A Closer Look at the Worst-case Behavior of Multi-armed Bandit Algorithms

Anand Kalvit
(Joint work with **Assaf Zeevi**)

Columbia University, New York

NeurIPS 2021

Anand Kalvit

Assaf Zeevi

The canonical 2-armed bandit (2-MAB) revisited

- Two arms with means μ_1, μ_2.
- Gap $\Delta := \mu_1 - \mu_2 > 0$.

The canonical 2-armed bandit (2-MAB) revisited

- Two arms with means μ_1, μ_2.
- Gap $\Delta := \mu_1 - \mu_2 > 0$.
- Reward sequence for arm $i \in \{1, 2\}$: $\{X_{i,j} : j = 1, 2, ...\}$.
- $X_{i,j}$'s are independent and bounded in $[0, 1]$.

The canonical 2-armed bandit (2-MAB) revisited

- Two arms with means μ_1, μ_2.
- Gap $\Delta := \mu_1 - \mu_2 > 0$.
- Reward sequence for arm $i \in \{1, 2\}$: $\{X_{i,j} : j = 1, 2, ...\}$.
- $X_{i,j}$'s are independent and bounded in $[0, 1]$.
- **Goal.** Maximize cumulative expected payoffs over n plays.
- **Question.** What should inform the sequence of arm-pulls?

The canonical 2-armed bandit (2-MAB) revisited

- Policy $\pi := \{\pi_t; t = 1, ..., n\}$ prescribes arm $\pi_t \in \{1, 2\}$ at time t.

The canonical 2-armed bandit (2-MAB) revisited

- Policy $\pi := \{\pi_t; t = 1, ..., n\}$ prescribes arm $\pi_t \in \{1, 2\}$ at time t.
- Cumulative regret of policy π after n samples is given by

$$R_n^\pi := \sum_{t=1}^{n} \left[\mu_1 - X_{\pi_t, N_{\pi_t}(t)}\right],$$

where $N_{\pi_t}(t)$ indicates the number of pulls of arm π_t until time t.

The canonical 2-armed bandit (2-MAB) revisited

- Policy $\pi := \{\pi_t; t = 1, ..., n\}$ prescribes arm $\pi_t \in \{1, 2\}$ at time t.
- Cumulative regret of policy π after n samples is given by

$$R_n^\pi := \sum_{t=1}^{n} \left[\mu_1 - X_{\pi_t, N_{\pi_t}(t)} \right],$$

where $N_{\pi_t}(t)$ indicates the number of pulls of arm π_t until time t.

- The goal is minimization of the **expected cumulative regret**, i.e.,

$$\inf_{\pi \in \Pi} \mathbb{E} R_n^\pi,$$

where Π is the set of non-anticipating policies
(A "good" policy has $o(n)$ regret, i.e., long-run-average optimality.).

Well-known algorithms for the problem

- Plethora of available algorithms.
- **Forced sampling-based:** $\underbrace{\text{Explore-then-Commit, } \epsilon_n\text{-Greedy}}_{non-adaptive\ (\Delta-dependent)}$, etc.
- **Posterior sampling-based:** $\underbrace{\text{Thompson Sampling and variants}}_{adaptive\ (\Delta-independent)}$, etc.
- **Optimism-based:** $\underbrace{\textbf{UCB} \text{ and variants}}_{adaptive\ (\Delta-independent)}$, etc.

Upper Confidence Bounds: The Optimism principle

UCB(ρ): UCB with exploration coefficient ρ

At time $t+1$, play an arm $\pi_{t+1} \in \{1,2\}$ according to

$$\pi_{t+1} \in \arg\max_{i \in \{1,2\}} \left(\bar{X}_i(t) + \sqrt{\frac{\rho \log t}{N_i(t)}} \right).$$

Here,

1. $\bar{X}_i(t)$ denotes the empirical mean reward from arm i at time t^+, i.e.,

$$\bar{X}_i(t) := \frac{\sum_{j=1}^{N_i(t)} X_{i,j}}{N_i(t)}.$$

2. $\rho = 2$ **corresponds to classical UCB1**.

Achievable regret in 2-MAB

- **Instance-dependent bounds** (Fixed Δ, large n) [**Easy problems**]:

$$\mathbb{E}R_n^\pi \leqslant \frac{C_1 \rho \log n}{\Delta} + \frac{C_2 \Delta}{\rho - 1} \qquad \text{for } \pi = \text{UCB with } \rho > 1.$$

$$\mathbb{E}R_n^\pi = \Omega\left(\frac{\log n}{\Delta}\right) \qquad \text{(L.B. for any policy } \pi\text{).}$$

Achievable regret in 2-MAB

- **Instance-dependent bounds** (Fixed Δ, large n) [**Easy problems**]:

$$\mathbb{E}R_n^\pi \leqslant \frac{C_1 \rho \log n}{\Delta} + \frac{C_2 \Delta}{\rho - 1} \qquad \text{for } \pi = \text{UCB with } \rho > 1.$$

$$\mathbb{E}R_n^\pi = \Omega\left(\frac{\log n}{\Delta}\right) \qquad \text{(L.B. for any policy } \pi\text{).}$$

- **Minimax bounds** (Fixed n, worst-case Δ) [**Hard problems**]:

$$\mathbb{E}R_n^\pi \leqslant C_\rho \sqrt{n \log n} \qquad \text{for } \pi = \text{UCB with } \rho > 1.$$

$$\mathbb{E}R_n^\pi = \Omega\left(\sqrt{n}\right) \qquad \text{(L.B. for any policy } \pi\text{).}$$

Achievable regret in 2-MAB

- **Instance-dependent bounds** (Fixed Δ, large n) [**Easy problems**]:

$$\mathbb{E}R_n^\pi \leqslant \frac{C_1 \rho \log n}{\Delta} + \frac{C_2 \Delta}{\rho - 1} \qquad \text{for } \pi = \text{UCB with } \rho > 1.$$

$$\mathbb{E}R_n^\pi = \Omega\left(\frac{\log n}{\Delta}\right) \qquad \text{(L.B. for any policy } \pi\text{).}$$

- **Minimax bounds** (Fixed n, worst-case Δ) [**Hard problems**]:

$$\mathbb{E}R_n^\pi \leqslant C_\rho \sqrt{n \log n} \qquad \text{for } \pi = \text{UCB with } \rho > 1.$$

$$\mathbb{E}R_n^\pi = \Omega\left(\sqrt{n}\right) \qquad \text{(L.B. for any policy } \pi\text{).}$$

- **Note:** Thompson Sampling also has similar guarantees, to wit, $\mathcal{O}\left(\frac{\log n}{\Delta}\right)$ and $\mathcal{O}\left(\sqrt{n \log n}\right)$ respectively.

How about the distribution of arm-pulls?

- How well do we understand the distribution of $\frac{N_1(n)}{n}$?

How about the distribution of arm-pulls?

- How well do we understand the distribution of $\frac{N_1(n)}{n}$?
- **Existing results offer limited insight.**
- E.g., if $\Delta \gg 0$, then first-order optimal algorithms guarantee

$$\frac{N_1(n)}{n} \Rightarrow 1 \quad \text{as } n \to \infty.$$

How about the distribution of arm-pulls?

- How well do we understand the distribution of $\frac{N_1(n)}{n}$?
- **Existing results offer limited insight.**
- E.g., if $\Delta \gg 0$, then first-order optimal algorithms guarantee

$$\frac{N_1(n)}{n} \Rightarrow 1 \quad \text{as } n \to \infty.$$

- But, what happens to $\frac{N_1(n)}{n}$ as $\Delta \to 0$?

Distribution of arm-pulls as $\Delta \to 0$

- Why bother about $\frac{N_1(n)}{n}$ as $\Delta \to 0$?

Distribution of arm-pulls as $\Delta \to 0$

- Why bother about $\frac{N_1(n)}{n}$ as $\Delta \to 0$?
- Consider a 2-MAB with $\Delta = 0$ and Bernoulli(0.5) rewards.

- Why bother about $\frac{N_1(n)}{n}$ as $\Delta \to 0$?
- Consider a 2-MAB with $\Delta = 0$ and Bernoulli(0.5) rewards.

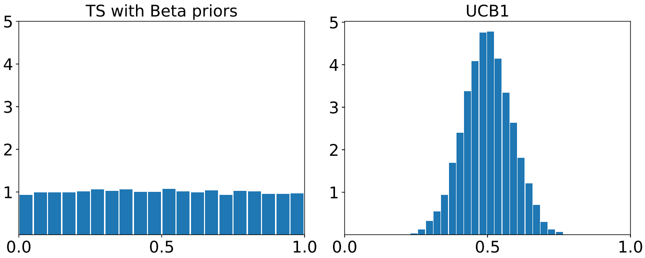

Figure: Empirical distribution of $\frac{N_1(n)}{n}$ after $n = 10^4$ pulls [$N = 10^5$ experiments].

Distribution of arm-pulls as $\Delta \to 0$

- Why bother about $\frac{N_1(n)}{n}$ as $\Delta \to 0$?
- Consider a 2-MAB with $\Delta = 0$ and Bernoulli(0.5) rewards.

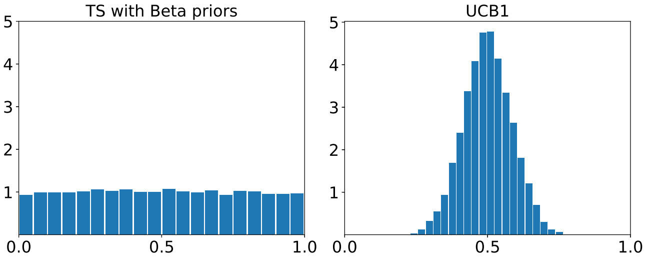

Figure: Empirical distribution of $\frac{N_1(n)}{n}$ after $n = 10^4$ pulls [$N = 10^5$ experiments].

- **Fairness**: "Similar" arms should get "similar" traffic w.h.p.
- **Ex post inference**: Clinical trials of 2 "similarly" efficacious vaccines!
- **The Countable-armed Bandit problem [KZ'20]**.

The curious case of $\Delta = 0$

TS-BP: Thompson Sampling with Beta priors, Bernoulli likelihoods

At time $t+1$, play an arm $\pi_{t+1} \in \{1, 2\}$ according to

$$\pi_{t+1} \in \arg\max_{i \in \{1,2\}} \mathcal{B}_{i,t}\left(S_i^t, F_i^t\right).$$

The curious case of $\Delta = 0$

TS-BP: Thompson Sampling with Beta priors, Bernoulli likelihoods

At time $t+1$, play an arm $\pi_{t+1} \in \{1,2\}$ according to

$$\pi_{t+1} \in \arg\max_{i \in \{1,2\}} \mathcal{B}_{i,t}\left(S_i^t, F_i^t\right).$$

[Theorem] "Instability" of TS-BP

In a 2-MAB with $\Delta = 0$, there exists a pair of instances (ν_1, ν_2) s.t.

- On ν_1, $\frac{N_1(n)}{n} \Rightarrow \frac{1}{2}$ as $n \to \infty$.
- On ν_2, $\frac{N_1(n)}{n} \Rightarrow$ Uniform on $[0,1]$ as $n \to \infty$.

General Δ: Distribution of arm-pulls under UCB

[Theorem] Sampling asymptotics for UCB with $\rho > 1$

In a 2-MAB with gap Δ, the following holds as $n \to \infty$:

$$\frac{N_1(n)}{n} \Rightarrow \begin{cases} 1 & \text{if } \Delta = \omega\left(\sqrt{\frac{\log n}{n}}\right), \\ \lambda^*_\rho(\theta) & \text{if } \Delta \sim \sqrt{\frac{\theta \log n}{n}} \text{ for some fixed } \theta \geqslant 0, \\ \frac{1}{2} & \text{if } \Delta = o\left(\sqrt{\frac{\log n}{n}}\right). \end{cases}$$

$\lambda^*_\rho(\theta)$ is deterministic and can be characterized in closed-form!

General Δ: Distribution of arm-pulls under UCB

[Theorem] Sampling asymptotics for UCB with $\rho > 1$

In a 2-MAB with gap Δ, the following holds as $n \to \infty$:

$$\frac{N_1(n)}{n} \Rightarrow \begin{cases} 1 & \text{if } \Delta = \omega\left(\sqrt{\frac{\log n}{n}}\right), \\ \lambda_\rho^*(\theta) & \text{if } \Delta \sim \sqrt{\frac{\theta \log n}{n}} \text{ for some fixed } \theta \geqslant 0, \\ \frac{1}{2} & \text{if } \Delta = o\left(\sqrt{\frac{\log n}{n}}\right). \end{cases}$$

$\lambda_\rho^*(\theta)$ is deterministic and can be characterized in closed-form!

Recall: Thompson Sampling may result in a non-degenerate limit!

Worst-case behavior of UCB

[Theorem] Minimax regret of UCB with $\rho > 1$

In a 2-MAB, the worst-case regret of UCB follows the sharp asymptotic

$$\mathbb{E}R_n^{\pi} \sim f(\rho)\sqrt{n \log n}.$$

The constant $f(\rho)$ can be characterized in closed-form!
(**Note:** The information-theoretic optimal minimax rate is $\Theta\left(\sqrt{n}\right)$.)

Worst-case behavior of UCB

[Theorem] Minimax regret of UCB with $\rho > 1$

In a 2-MAB, the worst-case regret of UCB follows the sharp asymptotic

$$\mathbb{E}R_n^\pi \sim f(\rho)\sqrt{n \log n}.$$

The constant $f(\rho)$ can be characterized in closed-form!
(**Note:** The information-theoretic optimal minimax rate is $\Theta\left(\sqrt{n}\right)$.)

Remark: Previous best result for UCB was $\mathcal{O}\left(\sqrt{n \log n}\right)$ minimax regret.

Diffusion-scale analysis of bandits

- **Information-theoretic hardest instances have $\boldsymbol{\Delta \asymp \frac{1}{\sqrt{n}}}$.**
- **Analogous to the "heavy-traffic/QED" regime in queuing, where 1 - traffic intensity $\asymp \frac{1}{\sqrt{n}}$.**
- The queuing problem admits well-known diffusion limits.
- Can similar results be established also for bandits?

Diffusion approximation for UCB

[Theorem] Diffusion limit regret of UCB with $\rho > 1$

In a 2-MAB with gap $\Delta \sim \frac{c}{\sqrt{n}}$, the following holds under UCB as $n \to \infty$:

$$\left(\frac{R^{\pi}_{\lfloor nt \rfloor}}{\sqrt{n}} \right)_{t \in [0,1]} \Rightarrow \left(\frac{ct}{2} + \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}} B(t) \right)_{t \in [0,1]},$$

where $\left\{ \sigma_i^2 : i = 1, 2 \right\}$ are the reward variances, and $B(t)$ is a standard Brownian motion in $\mathbb{R}$.

Diffusion approximation for UCB

[Theorem] Diffusion limit regret of UCB with $\rho > 1$

In a 2-MAB with gap $\Delta \sim \frac{c}{\sqrt{n}}$, the following holds under UCB as $n \to \infty$:

$$\left(\frac{R^{\pi}_{\lfloor nt \rfloor}}{\sqrt{n}} \right)_{t\in[0,1]} \Rightarrow \left(\frac{ct}{2} + \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}} B(t) \right)_{t\in[0,1]},$$

where $\{\sigma_i^2 : i = 1, 2\}$ are the reward variances, and $B(t)$ is a standard Brownian motion in $\mathbb{R}$.

Note: For Thompson Sampling, the diffusion limit is characterized by the solution(s) to a SDE ([**Wager & Xu, 2021**],[**Fan & Glynn, 2021**]).

Bibliography

- **[KZ'20]** A. Kalvit and A. Zeevi, "From Finite to Countable-armed Bandits," `NeurIPS 2020`.
- **[Wager & Xu, 2021]** S. Wager and K. Xu, "Diffusion Asymptotics for sequential experiments," `arXiv preprint arXiv:2101.09855`.
- **[Fan & Glynn, 2021]** L. Fan and P. Glynn, "Diffusion Approximations for Thompson Sampling," `arXiv preprint arXiv:2105.09232`.