Nash Social Welfare for Multi Armed Bandits: Trajectory-wise Expected and High Probability Regret

Avishek Ghosh
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay
avishek_ghosh@iitb.ac.in

Abstract

We study fair multi-armed bandits under the Nash Social Welfare (NSW) objective, which measures performance via the geometric mean of accumulated rewards as a fairness-respecting performance metric. Existing formulations define Nash regret as $\mathrm{NR}_T = \mu^\star - (\prod_{t=1}^T \mathbb{E}\mu_{I_t})^{1/T}$, where μ_{I_t} is the mean reward of the recommended arm I_t and T is the learning horizon. We observe that, since ensemble Nash regret operates on per-round marginal expectations before applying the geometric mean, it does not capture the joint distribution of rewards across rounds — leaving an aspect of the NSW fairness motivation unaddressed at the trajectory level.

To address this, we propose *trajectory-wise Nash regret* $\widetilde{\mathrm{NR}}_T = \mu^\star - \mathbb{E}[(\prod_{t=1}^T \mu_{I_t})^{1/T}]$, where the geometric mean is computed over complete sample paths before taking expectations. This formulation captures the fairness properties of NSW more faithfully and requires controlling the joint distribution of rewards across rounds, rather than merely per-round marginals. By Jensen's inequality applied to the concave geometric mean, $\widetilde{\mathrm{NR}}_T \geq \mathrm{NR}_T$, making it a strictly stronger metric. We additionally introduce *high probability Nash regret* $\widehat{\mathrm{NR}}_T = \mu^\star - (\prod_t \mu_{I_t})^{1/T}$, obtaining the first high probability regret bounds in the fair bandits literature. We propose a two-phase algorithm, Round Robin Nash Confidence Bound (`RR-NCB`), combining structured round robin exploration with a Nash confidence bound index policy. We establish that $\widetilde{\mathrm{NR}}_T \leq \widetilde{\mathcal{O}}(\sqrt{k \log T/T})$ and $\widehat{\mathrm{NR}}_T \leq \widetilde{\mathcal{O}}(\sqrt{k \log(kT/\delta)/T})$ with probability $1-\delta$, matching the optimal $\widetilde{\mathcal{O}}(\sqrt{k/T})$ scaling despite operating under strictly stronger metrics. Optimality follows from a lower bound chain via AM-GM and standard k-armed bandit minimax arguments. We validate our theoretical findings through simulations.

1 Introduction

We consider the canonical framework of multi-armed bandit where an agent is interacting with the environment with k arms for T rounds. At every time instant, the agent pulls an arm out of these and receives a random reward. The rewards received by the agent are iid across rounds with mean reward corresponding to arms $\{1, \ldots, k\}$ be denoted by $\mu_1, \ldots, \mu_k$ apriori unknown. We assume $\mu_i \in (0, 1]$ for all $i \in [k]$ and $\mu^\star := \max_{i \in [k]} \mu_i$. Without loss of generality, fix $i^\star \in \arg\max_i \mu_i$. Typically, to characterize the performance of a learning algorithm, one considers the notion of cumulative or average regret given by the difference between the mean reward of best arm μ^* and the arithmetic mean of accumulated rewards over rounds $\frac{1}{T}\sum_{i=1}^T \mathbb{E}\,\mu_{I_t}$, where $\{\mu_{I_t}\}_{i=1}^T$ are the recommendations of the leaning algorithm (Lattimore and Szepesvári [2020], Bubeck and Cesa-Bianchi [2012]).

However, in welfare economics, it is well understood that the arithmetic mean overlooks the fairness aspect completely (Taylor [2004], Kaneko and Nakamura [1979], Kaneko [1981], Caragiannis et al.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

[2019]). This becomes critical in certain applications like clinical trials and drug recommendation. A natural alternative is to use axiomatically-justified welfare function namely Nash Social Welfare (NSW), where one takes geometric mean (instead of arithmetic mean) of the accumulated rewards. As discussed in Taylor [2004], Caragiannis et al. [2019] NSW satisfies several fundamental axioms like scale invariance, symmetry, independence of unconcerned agents and the Pigou-Dalton transfer principle (Robinhood principle). In a nutshell, in our context, the Pigou-Dalton principle implies that NSW will increase in the event where a well-off arm (with high reward) transfers δ reward to a low reward (value) arm. At the same time, if the relative increase in the value of low reward arm is significantly less than the relative decrease in the value of the well-off arm, NSW does not favor such a transfer, striking a balance between fairness and efficiency. In Kaneko and Nakamura [1979], it is argued that NSW is the unique welfare function satisfying all the above-mentioned axioms.

To this end Barman et al. [2023] introduces the notion of *fair bandits*, where the goal is to minimize the Nash regret given by $\mathrm{NR}_T = \mu^\star - [(\prod_{t=1}^T \mathbb{E}\,\mu_{I_t})^{1/T}]$. Note that this formulation respects the fairness aspect as it considers the geometric mean of the accumulated reward. A learning algorithm with (optimal) regret scaling of $\tilde{\mathcal{O}}(\sqrt{K/T})$ is given for bounded (between $[0,1]$) rewards. Later Sarkar et al. [2025] extended this result to capture sub-Gaussian rewards and obtain similar regret. Moreover Krishna et al. [2025] study the p mean regret, where the ℓ_p norm of the accumulated reward is considered as a fair metric. Note that when $p \to 0$, this metric boils down to the NSW.

In all the above mentioned papers, the Nash Social Welfare function is defined by taking the expectation over the algorithm's randomness *first* and then applying the geometric mean. More formally, at every round t, one computes the ensemble average $\mathbb{E}[\mu_{I_t}]$ across sample paths, yielding a sequence of non-random quantities $(\mathbb{E}[\mu_{I_1}], \ldots, \mathbb{E}[\mu_{I_T}])$, to which the geometric mean is then applied. The distinction between this ensemble NSW and the arithmetic mean therefore reduces to the choice of aggregation function on this *same* non-random sequence: $f_1(z_1, \ldots, z_T) = \frac{1}{T}\sum_{i=1}^T z_i$ versus $f_2(z_1, \ldots, z_T) = \left(\prod_{i=1}^T z_i\right)^{1/T}$. Since these are related by a log transform, $\log f_2(z_1, \ldots, z_T) = f_1(\log z_1, \ldots, \log z_T)$, optimizing f_2 over policies is equivalent to optimizing f_1 on log-transformed marginals. This suggests that, at the level of per-round marginal expectations, ensemble Nash regret and the arithmetic mean objective may share more structural similarity than the NSW motivation implies — and crucially, neither captures the *joint* distribution of rewards across rounds, which is where trajectory-level fairness resides.

To address this, we propose a modified NSW that moves the expectation operator *outside* the geometric mean: $\widetilde{\mathrm{NSW}}_T = \mathbb{E}[(\prod_{t=1}^T \mu_{I_t})^{1/T}]$, where the geometric mean is computed over a complete sample path first, and then averaged across trajectories. Unlike the ensemble formulation, this requires controlling the joint distribution of $(\mu_{I_1}, \ldots, \mu_{I_T})$ rather than merely the per-round marginals $\mathbb{E}[\mu_{I_t}]$, and no deterministic log-transform reduces it to an arithmetic mean.

By Jensen's inequality applied to the concave geometric mean function, we have $\widetilde{\mathrm{NSW}}_T \leq \mathrm{NSW}_T$. Since we look at the complete trajectory, the corresponding regret is defined as Trajectory-wise Nash regret and the we call the original Nash regret as Ensemble Nash regret. We have,

$$\widetilde{\mathrm{NR}}_T \;=\; \mu^\star \;-\; \mathbb{E}[(\prod_{t=1}^T \mu_{I_t})^{1/T}] \geq \mathrm{NR}_T.$$

$\widetilde{\mathrm{NR}}_T$ is a stronger metric compared to NR_T and upper bounds on NR_T does not translate.

Moreover, in this paper we also study the random variable $(\prod_{t=1}^T \mu_{I_t})^{1/T}$ trajectory-wise and obtain high probability regret bound. There are important settings where averaging is not appropriate. In a clinical trial, for instance, each patient population experiences a single, irreversible sequence of treatment decisions — there is no ensemble of parallel runs to average over. What matters is the welfare of the actual trajectory, not its expectation. To the best of our knowledge, this is the first work to obtain a high probability regret on fair bandits. We define

$$\widehat{\mathrm{NR}}_T := \mu^\star - \left(\prod_t \mu_{I_t}\right)^{1/T}.$$

1.1 Summary of Contributions

We summarize our contribution in Table 1.1. We mention the details here.

	Ensemble Regret	Trajectory-wise Regret	High Prob. Regret	Regret Scaling
Barman et al. [2023]	✓	✗	✗	$\tilde{\mathcal{O}}(\sqrt{K/T})$
Krishna et al. [2025]	✓	✗	✗	$\tilde{\mathcal{O}}(\sqrt{K/T})$
Sarkar et al. [2025]	✓	✗	✗	$\tilde{\mathcal{O}}(\sqrt{K/T})$
This paper	✓	✓	✓	$\tilde{\mathcal{O}}(\sqrt{K/T})$

Table 1: Comparison of regret guarantees across methods. Barman et al. [2023], Sarkar et al. [2025], Krishna et al. [2025] only works with ensemble regret, whereas we explicitly handle trajectory-wise and high probability regret. Since $\widetilde{\mathrm{NR}}_T \geq \mathrm{NR}_T$, our framework also gives an upper bound for ensemble regret.

- *Matching Regret Upper Bound:* We propose a learning algorithm, namely Round Robbin Nash Confidence Bound (`RR-NCB`) which uses round robbin exploration along with an index policy governed by the Nash confidence bound (see Algorithm 4) for details). We show that

$$\widetilde{\mathrm{NR}}_T \leq \tilde{\mathcal{O}}\big(\sqrt{\tfrac{k\log T}{T}}\big) \quad \text{and} \quad \Pr\Big[\widehat{\mathrm{NR}}_T \leq \tilde{\mathcal{O}}\big(\sqrt{\tfrac{k\log(8kT/\delta)}{T}}\big)\Big] \geq 1-\delta,$$

  where $\tilde{\mathcal{O}}(.)$ hides other log factors. Note that the regret scaling is $\tilde{\mathcal{O}}(\sqrt{k/T})$ (ignoring log factor), which matches the regret upper bound of Barman et al. [2023], Sarkar et al. [2025] and Krishna et al. [2025] (for $p \to 0$). This implies that although $\widetilde{\mathrm{NR}}_T \geq \mathrm{NR}_T$, we were able to obtain the same regret scaling. Hence, the trajectory-wise Nash regret, which is a stronger metric comes with no additional cost in terms of regret. The same conclusion holds for high probability regret bound as well. To the best our knowledge no previous work has obtained high probability Nash regret.
- *Matching Regret Lower Bound:* Invoking the Jensen's inequality, we have the following chain of inequalities:

$$\widetilde{\mathrm{NR}}_T \geq \mathrm{NR}_T \geq \mu^* - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\,\mu_{I_{I_t}} \geq \Omega(\sqrt{k/T}),$$

  where the second inequality comes from AM-GM inequality and the third inequality is the standard lower bound for k-armed bandits (see Lattimore and Szepesvári [2020], Bubeck and Cesa-Bianchi [2012]). Hence, the regret achieved by `RR-NCB` is optimal. Similar argument holds for the high probability bound as well. We have

$$\widehat{\mathrm{NR}}_T \geq \mu^* - \frac{1}{T}\sum_{t=1}^{T} \mu_{I_{I_t}} \geq \Omega\big(\sqrt{k/T}\log(1/\delta)\big),$$

  with probability at least $1-\delta$. Note that this is a sample path-wise (trajectory-wise) regret. Hence `RR-NCB` yields optimal regret for both Trajectory-wise $\widetilde{\mathrm{NR}}_T$ and high probability $\widehat{\mathrm{NR}}_T$ Nash regret.
- *Simulations:* We validate out theoretical finding through simulations. In particular we show that trajectory-wise average $\widetilde{\mathrm{NR}}_T$ yields worse regret compared to NR_T as predicted by theory. However, we show that by playing `RR-NCB`, the gap can be made small. The details are deferred to Section 7.

1.2 Technical Novelty

The analysis of Trajectory-wise and high probability Nash regret comes with additional technical challenges. Here we succinctly summarize a few of them.

- *Trajectory-wise Jensen-on-exp identity:* The geometric mean satisfies an *exact* trajectory wise identity:

$$\Big(\prod_t \mu_{I_t}\Big)^{1/T} \;=\; \mu^\star \cdot \exp(-\widehat{R}_T) \quad \text{on every sample path,}$$

where $\widehat{R}_T = (1/T)\sum_t \log(\mu^\star/\mu_{I_t})$ is the realized cumulative log-regret. Now, we use Jensen's inequality on the exponential function. Combined with the identity $1 - e^{-x} \le x$, we obtain $\widetilde{\mathrm{NR}}_T \le \mu^\star \bar{R}_T$ where $\bar{R}_T = \mathbb{E}[\widehat{R}_T]$.

On the other hand, for high probability regret, we can skip the Jensen step and directly invoke $1 - e^{-x} \le x$ to obtain $\widehat{\mathrm{NR}}_T \le \mu^\star \widehat{R}_T$. Hence, the high-probability analysis of $\widehat{\mathrm{NR}}_T$ is cleaner than the in-expectation analysis of $\widetilde{\mathrm{NR}}_T$. By contrast, controlling $\widehat{NR}_T$ via concentrating $(\prod_t \mu_{I_t})^{1/T}$ around its expectation (the naive route) would be substantially more involved.

- *Unified analysis and Log regret as unifying intermediate:* We argue that both regret are reduced to the same intermediate object — the cumulative log-regret. This enables us to give a unified analysis that holds for both trajectory-wise and high probability Nash regret.
- *Round Robbin Exploration:* In Barman et al. [2023], Sarkar et al. [2025], thanks to the expectation inside geometric mean, a simple uniform exploration gives $\mathbb{E}\,\mu_{I_t} \ge \mu^*/k$ trivially. However, such a lower bound cannot be applied in out setup. Hence, in our learning algorithm `RR-NCB`, we design the first phase as round robbin exploration. We design the length of such exploration carefully and include a stopping criteria so that every arm gets pulled sufficiently and at the same time the length of exploration period is not too large (otherwise the regret would be impacted).
- *Good Event and bad arms:* We define a good event E judiciously and show (in Lemma A.7 that on E, bad arms (with mean reward $\mu_j \le \mu^\star/64$) are never played in Phase 2. This crucially controls the regret in Phase 2. Moreover, under E, the mean estimate of the good arms (with $\mu_j > \mu^\star/64$) stay close to the actual mean, incurring small regret.

1.3 Motivating Applications of Trajectory-wise Regret

Clinical Trials: The trajectory-level view is more appropriate when we have *non-replayability* like in clinical trials (the motivating example of Barman et al. [2023], Sarkar et al. [2025]). Here, each patient is treated exactly once. The sequence of treatments administered is a single trajectory. A metric that averages over multiple instantiations of the algorithm's randomization does not reflect what actually happens to the population of patients in any given run. In such cases, a high probability per-trajectory regret is of importance, which our framework provides (see high probability Nash regret, Theorem 5.5).

Risk-Sensitive Control: In classical risk-sensitive control (Howard and Matheson [1972], Whittle [1990] as well as the Linear-Exponential-Quadratic-Gaussian framework (Biswas and Borkar [2023]) the objective is to minimize the expected cumulative cost given by $\mathbb{E}(\exp(-\theta C))$, were θ is the risk sensitive parameter and C is the cost incurred in one trajectory. As discussed in Biswas and Borkar [2023], the expectation outside encodes risk sensitivity, and reducing it to $\exp(-\theta\,\mathbb{E}\,C)$ collapses into risk neutral setup.

Portfolio Optimization and Kelly Criteria: Consider a sequential investment problem where at each round t, a fraction of wealth is allocated to asset I_t yielding a (multiplicative) return R_t. After T rounds, terminal wealth is $W_T = \prod_{t=1}^T R_t$. The Kelly criterion Kelly [1956] maximizes $\mathbb{E}\,W_T$ which is equivalent to trajectory-wise expected regret.

Ergodicity Economics: Peters [2019] argues that for non-ergodic multiplicative processes, the time-average growth rate $\lim_{T\to\infty}(\prod_t X_t)^{1/T}$ differs from the ensemble average $\mathbb{E}[X_t]$, and rational agents should optimize the former.

Evolutionary Biology: The geometric mean fitness principle (Dempster [1955], Lewontin and Cohen [1969]) states that long-run reproductive success in varying environments is governed by $\mathbb{E}[(\prod_t W_t)^{1/T}]$, not $(\prod_t \mathbb{E}[W_t])^{1/T}$, where W_t is the fitness in generation t. Strategies producing occasional near-zero fitness are selected against even if their arithmetic mean fitness is high.

2 Other Related Work

Fairness in Online Learning and Bandits: Joseph et al. [2016] initiates meritocratic fairness in multi-armed bandits, requiring that a worse arm is never preferred over a better one, with tight

regret bounds. Joseph et al. [2018] extends this to contextual bandits under a Rawlsian notion with polynomial regret bounds. Gillen et al. [2018] studies individual fairness in linear contextual bandits with an unknown Mahalanobis similarity metric, learning constraints from weak regulatory feedback. Wang et al. [2021] formalizes exposure fairness in stochastic bandits, designing algorithms that jointly minimize reward and fairness regret.

Nash Regret in Other Settings: Building on Barman et al. [2023], Sawarni et al. [2023] extends Nash regret to stochastic linear bandits with tight upper bounds via log-transformed reward maximization. Sarkar et al. [2026] resolves the open problem of dimension-optimal Nash regret in linear bandits and initiates the study of p-mean regret, proposing FAIRLINBANDIT. Zhang et al. [2024] extends the NSW objective to multi-agent bandits with sublinear regret guarantees and matching lower bounds in both stochastic and adversarial settings.

Axiomatic Fairness in Mathematical Economics: Nash [1950] introduces the Nash bargaining solution via axioms of symmetry, scale invariance, independence of irrelevant alternatives, and Pareto optimality. Kaneko and Nakamura [1979] proves NSW is the *unique* social welfare function satisfying classical social choice axioms without interpersonal utility comparisons. Taylor [2004] formalizes the Pigou-Dalton transfer principle, positioning NSW between utilitarian and egalitarian objectives. Caragiannis et al. [2019] shows maximum Nash welfare simultaneously achieves EF1 and Pareto optimality. Cousins [2023] characterizes the p-mean welfare family in a machine learning context, identifying NSW as its unique fair member.

3 Problem Formulation

We work with the canonical stochastic bandit model. where the sample space is $\Omega = [k]^T \times [0,1]^{k\times T}$ equipped with the product σ-algebra. An element $\omega = ((I_1, \ldots, I_T), (Y_{i,s})_{i\in[k],s\in[T]})$ specifies the algorithm's pull sequence and a $Y_{i,s}$ is the reward of arm i on its s-th pull. Throughout, we assume that rewards $Y_{i,s}$ are in $[0,1]$ for all $i \in [k]$ and $s \in [T]$. Moreover, the rewards obtained from each arm are i.i.d. with with mean μ_i for the i-th arm. Recall that we assume $\mu_i \in (0,1]$ for all $i \in [k]$.

Any learning algorithm $\mathcal{A}$ recommends arm I_t for the learner at time t based on past history. As mentioned in Section 1, the objective is to characterize the *trajectory-wise Nash regret* as well as the *high probability Nash regret* given respectively by

$$\widetilde{\mathrm{NR}}_T \;=\; \mu^\star \;-\; \mathbb{E}[(\prod_{t=1}^{T} \mu_{I_t})^{1/T}] \geq \mathrm{NR}_T, \quad \widehat{\mathrm{NR}}_T := \mu^\star - \big(\prod_{t} \mu_{I_t}\big)^{1/T}.$$

First, we show that a positive floor assumption on μ_i is needed for tractability.

Assumption 3.1 (Positivity floor)**.** There exists a constant $\mu_{\min} \in (0, \mu^\star]$ such that $\mu_i \geq \mu_{\min}$ for every $i \in [k]$.

We now argue that such assumption is unavoidable for handling the trajectory-wise and high probability Nash regret. The following proposition formally shows that any algorithm $\mathcal{A}$ which pulls every arm with positive probability at least once in T rounds, the modified Nash regret cannot go to zero, and hence learning is not possible.

Proposition 3.2. *Fix $k \geq 2$. For any algorithm $\mathcal{A}$ that pulls every arm with positive probability at least once over T rounds, there exists a bandit instance ν with $\mu^\star = 1$ and $\min_i \mu_i = 0$ on which*

$$\widetilde{\mathrm{NR}}_T(\mathcal{A}, \nu) \;\geq\; \mu^\star \cdot \Pr\big[\exists\, t \leq T :\, I_t = i_0\big],$$

where i_0 is the zero-mean arm. In particular, no learning algorithm achieves $\widetilde{\mathrm{NR}}_T \to 0$ in T uniformly over instances allowing zero-mean arms.

Remark 3.3 (Assumption 3.1 for High Probability Regret). Without Assumption 3.1, on the event that an algorithm pulls a zero-mean arm at least once with non-zero probability, $\widehat{\mathrm{NR}}_T = \mu^\star$ exactly. Hence, no high-probability bound below $\mu^\star$ is possible without the floor.

Remark 3.4 (Such Assumption is not required for ensemble Nash regret of Barman et al. [2023], Sarkar et al. [2025]). This pathology disappears for the ensemble NR_T, where the expectation is taken *before* the geometric mean. Hence, a single zero-mean round is averaged out into a $\Theta(1/T)$ contribution to $\mathbb{E}[\mu_{I_t}]$.

Algorithm 1 Round Robbin Nash Confidence Bound (`RR-NCB`)

Require: Number of arms k, horizon T, confidence parameter $L > 0$; constants $c = 3$.

Phase 1 (Round-robin exploration).
$t \leftarrow 0$
while $\max_{i\in[k]}\ n_i \cdot \widehat{\mu}_i\ \leq\ 420\, c^2 L$ **do**
  Pull each arm $i \in [k]$ in round robbin fashion.
  Update n_i and $\widehat{\mu}_i$.
  $t \leftarrow t + 1$
end while
Let τ denote the round at which the stopping condition first holds.
Phase 2 (NCB selection).
for $t = \tau + 1, \ldots, T$ **do**
  Set $\mathrm{NCB}_i(t) := \widehat{\mu}_i + 2c\sqrt{2\widehat{\mu}_i L/n_i}$ for each arm i.
  Pull $I_t \in \arg\max_i \mathrm{NCB}_i(t)$
  Observe X_t; update $n_{I_t}, \widehat{\mu}_{I_t}$.
end for

Remark 3.5 (No Impact on Regret (order-wise)). We see (in Theorems 5.1, 5.5 that the floor enters our bound only logarithmically and only in the lower-order $1/T$ term. Hence, such assumption (i.e., a low value of $\mu_{\min}$ does not impact the regret scaling.

Remark 3.6 (Knowledge of $\mu_{\min}$ not needed). We see in Section 4, that the knowledge of $\mu_{\min}$ is not required to run the learning algorithm. This is just for theoretical tractability.

4 Algorithm

We propose the Round Robbin Nash Confidence Bound algorithm (`RR-NCB`), parameterized by a confidence parameter $L > 0$ in place of $\log T$. Set $c = 3$ and $S := c^2 L/\mu^\star$ throughout.

We use the following notation throughout. The algorithm observes $Y_{I_t, n_{I_t}(t)}$ at round t, where $n_i(t) := \sum_{s=1}^{t} \mathbf{1}\{I_s = i\}$. We write $\widehat{\mu}_{i,s} := \frac{1}{s}\sum_{j=1}^{s} Y_{i,j}$ for the empirical mean of arm i from its first s rewards, and abbreviate $n_i = n_i(t)$, $\widehat{\mu}_i = \widehat{\mu}_{i,n_i(t)}$ when t is clear from context.

We will instantiate Algorithm 1 twice: (a) with $L = \log T$, for the in-expectation bound on $\widetilde{\mathrm{NR}}_T$, and (b) with $L = L_\delta := \log(8kT/\delta)$, for the high-probability bound on $\widehat{\mathrm{NR}}_T$.

The algorithm is a two-phase procedure for the k-armed stochastic bandit problem over a known horizon T, parameterized by a confidence parameter $L > 0$ and a constant $c = 3$. We modify the Nash Confidence Bound (NCB) algorithm of Barman et al. [2023], adapted to support both in-expectation and high-probability Nash regret guarantees through a single unified design.

Phase 1: Round-robin exploration. The algorithm begins by exploring all arms uniformly. In each pass, it pulls every arm $i \in \{1, \ldots, k\}$ exactly once in a round robbin fashion, then checks whether the stopping condition

$$\max_i n_i \cdot \widehat{\mu}_i > 420\, c^2 L$$

is satisfied, where n_i is the total number of times arm i has been pulled so far and $\widehat{\mu}_i$ is its empirical mean. If the condition is not met and the horizon has not been reached, another pass begins. This round-robin schedule ensures that all arms have been pulled the same number of times. The stopping time τ is the first round at which the condition is met.

The stopping condition serves a dual purpose. First, it guarantees that Phase 1 does not end too early: the threshold $420\, c^2 L$ is large enough that, on the good event, no arm can cross it before round $192\, kS$ (where $S := c^2 L/\mu^\star$), ensuring every arm accumulates at least $192\, S$ samples. This sample-count floor is what makes Phase 2's concentration-based analysis viable. Second, the condition guarantees that Phase 1 does not run too long: the optimal arm's product $n_{i^\star}\widehat{\mu}_{i^\star}$ must cross the threshold by round $484\, kS$, keeping Phase 1's contribution to the regret bounded by $O(kL\log(1/\mu_{\min})/T)$.

Phase 2: NCB selection. From round $\tau + 1$ to T, the algorithm selects arms greedily by an optimistic index. For each arm i, it computes the Nash Confidence Bound (NCB) given by

$$\mathrm{NCB}_i := \widehat{\mu}_i + 2c\sqrt{2\,\widehat{\mu}_i\, L/n_i}$$

and pulls the arm with the largest NCB_i. The index scales with $\sqrt{\widehat{\mu}_i/n_i}$ rather than $\sqrt{1/n_i}$ similar to Barman et al. [2023]. This ensures that arms whose empirical means are small automatically receive tighter confidence widths, preventing over-exploration. This variance-sensitive design is the key to achieving $O(\sqrt{k/T})$ Nash regret rather than $O(\sqrt{k \log k/T})$. After each pull, the selected arm's count and empirical mean are updated, and the procedure repeats until round T.

5 Main Results

In this section, we obtain the main result of the paper. Specifically, we show that `RR-NCB` yields trajectory-wise as well as high probability Nash regret guarantee. We formally state the results now.

Theorem 5.1 (Trajectory-wise Expected Regret)**.** *Suppose Assumption 3.1 holds and we run Algorithm 1 with $L = \log T$, $c = 3$ and horizon T. The Trajectory-wise expected regret $\widetilde{\mathrm{NR}}_T$ satisfies*

$$\widetilde{\mathrm{NR}}_T \;\le\; C_1\sqrt{\tfrac{k\log T}{T}} + C_2\tfrac{k\log T\log(1/\mu_{\min})}{T} + \tfrac{3k\log(1/\mu_{\min})}{T^2}$$

for absolute constants $C_1, C_2 > 0$.

The proof is deferred to the Appendix. A few remarks are in order.

Remark 5.2. (Matching regret with Barman et al. [2023], Sarkar et al. [2025] The regret $\widetilde{\mathrm{NR}}_T$ consists of three terms, out of which the dominating term is $\tilde{\mathcal{O}}(\sqrt{k/T})$. The regret scaling matches with the (ensemble) Nash regret of Barman et al. [2023], Sarkar et al. [2025]. This shows that although we are working with a stronger notion of regret (as $\widetilde{\mathrm{NR}}_T \ge \mathrm{NR}_T$), we obtain the same result (order-wise). In other words, trajectory-wise or sample path based analysis comes with no additional cost (in terms of regret).

Remark 5.3. (Matching Regret Lower Bound) Since $\widetilde{\mathrm{NR}}_T \ge \mathrm{NR}_T$, a simple application of AM-GM inequality shows that $\widetilde{\mathrm{NR}}_T = \Omega(\sqrt{k/T})$, which implies that `RR-NCB` is order-wise optimal.

Remark 5.4. (Dependence on $\mu_{\min}$) We see that although $\widetilde{\mathrm{NR}}_T$ depends on the positivity floor $\mu_{\min}$, the dependence is present only in minor terms (i.e., the terms scaling with $1/T$ and $1/T^2$). Hence, the influence of $\mu_{\min}$ on regret is minor. In Section 7, we validate this finding through simulations.

We now present the high probability Nash regret. We have the following result.

Theorem 5.5 (High Probability Regret)**.** *Let $\delta \in (0,1)$ and we run Algorithm 1 with $L = \log(8kT/\delta)$. Under Assumption 3.1, we have*

$$\Pr\left[\,\widehat{\mathrm{NR}}_T \;\le\; C'\sqrt{\tfrac{k\log(kT/\delta)}{T}} \;+\; C''\,\tfrac{k\,\log(kT/\delta)\,\log(1/\mu_{\min})}{T}\,\right] \;\ge\; 1-\delta,$$

for absolute constants $C', C'' > 0$.

The formal proof is deferred to the Appendix. We collect a few remarks here.

Remark 5.6. (Matching regret upper and lower bounds) Similar to above, the high probability regret is $\tilde{\mathcal{O}}(\sqrt{k/T})$ which matches Barman et al. [2023], Sarkar et al. [2025] and is optimal (via same AM-GM inequality).

Remark 5.7. (Novel high probability Nash regret) We point out that this is the first result obtaining high probability Nash regret, which is of importance in safety critical applications.

6 Proof Sketch

Recall that we have k-armed stochastic bandit with means $\mu_i \in [\mu_{\min}, 1]$, $\mu^\star = \max_i \mu_i$, horizon T. We bound two variants of Nash regret:

$$\widetilde{\mathrm{NR}}_T := \mu^\star - \mathbb{E}\big[(\textstyle\prod_t \mu_{I_t})^{1/T}\big], \qquad \widehat{\mathrm{NR}}_T := \mu^\star - (\textstyle\prod_t \mu_{I_t})^{1/T}.$$

Our proposed algorithm, RR-NCB (given in 1) has 2 phases. In Phase I agent pull arms in a round-robin fashion until $\max_i n_i \widehat{\mu}_i > 420\, c^2 L$ ($c = 3$). In Phase II, we play the arm with maximum index given by $\mathrm{NCB}_i = \widehat{\mu}_i + 2c\sqrt{2\widehat{\mu}_i L/n_i}$ in Phase 2. We set $L = \log T$ for $\widetilde{\mathrm{NR}}_T$, $L = L_\delta = \log(8kT/\delta)$ for $\widehat{\mathrm{NR}}_T$, and $S = c^2 L/\mu^\star$. We now provide step by step reduction.

Step 1: Reduction to Trajectory-wise log-regret: The first step in the proof is to reduce the variants of Nash regret to log-regret trajectory-wise. Since $(\prod_t \mu_{I_t})^{1/T} = \exp\big(\log \mu^\star - \widehat{R}_T\big)$ trajectory-wise, where $\widehat{R}_T := \frac{1}{T}\sum_t \log(\mu^\star/\mu_{I_t}) \geq 0$, the inequality $1 - e^{-x} \leq x$ gives *trajectory-wise* $\widehat{\mathrm{NR}}_T \leq \mu^\star \widehat{R}_T$. For the in-expectation case, we apply Jensen inequality on the exponential convex function, which gives $\widetilde{\mathrm{NR}}_T \leq \mu^\star \bar{R}_T$, where $\bar{R}_T = \mathbb{E}[\widehat{R}_T]$.

Step 2: Define a Good Event E: We define $E := E_2 \cap E_3$ that supports the entire analysis. The definitions are parameterized by $L > 0$, with $S := c^2 L/\mu^\star$ with $c = 3$.

E_2 **(concentration of high-mean arms):** For all arms i with $\mu_i > \mu^\star/64$ and all sample counts $s \in [64S, T]$, $\quad |\widehat{\mu}_{i,s} - \mu_i| \;\leq\; c\sqrt{\frac{\mu_i L}{s}}$.

E_3 **(low-mean arms stay low):** For all arms j with $\mu_j \leq \mu^\star/64$ and all sample counts $s \in [64S, T]$, $\quad \widehat{\mu}_{j,s} \;<\; \frac{\mu^\star}{32}$.

We show that on good event E, we have

$$\Pr[E] \geq 1 - 3k/T^2 \quad \text{for} \quad L = \log T, \quad \Pr[E] \geq 1 - \delta \quad \text{for} \quad L = L_\delta$$

Step 3: Regret split into Phase 1 and Phase 2:

- *Phase 1* ($t \leq \tau$): each $\log(\mu^\star/\mu_{I_t}) \leq \log(1/\mu_{\min})$, so the regret contribution is at most $\mu^\star \cdot \tau \log(1/\mu_{\min})/T$.
- *Phase 2* ($t > \tau$): We show (in Lemma A.7 that on event E, bad arms are never played in Phase 2. In particular every arm j with $\mu_j \leq \mu^\star/64$ is never pulled in Phase 2. The condition on pulled arm $\mu_{I_t} \geq \mu^\star/64$ lets us linearize, $\log(\mu^\star/\mu_i) \leq 64(\mu^\star - \mu_i)/\mu^\star$, reducing Phase 2's log-regret to $(64/T)\sum_{t>\tau}(\mu^\star - \mu_{I_t})$. This is the standard cumulative gap and is handled via concentration arguments.

Step 4: Bound on Phase 1 Termination: Two auxiliary lemmas pin down the duration of Phase I:

- Lemma A.3: At $\tau \leq 192kS$, $n\widehat{\mu}_{i,n} < 420\, c^2 L$ (proved by splitting on $\mu_i \lesseqgtr \mu^\star/64$ and using E_2, E_3). Hence, Phase-1 hasn't stopped yet.
- Lemma A.4: At $\tau \geq 484kS$, $n\widehat{\mu}_{i^\star,n} > 420\, c^2 L$ (using E_2 at the optimal arm). So, Phase 1 has definitely stopped.

Step 5: $\sqrt{kT}$ NCB analysis. We characterize the per-pull gap in Lemma A.8. In particular,

$$\mu^\star - \mu_i \leq 4c\sqrt{\mu^\star L/(T_i - 1)}$$

for any arm pulled in Phase 2. This is obtained by exploiting Lemma A.6 ($\mathrm{NCB}_{i^\star} \geq \mu^\star$) along with the event E_2. Using this, we show that the j-th pull of arm i in Phase 2 yields a maximum regret of $4c\sqrt{\mu^\star L/j}$. Summing over all the pulls and applying the Cauchy-Schwartz inequality, we obtain

$$\sum_{t>\tau}(\mu^\star - \mu_{I_t}) \leq 8c\sqrt{kT\mu^\star L}.$$

Step 5: Combine. Phase 1 contributes $\leq C_1 kL\log(1/\mu_{\min})/T$, Phase 2 contributes $\leq C_2\sqrt{kL/T}$. Multiplying by $\mu^\star \leq 1$ and adding:

$$\widetilde{\mathrm{NR}}_T,\ \widehat{\mathrm{NR}}_T \;=\; O\Big(\sqrt{\tfrac{kL}{T}} \;+\; \tfrac{kL\log(1/\mu_{\min})}{T}\Big).$$

For the Trajectory-wise expected regret, we also calculate the regret off event E. By using the high probability of E, this term yields a minor $\mathcal{O}(1/T^2)$ term.

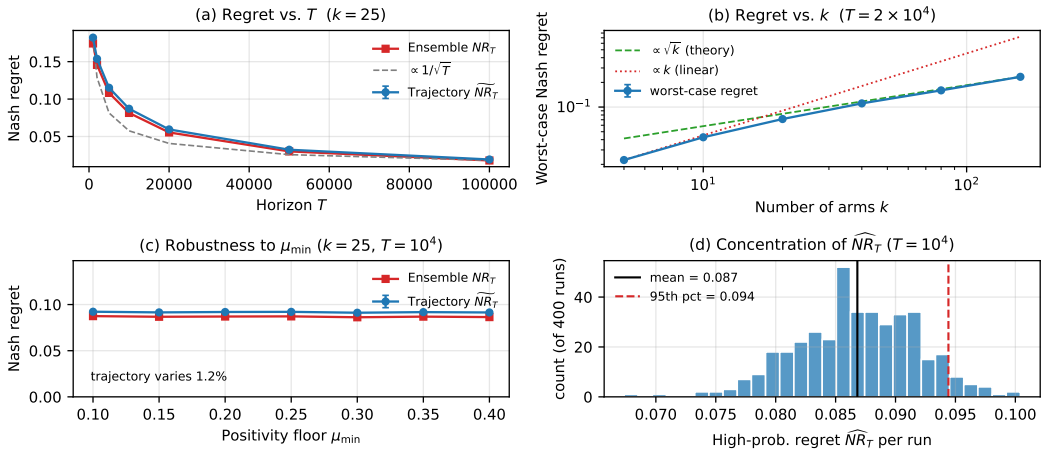


Figure 1: Nash-regret simulations for RR-NCB (200 seeds, 95% CIs). **(a)** Regret vs. horizon T. **(b)** Worst-case trajectory-wise regret vs. number of arms k (log–log). **(c)** Trajectory and ensemble regret vs. the floor $\mu_{\min}$. **(d)** Histogram of the single-run regret $\widehat{NR}_T$ over 400 runs.

7 Simulations

Our theory makes three concrete promises. First, the trajectory-wise regret should match the ensemble regret in order, even though it is a strictly harder target. Second, it should grow like $\sqrt{k}$ in the number of arms, not linearly. Third, it should barely notice the positivity floor $\mu_{\min}$. We test each promise in turn. We also check a fourth point: that the high-probability regret really concentrates, since this is the first such guarantee in the fair bandits setting. All four experiments appear in Figure 1.

We use stochastic bandits with Bernoulli rewards and a best arm of mean $\mu^\star = 0.9$. Every configuration is averaged over 200 independent seeds, and we report 95% confidence intervals. We track three notions of Nash regret, and it helps to say them in plain terms. The *ensemble* regret $NR_T = \mu^\star - \left(\prod_t \mathbb{E}[\mu_{I_t}]\right)^{1/T}$ of Barman et al. [2023] averages the pulled-arm means across runs first, and then takes the geometric mean; it describes the *average* run. The *trajectory-wise* regret $\widetilde{NR}_T = \mu^\star - \mathbb{E}\left[(\prod_t \mu_{I_t})^{1/T}\right]$ takes the geometric mean along each run first, and then averages; it describes a *typical* run. The *high-probability* regret $\widehat{NR}_T = \mu^\star - (\prod_t \mu_{I_t})^{1/T}$ uses a single run with no averaging at all; it is what one real deployment actually experiences. Throughout, the geometric means are taken over the true pulled-arm means μ_{I_t}, exactly as in the theory, and not over the raw realized rewards.

7.1 Results and insights

Regret vs. the horizon (Fig. 1a). Both regrets fall toward zero as T grows. So the algorithm learns, and its regret is sub-linear. The trajectory-wise curve sits above the ensemble curve at every horizon. This ordering is not an accident. Moving the expectation outside the geometric mean can only lower the welfare, by Jensen's inequality, so $\widetilde{NR}_T \geq NR_T$ must hold on every instance. What matters is how far apart they are. The gap is small, and it shrinks as T grows, from about 0.008 at $T = 10^3$ to about 0.001 at $T = 10^5$. Both curves also hug the dashed $1/\sqrt{T}$ reference, and a log–log fit gives a slope of -0.49. This is the $\widetilde{O}(\sqrt{k \log T/T})$ rate of Theorem 5.1. The takeaway is the one we hoped for: the stronger per-trajectory guarantee comes at no extra cost in the rate.

Regret vs. the number of arms (Fig. 1b). This is the plot that is easy to misread. If one fixes a single instance and a modest horizon, the regret can look like it grows linearly in k. The reason is the warm-up. Phase 1 explores every arm, so it contributes an $O(k \log T/T)$ term that really is linear in k. At a small horizon this term is large and hides the leading $\sqrt{k}$ behaviour. Theorem 5.1 is a worst-case statement, so we test it the right way. For each k, we search a grid of gaps and report

the hardest instance we find. Plotted on a log–log scale, the resulting curve runs parallel to the $\sqrt{k}$ reference line, and pulls away from the linear one as k grows. The fitted exponent is 0.57, close to the predicted 0.5. So the growth really is $\sqrt{k}$. The earlier linear look was a finite-horizon artifact, not a true rate.

Robustness to the floor (Fig. 1c). The bound carries a term that depends on $\mu_{\min}$, and a natural worry is that a tiny floor could ruin performance. It does not. We hold every arm fixed and slide the single worst arm from 0.1 up to 0.4. The trajectory-wise regret hardly moves. Its relative change over the whole range is only 1.2%, which is inside the error bars. This matches the theory cleanly. The floor enters only through the lower-order term $O\big(k \log T \, \log(1/\mu_{\min})/T\big)$, and it appears there only as $\log(1/\mu_{\min})$. So even a very small floor costs almost nothing. In short, the positivity assumption is a technical device, not a practical limitation.

Concentration of the single-run regret (Fig. 1d). Our high-probability bound promises more than a good average. It promises that one run, on its own, stays close to optimal. To check this, we draw 400 runs and plot the histogram of their single-run regret $\widehat{NR}_T$. The spread is narrow and there is no heavy tail. The 95th percentile sits at 0.094, within 9% of the mean at 0.087. So an unlucky run is barely worse than a typical one. This is exactly what the bound guarantees, and to our knowledge it is the first time a high-probability Nash-regret guarantee has been verified in simulation.

Reproducibility. RR-NCB uses the variance-aware index $\widehat{\mu}_i + 1.5\,\sqrt{\widehat{\mu}_i \log T / n_i}$, run after a round-robin warm-up of $\lceil \log T \rceil$ pulls per arm. We use the small constant 1.5 rather than the conservative theoretical value, which only inflates the lower-order term. The reward distributions and arm means are stated with each panel above.

8 Conclusion and Open Problems

In this paper we study fair multi-armed bandits from a trajectory (sample path) point of view (both in expectation and high probability). We propose and analyze a learning algorithm which yields optimal regret guarantees which we validate through simulations.

We end this paper with a few open problems. Until now, all the guarantees on Nash regret are instance independent. Can we obtain instance dependent optimal logarithmic Nash regret? Moreover, how does the trajectory-wise Nash regret scale in the presence of multiple agents (in a cooperative or competitive setup)? Furthermore, for structured bandits (like linear, kernelized), can we obtain regret guarantees for trajectory-wise Nash regret? We keep these questions for our future endeavors.

References

Siddharth Barman, Arindam Khan, Arnab Maiti, and Ayush Sawarni. Fairness and welfare quantification for regret in multi-armed bandits. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i6.25829. URL `https://doi.org/10.1609/aaai.v37i6.25829`.

Anup Biswas and Vivek S Borkar. Ergodic risk-sensitive control—a survey. *Annual Reviews in Control*, 55:118–141, 2023.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

Ioannis Caragiannis, David Kurokawa, Hervé Moulin, Ariel D. Procaccia, Nisarg Shah, and Junxing Wang. The unreasonable fairness of maximum nash welfare. *ACM Trans. Econ. Comput.*, 7(3), September 2019. ISSN 2167-8375. doi: 10.1145/3355902. URL `https://doi.org/10.1145/3355902`.

Cyrus Cousins. Revisiting fair-PAC learning and the axioms of cardinal welfare. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 7422–7442. PMLR, 2023.

Everett R. Dempster. Maintenance of genetic heterogeneity. *Cold Spring Harbor Symposia on Quantitative Biology*, 20:25–32, 1955. doi: 10.1101/SQB.1955.020.01.005.

Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Ronald A. Howard and James E. Matheson. Risk-sensitive markov decision processes. *Management Science*, 18(7):356–369, 1972. ISSN 00251909, 15265501. URL http://www.jstor.org/stable/2629352.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 325–333, 2016.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, Seth Neel, and Aaron Roth. Fair algorithms for infinite and contextual bandits. *arXiv preprint arXiv:1610.09559*, 2018.

Mamoru Kaneko. The nash social welfare function for a measure space of individuals. *Journal of Mathematical Economics*, 8(2):173–200, 1981. ISSN 0304-4068. doi: https://doi.org/10.1016/0304-4068(81)90021-5. URL https://www.sciencedirect.com/science/article/pii/0304406881900215.

Mamoru Kaneko and Kenjiro Nakamura. The nash social welfare function. *Econometrica*, 47(2): 423–435, 1979. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1914191.

J. L. Kelly. A new interpretation of information rate. *The Bell System Technical Journal*, 35(4): 917–926, 1956. doi: 10.1002/j.1538-7305.1956.tb03809.x.

Anand Krishna, Philips George John, Adarsh Barik, and Vincent YF Tan. p-mean regret for stochastic bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17966–17973, 2025.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

R. C. Lewontin and D. Cohen. On population growth in a randomly varying environment. *Proceedings of the National Academy of Sciences*, 62(4):1056–1060, 1969. doi: 10.1073/pnas.62.4.1056.

John F. Nash. The bargaining problem. *Econometrica*, 18(2):155–162, 1950.

Ole Peters. The ergodicity problem in economics. *Nature Physics*, 15:1216–1221, 2019. doi: 10.1038/s41567-019-0732-0.

Dhruv Sarkar, Nishant Pandey, and Sayak Ray Chowdhury. Revisiting social welfare in bandits: Ucb is (nearly) all you need. *arXiv preprint arXiv:2510.21312*, 2025.

Dhruv Sarkar, Nishant Pandey, and Sayak Ray Chowdhury. Improved algorithms for nash welfare in linear bandits. *arXiv preprint arXiv:2601.22969*, 2026.

Ayush Sawarni, Soumyabrata Pal, and Siddharth Barman. Nash regret guarantees for linear bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 33288–33318, 2023.

Alan D. Taylor. *Public Choice*, 119(3/4):468–470, 2004. ISSN 00485829, 15737101. URL http://www.jstor.org/stable/30026057.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018.

Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019.

Lequn Wang, Wen Sun, and Thorsten Joachims. Fairness of exposure in stochastic bandits. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 10686–10696. PMLR, 2021.

Peter Whittle. A risk-sensitive maximum principle. *Systems & Control Letters*, 15(3):183–192, 1990.

Mengxiao Zhang, Ramiro Deo-Campo Vuong, and Haipeng Luo. No-regret learning for fair multi-agent social welfare optimization. *arXiv preprint arXiv:2405.20678*, 2024.

A Proofs

A.1 Setup and Useful Lemmas

Without loss of generality, throughout the proof we assume

$$\mu^\star \geq 512\sqrt{kL/T},$$

which we call the warm-up condition. When this bound fails, our results hold deterministically with

$$\widetilde{\mathrm{NR}}_T \leq \mu^\star \leq 512\sqrt{kL/T}.$$

Similar bounds holds for $\widehat{\mathrm{NR}}_T$. This crucially leverages the non-negativity of the rewards.

A.2 The Good Event

We define a good event $E := E_2 \cap E_3$ that supports the entire analysis. The definitions are parameterized by $L > 0$, with $S := c^2 L/\mu^\star$ as before. Recall $c = 3$.

E_2 **(concentration of high-mean arms):** For all arms i with $\mu_i > \mu^\star/64$ and all sample counts $s \in [64S, T]$,

$$|\widehat{\mu}_{i,s} - \mu_i| \;\leq\; c\sqrt{\tfrac{\mu_i L}{s}}.$$

E_3 **(low-mean arms stay low):** For all arms j with $\mu_j \leq \mu^\star/64$ and all sample counts $s \in [64S, T]$,

$$\widehat{\mu}_{j,s} \;<\; \tfrac{\mu^\star}{32}.$$

A.3 Round-robin sampling

Under the round-robin schedule of Algorithm 1, after any number of rounds $r \geq k$ during Phase 1, each arm i has been pulled either $\lfloor r/k \rfloor$ or $\lceil r/k \rceil$ times. For the sake of simplicity and clarity of exposition we assume that τ is a multiple of k (otherwise we just use ceil and floor which will only impact the final regret in terms of universal constants). Hence, at round τ every arm has count τ/k exactly. We use this throughout below.

A.4 Probability of the good event

Lemma A.1 (Good event probability)**.** $\Pr[E^c] \leq 3kT\exp(-c^2 L/3)$.

Proof. We bound $\Pr[E_2^c]$ and $\Pr[E_3^c]$ separately and union-bound.

E_2*:* For an arm i with $\mu_i > \mu^\star/64$ and $s \in [64S, T]$, by Lemma E.1 with $\nu = \mu_i$ and $\delta := c\sqrt{L/(\mu_i s)}$, we have

$$\Pr\left[|\widehat{\mu}_{i,s} - \mu_i| \geq c\sqrt{\tfrac{\mu_i L}{s}}\right] \leq 2\exp\Big(-\tfrac{\delta^2 s\mu_i}{3}\Big) = 2\exp\Big(-\tfrac{c^2 L}{3}\Big).$$

We require $\delta \leq 1$ for Lemma E.1 to apply, i.e., $s \geq c^2 L/\mu_i$. Since $\mu_i > \mu^\star/64$, $c^2 L/\mu_i < 64c^2 L/\mu^\star = 64S$, and $s \geq 64S$ suffices.

Union-bounding over k arms and T sample counts: $\Pr[E_2^c] \leq 2kT\exp(-c^2 L/3)$.

E_3*:* For an arm j with $\mu_j \leq \mu^\star/64$ and $s \in [64S, T]$, by Lemma E.2 with $\nu_H = \mu^\star/64$ (note $(1+1)\nu_H = \mu^\star/32$), we have

$$\Pr\Big[\widehat{\mu}_{j,s} \geq \tfrac{\mu^\star}{32}\Big] \leq \exp\Big(-\tfrac{1\cdot s\cdot \mu^\star/64}{3}\Big) = \exp\Big(-\tfrac{s\mu^\star}{192}\Big).$$

For $s \geq 64S$, we have $s\mu^\star/192 \geq 64c^2 L/192 = c^2 L/3$. So the bound becomes $\exp(-c^2 L/3)$.

Union over k arms and T sample counts: $\Pr[E_3^c] \leq kT\exp(-c^2 L/3)$.

Adding the two bounds: $\Pr[E^c] \leq 3kT\exp(-c^2 L/3)$. □

Corollary A.2. *For $c = 3$:*

1. *With $L = \log T$ and $T \geq 3k$, $\Pr[E] \geq 1 - 3k/T^2$.*
2. *With $L = L_\delta := \log(8kT/\delta)$ and $\delta \in (0, 1)$, $\Pr[E] \geq 1 - \delta$.*

Proof. (1) With $L = \log T$ and $c = 3$: $\exp(-c^2 L/3) = \exp(-3 \log T) = T^{-3}$. So $\Pr[E^c] \leq 3k/T^2$.

(2) With $L = \log(8kT/\delta)$: $\exp(-c^2 L/3) = (8kT/\delta)^{-3}$. So

$$\Pr[E^c] \leq \tfrac{3kT\delta^3}{(8kT)^3} = \tfrac{3\delta^3}{512(kT)^2} \leq \delta,$$

for any $\delta \in (0, 1)$ and $kT \geq 1$. □

A.5 Analysis of Phase 1

We now characterize the number of pulls made in Phase-I. The Phase 1 stopping rule activates when $\max_i n_i \widehat{\mu}_i > 420c^2 L$. To pin down τ, we need to know when $n_i \widehat{\mu}_i$ has not yet crossed the threshold and when it must have. The next two lemmas provide both bounds.

Lemma A.3. *On the event E, for any arm $i \in [k]$ and any sample count $n \leq 192S$,*

$$n\,\widehat{\mu}_{i,n} \;<\; 210\,c^2 L < 420c^2 L.$$

Proof. Since $Y_{i,j} \in [0, 1]$ (non-negative), the sum $n\widehat{\mu}_{i,n} = \sum_{j=1}^{n} Y_{i,j}$ is non-decreasing in n. So it suffices to bound the case $n = N$ where $N = 192S$. We have

$$n\widehat{\mu}_{i,n} \;\leq\; N\widehat{\mu}_{i,N}. \tag{1}$$

Two cases on μ_i.

Case 1: $\mu_i \leq \mu^\star/64$. Since $N = 192S \geq 64S$, E_3 applies and gives $\widehat{\mu}_{i,N} < \mu^\star/32$. Hence

$$N\widehat{\mu}_{i,N} \;<\; 192S \cdot \tfrac{\mu^\star}{32} \;=\; 6S\mu^\star \;=\; 6c^2 L.$$

Case 2: $\mu_i > \mu^\star/64$. Since $N \geq 64S$, E_2 applies:

$$\widehat{\mu}_{i,N} \;\leq\; \mu_i + c\sqrt{\tfrac{\mu_i L}{N}} \;\leq\; \mu^\star + c\sqrt{\tfrac{\mu^\star L}{N}} \;=\; \mu^\star + c\sqrt{\tfrac{\mu^\star L \cdot \mu^\star}{192c^2 L}} \;=\; \mu^\star\Big(1 + \tfrac{1}{\sqrt{192}}\Big).$$

Hence, invoking the definition of N, we have $N\widehat{\mu}_{i,N} < 210c^2 L$.

In both cases $N\widehat{\mu}_{i,N} < 210c^2 L$, so (1) gives $n\widehat{\mu}_{i,n} < 210c^2 L$. □

Lemma A.4. *On the event E, for any sample count $n \geq 484S$,*

$$n\,\widehat{\mu}_{i^\star,n} \;\geq\; 462\,c^2 L > 420c^2 L.$$

Proof. By monotonicity, $n\widehat{\mu}_{i^\star,n} \geq M\widehat{\mu}_{i^\star,M}$ where $M = 484S$. Since $\mu_{i^\star} = \mu^\star > \mu^\star/64$ and $M \geq 64S$, E_2 applies and we have

$$\widehat{\mu}_{i^\star,M} \;\geq\; \mu^\star - c\sqrt{\tfrac{\mu^\star L}{M}} \;=\; \mu^\star - c\sqrt{\tfrac{\mu^\star L \cdot \mu^\star}{484c^2 L}} \;=\; \mu^\star\Big(1 - \tfrac{1}{\sqrt{484}}\Big) \;=\; \tfrac{21}{22}\mu^\star.$$

Hence $M\widehat{\mu}_{i^\star,M} \geq 462c^2 L$. □

A.5.1 Bounding the Phase 1 stopping time

We use the above 2 Lemmas to bound τ under event E. We have

$$192kS \leq \tau \leq 484kS (= T_e).$$

Corollary A.5 (Phase 1 sample counts)**.** *On E, at the end of Phase 1, every arm i has been pulled exactly τ/k times, with $\tau/k > 192S > 64S$. In particular, every arm satisfies $n_i(\tau) > 192c^2 L/\mu^\star$.*

A.6 Analysis of Phase 2: NCB-Based Selection

Throughout this section we work on the good event E, and we denote $S = c^2 L/\mu^\star$ as before. By Corollary A.5, on E every arm i has $n_i > 192S > 64S$ before the start of Phase 2 (the $\geq 64S$ form is what we use most often as it is defined in the event E).

Lemma A.6 (Optimal arm's NCB stays above $\mu^\star$)**.** *On the event E and assuming $\mu^\star \geq 512\sqrt{kL/T}$, for every Phase 2 round $t > \tau$,*

$$\mathrm{NCB}_{i^\star}(t) \;\geq\; \mu^\star.$$

Proof. Fix $t > \tau$ and let $n^\star := n_{i^\star}(t-1) \geq 64S$ (Corollary A.5).

Since $\mu^\star > \mu^\star/64$ and $n^\star \geq 64S$, E_2 gives $|\widehat{\mu}_{i^\star} - \mu^\star| \leq c\sqrt{\mu^\star L/n^\star}$. In particular,

$$\widehat{\mu}_{i^\star} \;\geq\; \mu^\star - c\sqrt{\tfrac{\mu^\star L}{n^\star}} \;\Longrightarrow\; c\sqrt{\tfrac{\mu^\star L}{n^\star}} \geq \mu^\star - \widehat{\mu}_{i^\star}. \tag{2}$$

We want to show $\widehat{\mu}_{i^\star} + 2c\sqrt{2\widehat{\mu}_{i^\star} L/n^\star} \geq \mu^\star$, equivalently

$$2c\sqrt{\tfrac{2\widehat{\mu}_{i^\star} L}{n^\star}} \;\geq\; \mu^\star - \widehat{\mu}_{i^\star}. \tag{3}$$

So (3) is implied by

$$2c\sqrt{\tfrac{2\widehat{\mu}_{i^\star} L}{n^\star}} \;\geq\; c\sqrt{\tfrac{\mu^\star L}{n^\star}} \quad\Longleftrightarrow\quad 8\widehat{\mu}_{i^\star} \geq \mu^\star. \tag{4}$$

We verify $8\widehat{\mu}_{i^\star} \geq \mu^\star$. Using (2) and $n^\star \geq 64S = 64c^2 L/\mu^\star$:

$$c\sqrt{\tfrac{\mu^\star L}{n^\star}} \;\leq\; c\sqrt{\tfrac{\mu^\star L \cdot \mu^\star}{64c^2 L}} \;=\; \tfrac{\mu^\star}{8}.$$

Hence $\widehat{\mu}_{i^\star} \geq \mu^\star - \mu^\star/8 = 7\mu^\star/8$, so $8\widehat{\mu}_{i^\star} \geq 7\mu^\star \geq \mu^\star$. This establishes (4) and hence $\mathrm{NCB}_{i^\star}(t) \geq \mu^\star$. □

Lemma A.7 (Bad arms are never pulled in Phase 2)**.** *On the event E and assuming $\mu^\star \geq 512\sqrt{kL/T}$, every arm j with $\mu_j \leq \mu^\star/64$ is never pulled in Phase 2.*

Proof. Fix such an arm j. By Corollary A.5, $n_j \geq 64S$ at the start of Phase 2. Since $\mu_j \leq \mu^\star/64$, E_3 gives $\widehat{\mu}_j < \mu^\star/32$ throughout Phase 2 (note that n_j only increases, and E_3 is a uniform statement over $s \in [64S, T]$).

For any Phase 2 round t, with $n_j \geq 64S$:

$$\mathrm{NCB}_j(t) \;=\; \widehat{\mu}_j + 2c\sqrt{\tfrac{2\widehat{\mu}_j L}{n_j}} \;<\; \tfrac{\mu^\star}{32} + 2c\sqrt{\tfrac{2(\mu^\star/32)L}{64S}} \;=\; \tfrac{\mu^\star}{32} + 2c\sqrt{\tfrac{\mu^\star L \cdot \mu^\star}{1024\, c^2 L}} \;=\; \tfrac{\mu^\star}{32} + \tfrac{\mu^\star}{16} \;=\; \tfrac{3\mu^\star}{32}.$$

By Lemma A.6, $\mathrm{NCB}_{i^\star}(t) \geq \mu^\star$. Since $3\mu^\star/32 < \mu^\star \leq \mathrm{NCB}_{i^\star}(t)$, we have $\mathrm{NCB}_j(t) < \mathrm{NCB}_{i^\star}(t)$, so arm j is not selected at round t. This holds for every $t > \tau$, so j is never pulled in Phase 2. □

A.6.1 Per-pull gap for arms pulled in Phase 2

Lemma A.8 (Per-pull gap)**.** *On the event E and assuming $\mu^\star \geq 512\sqrt{kL/T}$, suppose arm i is pulled in Phase 2 at some round, and let T_i denote its total pull count at that moment (including the current pull). Then*

$$\mu^\star - \mu_i \;\leq\; 4c\sqrt{\tfrac{\mu^\star L}{T_i - 1}}.$$

Proof. By Lemma A.7, $\mu_i > \mu^\star/64$. By Corollary A.5, $T_i - 1 \geq 64S$ (since the prior pull count is at least the Phase 1 count, which is $\geq 64S$). Hence E_2 applies at $s = T_i - 1$.

At the moment of this pull, since arm i was selected by NCB,

$$\widehat{\mu}_i + 2c\sqrt{\tfrac{2\widehat{\mu}_i L}{T_i - 1}} \;=\; \mathrm{NCB}_i \;\geq\; \mathrm{NCB}_{i^\star} \;\geq\; \mu^\star, \tag{5}$$

the last inequality by Lemma A.6.

We upper-bound $\widehat{\mu}_i$ using E_2 (since $\mu_i > \mu^\star/64$) and $T_i - 1 \geq 64S$:

$$\widehat{\mu}_i \;\leq\; \mu_i + c\sqrt{\tfrac{\mu_i L}{T_i - 1}} \;\leq\; \mu^\star + c\sqrt{\tfrac{\mu^\star L}{T_i - 1}} \;\leq\; \mu^\star + c\sqrt{\tfrac{\mu^\star L \cdot \mu^\star}{64c^2 L}} \;=\; \mu^\star + \tfrac{\mu^\star}{8} \;=\; \tfrac{9}{8}\mu^\star.$$

Substituting this bound into (5):

$$\mu^\star \;\leq\; \widehat{\mu}_i + 2c\sqrt{\tfrac{2\cdot(9/8)\mu^\star L}{T_i - 1}} \;=\; \widehat{\mu}_i + 3c\sqrt{\tfrac{\mu^\star L}{T_i - 1}}.$$

Combining with the upper bound on $\widehat{\mu}_i$ in terms of μ_i from E_2:

$$\widehat{\mu}_i \;\leq\; \mu_i + c\sqrt{\tfrac{\mu_i L}{T_i - 1}} \;\leq\; \mu_i + c\sqrt{\tfrac{\mu^\star L}{T_i - 1}} \quad\Longrightarrow\quad \mu^\star \;\leq\; \mu_i + 4c\sqrt{\tfrac{\mu^\star L}{T_i - 1}},$$

which is the claimed bound after rearranging. □

B Proof of Trajectory-wise Expectation Bound: $\widetilde{\mathrm{NR}}_T$

We use the lemmas of Sections A.2–A.6 with $L = \log T$, so that $\Pr[E] \geq 1 - 3k/T^2$ (Corollary A.2).

B.1 Reduction via Jensen on exp

Lemma B.1. *For any $[k]$-valued random sequence $(I_t)_{t=1}^T$ with $\mu_{I_t} > 0$ a.s.,*

$$\widetilde{\mathrm{NR}}_T \;\leq\; \mu^\star \cdot \bar{R}_T, \qquad \textit{where} \qquad \bar{R}_T \;:=\; \tfrac{1}{T}\sum_{t=1}^{T} \mathbb{E}\Big[\log \tfrac{\mu^\star}{\mu_{I_t}}\Big] \geq 0.$$

Proof. Pathwise, $(\prod_t \mu_{I_t})^{1/T} = \exp\big(\frac{1}{T}\sum_t \log \mu_{I_t}\big)$. Since exp is convex, Jensen's inequality gives

$$\mathbb{E}\left[\exp\left(\tfrac{1}{T}\sum_t \log \mu_{I_t}\right)\right] \;\geq\; \exp\left(\tfrac{1}{T}\sum_t \mathbb{E}[\log \mu_{I_t}]\right) = \exp(\log \mu^\star - \bar{R}_T) = \mu^\star e^{-\bar{R}_T},$$

using

$$\mathbb{E}[\log \mu_{I_t}] = \log \mu^\star - \mathbb{E}[\log(\mu^\star/\mu_{I_t})].$$

Hence,

$$\widetilde{\mathrm{NR}}_T \leq \mu^\star(1 - e^{-\bar{R}_T}) \leq \mu^\star \bar{R}_T,$$

using $1 - e^{-x} \leq x$ for all non-negative x. □

B.2 Concluding the proof of Theorem 5.1

By Lemma B.1 it suffices to bound $\mu^\star \bar{R}_T$. We decompose in the following way:

$$\mu^\star \bar{R}_T \;=\; \underbrace{\tfrac{\mu^\star}{T}\,\mathbb{E}\left[\mathbf{1}_E \sum_{t\leq\tau} \log \tfrac{\mu^\star}{\mu_{I_t}}\right]}_{\text{(I)}} \;+\; \underbrace{\tfrac{\mu^\star}{T}\,\mathbb{E}\left[\mathbf{1}_E \sum_{t=\tau+1}^{T} \log \tfrac{\mu^\star}{\mu_{I_t}}\right]}_{\text{(II)}} \;+\; \underbrace{\tfrac{\mu^\star}{T}\,\mathbb{E}\left[\mathbf{1}_{E^c} \sum_{t=1}^{T} \log \tfrac{\mu^\star}{\mu_{I_t}}\right]}_{\text{(III)}}.$$

(I) Phase 1 on E. On E, we have $\tau \leq 484c^2 k \log T/\mu^\star$. Also, each pulled arm satisfies $\mu_{I_t} \geq \mu_{\min}$. Hence, since $\mu^\star \leq 1$, we have $\log(\mu^\star/\mu_{I_t}) \leq \log(1/\mu_{\min})$. We have

$$\text{(I)} \;\leq\; \tfrac{\mu^\star}{T} \cdot \tfrac{484c^2 k \log T}{\mu^\star} \cdot \log(1/\mu_{\min}) \;=\; \tfrac{484c^2 k \log T \log(1/\mu_{\min})}{T}.$$

(II) Phase 2 on E. On E, we have $\mu_{I_t} \geq \mu^\star/64$; otherwise the arm is never played (see Lemma A.7). The elementary inequality

$$\log \tfrac{\mu^\star}{\mu_i} \;\leq\; \tfrac{64\,(\mu^\star - \mu_i)}{\mu^\star} \quad \text{whenever } \mu_i \geq \mu^\star/64, \tag{6}$$

follows from $\log(1+u) \le u$ with $u = (\mu^\star - \mu_i)/\mu_i$ and $\mu_i \ge \mu^\star/64$. Substituting,

$$\text{(II)} \;\le\; \tfrac{\mu^\star}{T}\,\mathbb{E}\left[\mathbf{1}_E \sum_{t>\tau} \tfrac{64(\mu^\star - \mu_{I_t})}{\mu^\star}\right] \;=\; \tfrac{64}{T}\,\mathbb{E}\left[\mathbf{1}_E \sum_{t>\tau} (\mu^\star - \mu_{I_t})\right].$$

Let m_i be the Phase 2 pull count of arm i, and $T_i^{(j)}$ the prior pull count of arm i before its j-th Phase 2 pull. By Corollary A.5, on E in phase 1, every arm has $> 128S$ pulls. Hence $T_i^{(j)} > 128S + j - 1 \ge j$. Lemma A.8 (applied with total count $T_i = T_i^{(j)} + 1$, so $T_i - 1 = T_i^{(j)}$) gives

$$\mu^\star - \mu_{I_t} \;\le\; 4c\sqrt{\tfrac{\mu^\star \log T}{T_i^{(j)}}} \;\le\; 4c\sqrt{\tfrac{\mu^\star \log T}{j}}.$$

Summing up, we obtain

$$\sum_{t>\tau} (\mu^\star - \mu_{I_t}) \;\le\; 4c\sqrt{\mu^\star \log T} \sum_i \sum_{j=1}^{m_i} \tfrac{1}{\sqrt{j}} \;\le\; 8c\sqrt{\mu^\star \log T} \sum_i \sqrt{m_i} \;\le\; 8c\sqrt{kT\mu^\star \log T},$$

using $\sum_{j=1}^{m} 1/\sqrt{j} \le 2\sqrt{m}$ and Cauchy–Schwarz $\sum_i \sqrt{m_i} \le \sqrt{k \sum_i m_i} \le \sqrt{kT}$. Therefore

$$\text{(II)} \;\le\; \tfrac{64 \cdot 8c\sqrt{kT\mu^\star \log T}}{T} \;=\; 512c\sqrt{\tfrac{k\mu^\star \log T}{T}} \;\le\; 512c\sqrt{\tfrac{k \log T}{T}},$$

the last using $\mu^\star \le 1$.

(III) Off the good event. On E^c, $\log(\mu^\star/\mu_{I_t}) \le \log(1/\mu_{\min})$ for every t, and $\Pr[E^c] \le 3k/T^2$:

$$\text{(III)} \;\le\; \tfrac{\mu^\star}{T} \cdot T \log(1/\mu_{\min}) \cdot \Pr[E^c] \;\le\; \tfrac{3k \log(1/\mu_{\min})}{T^2}.$$

Combining. With all 3 bounds above, we obtain

$$\widetilde{\mathrm{NR}}_T \le C_1 \tfrac{k \log T \log(1/\mu_{\min})}{T} + C_2 \sqrt{\tfrac{k \log T}{T}} + \tfrac{3k \log(1/\mu_{\min})}{T^2}.$$

C High-Probability Bound on $\widehat{\mathrm{NR}}_T$

We use the lemmas of Sections A.2–A.6 with $L = L_\delta := \log(8kT/\delta)$, so that $\Pr[E] \ge 1 - \delta$ (Corollary A.2).

C.1 Trajectory-wise reduction

Lemma C.1 (Trajectory-wise Jensen reduction)**.** *On every sample path with $\mu_{I_t} > 0$ for all t,*

$$\widehat{\mathrm{NR}}_T \;\le\; \mu^\star \cdot \widehat{R}_T, \qquad \textit{where} \qquad \widehat{R}_T \;:=\; \tfrac{1}{T} \sum_{t=1}^{T} \log \tfrac{\mu^\star}{\mu_{I_t}}.$$

Proof. Pathwise (no expectation), $(\prod_t \mu_{I_t})^{1/T} = \exp(-\widehat{R}_T)\mu^\star$. Hence $\widehat{\mathrm{NR}}_T = \mu^\star(1 - e^{-\widehat{R}_T}) \le \mu^\star \widehat{R}_T$, using $1 - e^{-x} \le x$. □

C.2 Concluding the proof of Theorem 5.5

By Lemma C.1, on every sample path $\widehat{\mathrm{NR}}_T \le \mu^\star \widehat{R}_T$. We bound $\mu^\star \widehat{R}_T$ pathwise on E:

$$\mu^\star \widehat{R}_T \;=\; \underbrace{\tfrac{\mu^\star}{T} \sum_{t \le \tau} \log \tfrac{\mu^\star}{\mu_{I_t}}}_{\text{(I)}} \;+\; \underbrace{\tfrac{\mu^\star}{T} \sum_{t=\tau+1}^{T} \log \tfrac{\mu^\star}{\mu_{I_t}}}_{\text{(II)}}.$$

(I) Phase 1 on E. On E, $\tau \le 484c^2 k L_\delta/\mu^\star$. Each pulled arm has $\mu_{I_t} \ge \mu_{\min}$, so $\log(\mu^\star/\mu_{I_t}) \le \log(1/\mu_{\min})$:

$$\text{(I)} \;\le\; \tfrac{484c^2 k L_\delta \log(1/\mu_{\min})}{T}.$$

(II) Phase 2 on E. The argument is identical to that in Theorem 5.1, but pathwise (trajectory-wise) rather than in expectation. Using (6) and the per-pull gap (Lemma A.8, applied pathwise), then Cauchy–Schwarz $\sum_i \sqrt{m_i} \le \sqrt{kT}$ pathwise:

$$\text{(II)} \;\le\; 512c\sqrt{\tfrac{kL_\delta}{T}}.$$

Combining. On E (which has probability $\ge 1-\delta$):

$$\widehat{\mathrm{NR}}_T \le \mu^\star \widehat{R}_T \le C'\sqrt{\tfrac{kL_\delta}{T}} + C''\,\tfrac{kL_\delta \log(1/\mu_{\min})}{T}.$$

D Positivity floor

The following statement makes precise the claim, made informally above (after Assumption 3.1), that some such floor is unavoidable.

Proposition D.1. *Fix $k \ge 2$. For any algorithm $\mathcal{A}$ that pulls every arm with positive probability at least once over T rounds, there exists a bandit instance ν with $\mu^\star = 1$ and $\min_i \mu_i = 0$ on which*

$$\widetilde{\mathrm{NR}}_T(\mathcal{A}, \nu) \;\ge\; \mu^\star \cdot \Pr\big[\exists\, t \le T :\, I_t = i_0\big],$$

where i_0 is the zero-mean arm. In particular, no learning algorithm achieves $\widetilde{\mathrm{NR}}_T \to 0$ in T uniformly over instances allowing zero-mean arms.

Proof. On the event $\{\exists t : I_t = i_0\}$, $\prod_{t=1}^T \mu_{I_t} = 0$, so the geometric mean $(\prod_t \mu_{I_t})^{1/T} = 0$. Hence

$$\mathbb{E}\Big[\big(\textstyle\prod_t \mu_{I_t}\big)^{1/T}\Big] \;\le\; \mu^\star \cdot \Pr\big[\nexists\, t : I_t = i_0\big] \;=\; \mu^\star\,\big(1 - \Pr[\exists t : I_t = i_0]\big),$$

and the claimed lower bound follows. The second statement follows because any algorithm that fails to pull some arm i_0 with probability bounded below cannot be uniformly consistent (a no-pull on a possibly-optimal arm). □

Why a positivity floor is necessary in High Probability Regret? Without Assumption 3.1, on the (positive-probability) event that an algorithm pulls a zero-mean arm at least once, $\widehat{\mathrm{NR}}_T = \mu^\star$ exactly. Since any consistent learning algorithm must pull every arm with positive probability, no high-probability bound below $\mu^\star$ is possible without the floor.

E Auxiliary Lemmas

We collect the concentration tools used throughout. These are standard (see Vershynin [2018], Wainwright [2019]). We mention here for completeness.

E.1 Multiplicative Chernoff–Hoeffding bound

Lemma E.1 (Multiplicative Chernoff for $[0,1]$ random variables)**.** *Let $Y_1, \dots, Y_n$ be independent random variables in $[0,1]$. Set $\bar{Y} := \frac{1}{n}\sum_j Y_j$ and $\nu := \mathbb{E}[\bar{Y}]$. For any $\delta \in [0,1]$,*

$$\Pr[\bar{Y} \ge (1+\delta)\nu] \;\le\; \exp\!\big(-\tfrac{\delta^2 n\nu}{3}\big), \qquad \Pr[\bar{Y} \le (1-\delta)\nu] \;\le\; \exp\!\big(-\tfrac{\delta^2 n\nu}{2}\big).$$

In particular, the symmetric bound $\Pr[|\bar{Y} - \nu| \ge \delta\nu] \le 2\exp(-\delta^2 n\nu/3)$ holds.

Proof. By the Cramér–Chernoff method. Let $S := \sum_j Y_j$, so $\mathbb{E}[S] = n\nu$. For any $\lambda > 0$, $\Pr[S \ge (1+\delta)n\nu] \le e^{-\lambda(1+\delta)n\nu}\,\mathbb{E}[e^{\lambda S}]$. Since each $Y_j \in [0,1]$, $\mathbb{E}[e^{\lambda Y_j}] \le 1 + \mathbb{E}[Y_j](e^\lambda - 1) \le \exp(\mathbb{E}[Y_j](e^\lambda - 1))$, by the inequality $1 + x \le e^x$. Hence $\mathbb{E}[e^{\lambda S}] \le \exp(n\nu(e^\lambda - 1))$, and

$$\Pr[S \ge (1+\delta)n\nu] \;\le\; \exp\!\big(n\nu(e^\lambda - 1) - \lambda(1+\delta)n\nu\big).$$

Optimizing in λ ($\lambda^\star = \log(1+\delta)$) gives $\Pr[S \ge (1+\delta)n\nu] \le ((e^\delta)/(1+\delta)^{1+\delta})^{n\nu}$. Using the standard estimate $\log\big((1+\delta)^{1+\delta}\big) - \delta \ge \delta^2/3$ for $\delta \in [0,1]$ gives the upper-tail bound. The lower-tail bound (with denominator 2) follows analogously from $\log\big((1-\delta)^{1-\delta}\big) + \delta \ge \delta^2/2$. The symmetric bound is the union of the two. □

Lemma E.2 (One-sided extension to bounded mean)**.** *Let $Y_1, \dots, Y_n$ be independent random variables in $[0,1]$, and suppose $\mathbb{E}[\bar{Y}] \le \nu_H$ for some $\nu_H \in [0,1]$. For any $\delta \in [0,1]$,*

$$\Pr[\bar{Y} \ge (1+\delta)\nu_H] \;\le\; \exp\!\big(-\tfrac{\delta^2 n \nu_H}{3}\big).$$

Proof. We use the Cramér–Chernoff method directly. For any $\lambda > 0$,

$$\Pr[\bar{Y} \ge (1+\delta)\nu_H] \;\le\; e^{-\lambda(1+\delta)n\nu_H} \prod_{j=1}^{n} \mathbb{E}[e^{\lambda Y_j}].$$

Since each $Y_j \in [0,1]$, $\mathbb{E}[e^{\lambda Y_j}] \le 1 + \mathbb{E}[Y_j](e^\lambda - 1)$ (by convexity of $e^{\lambda y}$ on $[0,1]$ and the fact that $e^{\lambda\cdot 0}(1-y) + e^{\lambda\cdot 1} y = 1 + y(e^\lambda - 1)$ dominates $e^{\lambda y}$ pointwise). Using $\mathbb{E}[Y_j] \le \nu_H$ (here we use $\mathbb{E}[\bar{Y}] \le \nu_H$ summand-wise; if individual means exceed ν_H, replace $\mathbb{E}[Y_j]$ by the sum bound below) and $1 + x \le e^x$:

$$\prod_{j=1}^{n} \mathbb{E}[e^{\lambda Y_j}] \;\le\; \prod_{j=1}^{n} \exp(\mathbb{E}[Y_j](e^\lambda - 1)) \;=\; \exp\!\Big(n\bar{\mu}(e^\lambda - 1)\Big) \;\le\; \exp\!\Big(n\nu_H(e^\lambda - 1)\Big),$$

where $\bar{\mu} := \frac{1}{n}\sum_j \mathbb{E}[Y_j] \le \nu_H$ (and $e^\lambda - 1 \ge 0$ for $\lambda \ge 0$). Optimizing in λ at $\lambda^\star = \log(1+\delta)$ gives

$$\Pr[\bar{Y} \ge (1+\delta)\nu_H] \;\le\; \Big(\tfrac{e^\delta}{(1+\delta)^{1+\delta}}\Big)^{n\nu_H} \;\le\; \exp\!\Big(-\tfrac{\delta^2 n\nu_H}{3}\Big),$$

the last step using $(1+\delta)\log(1+\delta) - \delta \ge \delta^2/3$ for $\delta \in [0,1]$. □

E.2 A useful numerical inequality

Claim E.3. For all $x \in [0, 1/2]$ and $a \in [0,1]$,

$$(1-x)^a \;\ge\; 1 - 2ax.$$

Proof. The binomial series gives, for $x \in [0,1)$ and $a \in [0,1]$,

$$(1-x)^a = 1 - ax - ax\Big(\tfrac{1-a}{2!}x + \tfrac{(1-a)(2-a)}{3!}x^2 + \cdots\Big).$$

The bracketed series is non-negative termwise (each factor $j - a \in [0, j]$ for $j \ge 1$, and dividing by $j!$ keeps coefficients positive), and using $\frac{(1-a)(2-a)\cdots(j-a)}{(j+1)!} \le \frac{1}{j+1} \le 1$ in each term, the bracket is bounded above by

$$\tfrac{x}{2} + \tfrac{x^2}{3} + \tfrac{x^3}{4} + \cdots \;\le\; x + x^2 + x^3 + \cdots \;=\; \tfrac{x}{1-x} \;\le\; 1$$

for $x \le 1/2$. Hence $(1-x)^a \ge 1 - ax - ax \cdot 1 = 1 - 2ax$. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in introduction and abstract match the paper's contribution.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations and future problems.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide full set of assumptions and a complete and correct proof.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We disclose all the details for experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Simple proof of concept experiments.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify all the test details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Simple proof of concept experiment. Reporting mean is insightful enough.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Small scale experiment. Such information is not relevant.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We respect NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper has no immediate societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.