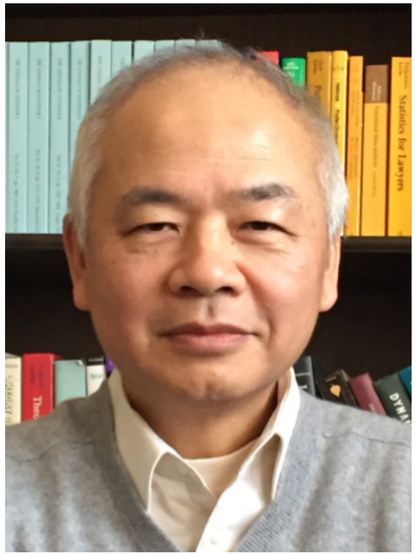
Automated Bandit A/B Testing: A Tale of Two Algorithms

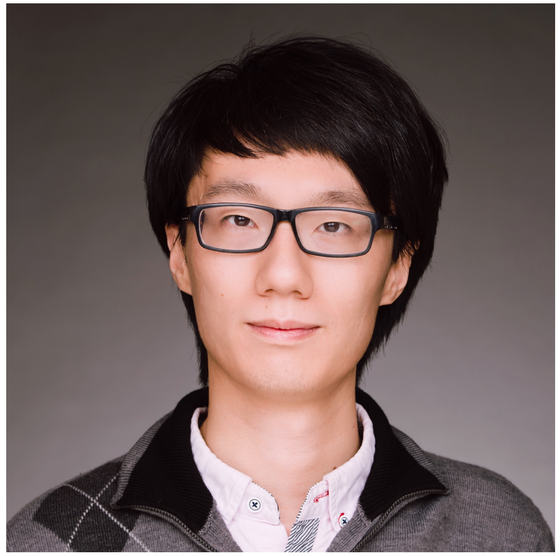
Koulik Khamaru

Rutgers University
Dept, of Statistics

RUTGERS

Collaborators



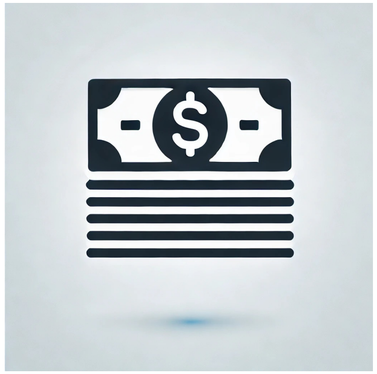


An experiment story:
Helping Syrian refugees find work

An Adaptive Targeted Field Experiment: Job Search Assistance for Refugees in Jordan

Stefano Caria, Grant Gordon, Maximilian Kasy, Simon Quinn, Soha Shami, Alexander Teytelboym

An experiment story:
Helping Syrian refugees find work



Cash
(Arm 1)

skill -training
(Arm 2)

Motivational support
(Arm 3)

How do we improve employment when resources are limited?

Digital health interventions

Provide targeted texts to promote exercise.

More examples of adaptive experiments

American Economic Review

Two goals:

1. Optimize resource allocation during experiment

2. Learn (Hypothesis testing) from data after experiment

Malaria
d Trial

geted treatment assignment policy, where the goal is to use a participant's survey responses to determine which charity to expose them to in a donation solicitation. The design balances two competing objectives: optimizing the outcomes for the subjects in the experiment ("cumulative regret minimization") and gathering data that will be most useful for policy learning, that is, for learning an assignment rule that will maximize welfare if used after the experiment ("simple regret minimization"). We evaluate alternative experimental designs by collecting pilot data and then conducting a simulation study. Next, we implement our selected algorithm. Finally, we perform a second simulation study anchored to the collected data that evaluates the benefits of the algorithm we chose. Our first result is that the value of a learned policy in this setting is higher when data is collected via a uniform randomization rather than collected adaptively using standard cumulative regret minimization or policy learning algorithms. We propose a

AMERICAN ECONOMIC REVIEW
VOL. 105, NO. 2, FEBRUARY 2015
(pp. 609–45)

The Multi-armed bandit framework

K arms with reward distributions $P_1, \ldots, P_K$ and $\mathbf{E}_{Y \sim P_a}[Y] = \mu_a$,

Thompson 1933, Robbins 1952

The Multi-armed bandit framework

K arms with reward distributions $P_1, \ldots, P_K$ and $\mathbf{E}_{Y \sim P_a}[Y] = \mu_a$,

- At round t,
  - Pull arm $A_t \in \sigma\left\{A_1, Y_1, \ldots, A_{t-1}, Y_{t-1}\right\}$, and observe reward Y_t
  - If $A_t = a$, We observe reward $Y_t = \mu_a + \epsilon_t$

Thompson 1933, Robbins 1952

The Multi-armed bandit framework

K arms with reward distributions $P_1, \ldots, P_K$ and $\mathbf{E}_{Y \sim P_a}[Y] = \mu_a$,

- At round t,
  - Pull arm $A_t \in \sigma\left\{A_1, Y_1, \ldots, A_{t-1}, Y_{t-1}\right\}$, and observe reward Y_t
  - If $A_t = a$, We observe reward $Y_t = \mu_a + \epsilon_t$

Minimize regret: $T\mu^\star - \mathbb{E}\left(\sum_{t=1}^{T} Y_t\right)$ $\mu^\star = \max_a \mu_a$

Thompson 1933, Robbins 1952

The linear contextual bandit

$$y_t = x_t^\top \mu + \epsilon_t \qquad x_i \in \mathcal{X}, \quad y_t \in \mathbb{R}$$

The linear contextual bandit

$$y_t = x_t^\top \mu + \epsilon_t \qquad x_i \in \mathcal{X}, \quad y_t \in \mathbb{R}$$

x_t depends on $\{x_1, y_1, x_2, y_2, \ldots, x_{t-1}, y_{t-1}\}$

The linear contextual bandit

$$y_t = x_t^\top \mu + \epsilon_t \quad x_t \in \mathcal{X}, \quad y_t \in \mathbb{R}$$

x_t depends on $\{x_1, y_1, x_2, y_2, \ldots, x_{t-1}, y_{t-1}\}$

K-armed bandit: $x_t \in \{e_1, \ldots, e_k\}$

$$\mu = \left(\mu_1, \mu_2, \ldots, \mu_k\right)^\top$$

The linear contextual bandit

$$y_t = x_t^\top \mu + \epsilon_t \qquad x_t \in \mathcal{X}, \quad y_t \in \mathbb{R}$$

$$a \in \mathbb{R}^d \qquad |\mathcal{X}| = K$$

The linear contextual bandit

$$y_t = x_t^\top \mu + \epsilon_t \qquad x_t \in \mathcal{X}, \quad y_t \in \mathbb{R}$$

$$a \in \mathbb{R}^d \qquad |\mathcal{X}| = K$$

Data set is not i.i.d.

The linear contextual bandit

$$y_t = x_t^\top \mu + \epsilon_t \qquad x_t \in \mathcal{X}, \quad y_t \in \mathbb{R}$$

$$a \in \mathbb{R}^d \qquad |\mathcal{X}| = K$$

Data set is not i.i.d.

Confidence interval for $a^\top \mu$ after bandit experiment.

The linear contextual bandit

$$y_t = x_t^\top \mu + \epsilon_t \qquad x_t \in \mathcal{X}, \quad y_t \in \mathbb{R}$$

x_t depends on $\{x_1, y_1, x_2, y_2, \ldots, x_{t-1}, y_{t-1}\}$

$a \in \mathbb{R}^d \qquad |\mathcal{X}| = K$

Data set is not i.i.d.

Confidence interval for $a^\top \mu$ after bandit experiment.

Prior works

Concentration bounds: Yadkori et al 2011, Shin et al. 2019, Smith et al. 2023, Lattimore 2022, Khamaru et al. 2021, …..

Debiasing methods: Zhang and Zhang 2013, Deshpande et al. 2017, 2019, Khamaru et al. 2021, Ying et al 2023, Lin et al. 2023, …..

Weighted Z-estimator: Zhang et al. 2020, Bibaut et al. 2021, Hadad et al. 2021, Lin et al. 2023, Syrgkanis and Zhan 2023, Zhan et al. 2021, ….

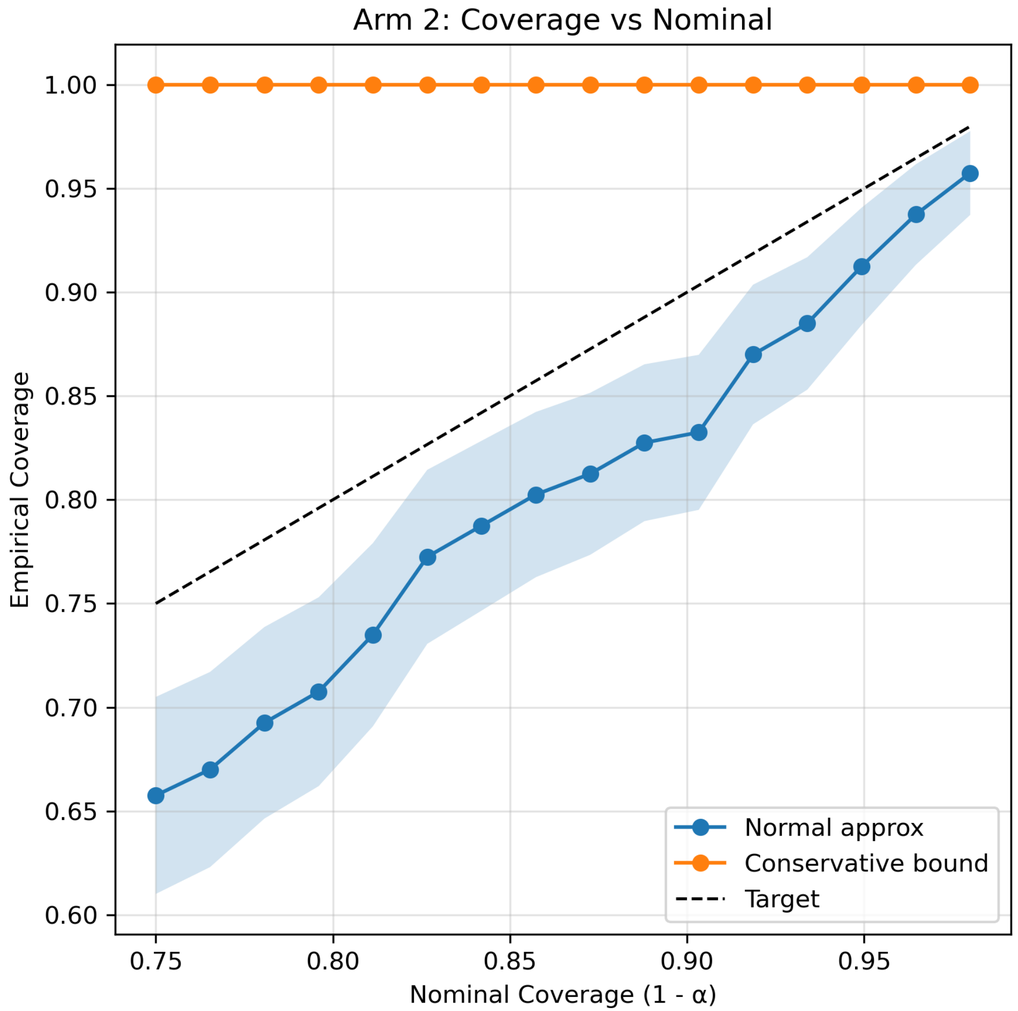
Prior works

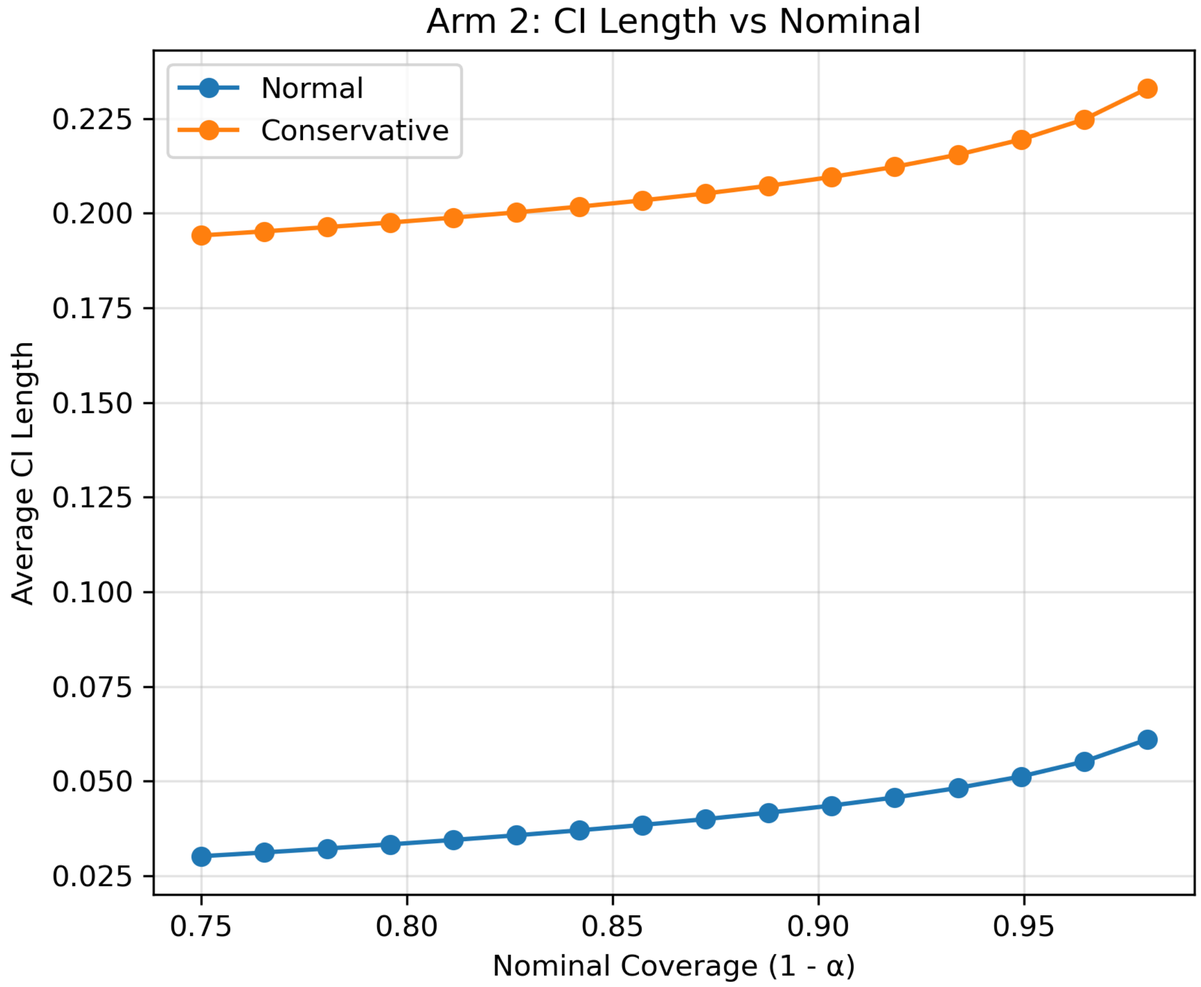
Concentration bounds: Yadkori et al 2011, Shin et al. 2019, Smith et al. 2023, Lattimore 2022, Khamaru et al. 2021, …..

Debiasing methods: Zhang and Zhang 2013, Deshpande et al. 2017, 2019, Khamaru et al. 2021, Ying et al 2023, Lin et al. 2023, …..

Weighted Z-estimator: Zhang et al. 2020, Bibaut et al. 2021, Hadad et al. 2021, Lin et al. 2023, Syrgkanis and Zhan 2023, Zhan et al. 2021, ….

Confidence Interval Comparison





Prior works

Flexible, but inference might be weak.

Our approach: Inference is much better when the bandit algorithm has structure.

Lai and Wei 1982

The Annals of Statistics
1982, Vol. 10, No. 1, 154–166

LEAST SQUARES ESTIMATES IN STOCHASTIC REGRESSION MODELS WITH APPLICATIONS TO IDENTIFICATION AND CONTROL OF DYNAMIC SYSTEMS

BY TZE LEUNG LAI[1] AND CHING ZONG WEI[2]

Columbia University and University of Maryland

Strong consistency and asymptotic normality of least squares estimates in stochastic regression models are established under certain weak assumptions on the stochastic regressors and errors. We discuss applications of these results to interval estimation of the regression parameters and to recursive on-line identification and control schemes for linear dynamic systems.

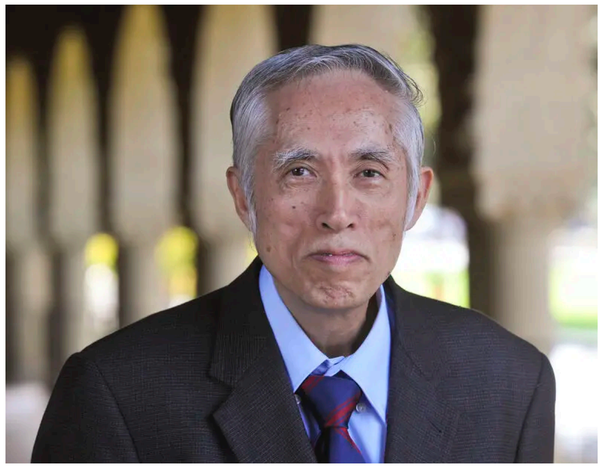
1. Introduction. Consider the multiple regression model

$$y_n = \beta_1 x_{n1} + \cdots + \beta_{p1} x_{np} + \varepsilon_n, \quad n = 1, 2, \cdots \tag{1.1}$$

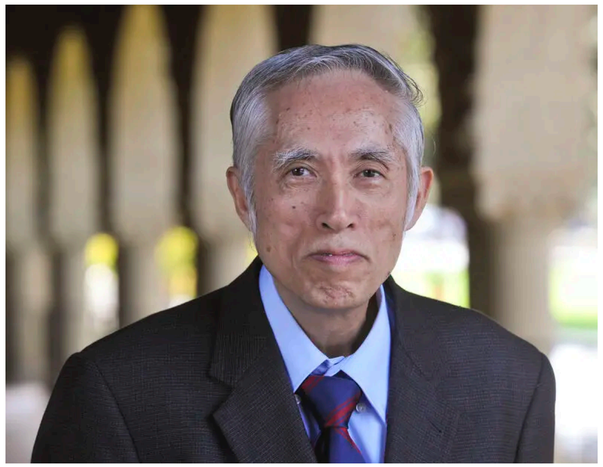
where the ε_n are unobservable random errors, $\beta_1, \cdots, \beta_p$ are unknown parameters, and y_n is the observed response corresponding to the design levels $x_{n1}, \cdots, x_{np}$. Let $\mathbf{x}_n = (x_{n1}, \cdots, x_{np})'$ and let $\mathbf{X}_n = (x_{ij})_{1\le i\le n, 1\le j\le p}$, $\mathbf{Y}_n = (y_1, \cdots, y_n)'$. Then

$$\mathbf{b}_n = (b_{n1}, \cdots, b_{np})' = (\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'\mathbf{Y}_n \tag{1.2}$$

denotes the least squares estimate of $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$ based on the observations $\mathbf{x}_1, y_1, \cdots, \mathbf{x}_n, y_n$, assuming that $\mathbf{X}_n\mathbf{X}_n'$ is nonsingular. Throughout the sequel we shall assume that $\{\varepsilon_n\}$ is a martingale difference sequence with respect to an increasing sequence of σ-fields $\{\mathcal{F}_n\}$; i.e., ε_n is $\mathcal{F}_n$-measurable and $E(\varepsilon_n | \mathcal{F}_{n-1}) = 0$ for every n. An important



Lai and Wei Stability



THEOREM 3. *Suppose that in the regression model* (1.1), $\{\varepsilon_n\}$ *is a martingale difference sequence with respect to an increasing sequence of* σ*-fields* $\{\mathcal{F}_n\}$ *such that* (1.6) *and* (4.1) *hold. Moreover, assume for each* n *that the design vector* $\mathbf{x}_n = (x_{n1}, \cdots, x_{np})'$ *at stage* n *is* $\mathcal{F}_{n-1}$*-measurable and that there exists a non-random positive definite symmetric matrix* $\mathbf{B}_n$ *for which*

$$\mathbf{B}_n^{-1}\left(\sum_1^n \mathbf{x}_i\mathbf{x}_i'\right)^{1/2} \to_P \mathbf{I}_p, \quad \text{and} \tag{4.2}$$

$$\max_{1\le i\le n} \|\mathbf{B}_n^{-1}\mathbf{x}_i\| \to_P 0. \tag{4.3}$$

Then the least squares estimate $\mathbf{b}_n$ *of* $\boldsymbol{\beta}$ *has an asymptotically normal distribution in the sense that*

$$\left(\sum_1^n \mathbf{x}_i\mathbf{x}_i'\right)^{1/2}(\mathbf{b}_n - \boldsymbol{\beta}) \to_D \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I}_p), \tag{4.4}$$

where $\to_D$ *denotes convergence in distribution.*

Lai and Wei Stability

THEOREM 3. *Suppose that in the regression model* (1.1), $\{\varepsilon_n\}$ *is a martingale difference sequence with respect to an increasing sequence of* σ*-fields* $\{\mathcal{F}_n\}$ *such that* (1.6) *and* (4.1) *hold. Moreover, assume for each* n *that the design vector* $\mathbf{x}_n = (x_{n1}, \cdots, x_{np})'$ *at stage* n *is* $\mathcal{F}_{n-1}$*-measurable and that there exists a non-random positive definite symmetric matrix* $\mathbf{B}_n$ *for which*

$$\mathbf{B}_n^{-1}\left(\sum_1^n \mathbf{x}_i\mathbf{x}_i'\right)^{1/2} \to_p \mathbf{I}_p, \quad \text{and} \tag{4.2}$$

(4.3)

Then the least squares estimate $\mathbf{b}_n$ *of* β *has an asymptotically normal distribution in the sense that*

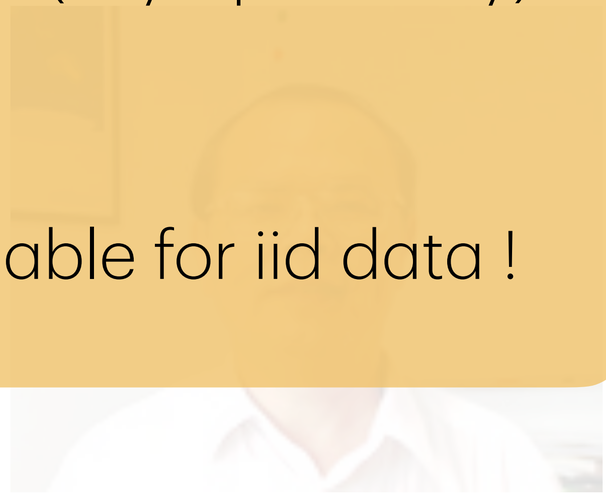
$$\left(\sum_1^n \mathbf{x}_i\mathbf{x}_i'\right)^{1/2}(\mathbf{b}_n - \beta) \to_D \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I}_p), \tag{4.4}$$

where $\to_D$ *denotes convergence in distribution.*

Under stability, non-iid data behaves like iid data (asymptotically) !

Under stability, we can use standard tools available for iid data !

(4.2) for bandit: $\xrightarrow{p} 1$ for all arms.



Stability for bandits

A bandit algorithm $\mathscr{A}$ is called weakly-stable if there exists non-random $n^{\star}_{a,T}(\mathscr{A})$

$$\frac{n_{a,T}(\mathscr{A})}{n^{\star}_{a,T}(\mathscr{A})} \xrightarrow{p} 1 \quad \text{for all } a \in [K]$$

For any stable bandit algorithm $\mathscr{A}$ we have

$$\frac{\sqrt{n_{a,T}}}{\hat{\sigma}} \left(\bar{\mu}_a - \mu_a\right) \xrightarrow{d} \mathcal{N}(0,1)$$

$$\bar{\mu}_a = \frac{\sum_{t=1}^{T} Y_t \cdot 1_{A_t=a}}{n_{a,T}}$$

$\hat{\sigma}^2$ is any consistent estimator of σ^2.

Stability for bandits

A bandit algorithm $\mathscr{A}$ is called weakly-stable if there exists non-random $n^{\star}_{a,T}(\mathscr{A})$

$$\frac{n_{a,T}(\mathscr{A})}{n^{\star}_{a,T}(\mathscr{A})} \xrightarrow{p} 1$$

For any stable bandit algorithm $\mathscr{A}$ we have

$$\frac{\sqrt{n_{a,T}}}{\hat{\sigma}}\left(\bar{\mu}_a - \mu_a\right) \xrightarrow{d} \mathcal{N}(0,1)$$

$$\left[\bar{\mu}_a - \frac{\hat{\sigma}\cdot z_{\alpha/2}}{\sqrt{n_{a,T}}},\quad \bar{\mu}_a + \frac{\hat{\sigma}\cdot z_{\alpha/2}}{\sqrt{n_{a,T}}}\right] \text{ is an asymptotically exact } 1-\alpha \text{ CI}.$$

Are popular bandit algorithms stable?

Upper Confidence Bound?

Thompson Sampling?

Thompson Sampling

- At round $t = 1, \ldots. T$
  - Sample $\theta_{a,t} \sim \mathcal{N}\left(\bar{\mu}_{a,t-1}, \frac{1}{1 + n_{a,t-1}}\right)$
  - Pick $A_t \sim \arg\max_a \theta_{a,t}$

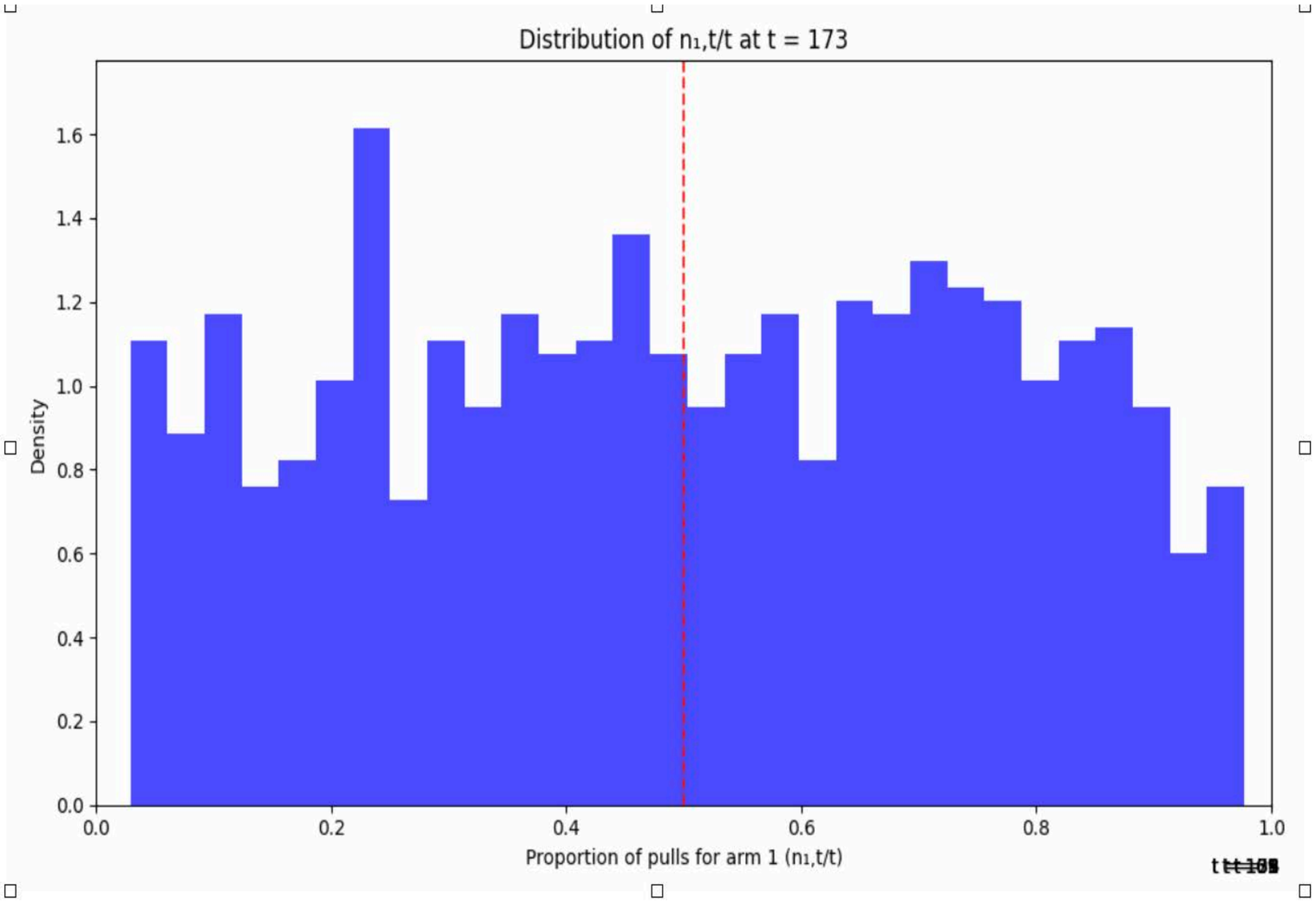
$n_{a,t-1}$ = # armpulls for arm a at $t - 1$

$\bar{\mu}_{a,t-1}$ = Sample mean of arm a at $t - 1$

Prior for both arms are Normal(0,1) .

Equal arm means $\mu_1 = \mu_2 = 1.0$

Thompson Sampling is not-stable



Kalvit and Zeevi 2021

Upper Confidence Bound

Algorithm 1 UCB algorithm

1: Pull once each of the K arms in the first K iterations.
2: **for** $t = K+1, \ldots, T-1$ **do**
3: Compute the UCB boundary

$$\mathrm{UCB}(a,t) := \bar{\mu}_{a,t} + \sqrt{\frac{2\log T}{n_{a,t}}}$$

4: Choose arm A_t given by

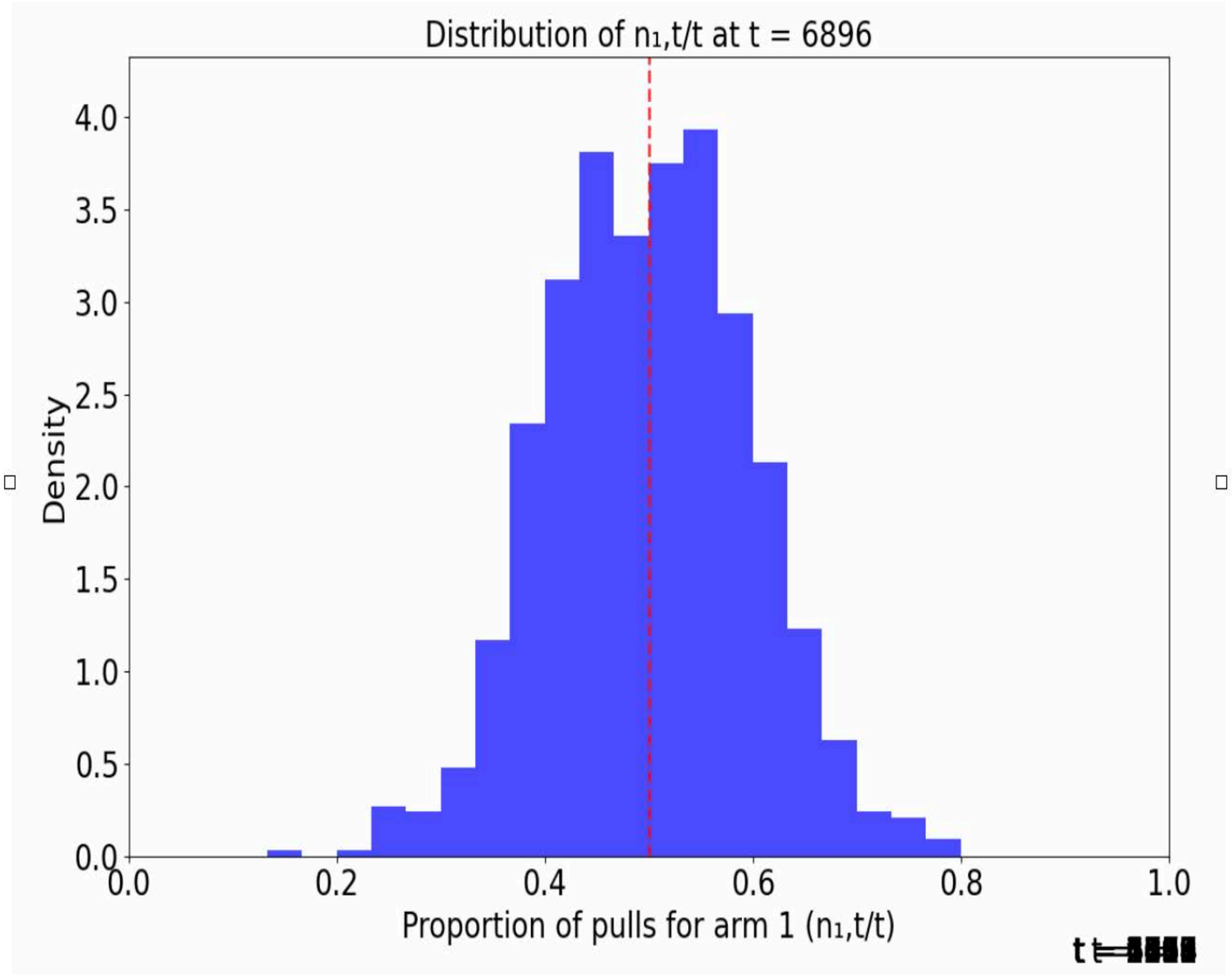
$$A_{t+1} = \arg\max_a \mathrm{UCB}(a,t)$$

5: **end for**

Two arm UCB with Gaussian rewards

Equal arm means $\mu_1 = \mu_2 = 1.0$

Upper confidence bound



Fan, Glynn' 22

UCB is stable

Theorem [KZ' 24, QKZ'24]

The UCB algorithm is stable:

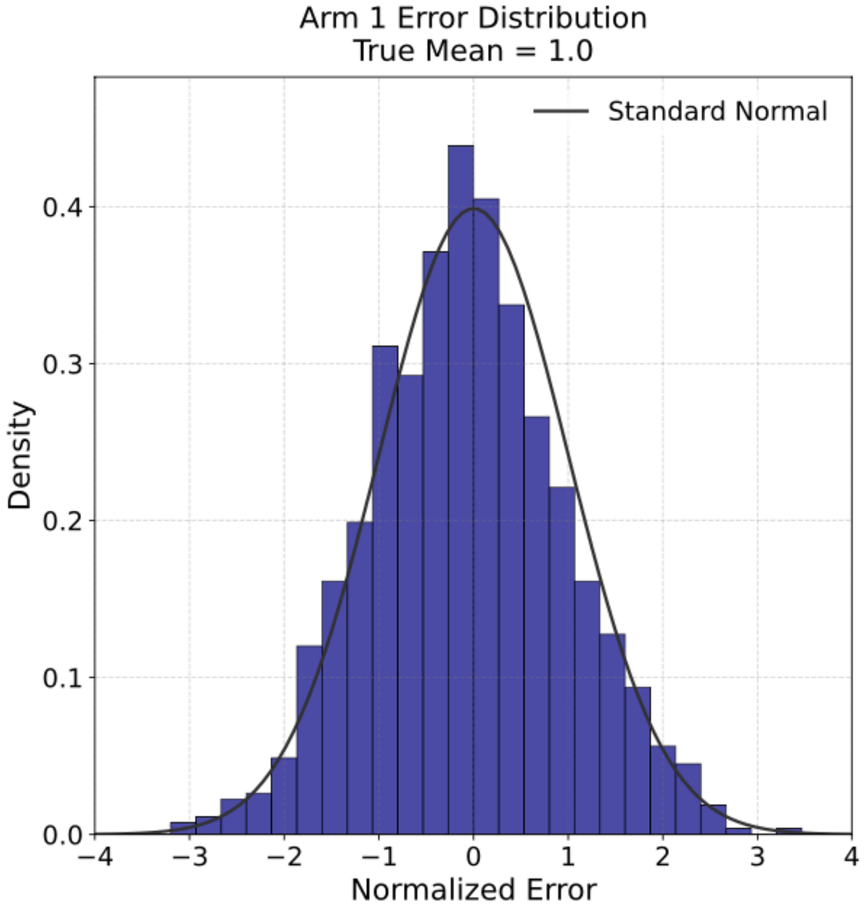
$$\frac{n_{a,T}}{\left(1/\sqrt{n^\star} + \sqrt{\Delta_a^2/2\log T}\right)^{-2}} \xrightarrow{p} 1$$

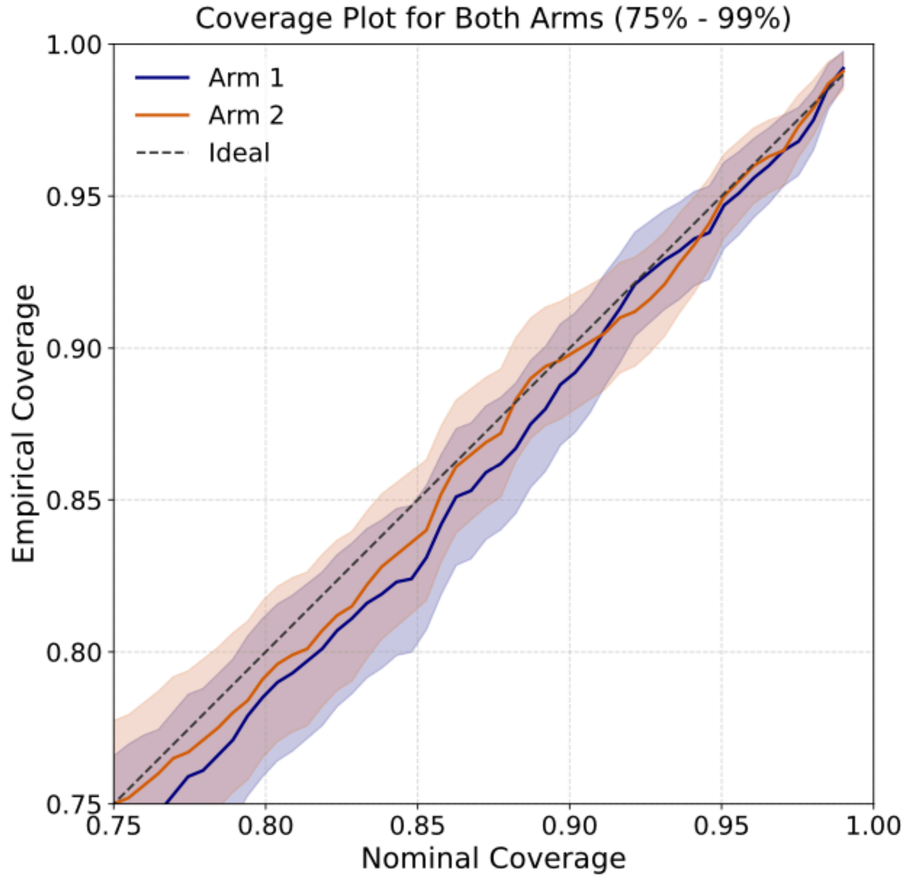
$\Delta_a = \mu^\star - \mu_a$, and $n^\star$ is unique solution to

$$\sum_a \frac{1}{\left(\sqrt{T/n^\star} + \sqrt{T\Delta_a^2/2\log T}\right)^{2}} = 1$$

Consequently, $$\frac{\sqrt{n_{a,T}}}{\hat{\sigma}}\left(\bar{\mu}_a - \mu_a\right) \xrightarrow{d} \mathcal{N}(0,1)$$

Simulation study





Precise regret bound

$$\mathbb{E}\left|\frac{n_{a,T}}{\left(1/\sqrt{n^\star}+\sqrt{\Delta_a^2/2\log T}\right)^{-2}}-1\right| \longrightarrow 0$$

$$\mathbb{E}n_{a,T} \approx \left(1/\sqrt{n^\star}+\sqrt{\Delta_a^2/2\log T}\right)^{-2}$$

$$\text{Regret} = \sum_{a=1}^{k}\Delta_a\mathbb{E}n_{a,T} \approx \sum_{a=1}^{K}\Delta_a\cdot\left(1/\sqrt{n^\star}+\sqrt{\Delta_a^2/2\log T}\right)^{-2}$$

A precise regret guarantee

Theorem [QKZ'24]

The regret of the UCB algorithm satisfies

$$\left| \frac{\text{Regret}}{\sum_{a=1}^{K} \Delta_a \cdot \left(1/\sqrt{n^\star} + \sqrt{\Delta_a^2 / 2 \log T} \right)^{-2}} - 1 \right| \lesssim \sqrt{\frac{2 \log \log T}{2 \log T}}$$

Worst case regret $\gtrsim \sqrt{KT \log T}$ (Not minimax optimal)

Rgeret $\to \sum_{\mu_a \neq \mu^\star} \frac{2 \log T}{\Delta_a}$ when $\Delta_a \gg \sqrt{\frac{K \log T}{T}}$.

Stability in Contextual bandit

Theorem [QKZ'24]

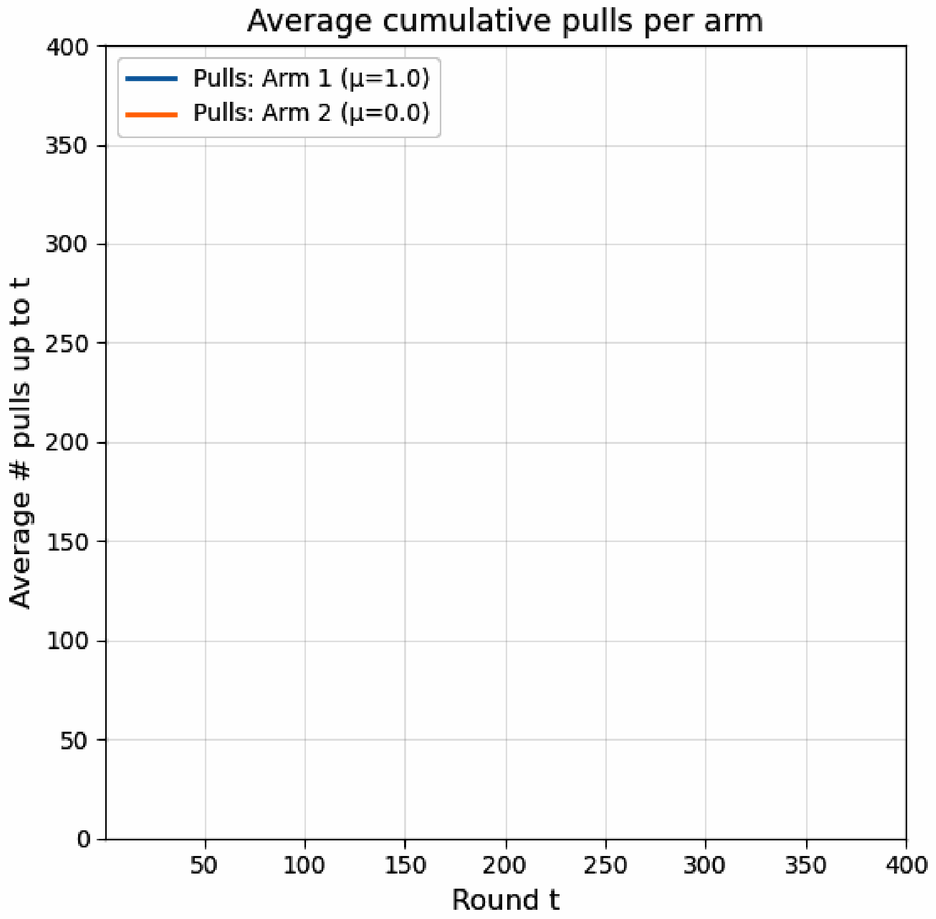
For a contextual bandit problem with $|\mathcal{X}| = K$ contexts, the UCB algorithm is stable, and consequently

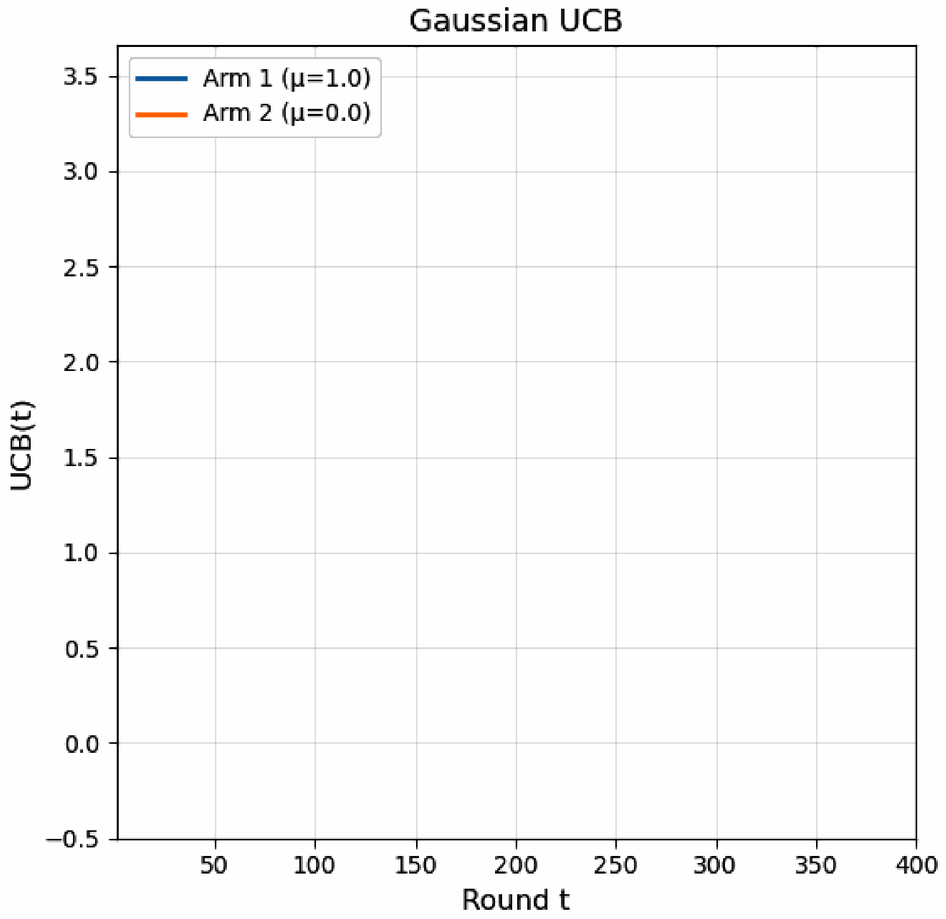
$$\left(\sum_{i=1}^{n} x_i x_i^\top\right)^{-1/2} \left(\bar{\mu}_{LS} - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 \mathbf{I}_d\right)$$

Rate of convergence of CLT can be derived.

A pictorial proof

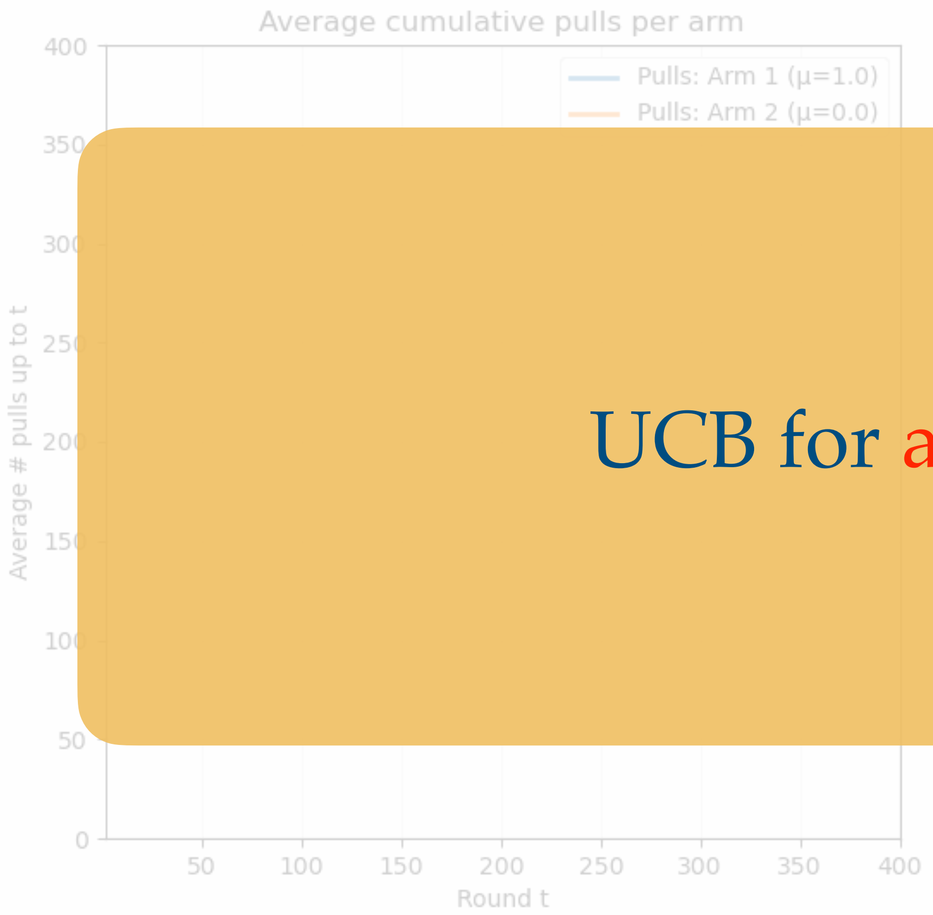
Stability of UCB





$$\mathrm{UCB}(a,t) = \bar{\mu}_{a,t-1} + \sqrt{\frac{2\log T}{n_{a,t-1}}}$$

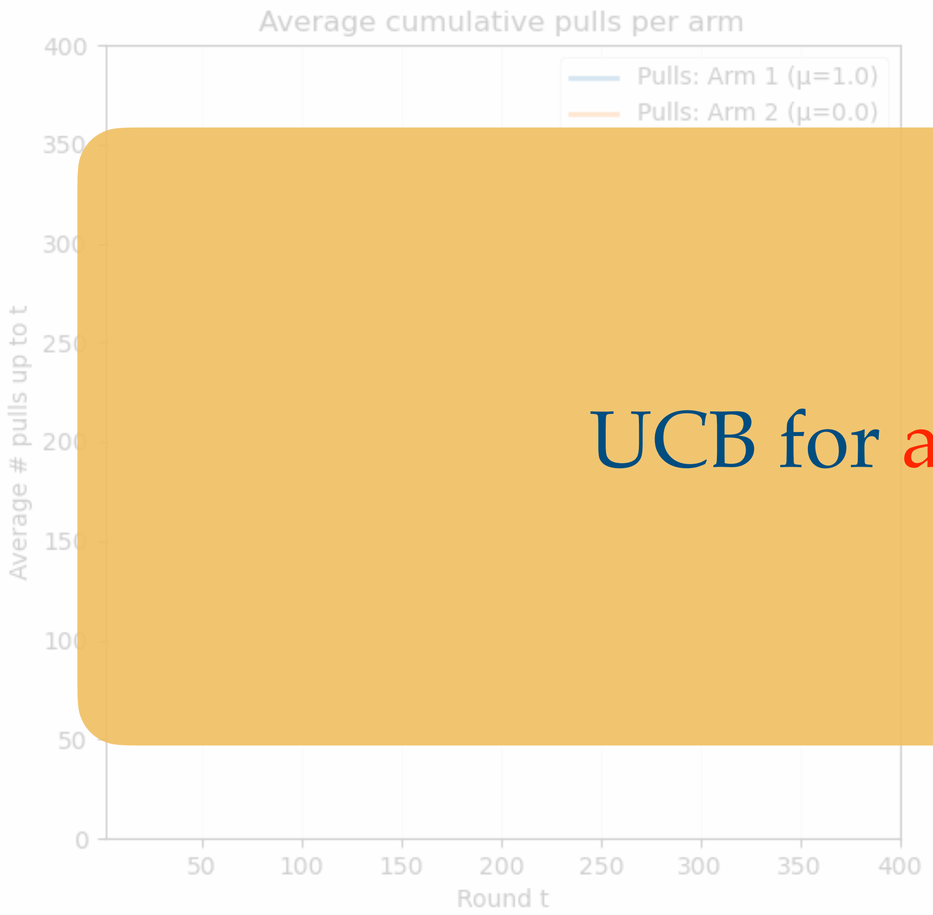
Stability of UCB



UCB for all arm have same limit

$$\mathrm{UCB}(a,t) = \bar{\mu}_{a,t-1} + \sqrt{\frac{2\log T}{n_{a,t-1}}}$$

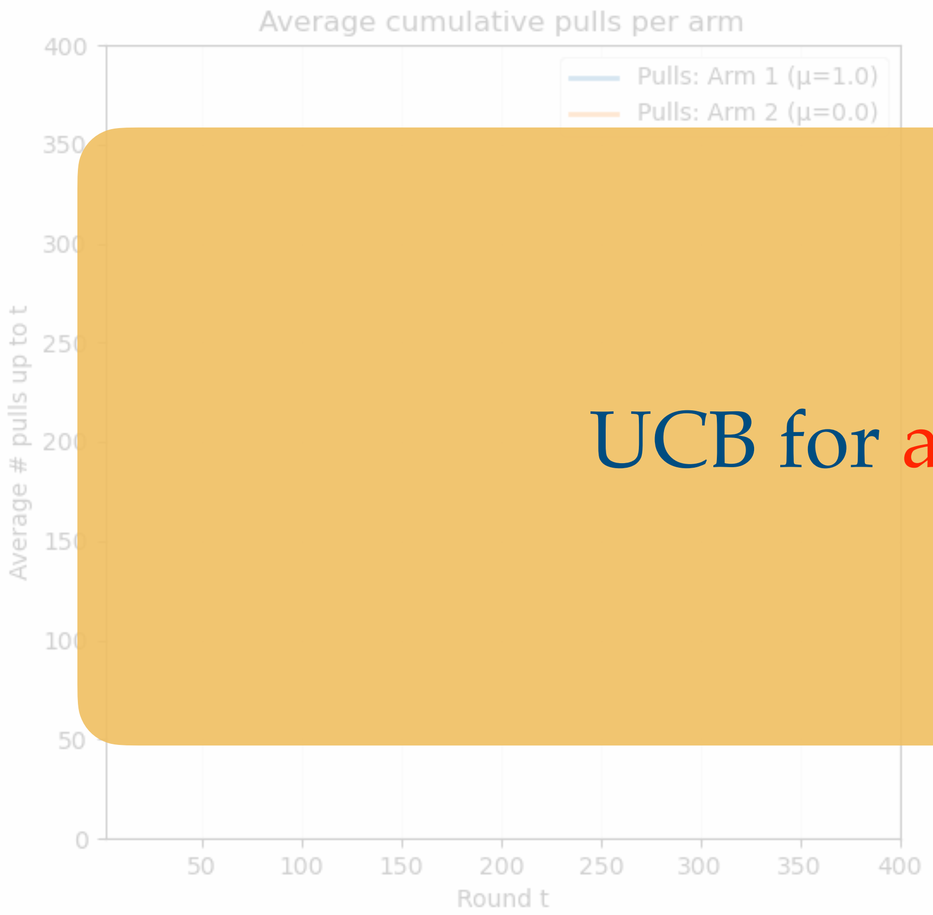
Stability of UCB



UCB for all arm have same limit

$$\mathrm{UCB}(a,t) = \bar{\mu}_{a,t-1} + \sqrt{\frac{2\log T}{n_{a,t-1}}}$$

Stability of UCB



UCB for all arm have same limit

$$\mathrm{UCB}(a,t) = \bar{\mu}_{a,t-1} + \sqrt{\frac{2\log T}{n_{a,t-1}}}$$

All UCB's are equal

$$\mathrm{UCB}(1,T) \approx \mathrm{UCB}(2,T) \approx \ldots\ldots\ldots \approx \mathrm{UCB}(k,T)$$

$$\bar{\mu}_1 + \sqrt{\frac{2\log T}{n_{1,T}}} \approx \bar{\mu}_2 + \sqrt{\frac{2\log T}{n_{2,T}}} \approx \ldots\ldots\ldots \approx \bar{\mu}_k + \sqrt{\frac{2\log T}{n_{k,T}}}$$

$$n_{1,T} + n_{2,T} + \ldots\ldots + n_{k,T} = T$$

Want to understand $\{n_{a,T}\}_{a\in[K]}$

UCB to noiseless-UCB

$$\text{UCB}(1,T) \approx \text{UCB}(2,T) \approx \ldots\ldots\ldots \approx \text{UCB}(k,T)$$

$$\mu_1 \pm \sqrt{\frac{2\log\log T}{n_{1,T}}} + \sqrt{\frac{2\log T}{n_{1,T}}} \approx \ldots\ldots\ldots \approx \mu_k \pm \sqrt{\frac{2\log\log T}{n_{k,T}}} + \sqrt{\frac{2\log T}{n_{k,T}}}$$

$$n_{1,T} + n_{2,T} + \ldots\ldots + n_{k,T} = T$$

Law of Iterated Logarithm

UCB to noiseless-UCB

$$\mathrm{UCB}(1,T) \approx \mathrm{UCB}(2,T) \approx \ldots\ldots\ldots \approx \mathrm{UCB}(k,T)$$

$$\mu_1 + \sqrt{\frac{2\log T}{n_{1,T}}} \approx \ldots\ldots\ldots \approx \mu_k + \sqrt{\frac{2\log T}{n_{k,T}}}$$

$$n_{1,T} + n_{2,T} + \ldots\ldots + n_{k,T} = T$$

K variables $\{n_{a,T}\}_{a\in[K]}$, K equations

UCB is stable

Theorem [KZ' 24, QKZ'24]

The UCB algorithm is stable:

$$\frac{n_{a,T}}{\left(1/\sqrt{n^\star} + \sqrt{\Delta_a^2/2\log T}\right)^{-2}} \xrightarrow{p} 1$$

$\Delta_a = \mu^\star - \mu_a$, and $n^\star$ is unique solution to

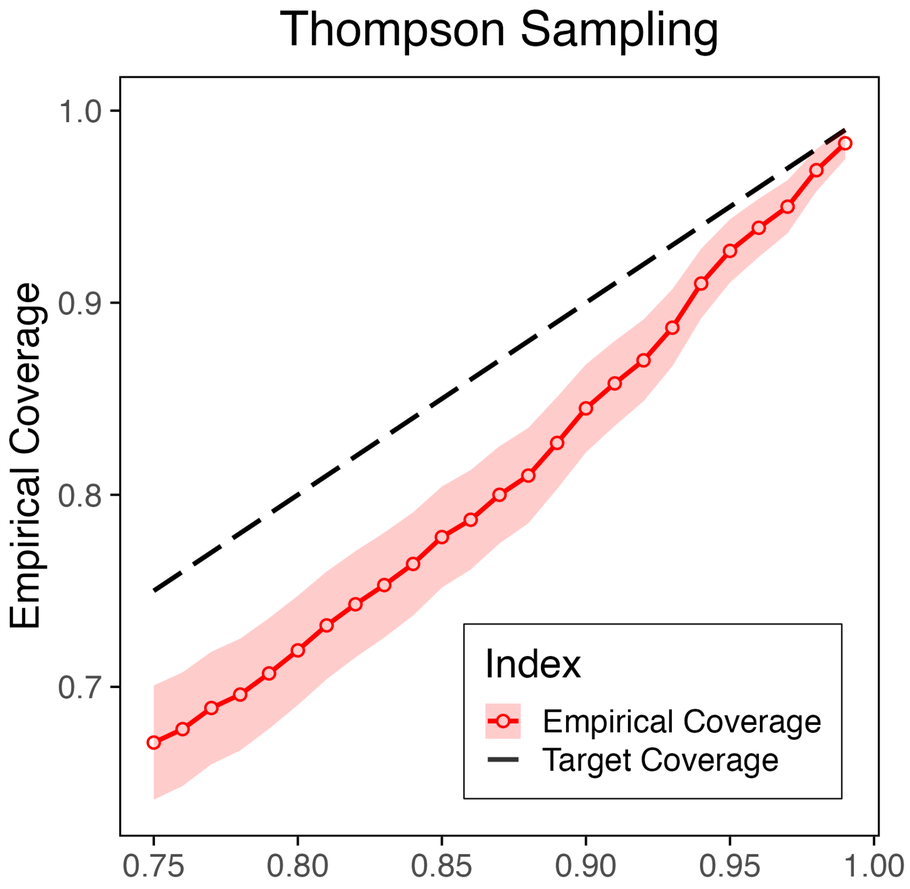
$$\sum_a \frac{1}{\left(\sqrt{T/n^\star} + \sqrt{T\Delta_a^2/2\log T}\right)^2} = 1$$

Consequently, $$\frac{\sqrt{n_{a,T}}}{\hat{\sigma}}\left(\bar{\mu}_a - \mu_a\right) \xrightarrow{d} \mathcal{N}(0,1)$$

How about Thompson Sampling?

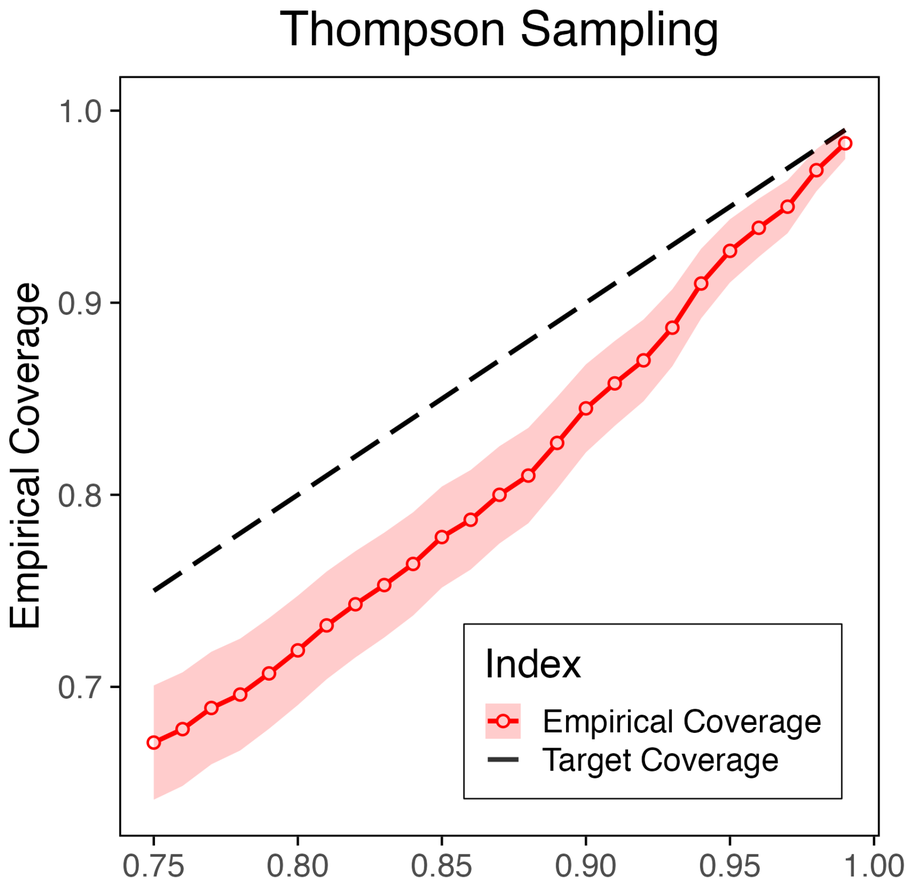
Can we stabilize it?

Thompson Sampling



- Sample $\theta_{a,t} \sim \mathcal{N}\left(\bar{\mu}_{a,t-1}, \frac{1}{1+n_{a,t-1}}\right)$
- Pick $A_t = \arg\max_a \theta_{a,t}$

Thompson Sampling

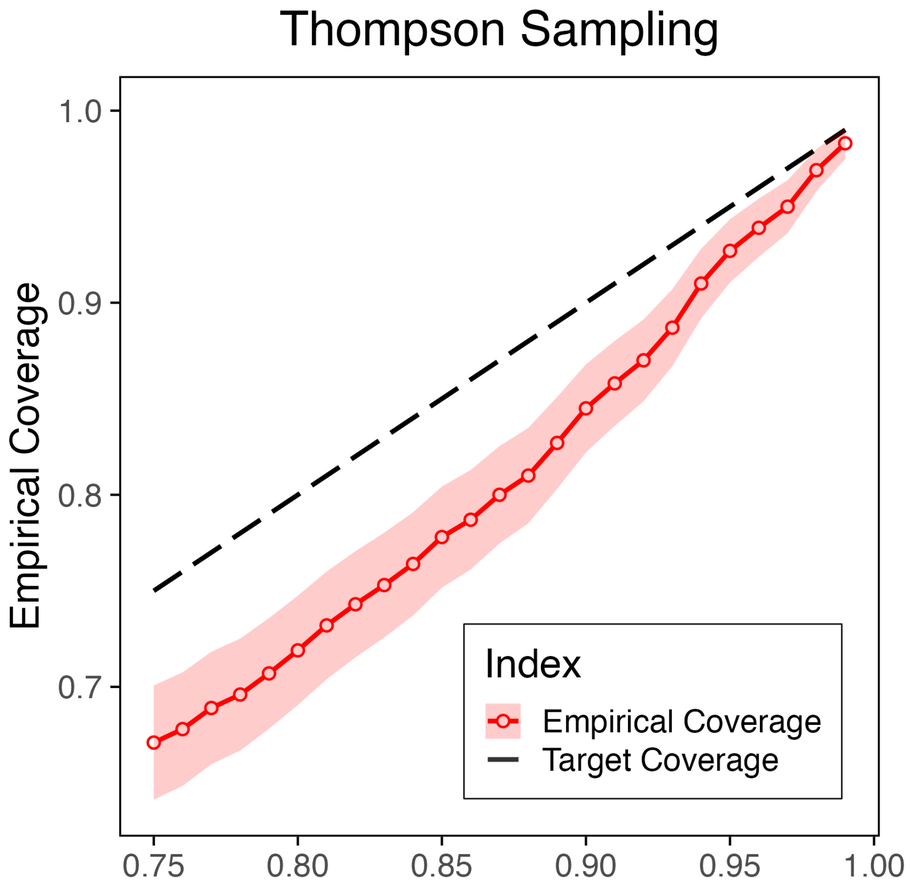


- Sample $\theta_{a,t} \sim \mathcal{N}\left(\bar{\mu}_{a,t-1}, \frac{1}{1+n_{a,t-1}}\right)$
- Pick $A_t = \arg\max_a \theta_{a,t}$

- $\theta_{a,t} = \bar{\mu}_{a,t-1} + \frac{Z_t}{\sqrt{1+n_{a,t-1}}}$
- Pick $A_t = \arg\max_a \theta_{a,t}$

Z_t is Gaussian

Thompson Sampling

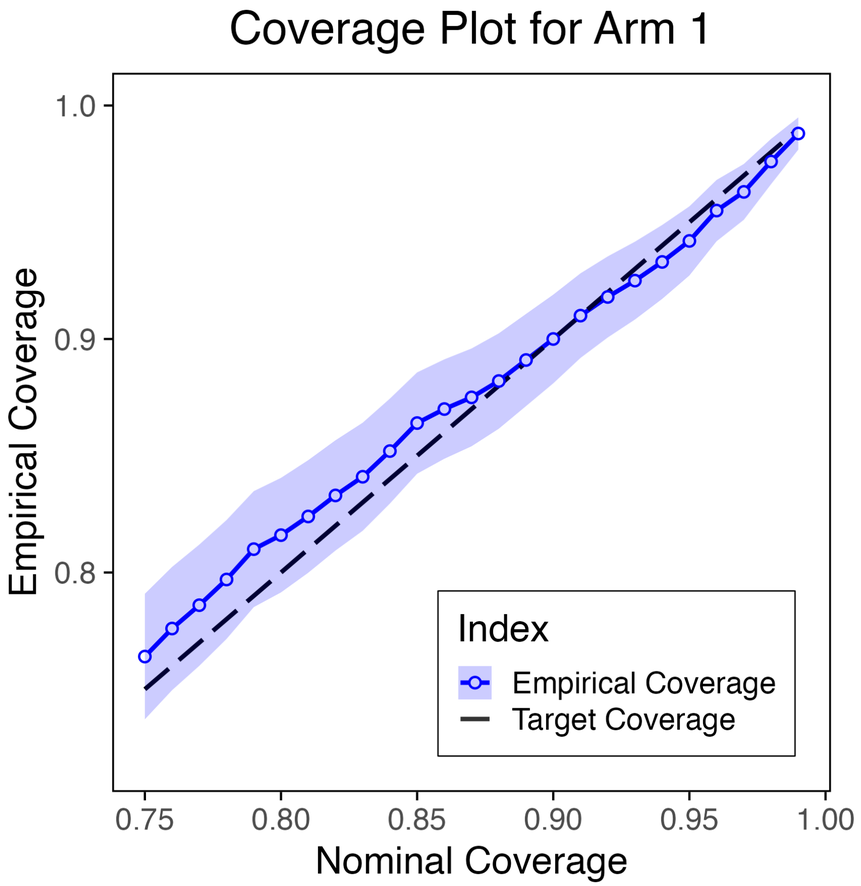


- Sample $\theta_{a,t} \sim \mathcal{N}\left(\bar{\mu}_{a,t-1}, \frac{1}{1+n_{a,t-1}}\right)$
- Pick $A_t = \arg\max_a \theta_{a,t}$

- $\theta_{a,t} = \bar{\mu}_{a,t-1} + \frac{\gamma_T Z_t}{\sqrt{1+n_{a,t-1}}}$
- Pick $A_t \sim \arg\max_a \theta_{a,t}$

$\sqrt{2\log\log T} \ll \gamma_T$

Stable Thompson Sampling [BSK'25]



- Sample $\theta_{a,t} \sim \mathcal{N}\left(\bar{\mu}_{a,t-1}, \dfrac{\gamma_T}{1+n_{a,t-1}}\right)$
- Pick $A_t \sim \arg\max_a \theta_{a,t}$

$K = 2$

$$\sqrt{2\log\log T} \ll \gamma_T \ll \sqrt{\log T}$$

Summary:

- Under stability, non-iid data behaves like iid data.
- Thompson Sampling is not stable, we can stabilize it.
- UCB is stable, a new proof technique.

References:

- Khamaru, Koulik, and Cun-Hui Zhang. "Inference with the Upper Confidence Bound Algorithm." *arXiv preprint arXiv:2408.04595*
- Qiyang Han, Khamaru, Koulik, and Cun-Hui Zhang. “UCB algorithms for multiarmed bandits: Precise regret and adaptive inference” *arXiv preprint arXiv:2412.06126* .
- Budhaditya Halder, Subhayan Pan, Koulik Khamatu. “Stable Thomopson Sampling: Valid Inference via Variance Inflation” *arXiv preprint arXiv:2505.23260* .

Stability in Contextual bandit

Theorem [QKZ'24]

For a contextual bandit problem with $|\mathcal{X}| = K$ contexts, the UCB algorithm is stable, and consequently

$$\left(\sum_{i=1}^{n} x_i x_i^\top\right)^{-1/2} \left(\bar{\mu}_{LS} - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 \mathbf{I}_d\right)$$

Rate of convergence of CLT can be derived.

Rate of convergence of CLT: A Berry Essen bound

Theorem [QKZ'24]

The UCB algorithm satisfies

$$\sup_{A,a} \left| \mathbb{P}(Z_a \in A) - P(\mathcal{N}(0,1) \in A) \right| \lesssim \left(\frac{2\log\log T}{2\log T} \right)^{\frac{1}{6}}$$

Here, $Z_a = \dfrac{\sqrt{n_{a,T}}}{\sigma} \left(\bar{\mu}_a - \mu_a \right)$

Rate of convergence of CLT: A Berry Essen bound

The UCB algorithm satisfies

$$\sup_{\alpha\in(0,1),a\in[K]} \left|\mathbb{P}(\mu_a \in \mathrm{CI}_{a,\alpha}) - (1-\alpha)\right| \lesssim \left(\frac{2\log\log T}{2\log T}\right)^{\frac{1}{6}}$$

Here, $\mathrm{CI}_{a,\alpha} = \left[\bar{\mu}_a \mp \frac{\hat{\sigma}\cdot z_{\alpha/2}}{\sqrt{n_{a,T}}}\right]$ and $\mathbb{P}(N(0,1) \geq z_{\alpha/2}) = \alpha/2$