welfare ($p \to 0$). Hence, these means provide a systematic tradeoff between fairness and economic efficiency. Studying Nash regret in broader settings—such as contextual or linear bandits—is a meaningful research direction as well.

Acknowledgments

Arindam Khan was partially supported by Pratiksha Trust Young Investigator Award, Google India Research Award, and Google ExploreCS Award. Siddharth Barman gratefully

[8] Recall that a bandit algorithm is said to be anytime iff does not need to know horizon of play T in advance.

acknowledges the support of a SERB Core research grant (CRG/2021/006165).

References

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

Barman, S.; Khan, A.; Maiti, A.; and Sawarni, A. 2022. Fairness and Welfare Quantification for Regret in Multi-Armed Bandits.

Bistritz, I.; Baharav, T.; Leshem, A.; and Bambos, N. 2020. My fair bandit: Distributed learning of max-min fairness with multi-player bandits. In *International Conference on Machine Learning*, 930–940. PMLR.

Caragiannis, I.; Kurokawa, D.; Moulin, H.; Procaccia, A. D.; Shah, N.; and Wang, J. 2019. The unreasonable fairness of maximum Nash welfare. *ACM Transactions on Economics and Computation (TEAC)*, 7(3): 1–32.

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.

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

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

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.

Lattimore, F.; Lattimore, T.; and Reid, M. D. 2016. Causal bandits: Learning good interventions via causal inference. *Advances in Neural Information Processing Systems*, 29.

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

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

Nash Jr, J. F. 1950. The bargaining problem. *Econometrica: Journal of the econometric society*, 155–162.

Patil, V.; Ghalme, G.; Nair, V.; and Narahari, Y. 2020. Achieving fairness in the stochastic multi-armed bandit problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5379–5386.

Peterson, M. 2017. *An introduction to decision theory*. Cambridge University Press.

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.

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

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

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.

Varian, H. R. 1974. Equity, envy, and efficiency. *Journal of Economic Theory*, 9(1): 63–91.

Young, H. P. 2004. *Strategic learning and its limits*. OUP Oxford.