Nash Regret and Beyond: Optimal Fairness Guarantees in Bandit Problems

Dhruv Sarkar
(IIT Kharagpur)

Sayak Ray Chowdhury
Assistant Professor
CSE, IIT Kanpur

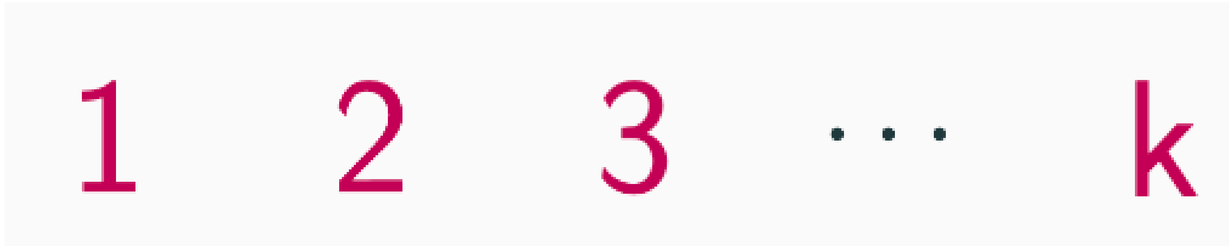
August 2026

Nishant Pandey
(IIT Kanpur)

Roadmap

1. Bandits, and what standard regret measures
2. Why averages hide unfairness: Nash social welfare
3. Optimal Nash regret — plain UCB is (nearly) all you need
4. Beyond Nash: the p-mean family
5. Nash regret in linear bandits
6. Open problems

Multi-armed Bandits

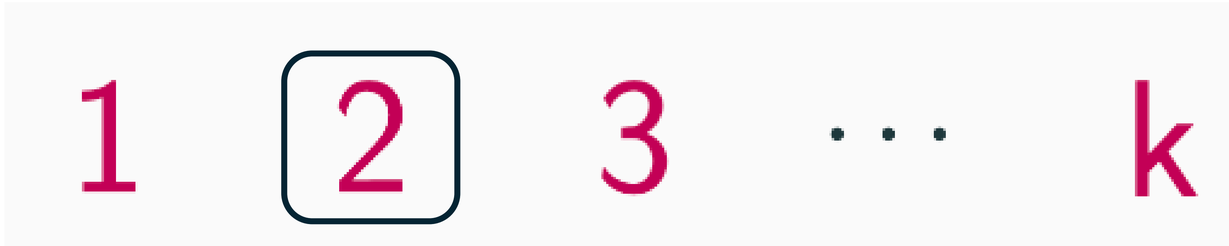


k arms with unknown parameters $\mu_1, \ldots, \mu_k$

Drug Trials (Thomson'33): k drugs administered among T agents.

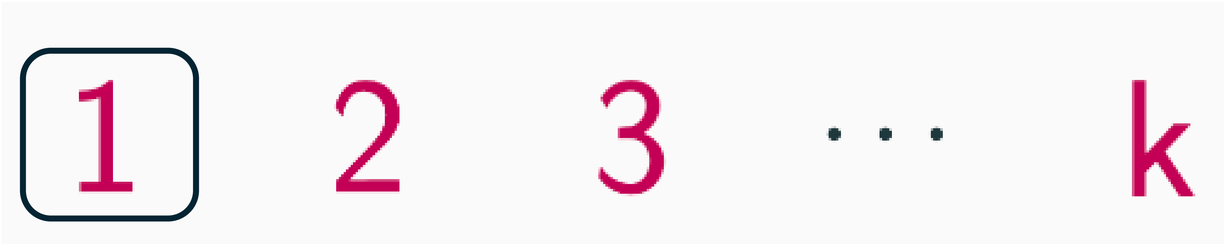
Display Ads (Schwartz'17): k ad configurations among T users.

Multi-armed Bandits



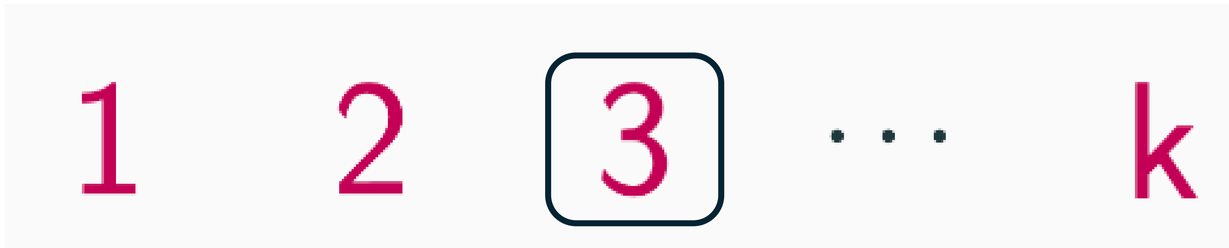
$$y_1 \sim \mu_2$$

Multi-armed Bandits



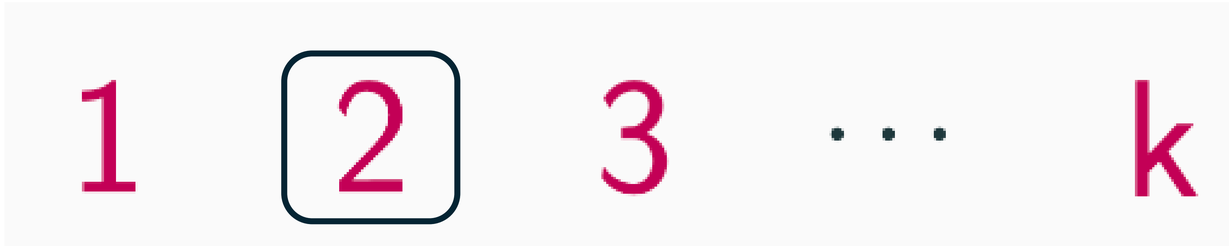
$$y_2 \sim \mu_1$$

Multi-armed Bandits



$$y_3 \sim \mu_3$$

Multi-armed Bandits



$$y_4 \sim \mu_2$$

Multi-armed Bandits

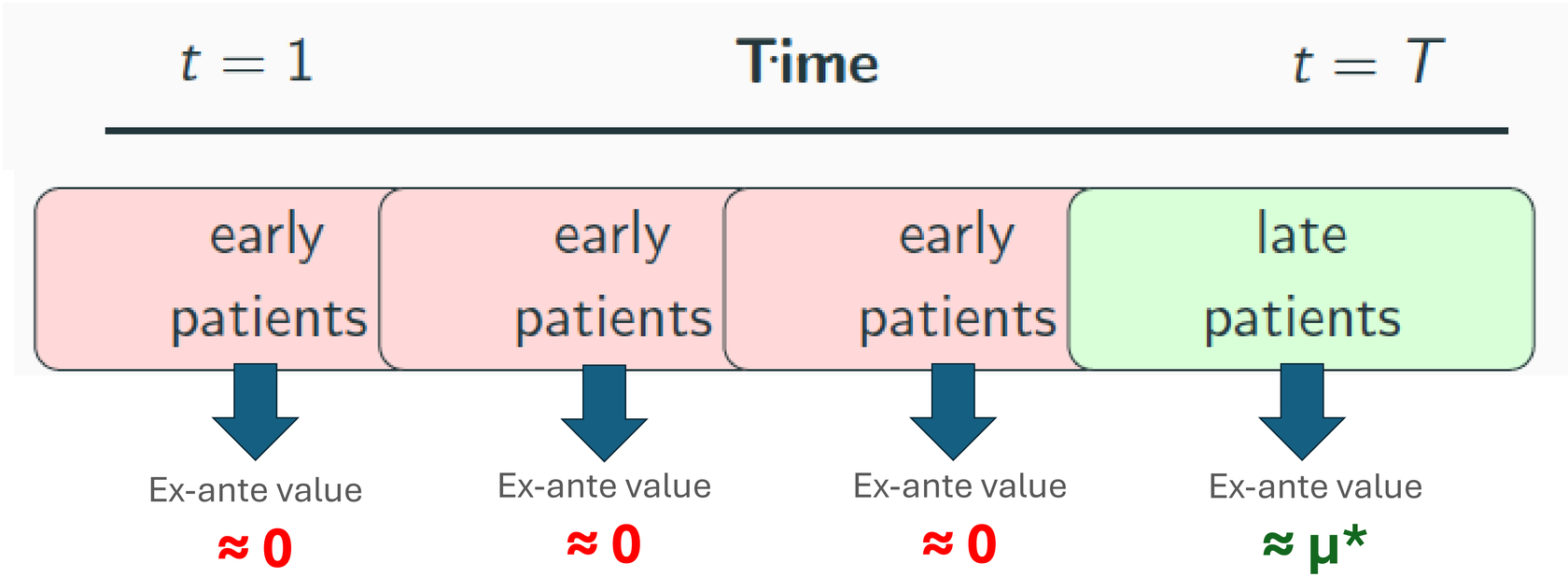
- Alg has sample access to k distributions (arms) for T rounds.
- Each round $t \in [T]$, alg selects an arm $I_t \in [k]$ and receives a random reward.
- Expected reward at round t: $\mathbb{E}[\mu_{I_t}]$.

$$\textbf{Benchmark: } \mu^* := \max_{i \in [k]} \mu_i$$

Regret of not knowing the benchmark

$$R_{\text{avg}}(T) := \mu^* - \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\mu_{I_t}]$$

Average Regret: Not a Fair Metric



Not fair for early patients, but average ex-ante value across T rounds can still be high

Same Average, Very Different Experience

Algorithm A — write off the early agents

$$v^A = \left(0,\, 0,\, \ldots,\, 0,\, \tfrac{1}{2},\, \tfrac{1}{2},\, \ldots,\, \tfrac{1}{2}\right)$$

$$\text{avg} = \tfrac{1}{4}$$

Algorithm B — spread the exploration out

$$v^B = \left(\tfrac{1}{4},\, \tfrac{1}{4},\, \tfrac{1}{4},\, \ldots,\, \tfrac{1}{4},\, \tfrac{1}{4},\, \tfrac{1}{4}\right)$$

$$\text{avg} = \tfrac{1}{4}$$

Identical average reward. Average regret cannot tell these apart
-- need a performance measure that can
-- a metric that is based on collective social welfare

Here v is the vector of per-round values.

Fairness by Design: Nash Social Welfare

Agents' values: $v_1, v_2, \ldots, v_T$

Welfare function: $\mathbb{R}_+^T \mapsto \mathbb{R}_+$

$$\min_t v_t \quad \leq \quad \left(\prod_t v_t\right)^{1/T} \quad \leq \quad \frac{1}{T}\sum_t v_t$$

Egalitarian Welfare

(Minimum)

Nash Social Welfare

(Geometric Mean)

Social Welfare

(Arithmetic Mean)

$$\mathrm{SW}\left(0, 0, \ldots, 0, \tfrac{1}{2}, \tfrac{1}{2}, \ldots, \tfrac{1}{2}\right) = \mathrm{SW}\left(\tfrac{1}{4}, \tfrac{1}{4}, \ldots, \tfrac{1}{4}\right)$$

$$\mathrm{NSW}\left(0, 0, \ldots, 0, \tfrac{1}{2}, \tfrac{1}{2}, \ldots, \tfrac{1}{2}\right) < \mathrm{NSW}\left(\tfrac{1}{4}, \tfrac{1}{4}, \ldots, \tfrac{1}{4}\right)$$

NSW satisfy social-choice axioms:

- Pigou-Dalton inequality aversion
- Scale invariance, Pareto efficiency, Symmetry, independent of irrelevant attributes

Whose Value, and When?

Each agent is handed a lottery: Algorithm's history-dependent choice of arm and then noisy reward

Agent's value = the value of that lottery, before it resolves:

$$v_t \;:=\; \mathbb{E}\big[\mu_{I_t}\big] \;=\; \sum_{i=1}^{k} \Pr\big[I_t = i\big]\,\mu_i$$

Why not the realized reward?

- No algorithm controls the noise — scoring realizations measures luck, not policy.
- Realized rewards can be unbounded below, so a product of realized rewards doesn't define welfare.
- A randomized rule is fair ex ante; ex post, somebody always loses the coin flip.

Why it never came up before

Arithmetic mean — the distinction is invisible:

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} y_t\right] \;=\; \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\mu_{I_t}\big]$$

Geometric mean — a choice has to be made:

$$\mathbb{E}\left[\Big(\prod_{t=1}^{T} y_t\Big)^{1/T}\right] \;\neq\; \Big(\prod_{t=1}^{T}\mathbb{E}\big[\mu_{I_t}\big]\Big)^{1/T}$$

Only the aggregator changes: arithmetic → geometric, over the same ex-ante values.

Nash Regret

Agents' values: $v_1, v_2, \ldots, v_T$

Welfare function: $\mathbb{R}_+^T \mapsto \mathbb{R}_+$

$$\min_t v_t \quad \leq \quad \left(\prod_t v_t\right)^{1/T} \quad \leq \quad \frac{1}{T}\sum_t v_t$$

Egalitarian Welfare

(Minimum)

Nash Social Welfare

(Geometric Mean)

Social Welfare

(Arithmetic Mean)

(Regret Notions)

$$R_{\text{Nash}}(T) := \mu^* - \left(\prod_{t=1}^{T} \mathbb{E}[\mu_{I_t}]\right)^{1/T} \geq R_{\text{avg}}(T) := \mu^* - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\mu_{I_t}]$$

Same ex-ante values, same benchmark. Only the aggregator changes

What a Low Nash Regret Promises and What It Doesn't

Fix the horizon T up front. The goal is to minimize Nash Regret:

$$\mathrm{NSW}_T := \left(\prod_{t=1}^{T} \mathbb{E}[\mu_{I_t}]\right)^{1/T}, \qquad R_{\mathrm{Nash}}(T) = \mu^* - \mathrm{NSW}_T$$

Equivalently: keep the average log-shortfall small.

$$\frac{1}{T}\sum_{t=1}^{T} \log \frac{\mu^*}{\mathbb{E}[\mu_{I_t}]} = \log \frac{\mu^*}{\mu^* - R_{\mathrm{Nash}}(T)} \approx \frac{R_{\mathrm{Nash}}(T)}{\mu^*}$$

Consequence (Markov's inequality): only a vanishing fraction of agents can be starved by any fixed factor.

$$\frac{1}{T}\left|\left\{ t : \mathbb{E}[\mu_{I_t}] \le \mu^*/c \right\}\right| \le \frac{1}{\log c} \log \frac{\mu^*}{\mu^* - R_{\mathrm{Nash}}(T)} \approx \frac{1}{\log c} \cdot \frac{R_{\mathrm{Nash}}(T)}{\mu^*}$$

Does NOT promise: that every agent gets the best arm. Impossible — you have to explore.

DOES promise: nobody is written off.

$$\mathbb{E}[\mu_{I_s}] = 0 \text{ for a single } s \implies \mathrm{NSW} = 0 \implies R_{\mathrm{Nash}}(T) = \mu^*$$

Uniform exploration is not unfair — every exploring agent faces the same lottery:

$$\Pr[I_t = i] = \tfrac{1}{k} \implies \mathbb{E}[\mu_{I_t}] = \frac{1}{k}\sum_{i=1}^{k} \mu_i \ge \frac{\mu^*}{k}$$

Roadmap

1 Bandits, and what standard regret measures

2 Why averages hide unfairness: Nash social welfare

3 Optimal Nash regret — plain UCB is (nearly) all you need

4 Beyond Nash: the p-mean family

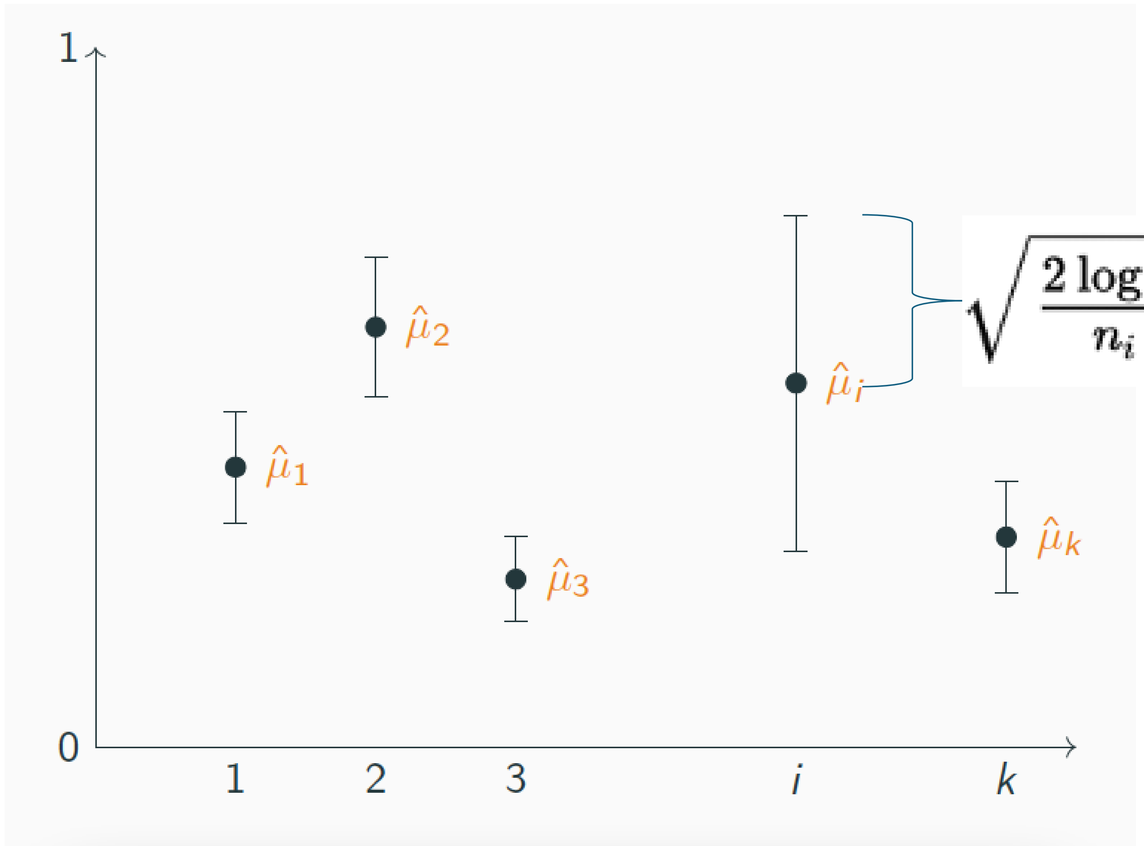
5 Nash regret in linear bandits

6 Open problems

UCB Algorithm for Bandits (Auer 2002)

By round $t \in [T]$, arm i sampled, say, n_i times and has empirical mean $\hat{\mu}_i$

Define $\mathrm{UCB}_i := \hat{\mu}_i + \sqrt{\frac{2 \log T}{n_i}}$



Choice of width $\varepsilon = \sqrt{\frac{2 \log T}{n}}$ guided by

Hoeffding's inequality

$$\Pr\{|\mu_i - \hat{\mu}_i| > \varepsilon\} \leq 2 \exp(-\varepsilon^2 n)$$

Select arm $I_t \in \arg\max_{1 \leq i \leq k} \mathrm{UCB}_i$

$\tilde{O}\left(\sqrt{\frac{k}{T}}\right)$ regret.

Nash-UCB (Barman et al. 2023)

Use $\mathrm{NCB}_i := \hat{\mu}_i + \sqrt{\frac{2\hat{\mu}_i \log T}{n_i}}$ instead of $\mathrm{UCB}_i := \hat{\mu}_i + \sqrt{\frac{2 \log T}{n_i}}$

Phase I: Perform uniform sampling for $\sqrt{T}$ rounds

Phase II: Select arm $I_t \in \arg\max_{1 \le i \le k} \mathrm{NCB}_i$

$\tilde{O}\left(\sqrt{\frac{k}{T}}\right)$ Nash Regret

Analysis via multiplicative-Chernoff bound: $\Pr\{|\mu_i - \hat{\mu}_i| > \delta \mu_i\} \le 2 \exp(-\delta^2 n \mu_i)$

Works for non-negative and bounded rewards only!!!

What happens when rewards are sub-Gaussian?

We propose: Go back to UCB!!!

Phase I: Perform uniform sampling for ~~$\sqrt{T}$ rounds~~ some random τ number of rounds

Phase II: Select arm $I_t \in \arg\max_{1 \le i \le k}$ ~~NCB_i~~ UCB_i

For sub-Gaussian rewards with unit variance:

$$\mathrm{UCB}_i := \hat{\mu}_i + \sqrt{\frac{2 \log T}{n_i}}$$

Stopping Rule:

Stop Phase I at the first time t when, for some arm i,

$$\hat{\mu}_i > 2\sqrt{\frac{2 \log T}{n_i}} \quad \text{and} \quad n_i > \frac{192 \log T}{\left(\hat{\mu}_i - 2\sqrt{\frac{2 \log T}{n_i}}\right)^2}.$$

→ Ensures: $\tau \approx \frac{k \log T}{(\mu^*)^2}$

Phase II starts once confidence is strong enough: UCB then delivers $\tilde{O}\left(\sqrt{\frac{k}{T}}\right)$ Nash Regret

Revisiting Social Welfare in Bandits: UCB is (Nearly) All You Need. Dhruv Sarkar, Nishant Pandey, C. (AISTATS 2026).

Roadmap

1 Bandits, and what standard regret measures

2 Why averages hide unfairness: Nash social welfare

3 Optimal Nash regret — plain UCB is (nearly) all you need

4 Beyond Nash: the p-mean family

5 Nash regret in linear bandits

6 Open problems

Beyond NSW: P-Mean Welfare

Generalized p-mean welfare, with $-\infty < p \leq 1$

$$\left(\frac{1}{T}\sum_t v_t^p\right)^{1/p}$$

$p \to -\infty$ | $p = 0$ | $p = 1$

$$\min_t v_t$$

Egalitarian Welfare

(Minimum)

$$\left(\prod_t v_t\right)^{1/T}$$

Nash Social Welfare

(Geometric Mean)

$$\frac{1}{T}\sum_t v_t$$

Social Welfare

(Arithmetic Mean)

Interpret p as a **fairness–utility knob**

Same Uniform exploration + UCB Works!!!

Minimal change in stopping rule

Stop Phase I at the first time t when, for some arm i,

$$\hat{\mu}_i > 2\sqrt{\frac{2\log T}{n_i}} \quad \text{and} \quad n_i > \frac{C_p\, 192 \log T}{\left(\hat{\mu}_i - 2\sqrt{\frac{2\log T}{n_i}}\right)^2}.$$

where $C_p = \begin{cases} 1 \text{ if } p \geq -1 \\ p^2 \text{ if } p < -1 \end{cases}$

P-mean regret bound*: $f(p,k) \cdot \tilde{O}\left(\sqrt{\frac{k}{T}}\right)$ where $f(p,k) = \begin{cases} 1 \text{ if } p \geq 0 \\ k^{|p|/2} \text{ if } p < 0 \end{cases}$

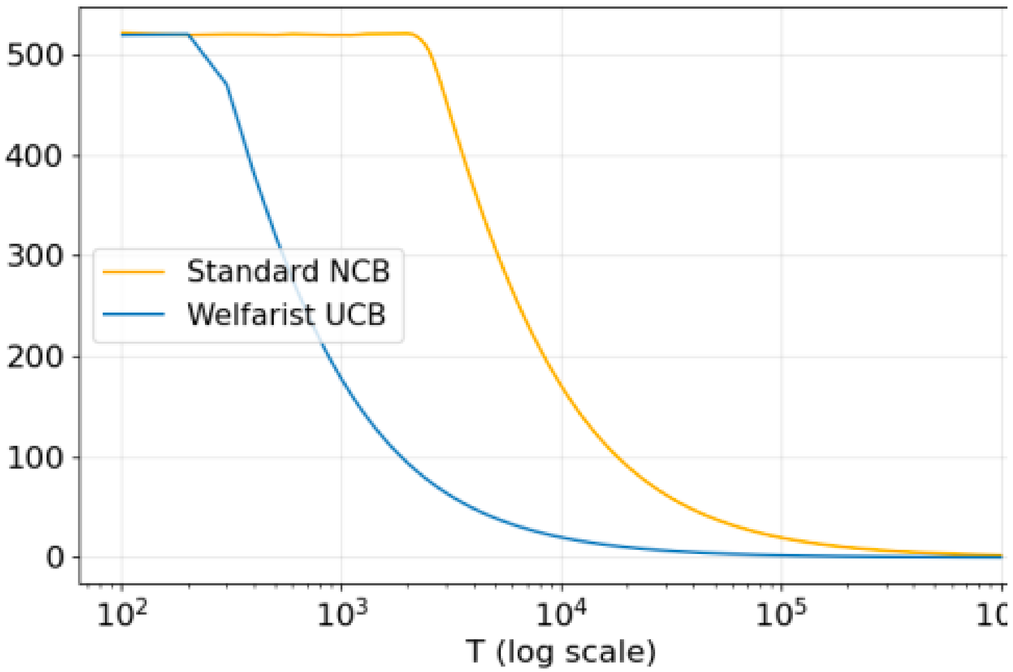
No free lunch!

Minimax optimal scaling**: $f(p,k) = \begin{cases} 1 \text{ if } p \geq -1 \\ k^{\frac{|p|-1}{2}} \text{ if } p < -1 \end{cases}$

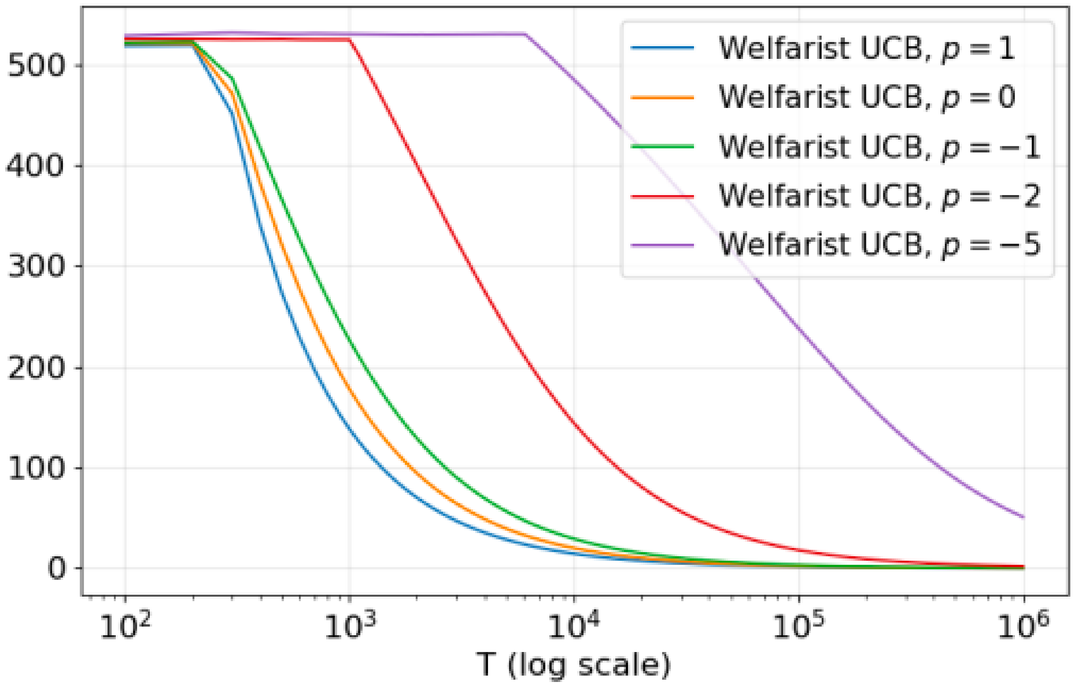
* Revisiting Social Welfare in Bandits: UCB is (Nearly) All You Need. Sarkar, Pandey, C. (AISTATS 2026).

** Price of Fairness in Bandits: A Tight Minimax Characterization. Sarkar, Dutta, C. (arxiv preprint 2026, under review).

Numerics



Nash regret comparison
for sub-Gaussian rewards



P-mean regret comparison
for sub-Gaussian rewards

* Revisiting Social Welfare in Bandits: UCB is (Nearly) All You Need. Sarkar, Pandey, C. (AISTATS 2026).

Roadmap

1 Bandits, and what standard regret measures

2 Why averages hide unfairness: Nash social welfare

3 Optimal Nash regret — plain UCB is (nearly) all you need

4 Beyond Nash: the p-mean family

5 Nash regret in linear bandits

6 Open problems

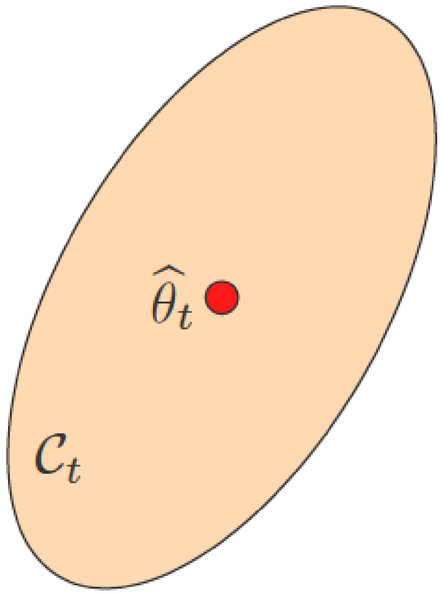
Linear Bandits (Abbasi-Yadkori et al. 2011)

Problem setting

- Each arm i is a vector $x_i \in \mathbb{R}^d$
- Playing arm i_t gives reward

$$y_t = \theta^\top x_{i_t} + \varepsilon_t$$

- $\theta \in \mathbb{R}^d$ is unknown and $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ is noise



LinUCB algorithm

- Build a point estimate (least squares estimate) $\widehat{\theta}_t$ and a confidence region (ellipsoid) $\mathcal{C}_t$
- Play the most optimistic action w.r.t. this ellipsoid

$$i_t = \operatorname*{argmax}_{1 \leqslant i \leqslant N} \max_{\theta \in \mathcal{C}_t} \left(\theta^\top x_i\right)$$

LinUCB achieves $\tilde{O}\left(\frac{d}{\sqrt{T}}\right)$ average regret

Nash Regret in Linear Bandits

Sawarni, Pal, Barman (23): Lin-NCB achieves $\tilde{O}\left(\frac{d^{5/4}}{\sqrt{T}}\right)$ Nash regret (only for positive rewards)

Open problem: $\tilde{O}\left(\frac{d}{\sqrt{T}}\right)$ Nash regret (for general sub-Gaussian rewards)

Our solution: Back to Lin-UCB

Phase I: D-optimal design based uniform exploration

Data-adaptive Phase I stopping rule

$$t \geq \frac{900\, p_a^2 \sigma^2 d^2 \log T}{\left(\max_{x \in \mathcal{X}} \langle x, \hat{\theta} \rangle - \sqrt{\frac{48 \sigma^2 d^2 \log T}{t}}\right)^2}$$

Phase II: Lin-UCB based controlled exploitation

Improved Algorithms for Nash Welfare in Linear Bandits. Dhruv Sarkar, Nishant Pandey, C. (ICML 2026).

Future Directions

1. Non-linear structure in rewards (e.g., GLMs, kernels,…)

Our hypothesis: Take any optimistic (e.g., UCB style) average regret minimization algorithm $\mathcal{A}$ for the given bandit problem and design:

1. A uniform arm exploration strategy in Phase I

2. A stopping rule using per-round regret and LCB index of $\mathcal{A}$

2. Notion of Fairness with ex-post rewards, anytime valid confidence sets