Lecture 6: Thompson Sampling for Bandits

ID 6002W: Online and Reinforcement Learning

Web-enabled M.Tech. program, IIT Madras

Suryanarayana Sankagiri

26th May, 2026

General Principles

For multi-armed bandits

- Minimising regret involves careful exploration-exploitation tradeoff

- Algorithms should use data to calibrate uncertainty about arms to make decisions

General Principles

For multi-armed bandits

- Minimising regret involves careful exploration-exploitation tradeoff

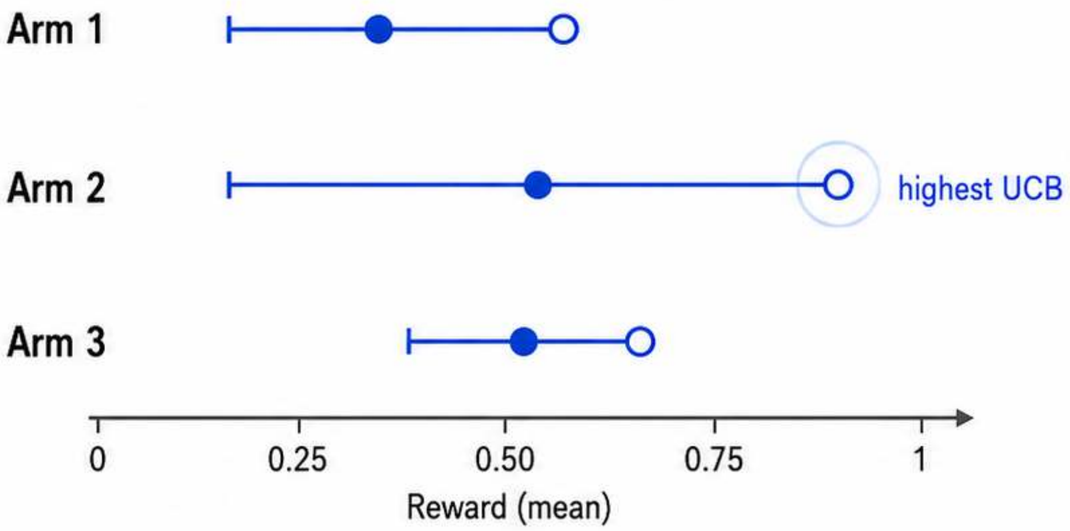
- Algorithms should use data to calibrate uncertainty about arms to make decisions

- UCB and Thompson Sampling differ in how they represent and use uncertainty:
  - UCB maintains *confidence intervals*
  - Thompson Sampling uses *belief distributions*

UCB vs Thompson Sampling

UCB

Choose the highest upper confidence bound

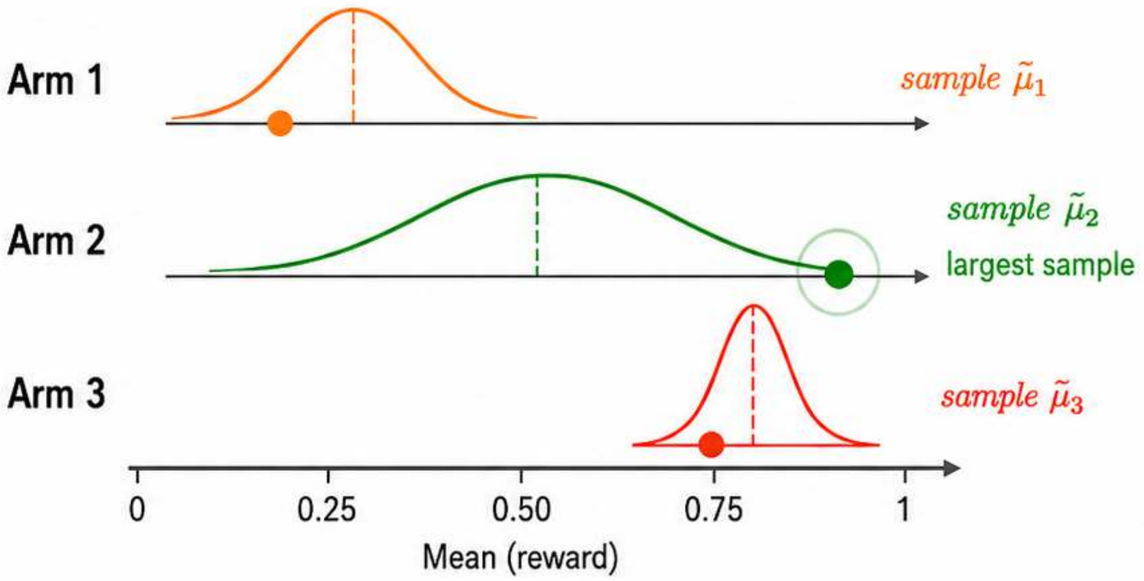


$$UCB_i(t) = \hat{\mu}_i(t) + bonus_i(t)$$

Deterministic: pick the arm with the largest optimistic value.

Thompson Sampling

Sample one plausible mean for each arm



$$sample\ \tilde{\mu}_i(t) \text{ for each arm}, \quad A_t = \arg\max_i \tilde{\mu}_i(t)$$

Randomized: pick the arm with the largest sampled plausible value.

Bayesian Viewpoint

For multi-armed bandits

The true arm mean is *fixed but unknown:*

$$\mu_i = \mathbb{P}(X = 1 \,|\, \textit{arm } i)$$

- With finite amount of i.i.d. data, we will always have some uncertainty about μ_i
- A belief distribution represents our uncertainty about μ_i

Bayesian Viewpoint

For multi-armed bandits

The true arm mean is *fixed but unknown:*

$$\mu_i = \mathbb{P}(X = 1 \,|\, \textit{arm } i)$$

- With finite amount of i.i.d. data, we will always have some uncertainty about μ_i
- A belief distribution represents our uncertainty about μ_i

Data does not change μ_i — data changes our *belief* about μ_i

With more data, the belief usually gets more *concentrated*

Bayes' Rule

For updating beliefs

$$p(\mu \mid \text{data}) = \frac{p(\mu)\mathbb{P}(\text{data} \mid \mu)}{\mathbb{P}(\text{data})}$$

For simplicity, we write

$$p(\mu \mid \text{data}) \propto p(\mu)\mathbb{P}(\text{data} \mid \ \mu)$$

Bayes' Rule

For updating beliefs

$$p(\mu \mid \text{data}) = \frac{p(\mu)\mathbb{P}(\text{data} \mid \mu)}{\mathbb{P}(\text{data})}$$

For simplicity, we write

$$p(\mu \mid \text{data}) \propto p(\mu)\mathbb{P}(\text{data} \mid \mu)$$

- $p(\mu)$: Prior belief
- $p(\mu \mid \text{data})$: Posterior belief
- $\mathbb{P}(\text{data} \mid \mu)$: Likelihood

Prior × Likelihood → Posterior

Bayes' Rule

For updating beliefs

$$p(\mu \mid \text{data}) \propto p(\mu)\mathbb{P}(\text{data} \mid \mu)$$

We can also write:

$$p(\mu \mid \text{all data}) = p(\mu \mid \text{new data, old data}) \propto p(\mu \mid \text{old data})\ \mathbb{P}(\text{new data} \mid \mu, \text{old data})$$

Bayes' Rule

For updating beliefs

$$p(\mu \mid \text{data}) \propto p(\mu)\mathbb{P}(\text{data} \mid \mu)$$

We can also write:

$$p(\mu \mid \text{all data}) = p(\mu \mid \text{new data, old data}) \propto p(\mu \mid \text{old data})\ \mathbb{P}(\text{new data} \mid \mu, \text{old data})$$

- $p(\mu \mid \text{old data})$: Prior belief
- $p(\mu \mid \text{all data})$: Posterior belief
- $\mathbb{P}(\text{new data} \mid \mu, \text{ old data}) = \mathbb{P}(\text{new data} \mid \mu)$ = likelihood

Posterior of one round becomes prior of the next round

i.i.d. data

Generating Samples from Beliefs

If we maintain a belief distribution $f_{i,t}(\mu)$ for each arm i at all times t,

We can draw a random sample $\tilde{\mu}_i(t) \sim f_{i,t}(\mu)$

- Sample is drawn through simulation on a computer (easy to do)

Generating Samples from Beliefs

If we maintain a belief distribution $f_{i,t}(\mu)$ for each arm i at all times t,

We can draw a random sample $\tilde{\mu}_i(t) \sim f_{i,t}(\mu)$

- Sample is drawn through simulation on a computer (easy to do)

Not a datapoint!

The sample $\tilde{\mu}_i(t)$ is a plausible value of μ_i (*not an empirical average*)

Generating Samples from Beliefs

If we maintain a belief distribution $f_{i,t}(\mu)$ for each arm i at all times t,

We can draw a random sample $\tilde{\mu}_i(t) \sim f_{i,t}(\mu)$

- Sample is drawn through simulation on a computer (easy to do)

Not a datapoint!

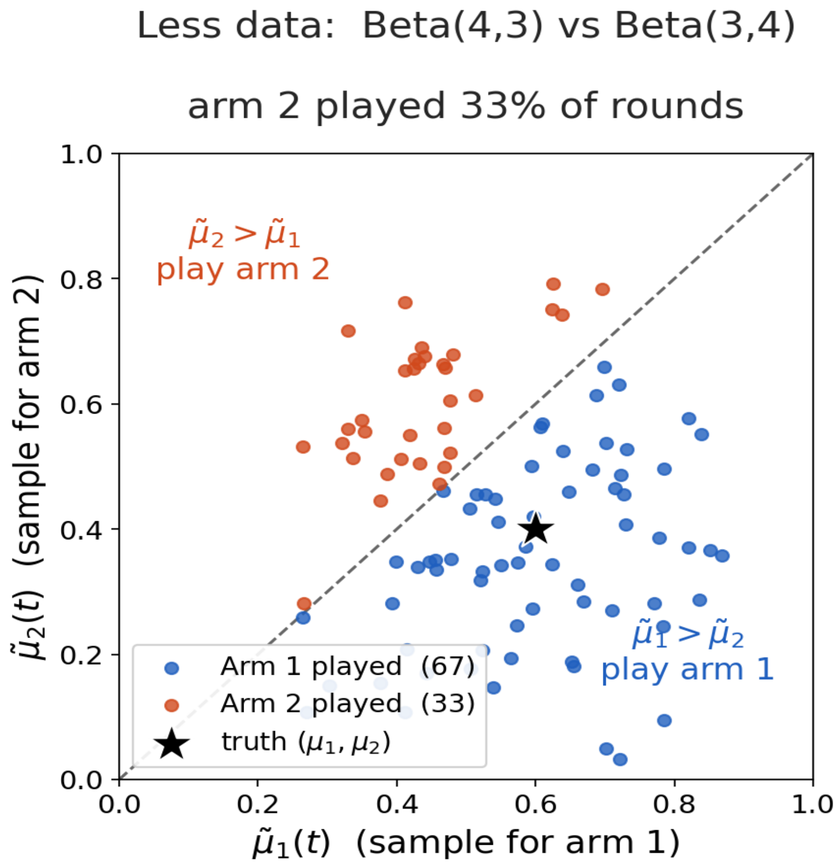
The sample $\tilde{\mu}_i(t)$ is a plausible value of μ_i (*not an empirical average*)

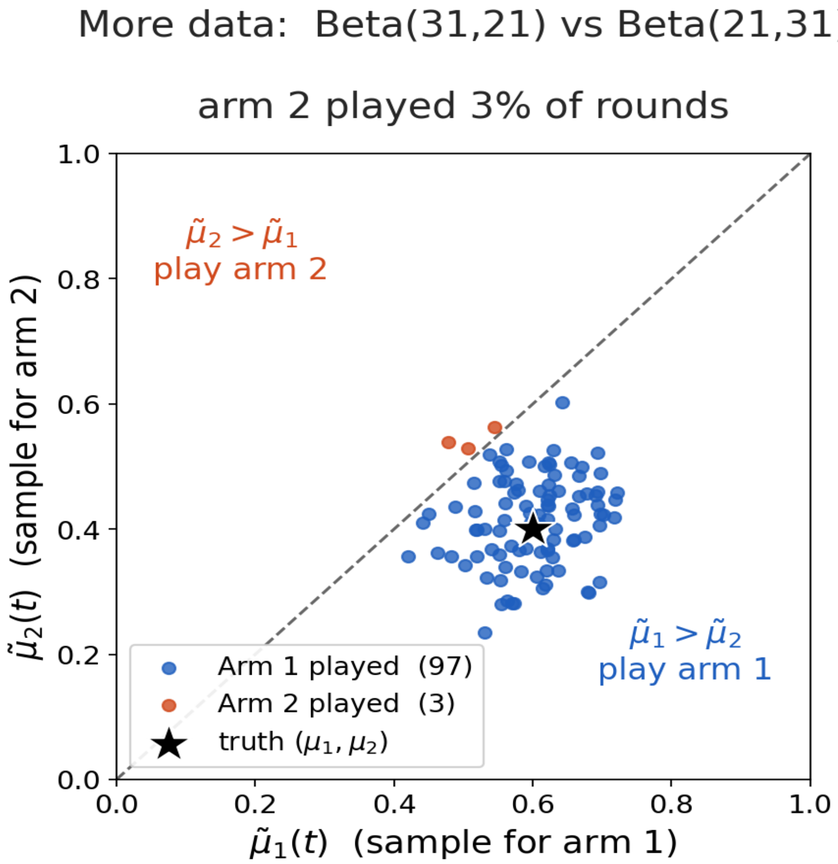
The set of samples $(\tilde{\mu}_1(t), \ldots, \tilde{\mu}_K(t))$ are a set of plausible values of $(\mu_1, \ldots, \mu_K)$

Thompson sampling rule: play as if the sampled world is the true world

Why Thompson Sampling?

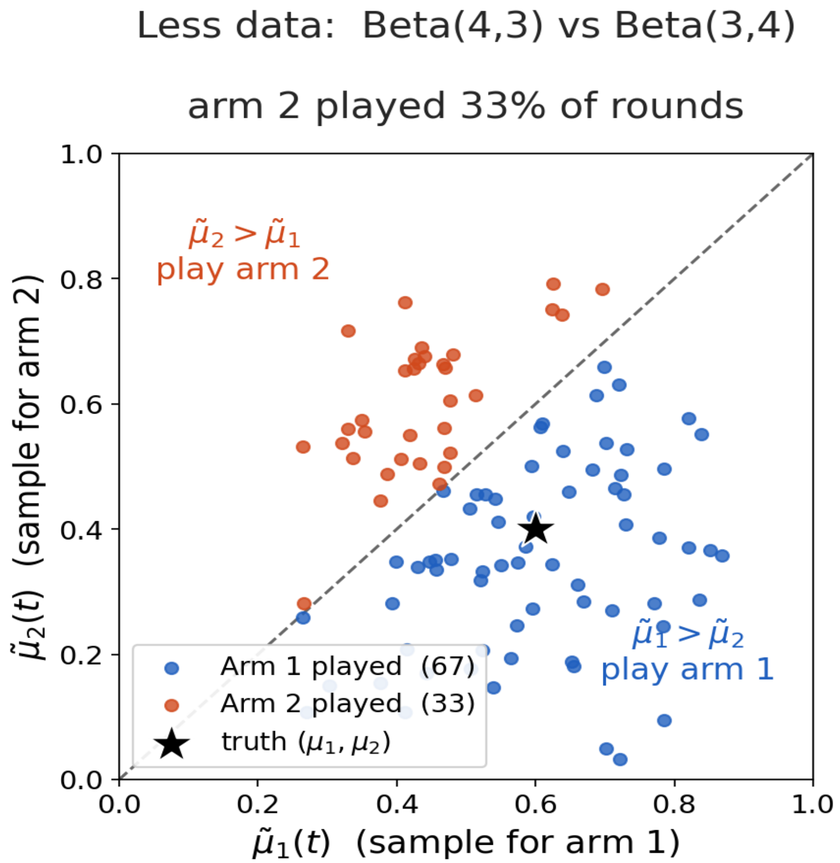
The set of samples $(\tilde{\mu}_1(t), \tilde{\mu}_2(t))$ are a set of plausible values of $(\mu_1 = 0.6,\ \mu_2 = 0.4)$

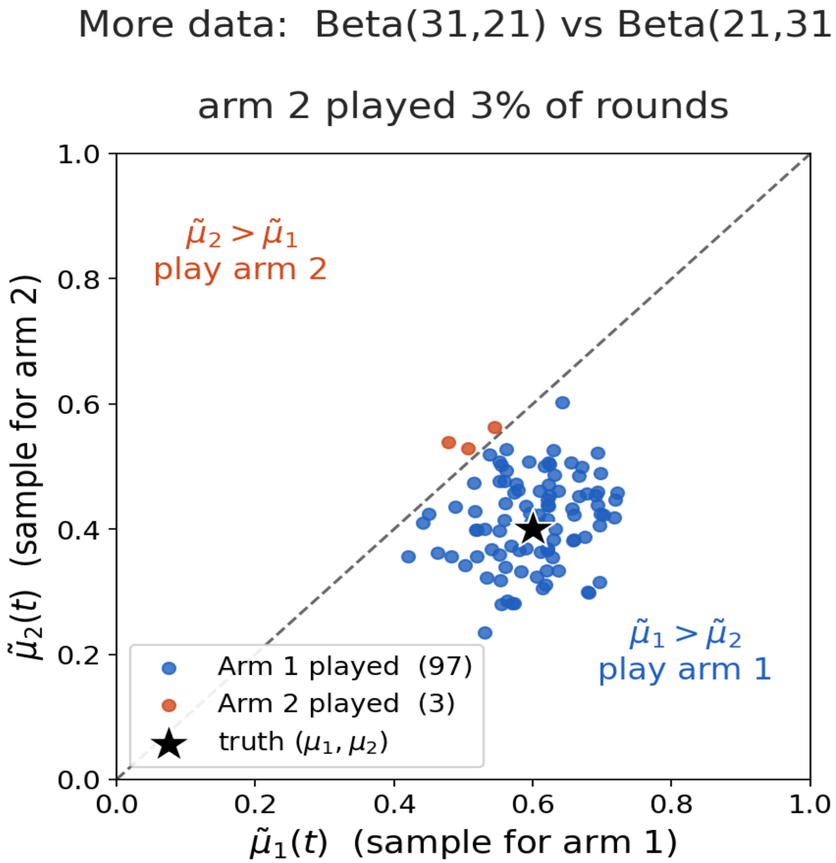




Why Thompson Sampling?

The set of samples $(\tilde{\mu}_1(t), \tilde{\mu}_2(t))$ are a set of plausible values of $(\mu_1 = 0.6,\ \mu_2 = 0.4)$





Thompson sampling rule: play as if the sampled world is the true world

Leans towards exploration with less data, leans towards exploitation with more data

Making This Concrete: Bernoulli Bandits

What Do We Know About an Arm?

Bernoulli reward setting

For arm i, we can keep track of:

$$S_i(t) = \text{number of successes (ones)}$$

$$F_i(t) = \text{number of failures (zeros)}$$

Empirical average:

$$\hat{\mu}_i(t) = \frac{S_i(t)}{S_i(t) + F_i(t)}$$

How can we construct a belief distribution from $(S_i(t), F_i(t))$?

The Beta Distribution

A belief over mean rewards

For arms with Bernoulli rewards, $\mu_i \in [0,1]$

$\Rightarrow$ the belief distribution of μ_i should also take values in $[0,1]$

We use

$$\mu_i \sim \text{Beta}(\alpha_i, \beta_i)$$

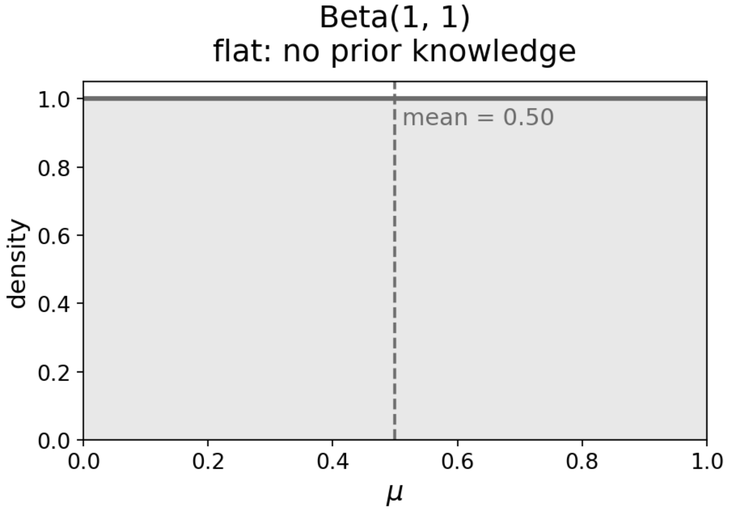
Where the density function is

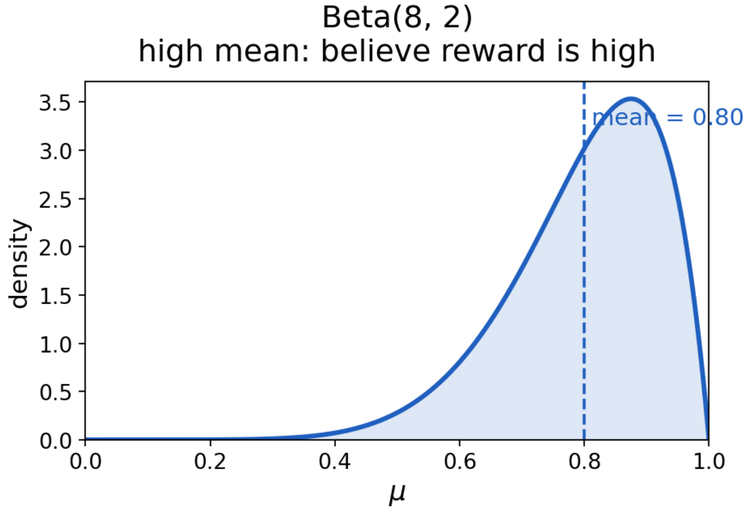
$$f_{\alpha,\beta}(\mu) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \;\forall\; \mu \in [0,1]$$

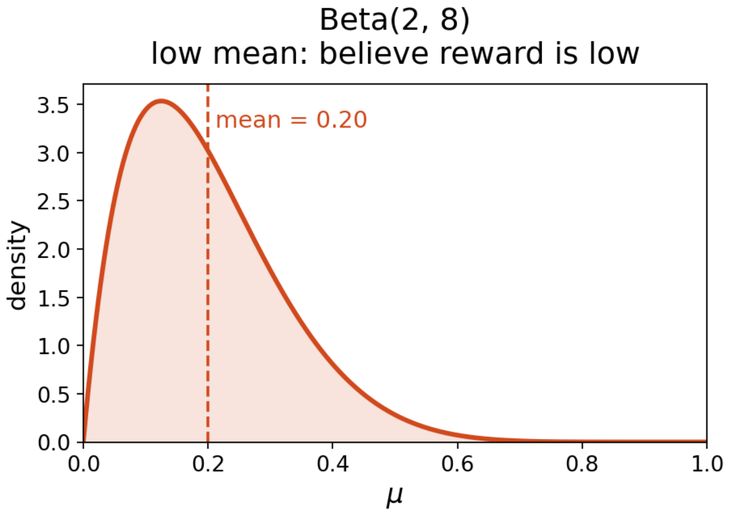
$$\tilde{\mu} \sim f_{\alpha,\beta}(\mu) \Rightarrow \mathbb{E}[\tilde{\mu}] = \frac{\alpha}{\alpha+\beta}, \quad \mathbb{V}[\mu] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

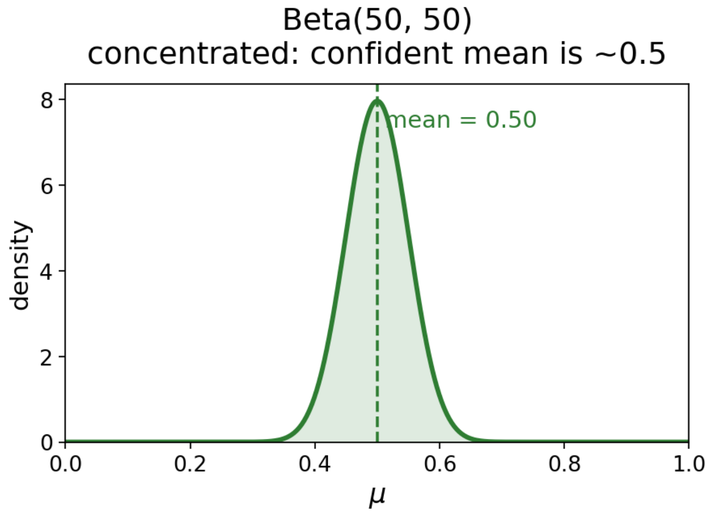
The Beta Distribution

Examples









Beta-Bernoulli Conjugacy

Updates become simplified!

For one pull of one arm:

$$X \in \{0,1\}, \qquad \mathbb{P}(X = 1 \mid \mu) = \mu, \qquad \mathbb{P}(X = 0 \mid \mu) = 1 - \mu$$

$\Rightarrow$ Likelihood for one data point (x)

$$\mathbb{P}(X = x \mid \mu) = \mu^x (1 - \mu)^{1-x}$$

Beta-Bernoulli Conjugacy

Updates become simplified!

For one pull of one arm:

$$X \in \{0,1\}, \qquad \mathbb{P}(X = 1 \mid \mu) = \mu, \qquad \mathbb{P}(X = 0 \mid \mu) = 1 - \mu$$

$\Rightarrow$ Likelihood for one data point (x)

$$\mathbb{P}(X = x \mid \mu) = \mu^x (1 - \mu)^{1-x}$$

Prior Beta distribution:

$$\mathbb{P}(\mu) \propto \mu^{\alpha-1} (1 - \mu)^{\beta-1}$$

Beta-Bernoulli Conjugacy

Updates become simplified!

For one pull of one arm:

$$X \in \{0,1\}, \qquad \mathbb{P}(X = 1 \mid \mu) = \mu, \qquad \mathbb{P}(X = 0 \mid \mu) = 1 - \mu$$

$\Rightarrow$ Likelihood for one data point (x)

$$\mathbb{P}(X = x \mid \mu) = \mu^x (1 - \mu)^{1-x}$$

Prior Beta distribution:

$$\mathbb{P}(\mu) \propto \mu^{\alpha-1}(1 - \mu)^{\beta-1}$$

Multiplying the two, we get that the posterior distribution is:

$$\mathbb{P}(\mu \mid x) \propto \mu^{\alpha+x-1}(1 - \mu)^{\beta+(1-x)-1}$$

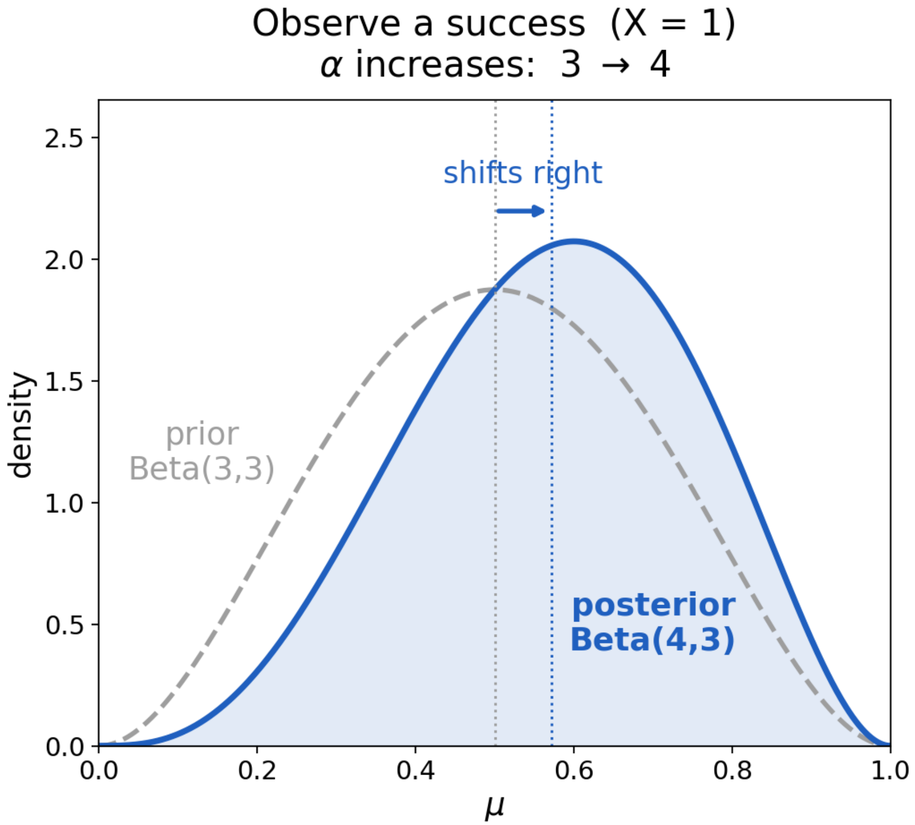
Posterior Update Illustration

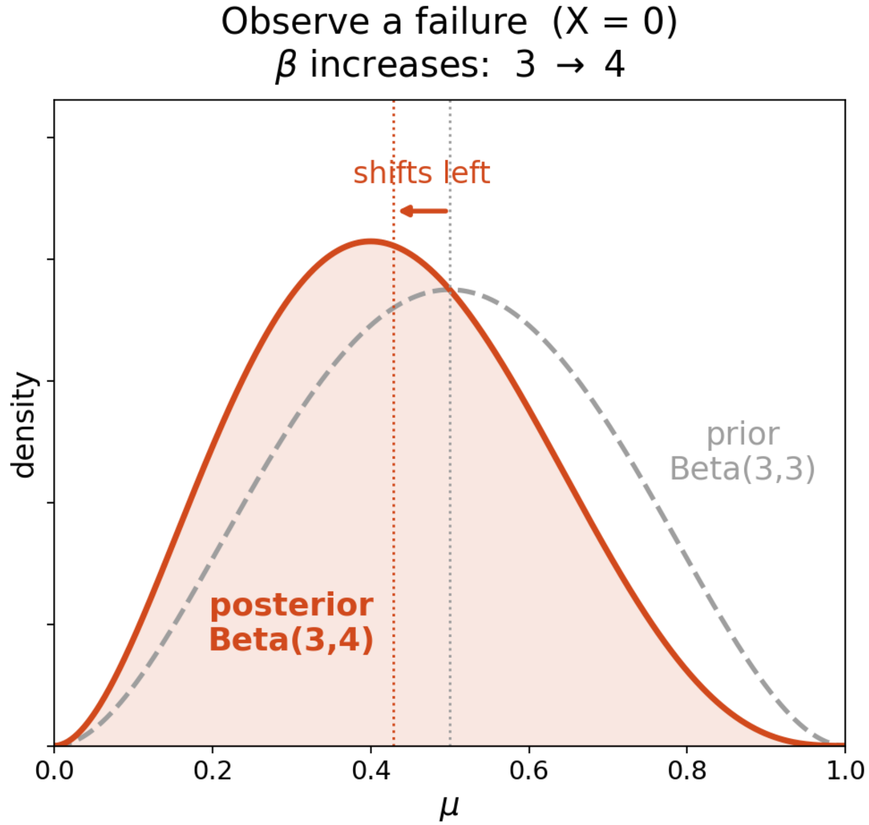
Interpretation:

- If $(X = 1)$: larger μ becomes more plausible
- If $(X = 0)$: smaller μ becomes more plausible

$$X = 0 : \text{Beta}(\alpha, \beta) \rightarrow \text{Beta}(\alpha, \beta + 1)$$

$$X = 1 : \text{Beta}(\alpha, \beta) \rightarrow \text{Beta}(\alpha + 1, \beta)$$





Thompson Sampling

Pseudocode

Initialise: $\alpha_i = \beta_i = 1$ for all arms i

For $t = 1,2,\ldots,T$

1. For each arm i:

   Sample $\tilde{\mu}_i \sim \text{Beta}(\alpha_i, \beta_i)$

2. Pull arm:

$$A_t = \arg \max_{i \in \{1,2,\ldots,K\}} \tilde{\mu}_i(t-1)$$

3. Observe reward $X_t \in \{0,1\}$ and update:

   if $X_t = 1$: α_{A_t}++, else: β_{A_t}++

Thompson Sampling

Pseudocode

Initialise: $\alpha_i = \beta_i = 1$ for all arms i

For $t = 1,2,\ldots,T$

1. For each arm i:

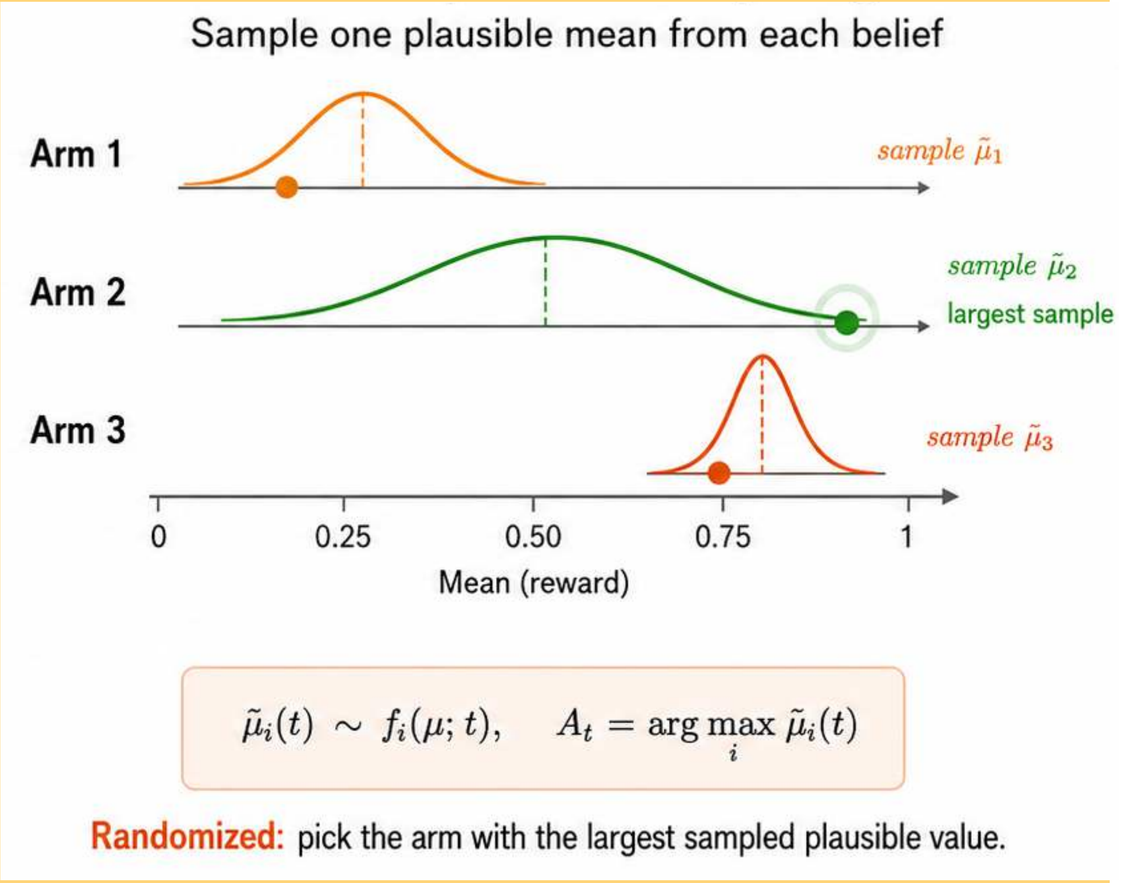
$$\text{Sample } \tilde{\mu}_i \sim \text{Beta}(\alpha_i, \beta_i)$$

2. Pull arm:

$$A_t = \arg \max_{i \in \{1,2,\ldots,K\}} \tilde{\mu}_i(t-1)$$

3. Observe reward $X_t \in \{0,1\}$ and update:

$$\text{if } X_t = 1\text{: } \alpha_{A_t}\text{++, else: } \beta_{A_t}\text{++}$$

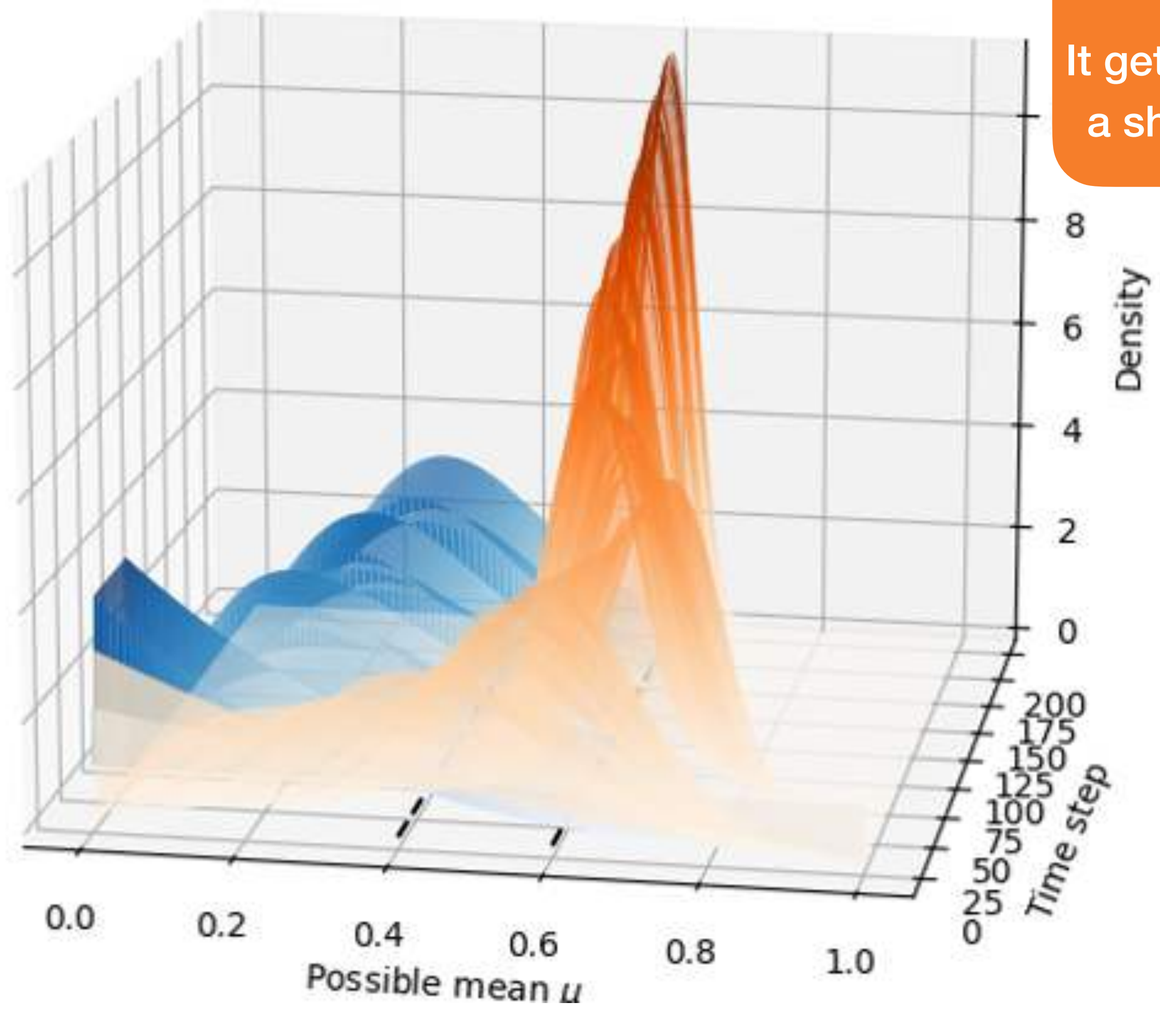


Empirical Results

Evolution of Posteriors

Arm 1 has true mean $\mu_1 = 0.4$

It gets pulled less often, so it has a sharper posterior at $T = 200$

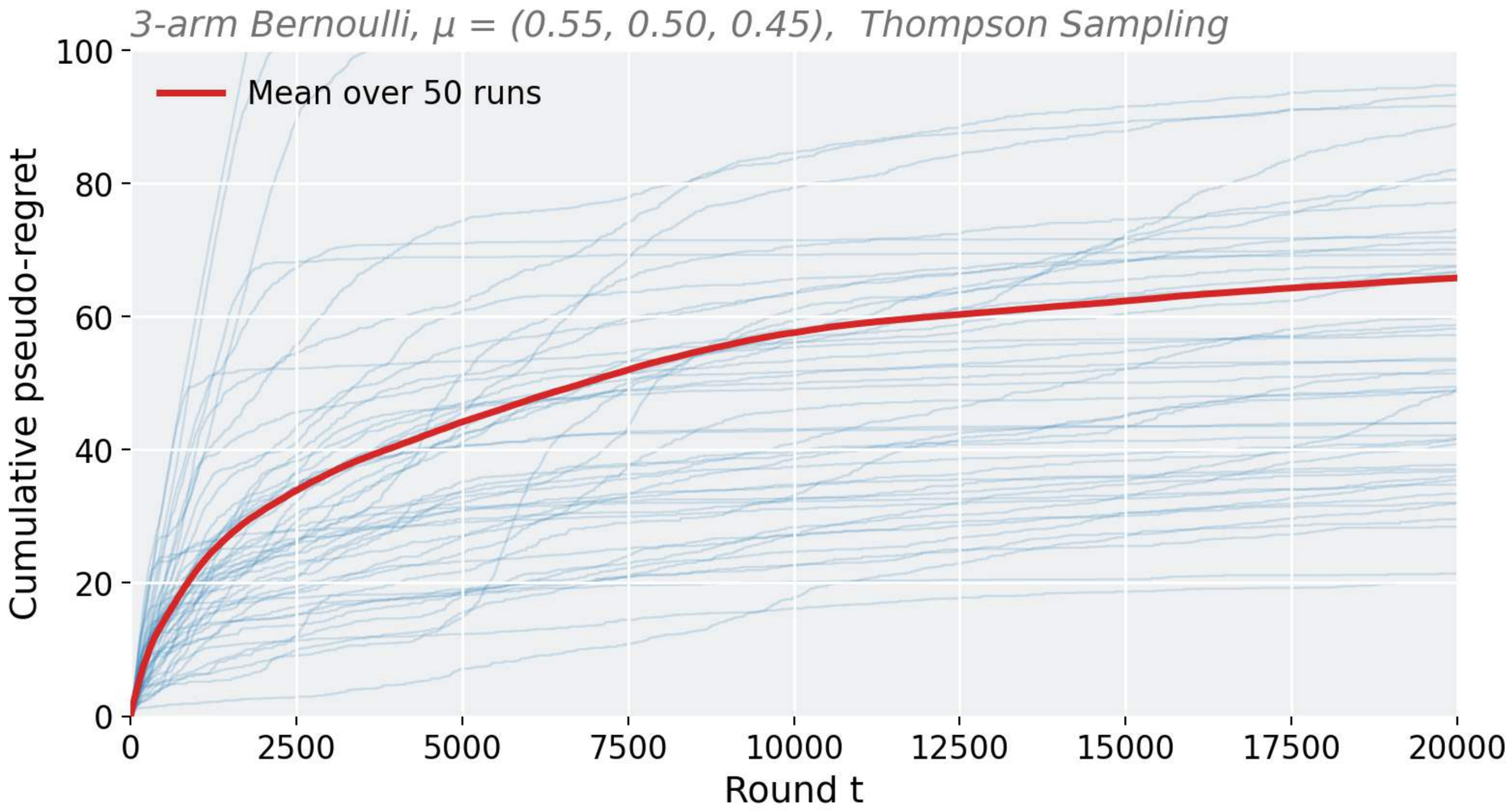


Arm 2 has true mean $\mu_2 = 0.4$

It gets pulled less often, so it has a wider posterior at $T = 200$

Empirical Performance

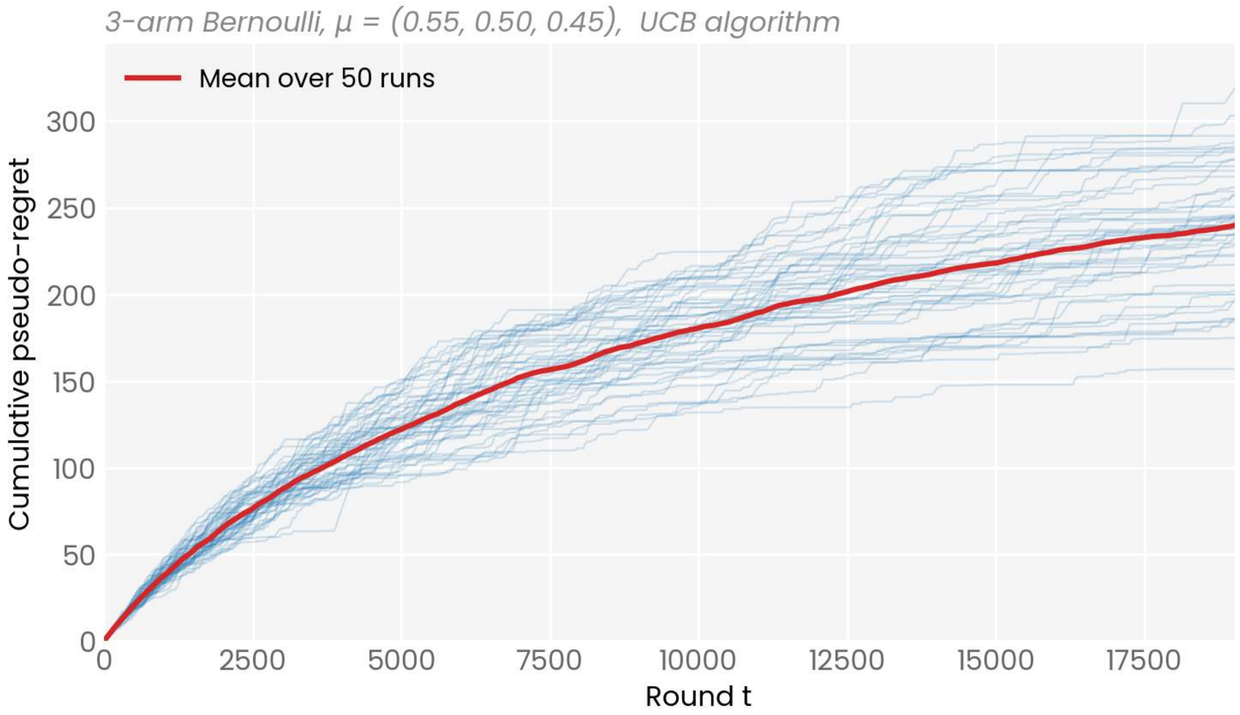
Cumulative regret curves



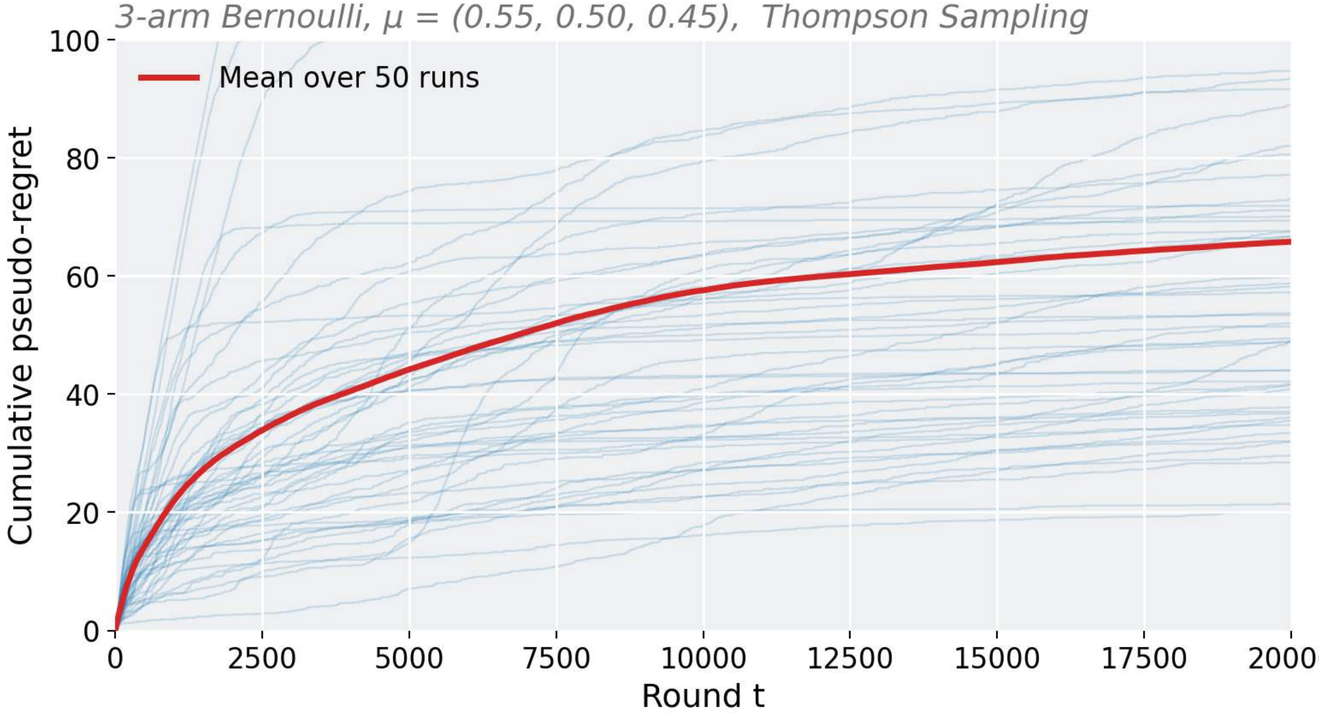
Empirical Performance

Cumulative regret curves

UCB



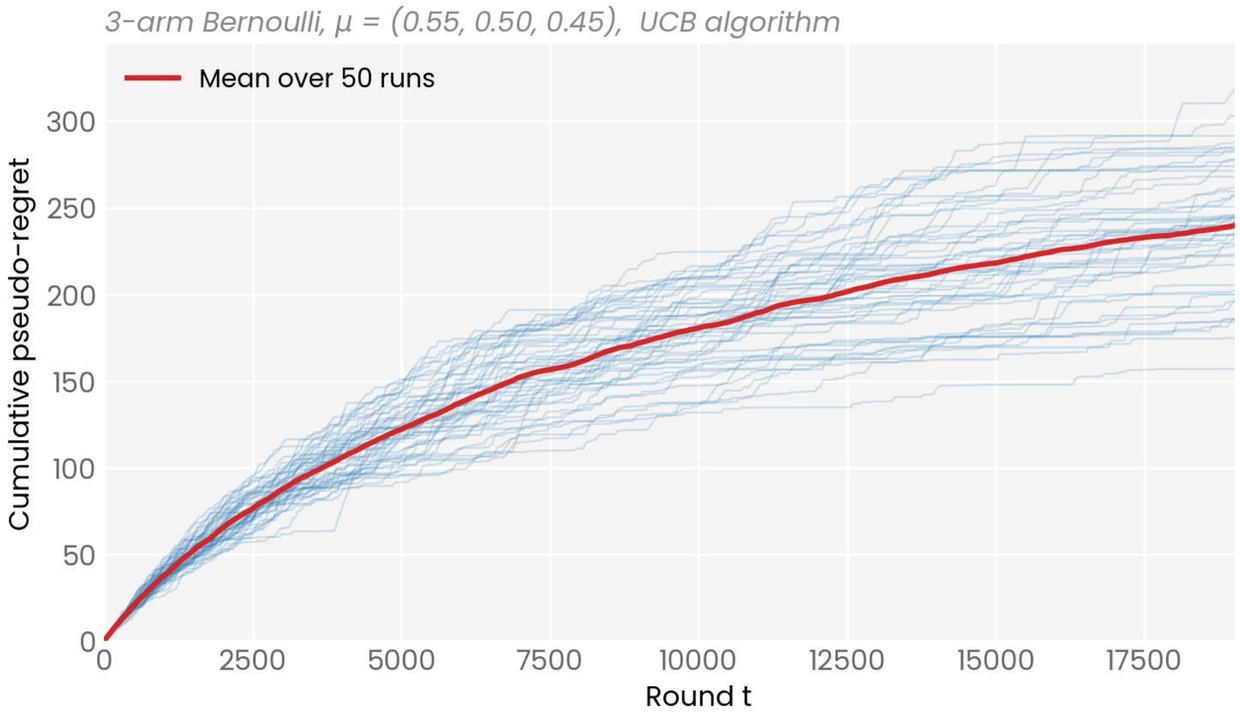
Thompson Sampling



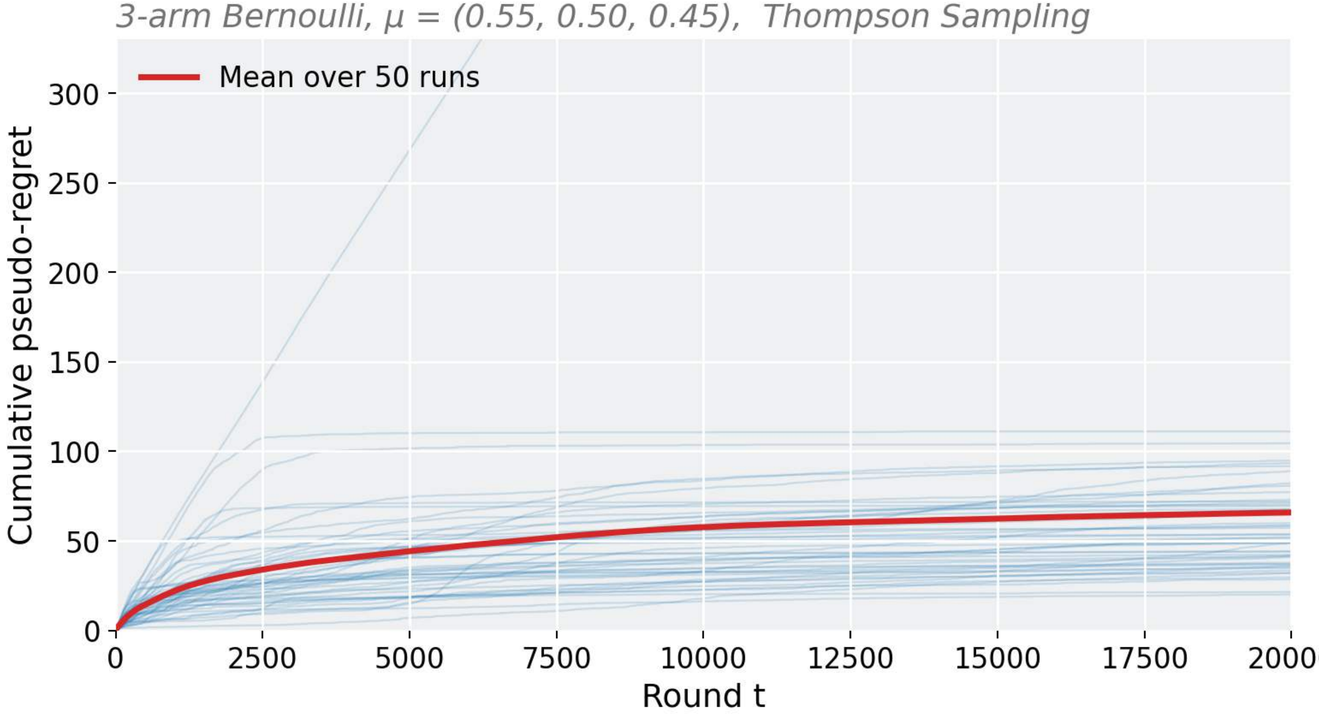
Empirical Performance

Cumulative regret curves

UCB



Thompson Sampling



Conclusion

Thompson sampling in one line

Maintain beliefs, Sample a plausible world, act greedily in it

Conclusion

Thompson sampling in one line

Maintain beliefs, Sample a plausible world, act greedily in it

Key Ingredients:

- Belief distribution for each arm, which captures uncertainty about μ_i
- Bayes' rule, used to update beliefs after each observed reward
- Posterior sampling selects arms that could plausibly be best

Conclusion

Thompson sampling in one line

Maintain beliefs, Sample a plausible world, act greedily in it

Why it works:

- Uncertain arms are explored naturally
- Confidently bad arms are rarely sampled as best
- As data grows, beliefs concentrate and exploration fades

Conclusion

Thompson sampling in one line

Maintain beliefs, Sample a plausible world, act greedily in it

Why it works:

- Uncertain arms are explored naturally
- Confidently bad arms are rarely sampled as best
- As data grows, beliefs concentrate and exploration fades

UCB = deterministic optimism Thompson Sampling = randomised optimism