Columbia Business School
AT THE VERY CENTER OF BUSINESS

What Does Thompson Sampling Optimize?

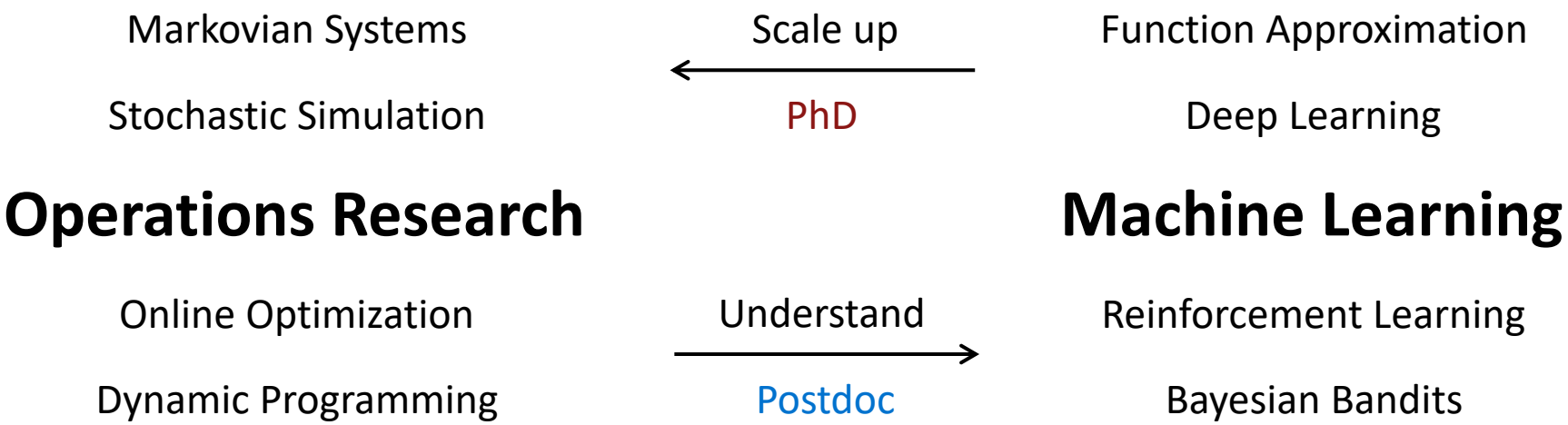
A Broader View of Thompson Sampling

Yanlin Qu

Postdoc, Columbia DRO | PhD, Stanford MS&E

Joint work with Hongseok Namkoong and Assaf Zeevi

Columbia Business School


Markovian Systems
Stochastic Simulation
Operations Research
Online Optimization
Dynamic Programming
Scale up
PhD
Understand
Postdoc
Function Approximation
Deep Learning
Machine Learning
Reinforcement Learning
Bayesian Bandits

Background

Multi-armed bandit

Multi-armed Bandit (MAB)

- Each time we pull an arm, we get a random reward, but we don't know in advance which arm is better.
- Shall we explore an unfamiliar option to learn more, or exploit the option that seems best so far?
- Balancing **exploration** and **exploitation** is the key.
- Clinical trials, recommender systems, etc.

A two-armed bandit
(credit: GPT-5 and me)

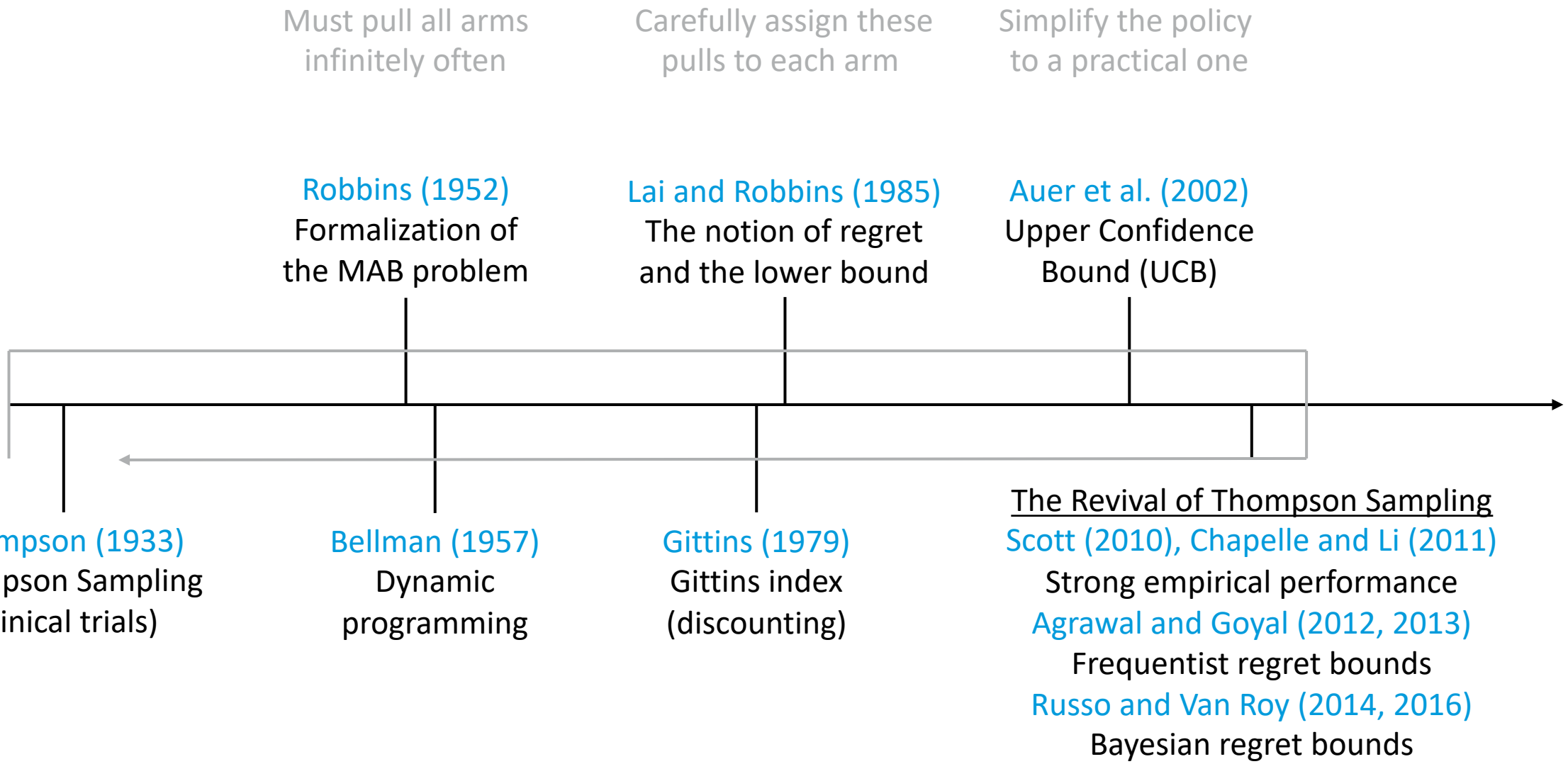
Bayesian Setting

Notations

- Environment paramter: θ
- Prior distribution (of θ): π_0
- Reward distribution (given θ): P_θ
- Reward vector: $R_0|\theta, R_1|\theta, ... \overset{\text{iid}}{\sim} P_\theta$
- Arm pulled: $A_0, A_1, ...$

Mechanism

- At the start, θ is sampled from π_0 and remains fixed thereafter.
- At the t-th round, R_t is independently sampled from P_θ, but only its A_t-th entry is observed.
- After t observations, prior π_0 becomes posterior π_t (Bayes' rule).


Must pull all arms infinitely often
Carefully assign these pulls to each arm
Simplify the policy to a practical one
Robbins (1952)
Formalization of the MAB problem
Lai and Robbins (1985)
The notion of regret and the lower bound
Auer et al. (2002)
Upper Confidence Bound (UCB)
Thompson (1933)
Thompson Sampling (clinical trials)
Bellman (1957)
Dynamic programming
Gittins (1979)
Gittins index (discounting)
The Revival of Thompson Sampling
Scott (2010), Chapelle and Li (2011)
Strong empirical performance
Agrawal and Goyal (2012, 2013)
Frequentist regret bounds
Russo and Van Roy (2014, 2016)
Bayesian regret bounds

ChatGPT 4.5 >

Which is better TS or UCB? Answer without explanation.

TS.

Do you know what objective UCB optimizes? Yes or no.

Yes.

Do you know what objective TS optimizes? Yes or no.

No.

Background

Thompson Sampling

Thompson Sampling (Thompson, 1933)

- Sample an environment (mean reward vector) from the posterior distribution

$$\theta' \sim \pi_t$$

- Select the arm that is perceived to be optimal in the sampled environment

$$A_t = \text{argmax}(\theta'_1, \theta'_2)$$

- Observe reward and update belief

$$\pi_t \xrightarrow{R_{A_t,t}} \pi_{t+1}$$

Bernoulli & Gaussian

Beta-Bernoulli

- Sample

  $(\theta_1', \theta_2') \sim \text{Beta}(\alpha_1, \beta_1) \times \text{Beta}(\alpha_2, \beta_2)$

- Select $A = \text{argmax}(\theta_1', \theta_2')$
- Observe Bernoulli R and update
  $\alpha_A \leftarrow \alpha_A + R$
  $\beta_A \leftarrow \beta_A + (1 - R)$

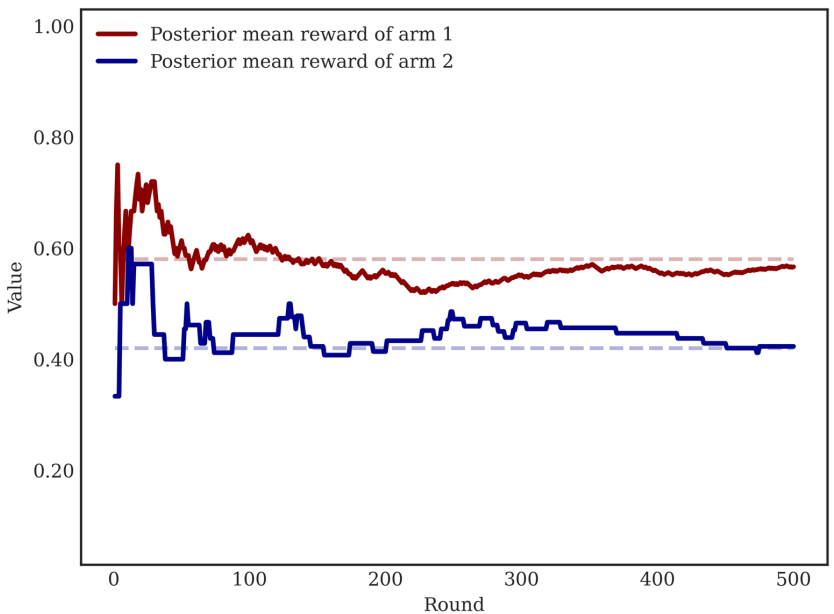
Gaussian-Gaussian

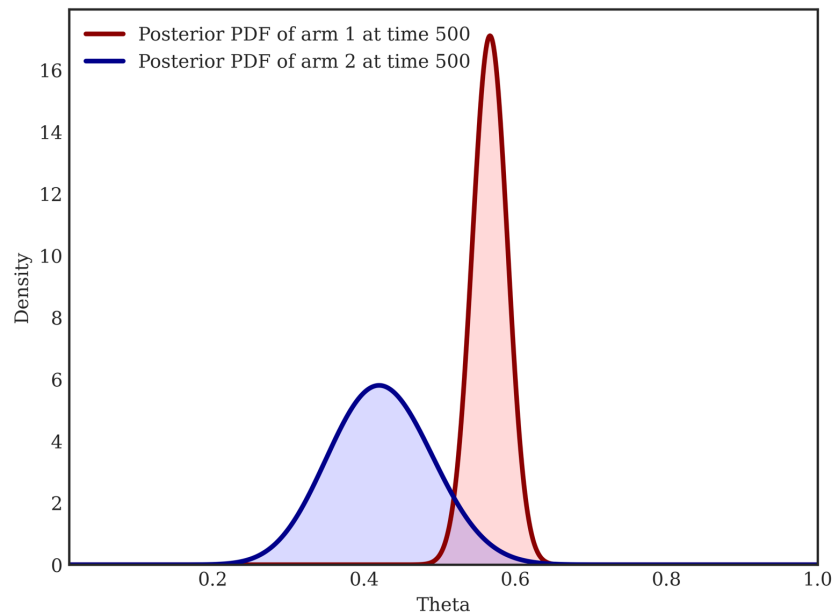
- Sample

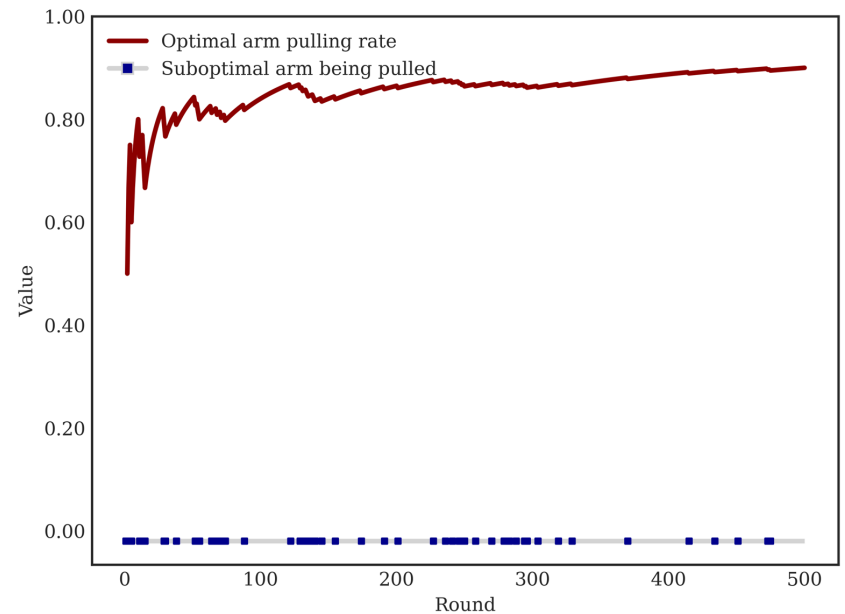
  $(\theta_1', \theta_2') \sim N(\mu_1, \sigma_1^2) \times N(\mu_2, \sigma_2^2)$

- Select $A = \text{argmax}(\theta_1', \theta_2')$
- Observe Gaussian R and update
  $\mu_A \leftarrow (\mu_A/\sigma_A^2 + R/\tau^2)/(1/\sigma_A^2 + 1/\tau^2)$
  $\sigma_A^2 \leftarrow 1/(1/\sigma_A^2 + 1/\tau^2)$

Thompson Sampling







Ber(0.58) vs. Ber(0.42)

Background

MDP Formulation

MDP Formulation (Gittins, 1979)

At the t-th round

- State: current belief π_t
- Action: arm pulled A_t
- Transition: updating π_t to π_{t+1} after observing the A_t-th entry of $R_t \sim P_{\theta'}$ where $\theta' \sim \pi_t$
- Reward: expected reward $\mathbb{E}_{\pi_t} R_{A_t,t}$

$\Leftrightarrow$

Mechanism

- At the start, θ is sampled from π_0 and remains fixed thereafter.
- At the t-th round, R_t is independently sampled from P_θ, but only its A_t-th entry is observed.
- After t observations, prior π_0 becomes posterior π_t (Bayes' rule).

Thompson vs. Gittins

Thompson Sampling

- Sample $\theta' \sim \pi_t$
- Select $A_t = \text{argmax}(\theta'_1, \theta'_2)$

It somehow achieves low regret

$$\mathbb{E}_{\pi_0}\left[\sum_{t=0}^{T-1}(\max(\theta_1, \theta_2) - \theta_{A_t})\right] = O(\sqrt{T})$$

across horizons (Russo and Van Roy 2016).

The Gittins index policy

- Compute Gittins indices $G(\pi_{k,t})$
- Select $A_t = \text{argmax}(G(\pi_{1,t}), G(\pi_{2,t}))$

It solves a Bellman equation to maximize

$$\mathbb{E}_{\pi_0}\left[\sum_{t=0}^{\infty}\gamma^t\theta_{A_t}\right]$$

but suffers linear regret (Rothschild 1974).

Thompson vs. Gittins

?

It somehow achieves low regret

$$\mathbb{E}_{\pi_0}\left[\sum_{t=0}^{T-1}(\max(\theta_1,\theta_2)-\theta_{A_t})\right]=O(\sqrt{T})$$

across horizons (Russo and Van Roy 2016).

$$V(\pi_t)=\max_{q_t}[q_t\cdot\mathbb{E}_{\pi_t}\theta+\gamma\mathbb{E}_{\pi_t,q_t}V(\pi_{t+1})]$$

$$\mathbb{E}_{\pi_t,q_t}V(\pi_{t+1})=q_t\cdot\mathbb{E}_{\pi_t}[V(\pi_{t+1})|A_t=\cdot]$$

It solves a Bellman equation to maximize

$$\mathbb{E}_{\pi_0}\left[\sum_{t=0}^{\infty}\gamma^t\theta_{A_t}\right]$$

but suffers linear regret (Rothschild 1974).

An Idea

Faithful Stationarization (Qu, Namkoong, Zeevi, 2025)

- Cumulative regret (of policy $Q : A_t | \pi_t \sim q_t$)

$$\begin{aligned}\mathcal{R}_T(Q;\pi_0) =& \mathbb{E}_{\pi_0}\left[\sum_{t=0}^{T-1} \mathbb{E}_{\pi_0}\left[\max(\theta_1,\theta_2) - \theta_{A_t}\middle|\pi_t\right]\right] \\ =& \mathbb{E}_{\pi_0}\left[\sum_{t=0}^{T-1} r(q_t;\pi_t)\right]\end{aligned}$$

- (Cumulative) squared regret

$$\mathcal{R}^2(Q;\pi_0) = \mathbb{E}_{\pi_0}\left[\sum_{t=0}^{\infty} r^2(q_t;\pi_t)\right]$$

Faithful Stationarization (Qu, Namkoong, Zeevi, 2025)

$$
\begin{aligned}
\mathcal{R}_T(Q;\pi_0) \leq & \mathbb{E}_{\pi_0}\left[\left(\sum_{t=0}^{T-1} 1\right)^{1/2}\left(\sum_{t=0}^{T-1} r^2(q_t;\pi_t)\right)^{1/2}\right] \\
& \leq \sqrt{T}\cdot\left(\mathbb{E}_{\pi_0}\left[\sum_{t=0}^{T-1} r^2(q_t;\pi_t)\right]\right)^{1/2} \\
& \leq \sqrt{\mathcal{R}^2(Q;\pi_0)\cdot T}
\end{aligned}
$$

$$
\boxed{\mathcal{R}^2(Q^{\mathsf{TS}};\pi_0) < \infty}
$$

Faithful Stationarization (Qu, Namkoong, Zeevi, 2025)

$$V(\pi_t) = \min_{q_t}[r^2(q_t;\pi_t) + \mathbb{E}_{\pi_t,q_t} V(\pi_{t+1})]$$

- Minimizing

$$\mathbb{E}_{\pi_0}\left[\sum_{t=0}^{\infty}(\mathbb{E}_{\pi_t}[\max(\theta_1,\theta_2) - \theta_{A_t}])^2\right]$$

- Intrinsic regret decay
- Not indexable: $q_{1,t} \in [0,1]$
- "Faithful": $\mathcal{R}_T(Q^{\mathsf{R2}};\pi_0) \le \sqrt{V(\pi_0)\cdot T}$

$$V(\pi_t) = \max_{q_t}[q_t \cdot \mathbb{E}_{\pi_t}\theta + \gamma\mathbb{E}_{\pi_t,q_t} V(\pi_{t+1})]$$

- Maximizing

$$\mathbb{E}_{\pi_0}\left[\sum_{t=0}^{\infty}\gamma^t\theta_{A_t}\right]$$

- Extrinsic discounting
- Indexable: $q_{1,t} \in \{0,1\}$
- Not "faithful": $\mathcal{R}_T(Q^{\mathsf{GI}};\pi_0) = \Theta(T)$

A Change of Variables

Online Optimization Form

$$q_t^{\mathrm{R2}} = \underset{q_t}{\operatorname{argmin}}[r^2(q_t;\pi_t) + \mathbb{E}_{\pi_t,q_t} V(\pi_{t+1})]$$

$$x_t = q_t \cdot \mathbb{E}_{\pi_t}\theta$$

$$x_t^{\mathrm{R2}} = \underset{x_t}{\operatorname{argmin}}[(\mathbb{E}_{\pi_t} \max(\theta_1,\theta_2) - x_t)^2 + \nu^{\mathrm{R2}}(\pi_t)x_t]$$

$$\nu^{\mathrm{R2}}(\pi_t) = \frac{\mathbb{E}_{\pi_t}[V(\pi_{t+1})|A_t=1] - \mathbb{E}_{\pi_t}[V(\pi_{t+1})|A_t=2]}{\mathbb{E}_{\pi_t}\theta_1 - \mathbb{E}_{\pi_t}\theta_2}$$

Thompson Sampling

We focus on the two-armed case $q_{1,t}^{\mathsf{TS}} = P_{\pi_t}(\theta_1 > \theta_2)$ as the K-armed case can be viewed as repeating the two-armed case K times (to determine K probabilities)

$$q_{1,t}^{\mathsf{TS}} = P_{\pi_t}(\theta_1 > \theta_2, ..., \theta_K) = P_{\pi_t}(\theta_1 > \theta_{-1}),$$

where $\theta_{-1} = \max\{\theta_2, ..., \theta_K\}$ can be viewed as a single competing arm against θ_1.

Theorem. *The online optimization form of Thompson Sampling is*

$$x_t^{\mathsf{TS}} = \underset{x_t}{\operatorname{argmin}}[(\mathbb{E}_{\pi_t} \max(\theta_1, \theta_2) - x_t)^2 + \nu^{\mathsf{TS}}(\pi_t)x_t],$$

where $\nu^{\mathsf{TS}}(\pi_t) = \mathrm{Cov}_{\pi_t}(\theta_1 - \theta_2, \mathrm{sign}(\theta_1 - \theta_2))$.

Thompson Sampling

The regularizer is the covariance between the following two fundamental quantities

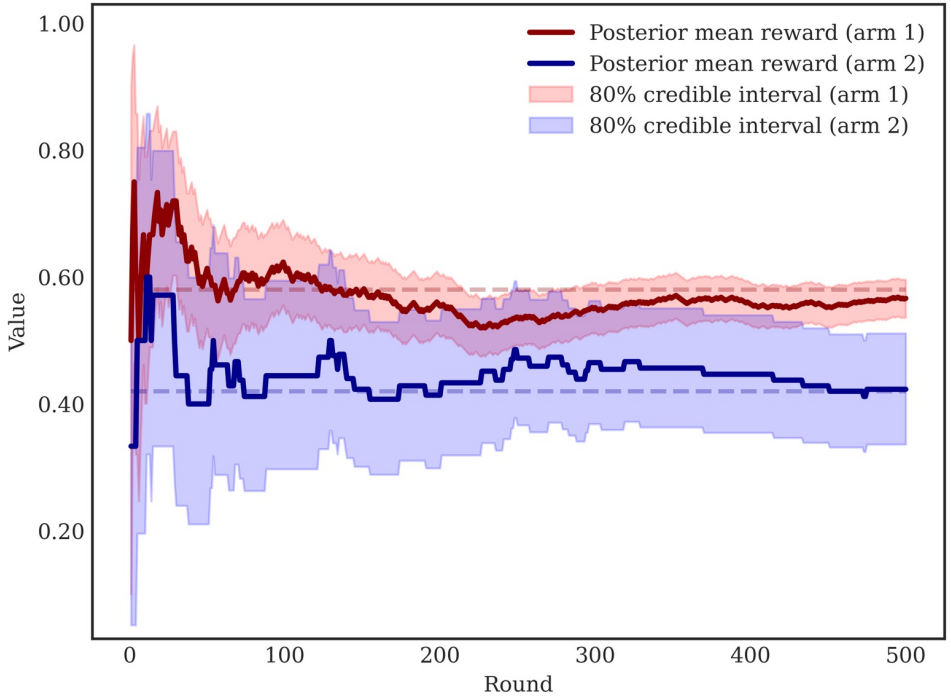
$$\Delta = \theta_1 - \theta_2 \text{ the reward gap between the two arms}$$
$$\Lambda = \text{sign}(\theta_1 - \theta_2) \text{ the identity of the optimal arm.}$$

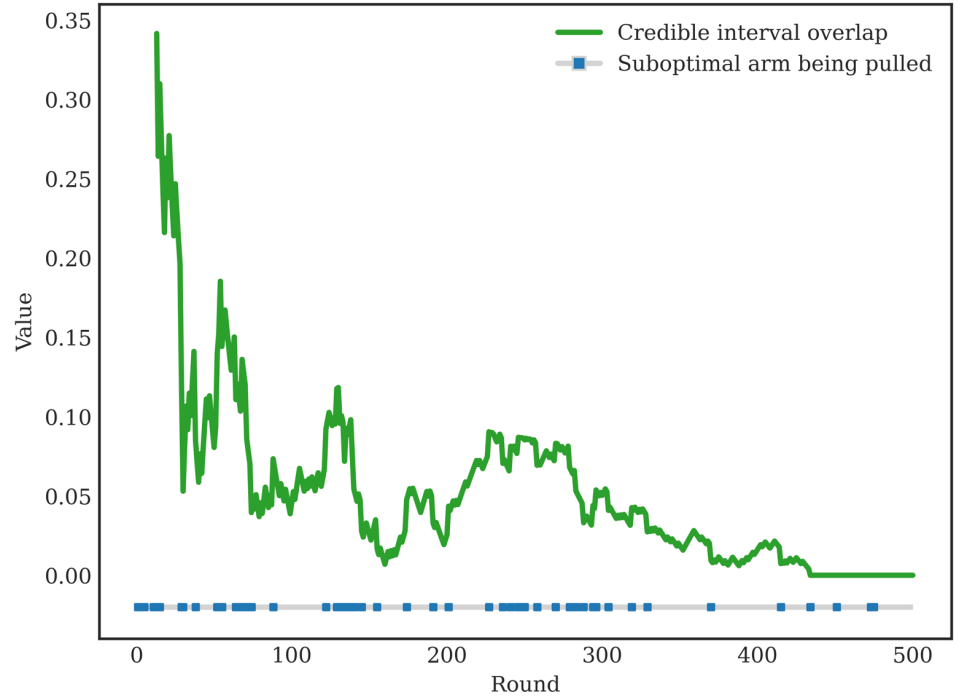
The study of "biserial" covariance dates back to Pearson (1909).

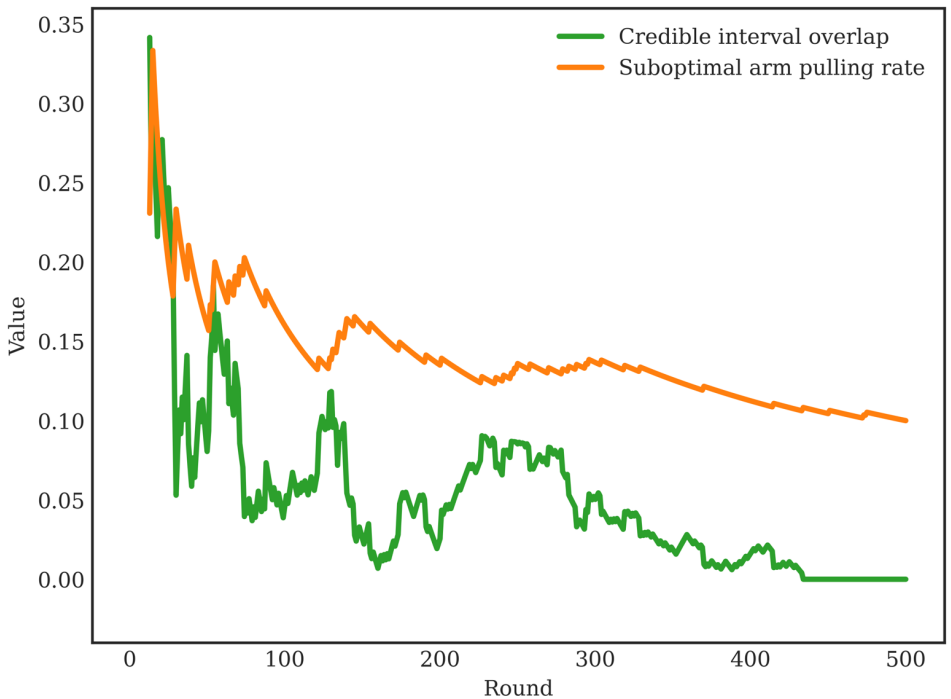
Proposition (Factorization). *If* $\text{Var}_{\pi_t}\Lambda = 0$, *then* $\text{Cov}_{\pi_t}(\Delta, \Lambda) = 0$. *Otherwise,*

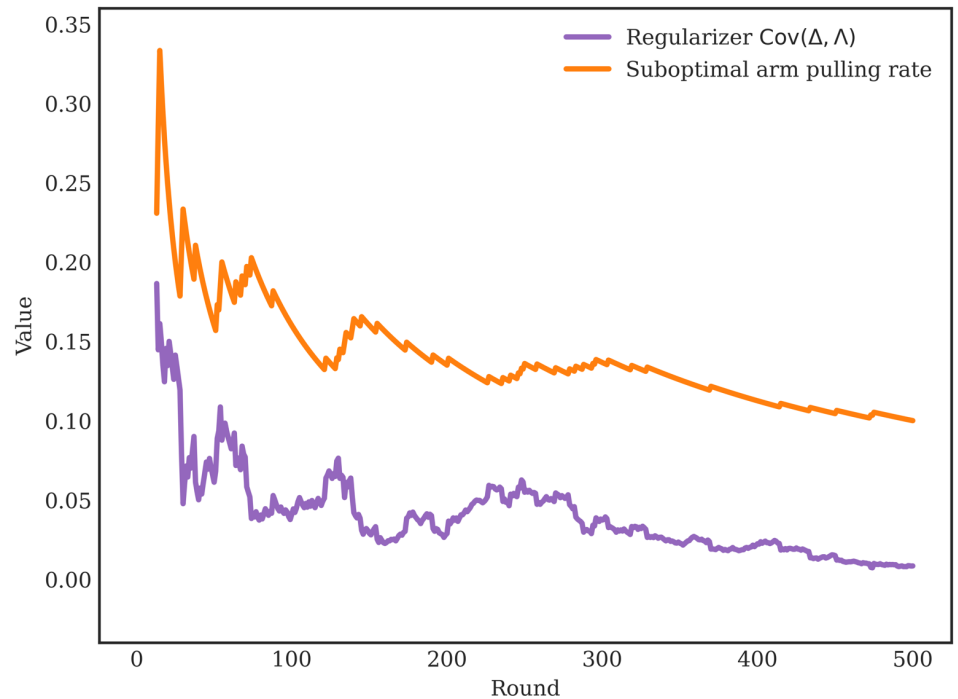
$$\frac{\text{Cov}_{\pi_t}(\Delta, \Lambda)}{\text{Var}_{\pi_t}\Lambda} = \frac{\mathbb{E}_{\pi_t}[\Delta|\Delta > 0] + \mathbb{E}_{\pi_t}[-\Delta|\Delta < 0]}{2}.$$

uncertainty measure — in the same unit as the reward


Posterior mean reward (arm 1)
Posterior mean reward (arm 2)
80% credible interval (arm 1)
80% credible interval (arm 2)
Value
Round


Credible interval overlap
Suboptimal arm being pulled
Value
Round


Credible interval overlap
Suboptimal arm pulling rate
Value
Round


Regularizer Cov(Δ, Λ)
Suboptimal arm pulling rate
Value
Round

A Duel

Thompson vs. Bellman

$$x_t^{\mathsf{TS}} = \underset{x_t}{\operatorname{argmin}}[(\mathbb{E}_{\pi_t} \max(\theta_1, \theta_2) - x_t)^2 + \boxed{\nu^{\mathsf{TS}}(\pi_t)} x_t]$$

$$x_t^{\mathsf{R2}} = \underset{x_t}{\operatorname{argmin}}[(\mathbb{E}_{\pi_t} \max(\theta_1, \theta_2) - x_t)^2 + \boxed{\nu^{\mathsf{R2}}(\pi_t)} x_t]$$

$$\boxed{\mathcal{R}_T(Q; \pi_0)} \leq \sqrt{\mathcal{R}^2(Q; \pi_0) \cdot T}$$

Round One: One Arm

Proposition (Closed-form solution). *When* $\theta_2 \equiv 0$ *and* $\mathbb{E}_{\pi_t}\theta_1 \neq 0$,

$$q_{1,t}^{\mathrm{R2}} = \min\left(\frac{\mathbb{E}_{\pi_t}[(\theta_1)_+]}{|\mathbb{E}_{\pi_t}\theta_1|}, 1\right).$$

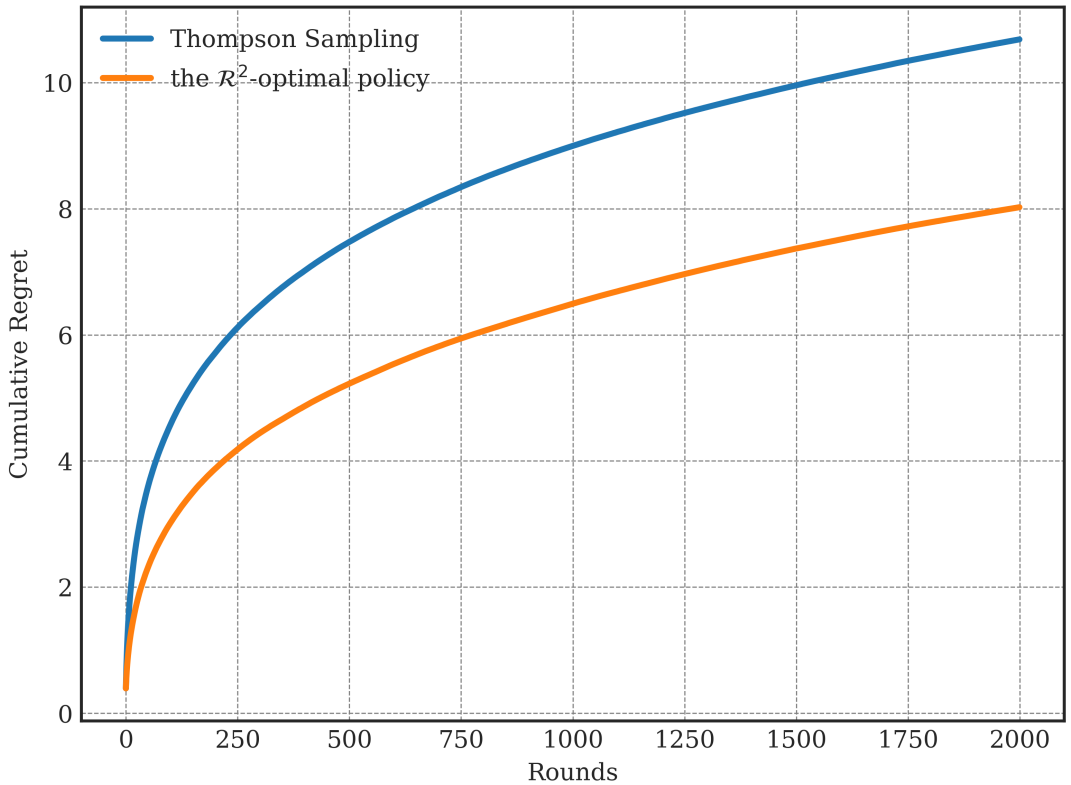
Proposition (Phase change). *When* $\theta_2 \equiv 0$ *and* $\theta_1 \sim N(\mu_t, \sigma_t^2)$ *under* π_t,

$$q_{1,t}^{\mathrm{R2}} = 1 \quad \Leftrightarrow \quad \mu_t/\sigma_t \geq \bar{x},$$

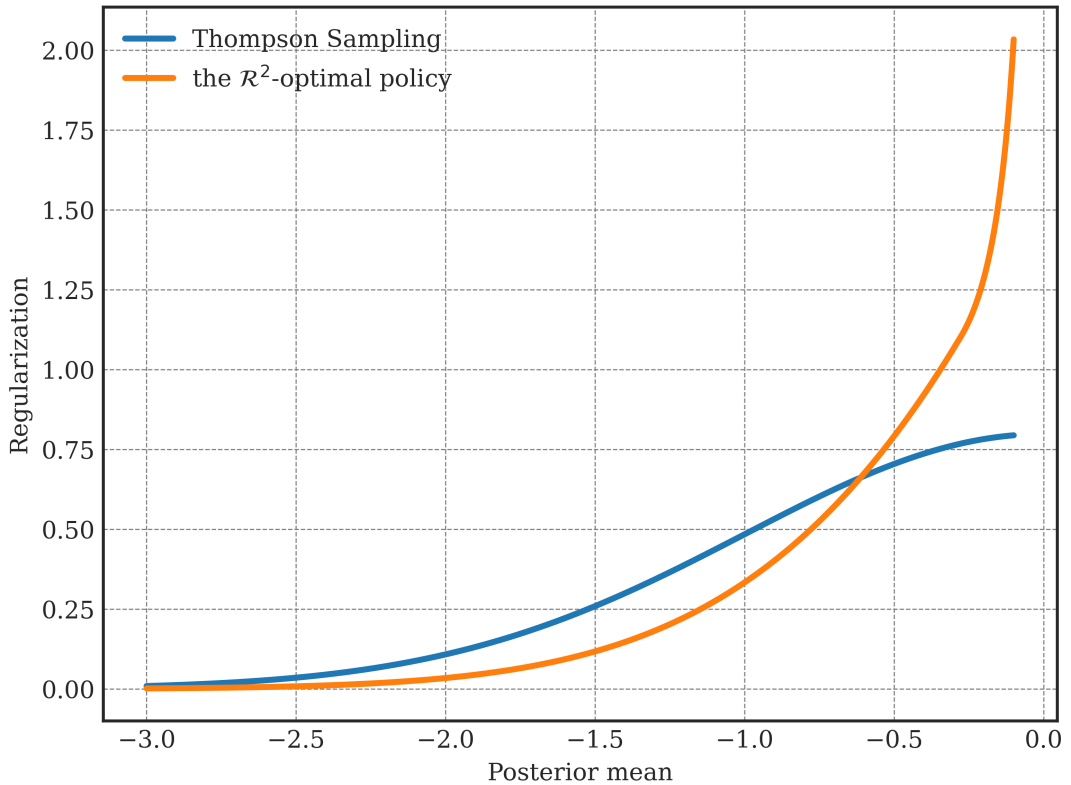
where $\bar{x} \approx -0.276$ *is the unique root of the increasing function* $x\Phi(x) + \phi(x) + x$.

"fair price" of exploring
one unit of uncertainty

Round One: One Arm



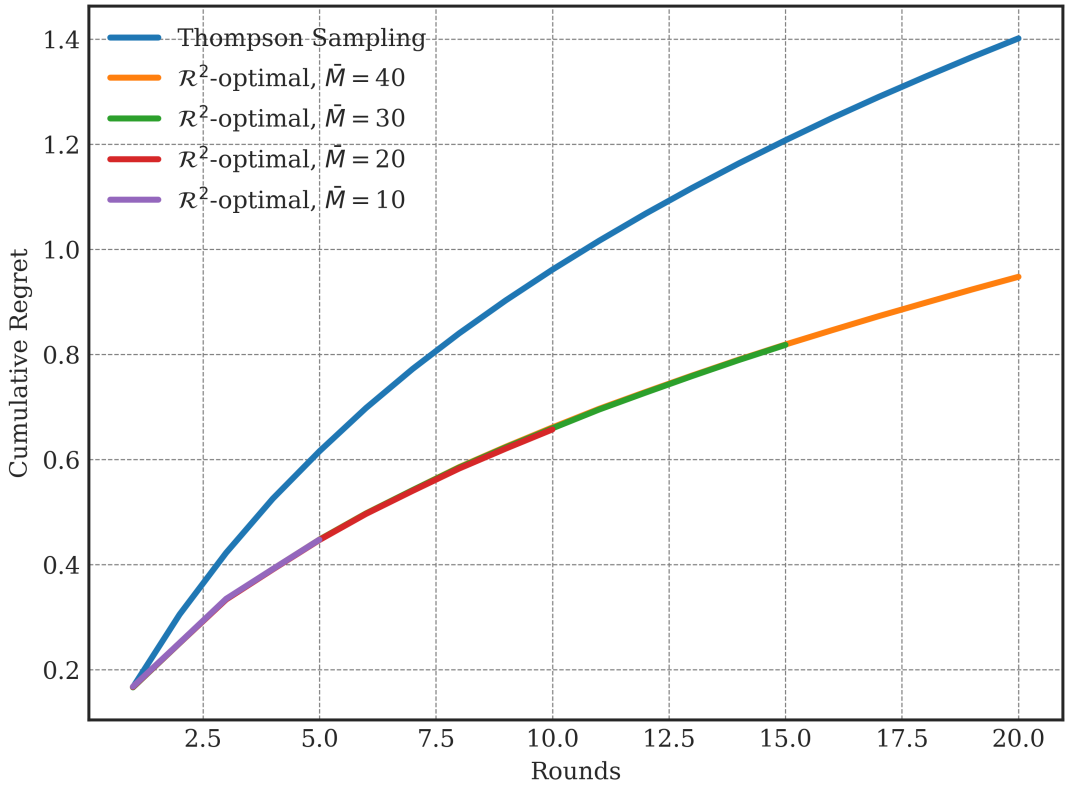
$N(0,1)$ vs. 0

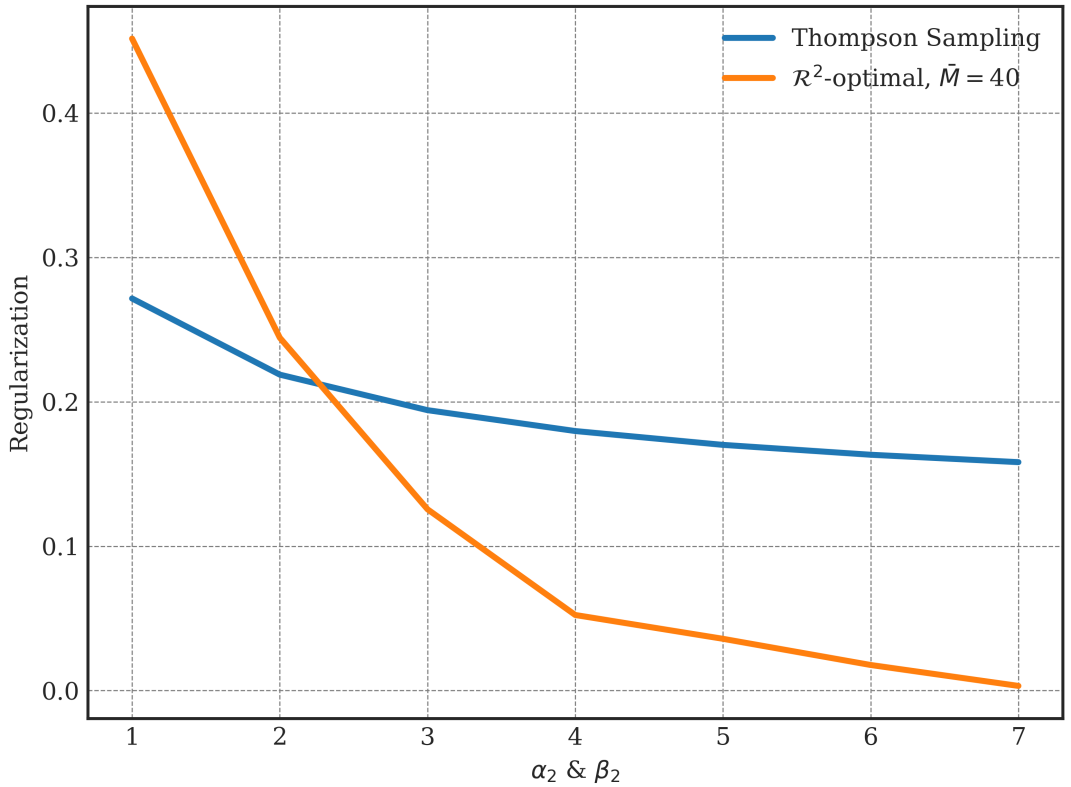


$N(-3,1) \longrightarrow N(0,1)$ vs. 0

Columbia Business School

Round Two: Two Arms





Beta(1, 1) vs. Beta(1, 1)

Beta(5, 4) vs. Beta(1, 1) ⟶ Beta(7, 7)

Columbia Business School

A Fix

Regularizer Shutdown

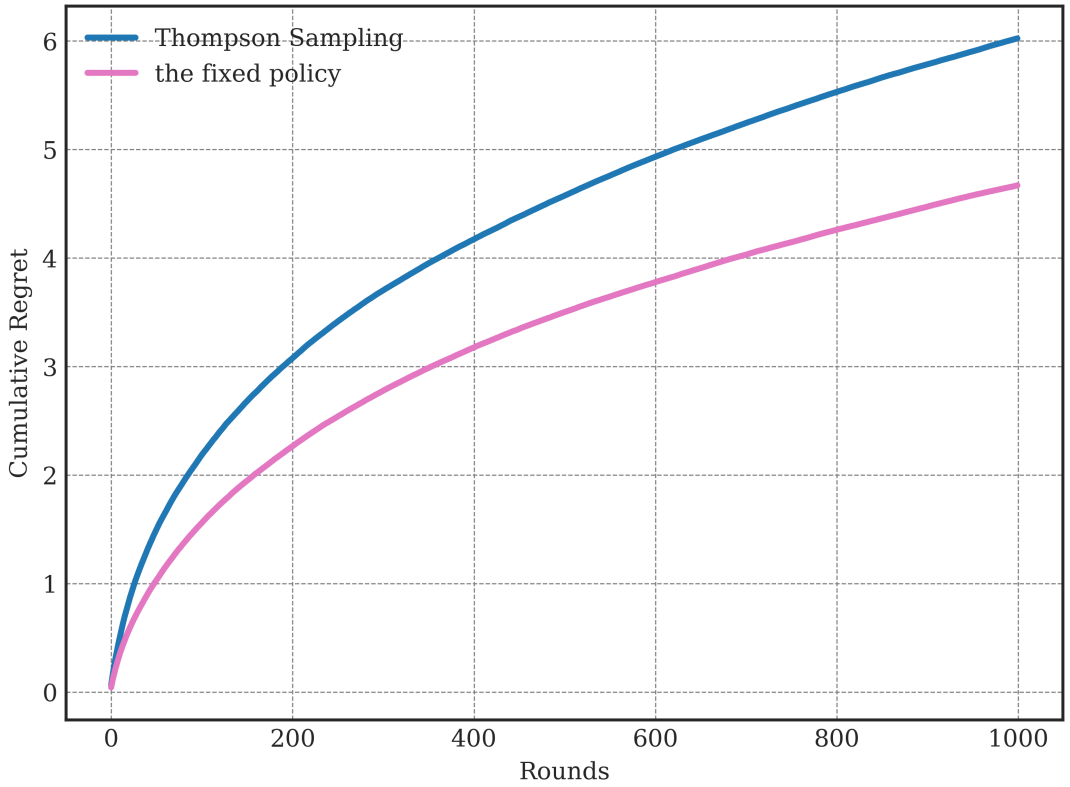
- ν^{TS} does not drop as fast as ν^{R2} to drop the overexplored underperforming arm.
- Why not shut it down? $\nu^{\mathsf{FIX}}(\pi_t) = (1 - s(\pi_t))\nu^{\mathsf{TS}}(\pi_t)$
- Shutdown criterion:

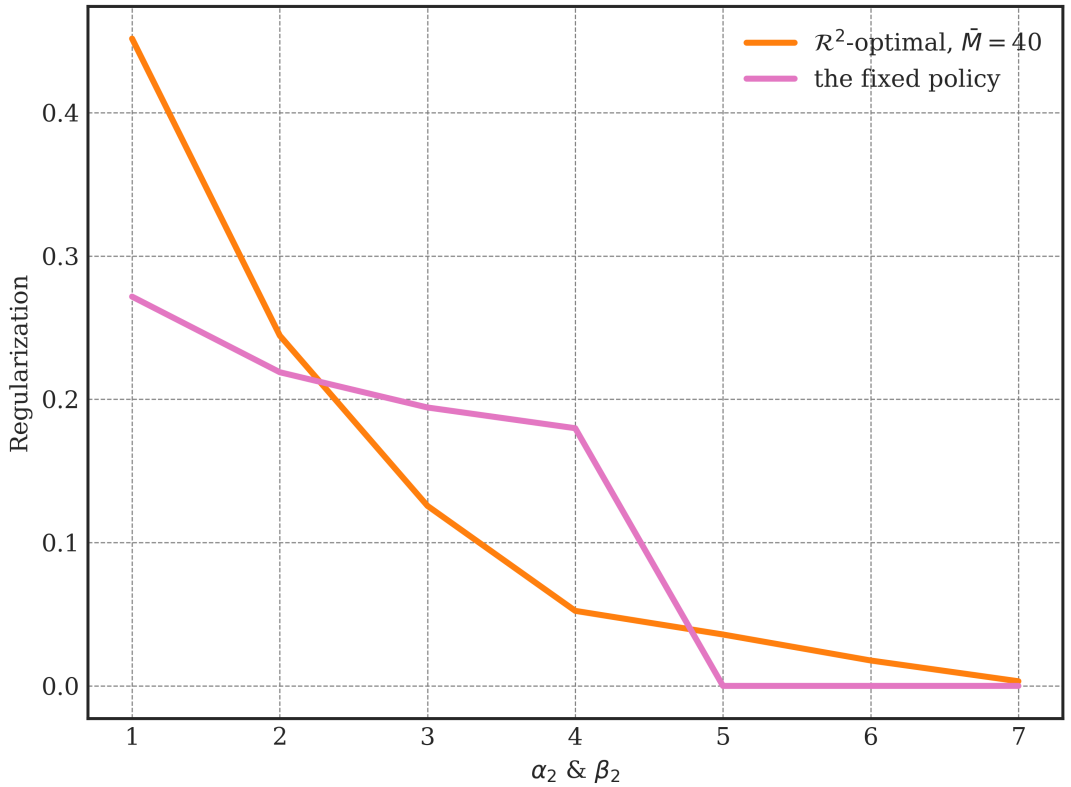
$$s(\pi_t) = I(\mathbb{E}_{\pi_t}\theta_1 > \mathbb{E}_{\pi_t}\theta_2)I(\mathcal{V}_1(\pi_t) > \mathcal{V}_2(\pi_t)) + I(\mathbb{E}_{\pi_t}\theta_2 > \mathbb{E}_{\pi_t}\theta_1)I(\mathcal{V}_2(\pi_t) > \mathcal{V}_1(\pi_t))$$

- "Information gain" (Russo and Van Roy 2014): $\mathcal{V}_k(\pi_t) = \mathsf{Var}_{\pi_t}\mathbb{E}_{\pi_t}(\theta_k|\Lambda)$

$$\boxed{\mathcal{R}^2(Q^{\mathsf{FIX}}; \pi_0) < \infty}$$

Regularizer Shutdown





$\text{Beta}(5, 4)$ vs. $\text{Beta}(500, 500)$

$\text{Beta}(5, 4)$ vs. $\text{Beta}(1, 1) \longrightarrow \text{Beta}(7, 7)$

Columbia Business School

A Broader View

Takeaway

Faithful Stationarization

$\mathcal{R}^2$

ν^{TS}

ν^{FIX}

ν^{IDS}

ν^{R2}

Online Optimization Form

0.276

Thank You

quyanlin.github.io

Qu, Y., Namkoong, H., & Zeevi, A. (2025), A Broader View of Thompson Sampling, arXiv:2510.07208

Columbia Business School

References

Agrawal S, Goyal N (2012) Analysis of Thompson sampling for the multi-armed bandit problem. *Conference on Learning Theory*, 39–1 (JMLR Workshop and Conference Proceedings).

Agrawal S, Goyal N (2013) Further optimal regret bounds for Thompson sampling. *Artificial Intelligence and Statistics*, 99–107 (PMLR).

Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47:235–256.

Bellman R (1957) *Dynamic Programming* (Princeton University Press).

Chapelle O, Li L (2011) An empirical evaluation of Thompson sampling. *Advances in Neural Information Processing Systems* 24.

Gittins JC (1979) Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 41(2):148–164.

Lai TL, Robbins H (1985) Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* 6(1):4–22.

Pearson K (1909) On a new method of determining correlation between a measured character A, and a character B, of which only the percentage of cases wherein B exceeds (or falls short of) a given intensity is recorded for each grade of A. *Biometrika* 7(1/2):96–105.

References

Robbins H (1952) Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* 58(5):527–535.

Rothschild M (1974) A two-armed bandit theory of market pricing. *Journal of Economic Theory* 9(2):185–202.

Russo D, Van Roy B (2014) Learning to optimize via information-directed sampling. *Advances in Neural Information Processing Systems* 27.

Russo D, Van Roy B (2014) Learning to optimize via posterior sampling. *Mathematics of Operations Research* 39(4):1221–1243.

Russo D, Van Roy B (2016) An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research* 17(68):1–30.

Scott SL (2010) A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry* 26(6):639–658.

Thompson WR (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3/4):285–294.