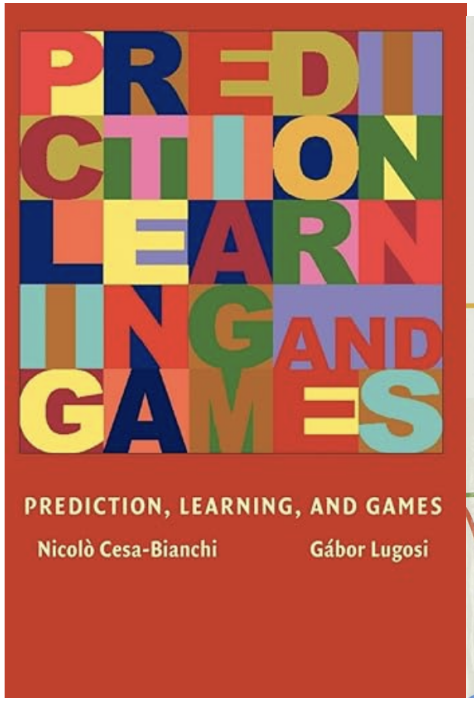
Summer School Uganda of the Mathematics of AI

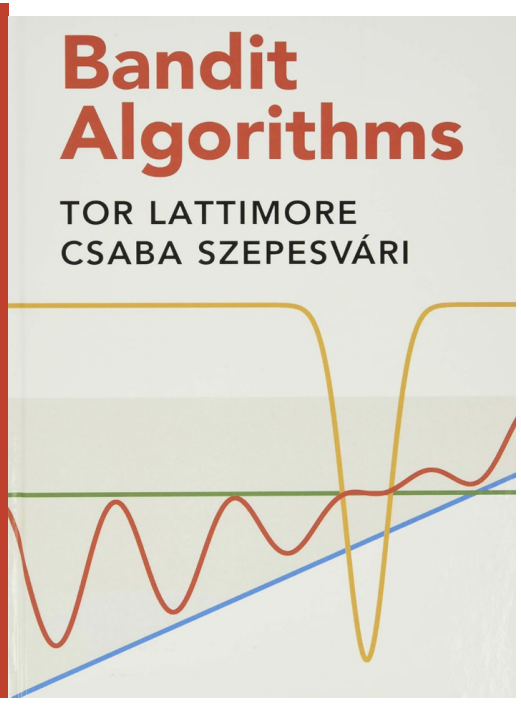
Jana de Wiljes

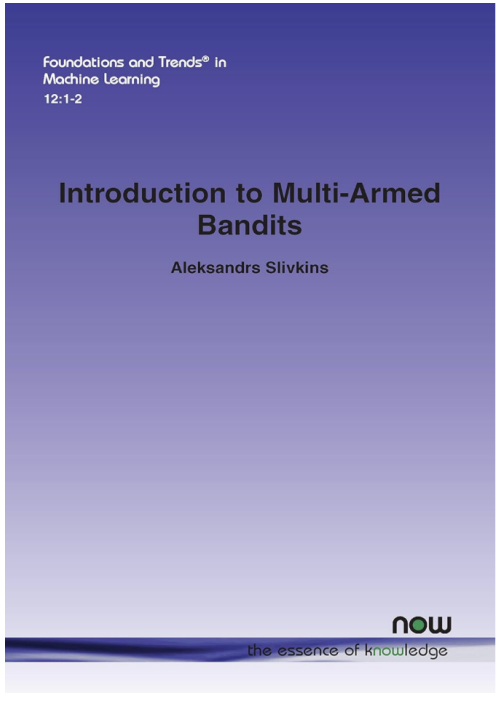
Summer School of the Mathematics of AI 2024

Kampala

References



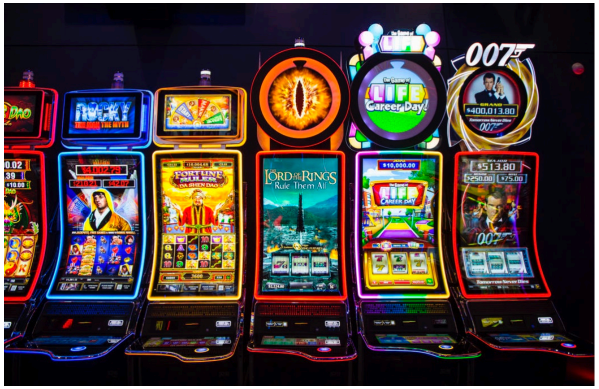




Good online Reference: https://emiliekaufmann.github.io

Multi-armed bandits

Choose from K options
to receive a
high reward and
to reduce loss after T rounds



Examples:

- Which advertising campaign generates the largest revenue?

a_1

a_2

a_3

a_4

- Which vaccine should enter next stage of clinical trials?

a_1

a_2

a_3

a_4

Multi-armed bandits

Stochastic K-Armed Bandit

A stochastic K-Armed Bandit is a collection of distributions

$$\nu = (\nu_a : a \in \mathcal{A})$$

where $\mathcal{A}$ is a set of actions (arms) and $|\mathcal{A}| = K$ and

$$\mu_a(\nu) = \int_{-\infty}^{\infty} x \nu_a(x) dx \tag{1}$$

Procedure: in each round $t \in \{1, \ldots, T\}$

1. learner chooses an action $A_t = a$
2. receives reward $R_t \sim \nu_a$ (independent from the past)

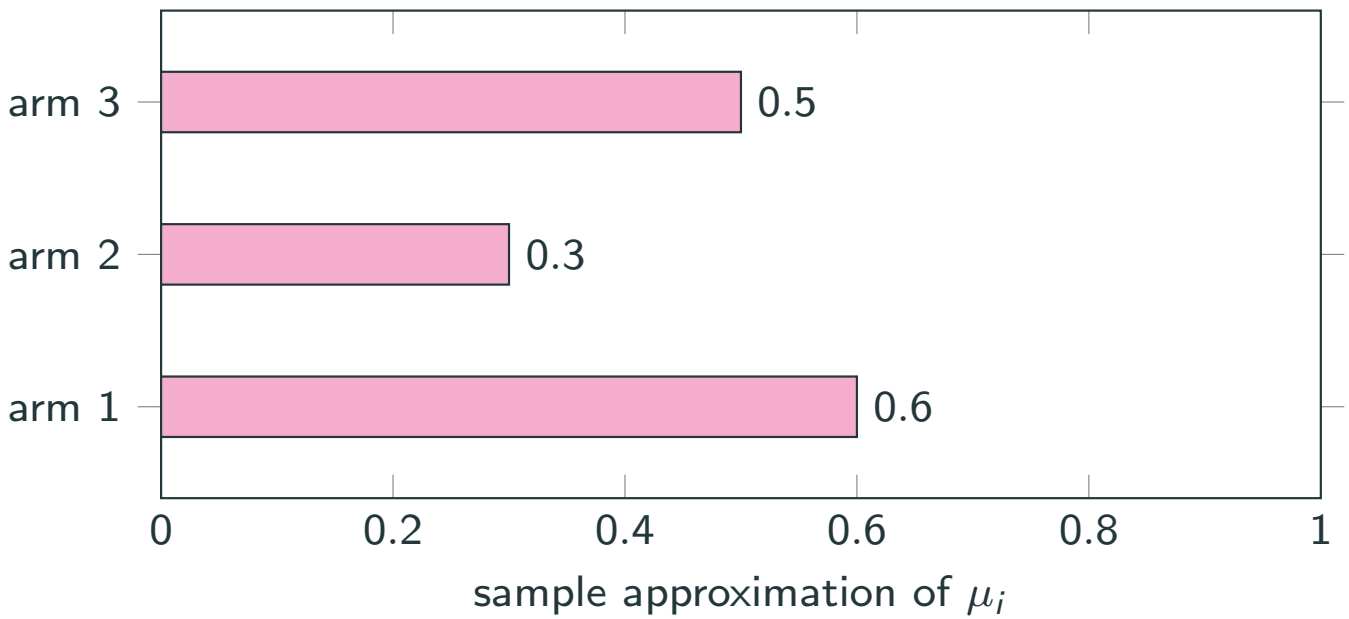
Bernoulli example

Bernoulli setting:

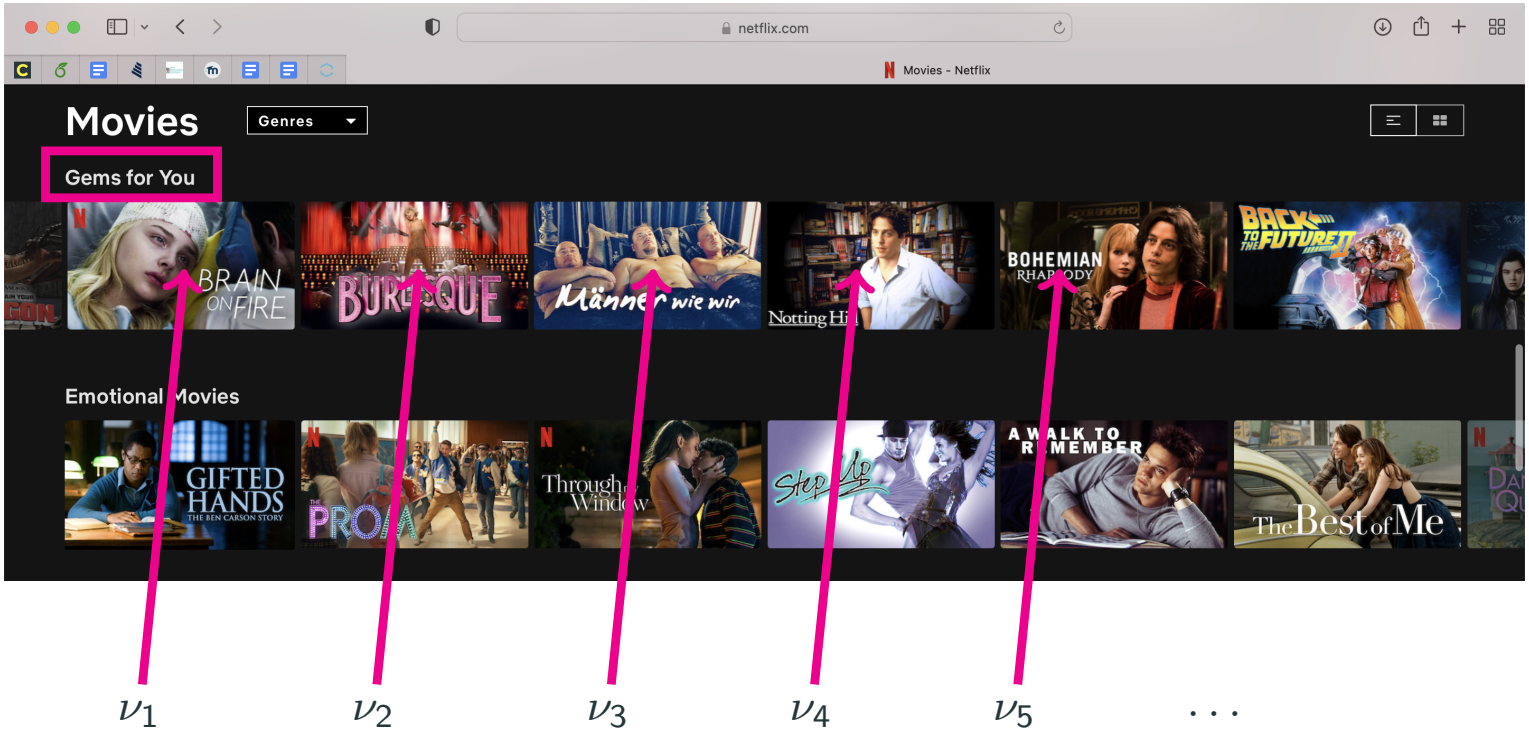
- $\{(\mathcal{B}(\mu_i))_i : \mu_i \in [0,1]\}$
- Reward when choosing **arm a** at **time t**

$$R_t = \begin{cases} 1 & \text{with probability } \mu_a \\ 0 & \text{with probability } 1 - \mu_a \end{cases}$$

Example $|\mathcal{A}| = 3$:



Example recommender system



For the t-th visit of the app/website

- recommend a movie a_t
- observe $R_t \sim \nu_{a_t}$ (e.g., a rating, number of clicks or of times watched)

Stochastic Bandit Problem

Let the the largest mean of all the arms be denoted

$$\mu^*(\nu) = \max_{a \in \mathcal{A}} \mu_a(\nu)$$

Regret

The T-period regret of the sequence of random actions $a_1, \ldots, a_T$ is the random variable

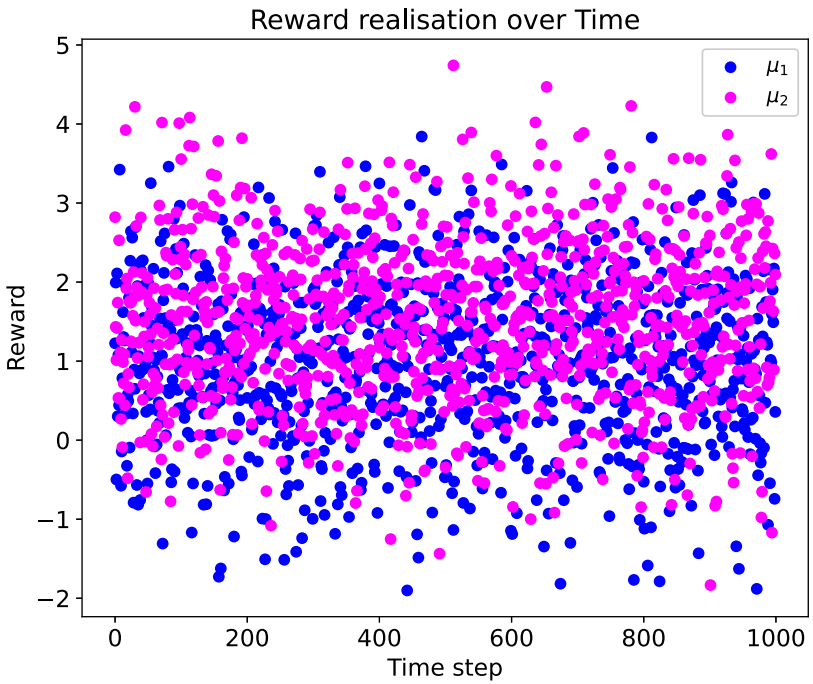
$$\mathcal{R}(\nu, T) = T\mu^*(\nu) - \mathbb{E}[\sum_{t=1}^{T} R_t] \tag{2}$$

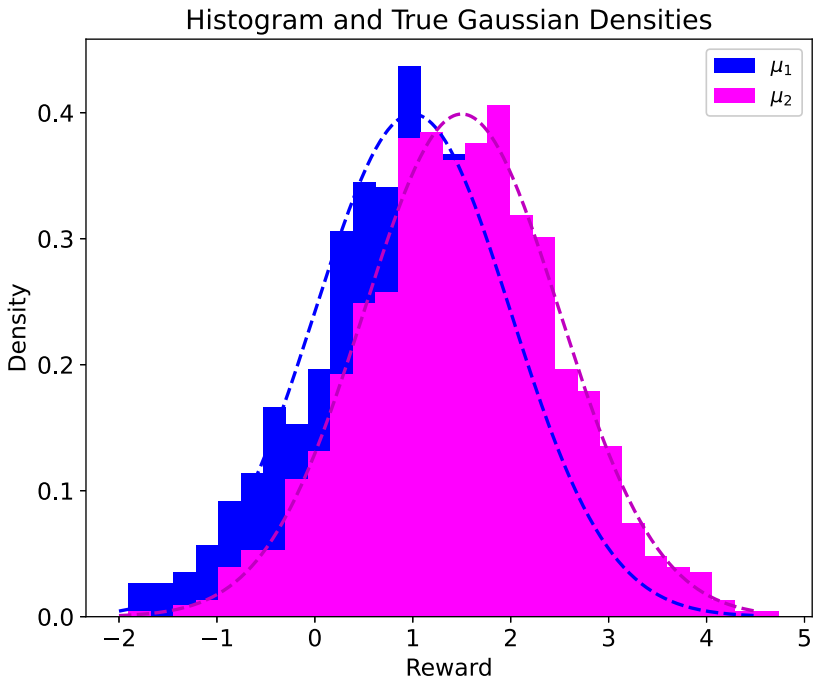
Goal: find a sequential sampling strategy

$$A_{t+1} = \pi_t(A_1, R_1, \ldots, A_t, R_t) \tag{3}$$

that minimises the regret

Structured stochastic bandits





Name	Symbol	Definition
Bernoulli	$\mathcal{E}_{\mathcal{B}}^k$	$\{(\mathcal{B}(\mu_i))_i : \mu \in [0,1]^k\}$
Uniform	$\mathcal{E}_{\mathcal{U}}^k$	$\{(\mathcal{U}(a_i, b_i))_i : a, b \in \mathbb{R}^k$ with $a_i \leq b_i$ for all $i\}$
Gaussian (known var.)	$\mathcal{E}_{\mathcal{N}}^k(\sigma^2)$	$\{(\mathcal{N}(\mu_i, \sigma^2))_i : \mu \in \mathbb{R}^k\}$
Gaussian (unknown var.)	$\mathcal{E}_{\mathcal{N}}^k$	$\{(\mathcal{N}(\mu_i, \sigma_i^2))_i : \mu \in \mathbb{R}^k$ and $\sigma^2 \in [0,\infty)^k\}$

Decomposition of the regret

We define:

- **action gap:** $\Delta_a(\nu) = \mu^* - \mu_a(\nu)$
- number of times action a was chosen by the learner

$$N_a(t) = \sum_{s=1}^{t} \mathbb{I}\{A_s = a\} \tag{4}$$

Decomposing the Regret

For any policy π and stochastic bandit environment ν with A finite and horizon $N \in \mathbb{N}$, the regret R_N of policy π in ν satisfies

$$\mathcal{R}(\nu, T) = \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}[N_a(T)]$$

Naive strategies

Uniform Exploration:

- choose each arm a for T/K times
- Exploration:

$$\mathcal{R}_\nu(\mathcal{A}, T) = \Big(\frac{1}{K} \sum_{a:\mu_a<\mu^*} \Delta_a\Big) T \tag{5}$$

Follow The Leader:

- Define the empirical estimate of the true unknown mean μ_a of an arm a at time t

$$\hat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} R_{a,s} \mathbb{I}(A_s = a) \tag{6}$$

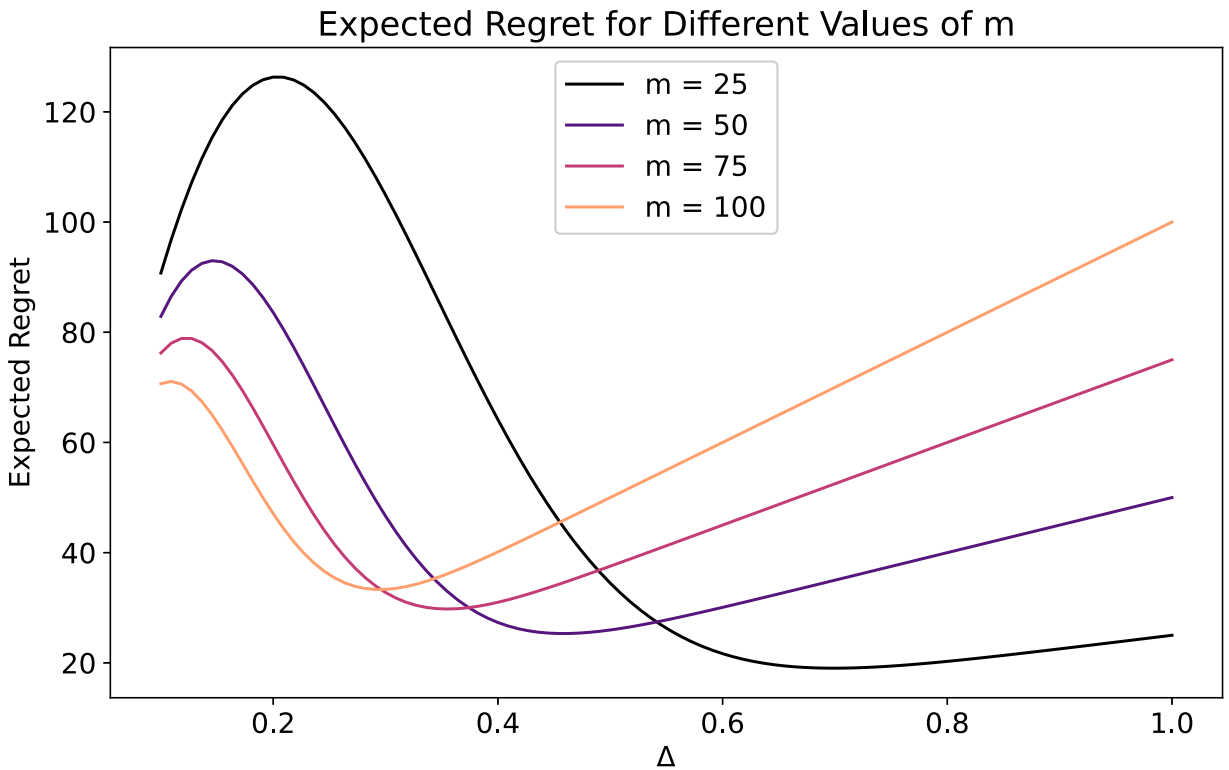
- chooses $a_{t+1} = \arg\max_{a\in\{1,\dots,K\}} \hat{\mu}_a(t)$
- Focus on exploitation but **no** exploration

Goal: develop an algorithm that balances Exploration and Exploitation

Explore-Then-Commit

Algorithm 1 Explore-Then-Commit

Initialization: Play each machine m times;
for $t = Km + 1 : T$ **do**
    Perform action $a_t = \arg\max_{a' \in \{1,\ldots,K\}} \hat{\mu}_{a'}(mK)$
end for



Variants of convergence

Let X be a random variable and $\{X_n\}_{n \in N}$ a sequence of random variables.

- $\{X_n\}$ converges to X almost surely, $X_n \xrightarrow{a.s.} X$, if

$$\mathbb{P}(\lim_{n \to \infty} X_n = X) = 1 \tag{7}$$

- $\{X_n\}$ converges to X in probability $X_n \xrightarrow{P} X$, if for any $\epsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}[|X_n - X| > \epsilon] = 0 \tag{8}$$

- $\{X_n\}$ converges to X in law (or in distribution), $X_n \xrightarrow{D} X$, if for any bounded continuous function f

$$\lim_{n \to \infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)] \tag{9}$$

Remark: $X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{P} X \implies X_n \xrightarrow{D} X$

Empirical estimates

Note:

- μ_n is an unbiased estimator, i.e., $\mathbb{E}[\mu_n] = \mu$
- $\mathbb{V}[\mu_n] = \frac{\mathbb{V}[X]}{n}$

Weak law of large numbers (WLLN): $\mu_n \xrightarrow{P} \mu$

Strong law of large numbers (SLLN): $\mu_n \xrightarrow{a.s.} \mu$

Central limit theorem (CLT): $\sqrt{n}(\mu_n - \mu) \xrightarrow{D} \mathcal{N}(0, \mathbb{V}[X])$

Note: (CLT) yields that $\frac{\sqrt{n}(\mu_n - \mu)}{\sigma}$ as $n \to \infty$ is a Gaussian with mean zero and unit variance. If $Z \sim N(0, 1)$, then

$$P(Z \geq u) = \int_u^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx.$$

Asymptotic confidence interval

For

$$P(Z \geq u) = \int_u^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx.$$

The integral has no closed-form solution, but is easy to bound:

$$\int_u^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \leq \int_u^\infty \frac{x}{u\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx = \frac{1}{\sqrt{2\pi}u^2} \exp\left(-\frac{u^2}{2}\right),$$

(The factor $\frac{x}{u}$ multiplies the density function, and it is always greater than or equal to 1 for $x \geq u$.)

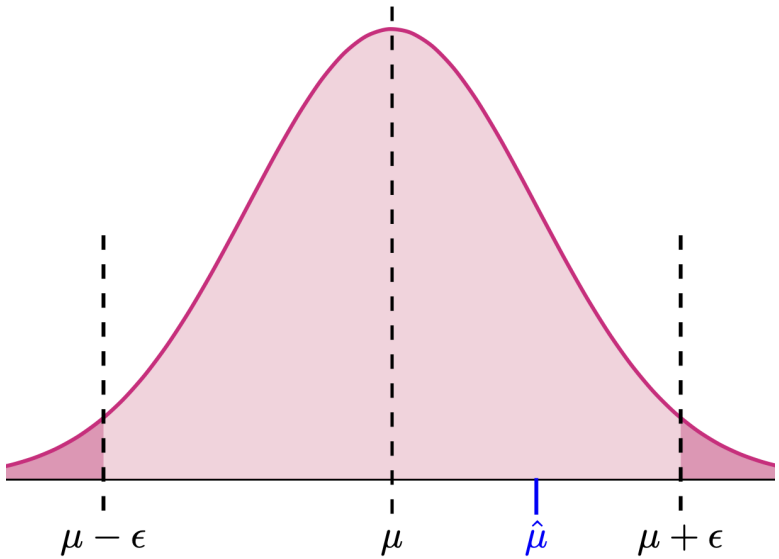
Problem: The asymptotic nature of the CLT makes it unsuitable for designing bandit algorithms.

Want: Non-asymptotic high probability bounds for μ_n and confidence interval

For non-asymptotic bounds, we can use concentration inequalities such as Hoeffding's or Bernstein's inequalities, which provide bounds on the probability that μ_n deviates from its expectation by a certain amount.

Understanding the tail probabilities

How accurately is the empirical estimate $\hat{\mu}$ approximating μ based on a set of samples?

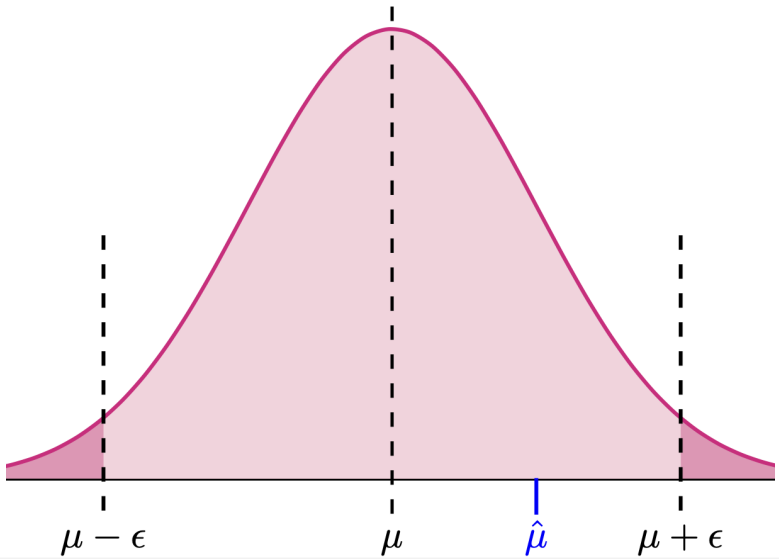


Goals:

- investigate tail probabilities of $\hat{\mu} - \mu$
- derive bounds on $\mathbb{P}(|\hat{\mu} - \mu| \geq \epsilon)$
- use this information to built new algorithms and derive bounds for regret

Concentration inequalities

Goal: investigate tail probabilities of $\hat{\mu} - \mu$ in other words $\mathbb{P}(|\hat{\mu} - \mu| \geq \epsilon)$



Markov inequality: For any positive random variable X and $\epsilon > 0$, the following holds:

$$\mathbb{P}(X \geq \epsilon) \leq \frac{\mathbb{E}(X)}{\epsilon} \qquad \text{(Markov)}$$

Proof:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x)\, dx = \int_{0}^{\infty} x f(x)\, dx = \int_{0}^{a} x f(x)\, dx + \int_{a}^{\infty} x f(x)\, dx \qquad (10)$$

$$\geq \int_{a}^{\infty} x f(x)\, dx \geq \int_{a}^{\infty} a f(x)\, dx = a \int_{a}^{\infty} f(x)\, dx = a \Pr(X \geq a) \qquad (11)$$

Concentration inequalities

Chebyshev inequality: Let X (integrable) be a random variable with finite non-zero variance σ^2 (and thus finite expected value μ).

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\mathbb{V}(X)}{\epsilon^2}$$

Proof: Apply Markov's inequality to the random variable $(X - \mu)^2$, this yields

$$\begin{aligned} \mathbb{P}(|X - \mu| \geq \epsilon) &= \mathbb{P}((X - \mu)^2 \geq \epsilon^2) && (12) \\ &\leq \frac{1}{\epsilon^2}\mathbb{E}[(X - \mu)^2] \quad \text{(Markov inequality)} && (13) \\ &= \frac{\mathbb{V}(X)}{\epsilon^2} && (14) \end{aligned}$$

□

Hoeffding inequality:

Let X be a centered random variable bounded in $[a, b]$. Then for any $s \in \mathbb{R}$

$$\mathbb{E}[e^{sX}] \leq e^{s^2(b-a)^2/8} \tag{15}$$

Hoeffding Inequality

Hoeffding inequality:

Let X be a centered random variable bounded in $[a, b]$. Then for any $s \in \mathbb{R}$

$$\mathbb{E}[e^{sX}] \leq e^{s^2(b-a)^2/8} \tag{16}$$

Reminder: convexity $f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$

Proof: From convexity of the exponential function, for any $a \leq x \leq b$,

$$e^{sx} \leq \underbrace{\frac{x-a}{b-a}}_{t} e^{sb} + \underbrace{\frac{b-x}{b-a}}_{(1-t)} e^{as} \tag{17}$$

where we use that

$$sx = s\Big(\frac{(x-a)}{b-a}b + \frac{(b-x)}{b-a}a\Big) \tag{18}$$

Recall that $\mathbb{E}[X] = 0$ and let $p = -a/(b-a)$ (note $a = -p(b-a)$) then

$$\mathbb{E}[e^{sX}] \leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} = (1-p)e^{sa} + pe^{sb} \tag{19}$$

Hoeffding Inequality

Continuation of proof:

$$= \left((1-p) + pe^{s(b-a)}\right)e^{sa} = (1-p+pe^{s(b-a)})e^{-ps(b-a)} \tag{20}$$

$$= (1-p+pe^{u})e^{-pu} = \exp(\phi(u)) \quad \text{with } u = s(b-a) \tag{21}$$

Note that

$$\phi(u) := -pu + \log(1-p+pe^{u}) \tag{22}$$

and corresponding derivative with respect to u is

$$\phi'(u) = -p + \frac{pe^{u}}{(1-p+pe^{u})} \tag{23}$$

and

$$\phi''(u) = \frac{p(1-p)e^{u}}{(1-p+pe^{u})^2}. \tag{24}$$

Hoeffding Inequality

Continuation of proof: Thus from Taylor's theorem, the exists a $\theta \in [0, u]$ such that

$$\phi(\theta) = \phi(0) + \theta\phi'(0) + \frac{u^2}{2}\phi''(\theta) \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8} \tag{25}$$

Note that $\phi(0) = \phi'(0) = 0$ and therefore it remains to maximize $\phi''(u)$. Substituting z for e^u we see that $\phi''(u)$ is concave for $z > 0$ as it is linear over quadratic. In order to determine the critical point of $\phi''(u)$ we compute

$$\begin{aligned}\frac{d}{dz} &= \frac{p(1-p)}{(1-p+pz)^2} - \frac{2p^2(1-p)z}{(1-p+pz)^3} = \frac{p(1-p)(1-p+p+z) - 2p^2(1-p)z)}{(1-p+pz)^3} \\ &= \frac{p^2(1-p)z - p^2(1-p)z - p^2(1-p)z + p(1-p)^2}{(1-p+pz)^3} = \frac{p(1-p)(1-p-zp)}{(1-p+pz)^3}\end{aligned}$$

The critical point is at $z = e^u = \frac{1-p}{p}$ and substituting yields

$$\phi''(u) \leq \frac{p(1-p)\cdot\frac{1-p}{p}}{(1-p+p\cdot\frac{1-p}{p})^2} = \frac{(1-p)^2}{4(1-p)^2} = \frac{1}{4} \tag{26}$$

Hoeffding Inequality

Continuation of proof: Inserting into the Taylor expansion leads to the bound

$$\phi(u) \leq \frac{1}{2}u^2 \cdot \frac{1}{4} = \frac{1}{8}s^2(b-a)^2 \tag{27}$$

This completes the proof of the lemma as we have

$$\mathbb{E}[e^{sX}] \leq e^{\phi(u)} \leq e^{\frac{s^2(b-a)^2}{8}} \tag{28}$$

□

Chernoff-Hoeffding Inequality

Chernoff-Hoeffding: Let $X_i \in [a_i, b_i]$ be n independent r.v. with mean $\mu_i = \mathbb{E}[X_i]$. Then

$$\mathbb{P}[|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu_i| > \epsilon] \leq 2\exp(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}) \tag{29}$$

Proof: Let $t > 0$ and $s > 0$

$$\mathbb{P}\Big(\sum_{i=1}^{n}(X_i - \mu_i) \geq t\Big) = \mathbb{P}(e^{s\sum_{i=1}^{n} X_i - \mu_i} \geq e^{st}) \tag{30}$$

$$\leq e^{-st}\mathbb{E}[e^{s\sum_{i=1}^{n} X_i - \mu_i}] \quad \text{(Markov inequality)} \tag{31}$$

$$= e^{-st}\prod_{i=1}^{n}\mathbb{E}[e^{s(X_i - \mu_i)}] \quad \text{(independent random variables)} \tag{32}$$

$$\leq e^{-st}\prod_{i=1}^{n} e^{s^2(b_i - a_i)^2/8} \quad \text{(Hoeffding inequality)} \tag{33}$$

$$= e^{-st + \sum_{i=1}^{n} s^2(b_i - a_i)^2/8} \tag{34}$$

If we choose $s = 4t/(\sum_{i=1}^{n}(b_i - a_i)^2)$ and $t = \epsilon n$ which are both larger than 0 the result follows. Similar arguments hold for $\mathbb{P}(\sum_{i=1}^{n} X_i - \mu_i \leq -t)$.

Bernstein Inequality

Idea: want to obtain sharper Inequalities using more information of the random variables (e.g., Hoeffding's inequality only uses that random variables are bounded)

Bernstein inequality: Let $X_1 \ldots, X_n$ be independent random variables with $\mathbb{P}[|X_i| \leq c]$ and $\mathbb{E}[X_i] = \mu$ then for any $\epsilon > 0$

$$\mathbb{P}[|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu_i| > \epsilon] \leq 2\exp(-\frac{n^2\epsilon^2}{2\sigma^2 + 2c\epsilon/3}) \tag{35}$$

where

$$\frac{1}{n}\sum_{t=1}^{n} \mathrm{Var}[X_i] = \sigma^2 \tag{36}$$

Auxiliary Lemma: Let X be a random variable with $|X| \leq c$ and $\mathbb{E}[X] = 0$ then for any $t > 0$

$$\mathbb{E}[e^{tX}] \leq \exp\left\{t^2\sigma^2\left(\frac{e^{tc} - 1 - tc}{(tc)^2}\right)\right\} \tag{37}$$

where $\mathrm{Var}[X] = \sigma^2$.

Note: If the variance of X_i is small, then we can get a sharper inequality from Bernstein's inequality.

Subgaussian Random Variables

Subgaussianity

A random variable X is σ-subgaussian if for all $\lambda \in \mathbb{R}$, it holds that

$$\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\lambda^2 \sigma^2 / 2\right)$$

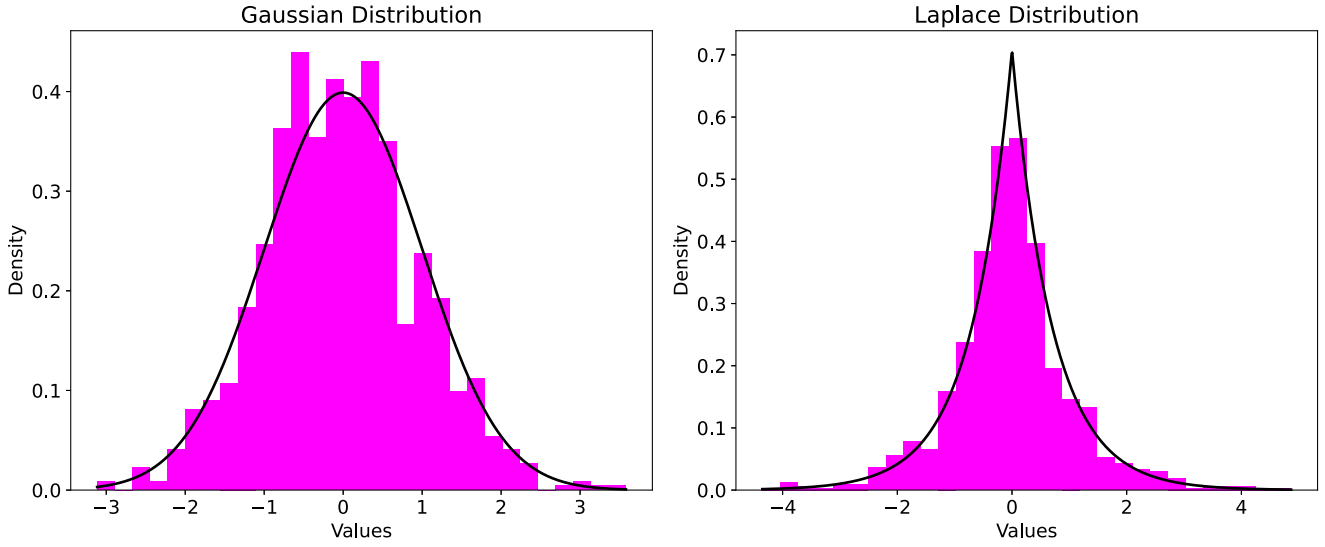
Theorem: If X is σ-subgaussian, then for any $\epsilon \geq 0$

$$\mathbb{P}(X \geq \epsilon) \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$$

Proof: Let $\lambda > 0$, then

$$\begin{aligned}
\mathbb{P}(X \geq \epsilon) &= \mathbb{P}(\exp(\lambda X) \geq \exp(\lambda \epsilon)) \\
&\leq \mathbb{E}[\exp(\lambda X)] \exp(-\lambda \epsilon) \quad \text{(Markov's inequality)} \\
&\leq \exp(0.5\lambda^2 \sigma^2 - \lambda \epsilon) \quad \text{(subgaussianity)} \\
&= \exp(-\epsilon^2 / 2\sigma^2) \quad \text{(choose } \lambda = \epsilon/\sigma^2\text{)}
\end{aligned}$$

Subgaussian Random Variables



Lemma: Suppose that X is σ-subgaussian and X_1 and X_2 are independent and σ_1 and σ_2-subgaussian, respectively, then:

(a) $\mathbb{E}[X] = 0$ and $Var(X) \leq \sigma^2$.

(b) cX is $|c|\sigma$-subgaussian for all $c \in \mathbb{R}$.

(c) $X_1 + X_2$ is $\sqrt{\sigma_1^2 + \sigma_2^2}$-subgaussian.

Note: It holds that $\hat{\mu} - \mu = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)$ is $\frac{\sigma}{\sqrt{n}}$-subgaussian.

Explore-Then-Commit

Algorithm 2 Explore-Then-Commit

Initialization: Play each machine m times;
for $t = Km + 1 : T$ **do**
    Perform action $a_t = \arg\max_{a' \in \{1,\dots,K\}} \hat{\mu}_{a'}(mK)$
end for

Analysis for two arms

Let $\mu_1 > \mu_2$ and let $\Delta := \mu_1 - \mu_2$

$$\begin{aligned}
\mathcal{R}_\nu(\mathcal{A}, T) &= \Delta \mathbb{E}[N_2(T)] = \Delta m + (T - 2m)\Delta \mathbb{P}(A_{Km+1} = 2) \\
&\leq \Delta m + \Delta T \times \mathbb{P}[\hat{\mu}_2(Km) \geq \hat{\mu}_1(Km)] \\
&\leq \Delta m + \Delta T \times \mathbb{P}[\underbrace{\hat{\mu}_2(Km) - \mu_2 - (\hat{\mu}_1(Km) - \mu_1)}_{\sqrt{2/m}-\text{subgaussian}} \geq \Delta] \\
&\leq \Delta m + \Delta T \times \exp(-m\Delta^2/4)
\end{aligned}$$

Optimal m

Derived bound for ETC-algorithm for two arms:

$$\mathcal{R}_\nu(\mathcal{A}, T) \leq m\Delta + T\Delta \exp\left(-\frac{m\Delta^2}{T}\right). \tag{6.4}$$

For large T, the quantity on the right-hand side of Eq. (6.4) is minimised, up to a possible rounding error, by choosing

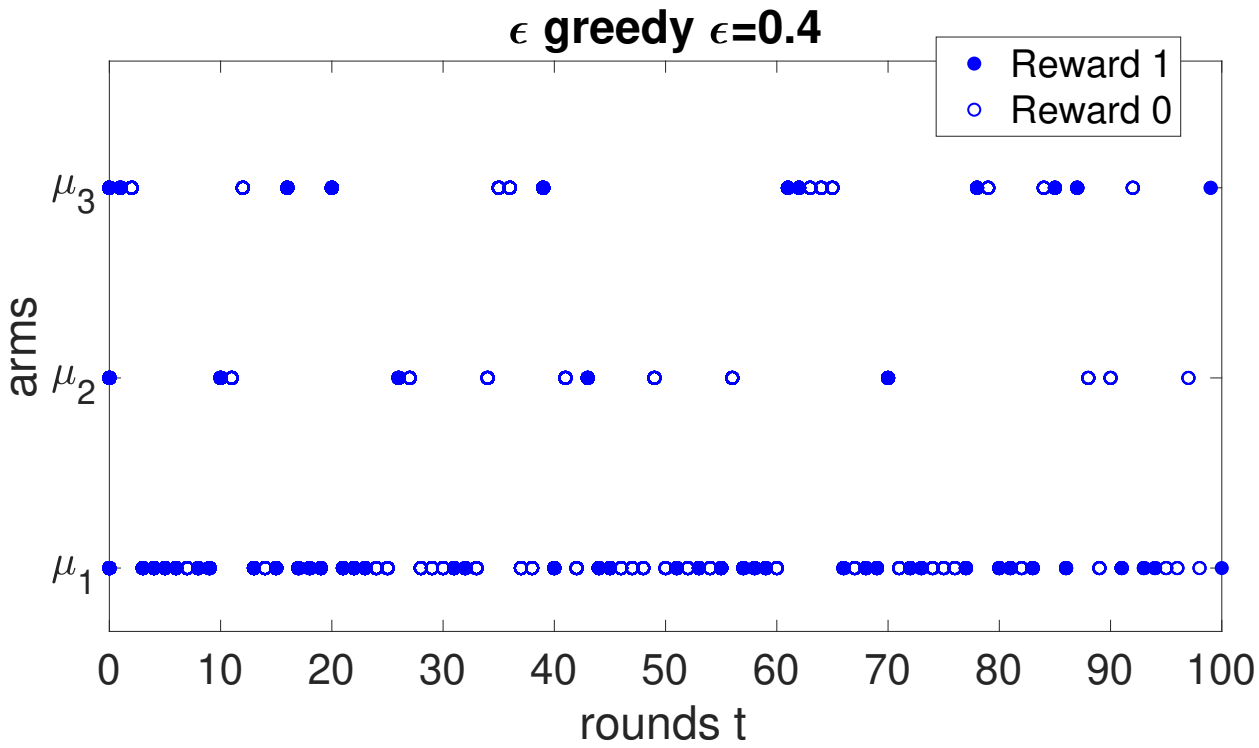
$$m = \max\left(1, \left\lfloor \frac{4}{\Delta^2} \log\left(\frac{T\Delta^2}{4}\right) \right\rfloor\right),$$

where $\lfloor \cdot \rfloor$ denotes the floor function, which rounds down to the nearest integer.

ϵ-greedy:

In each round t: choose arm according to

$$A_t = \begin{cases} a_k & \text{with probability } \epsilon/K \\ \arg\max_{a' \in \{1,\ldots,K\}} \hat{\mu}_{a'}(t) & \text{with probability } 1-\epsilon \end{cases}$$

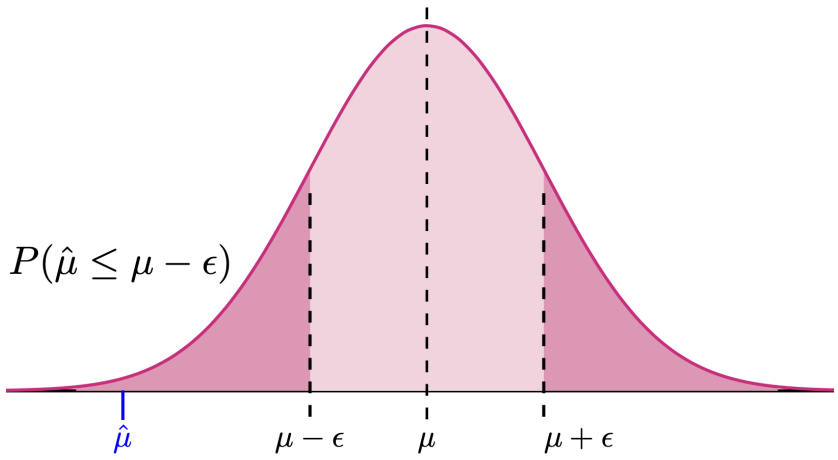


Confidence Bounds

Corollary: Let $X_i - \mu$ be independent and σ-subgaussian for all i. Then

$$\mathbb{P}(\hat{\mu} \geq \mu + \epsilon) \leq \underbrace{\exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right)}_{\delta}$$

$$\mathbb{P}(\hat{\mu} \leq \mu - \epsilon) \leq \underbrace{\exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right)}_{\delta}$$



for any $\epsilon \geq 0$.

Then we have

$$\hat{\mu} - \underbrace{\sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}}_{\epsilon} \leq \mu \leq \hat{\mu} + \underbrace{\sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}}_{\epsilon} \tag{38}$$

with probability at least $1 - \delta$

UCB

Algorithm 3 UCB(δ)

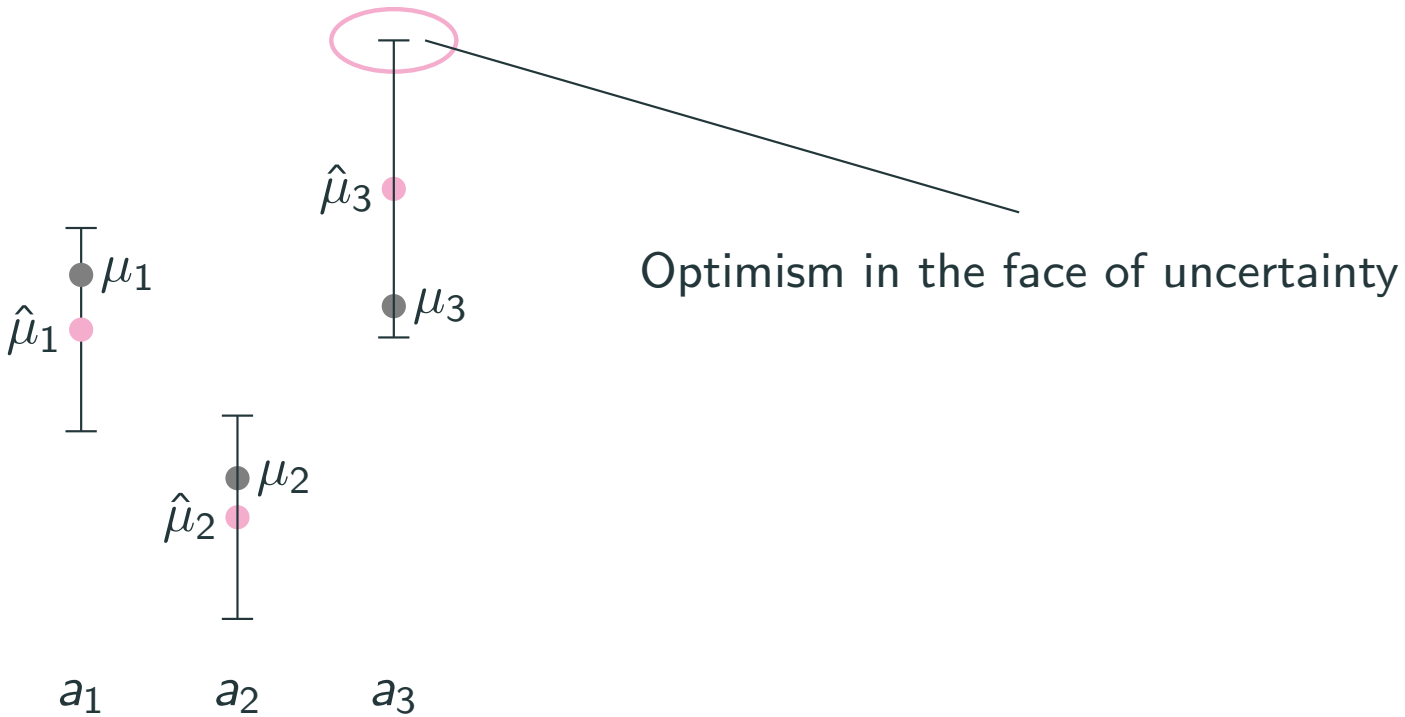
Initialization: Play each machine once;

for $t = 1, 2, 3....$ **do**

Perform action $a_{t+1} = \arg\max_{a \in \mathcal{A}} \hat{\mu}_a(t) + \sqrt{\frac{2 \log(1/\delta)}{N_t(a)}}$

Update $\hat{\mu}_{a+1}(t+1)$ and $N_{t+1}(a+1)$

end for



UCB1

Algorithm 4 UCB1

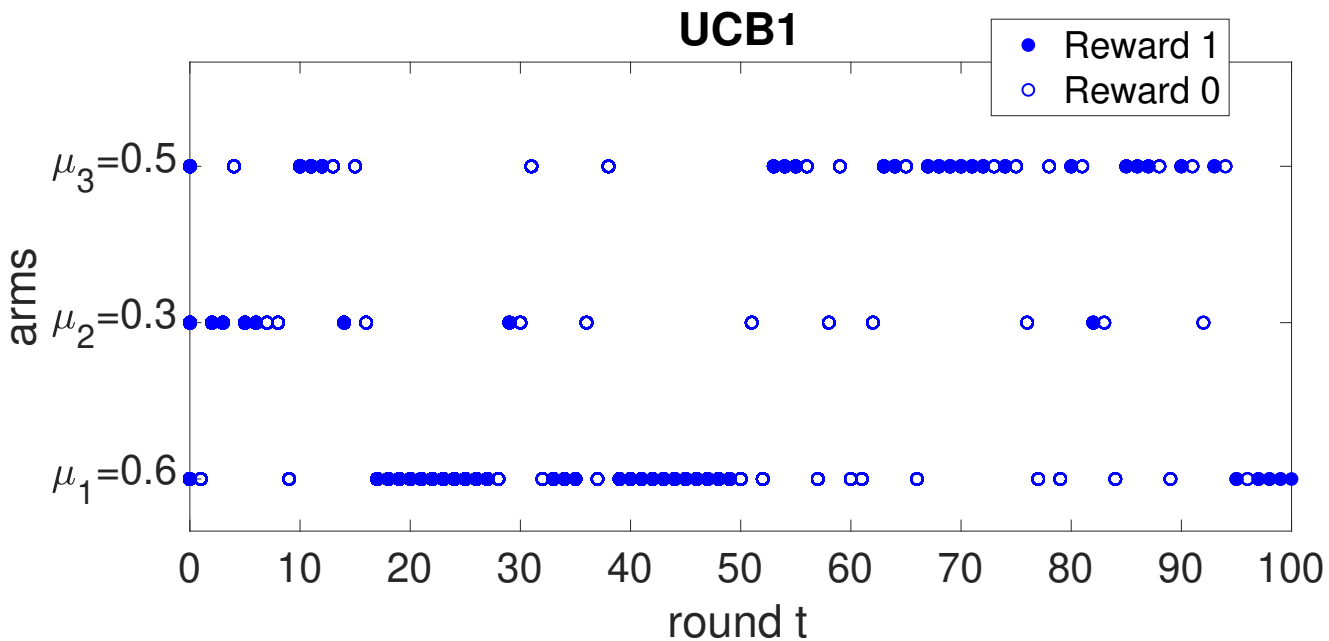
Initialization: Play each machine once;

for $t = 1, 2, 3....$ **do**

Perform action $a_{t+1} = \arg\max_{a \in \mathcal{A}} \hat{\mu}_a(t) + \sqrt{\frac{2\log(1/\delta)}{N_t(a)}}$

Update $\hat{\mu}_{a+1}(t+1)$ and $N_{t+1}(a+1)$

end for



Regret bound UCB

Idea: one potential option is to choose $1/\delta$ adaptively with to or according to the fixed horizon with $1/\delta = T^2$

Theorem: Consider the UCB1 on a stochastic k-armed 1-subgaussian bandit problem. For any horizon T, with $\delta = 1/T^2$, then

$$\mathcal{R}_T \leq 3\sum_{i=1}^{k} \Delta_i + \sum_{i:\Delta_i>0} \frac{16\log(T)}{\Delta_i} \tag{39}$$

Proof: Without loss of generality, we assume $\mu_1 = \mu^*$ and note that

$$\mathcal{R}_T = \sum_{a=1}^{k} \Delta_a \mathbb{E}[N_a(T)] \tag{40}$$

Define

$$G_a = \left\{\mu_1 < \min_{t\in[T]} UCB_1(t,\delta)\right\} \cap \left\{\hat{\mu}_{a,u_a} + \sqrt{\frac{2}{u_a}\log\left(\frac{1}{\delta}\right)} < \mu_1\right\} \tag{41}$$

where $u_a \in [T]$ is a constant to be chosen later.

Regret bound UCB

Proof: Aim is to show:

1. If G_a occurs, then arm a will be played at most u_a times: $N_a(T) \leq u_a$.
2. The complement event G_a^c occurs with low probability (governed in some way yet to be discovered by u_a)

In case these are true and since we know that $N_a(T) \leq T$ it follows that

$$\mathbb{E}[N_a(T)] = \mathbb{E}[\mathbb{I}\{G_a\}N_a(T)] + \mathbb{E}[\mathbb{I}\{G_a^c\}N_a(T)] \leq u_a + \mathbb{P}(G_a^c)T \tag{42}$$

Goal: Show that $\mathbb{P}(G_a^c)$ small and that $N_a(T) \leq u_a$. Under the assumption that G_a holds suppose that $N_a(T) > u_a$, i.e., there exists a round $t \in [T]$ where $N_a(t-1) = u_a$ and $A_t = a$. Using the definition of G_a

$$UCB_a(t-1,\delta) = \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_{t-1}(a)}} \tag{43}$$

$$= \hat{\mu}_{au_a} + \sqrt{\frac{2\log(1/\delta)}{u_a}} < \mu_1 < UCB_1(t-1,\delta) \tag{44}$$

Hence $A_t = \text{argmax}_j\, UCB_j(t-1,\delta) \neq a$ which is a contradiction.

Regret bound UCB

Let us know turn to upper bounding $\mathbb{P}(G_a^c)$. By its definition

$$G_a^c = \Big\{\mu_1 \geq \min_{t\in[T]} UCB_1(t,\delta)\Big\} \cup \Big\{\hat{\mu}_{a,u_a} + \sqrt{\frac{2}{u_a}\log\Big(\frac{1}{\delta}\Big)} \geq \mu_1\Big\} \tag{45}$$

The first of these sets is decomposed using the definition of $UCB_1(t,\delta)$,

$$\Big\{\mu_1 \geq \min_{t\in[T]} UCB_1(t,\delta)\Big\} \subset \Big\{\mu_1 \geq \min_{s\in[T]} \Big(\hat{\mu}_{1s} + \sqrt{\frac{2\log(1/\delta)}{s}}\Big)\Big\} \tag{46}$$

$$= \bigcup_{s\in[T]} \Big\{\mu_1 \geq \hat{\mu}_{1,s} + \sqrt{\frac{2\log(1/\delta)}{s}}\Big\} \tag{47}$$

Then using a union bound and the concentration bound for sums of independent subgaussian random variables we obtain

$$\mathbb{P}\Big(\mu_1 \geq \min_{t\in[T]} UCB_1(t,\delta)\Big) \leq \mathbb{P}\Big(\bigcup_{s\in[T]} \Big\{\mu_1 \geq \hat{\mu}_{1,s} + \sqrt{\frac{2\log(1/\delta)}{s}}\Big\}\Big) \tag{48}$$

$$\sum_{s=1}^{T} \mathbb{P}\Big(\mu_1 \geq \hat{\mu}_{1,s} + \sqrt{\frac{2\log(1/\delta)}{s}}\Big) \leq T\delta \tag{49}$$

Regret bound UCB

The next step is to bound the probability of the second set in (45). Assume that u_a is chosen large enough so that

$$\Delta_a - \sqrt{\frac{2\log(1/\delta)}{u_a}} \geq c\Delta_a \tag{50}$$

for some $c \in (0,1)$ to be chosen later. Then since $\mu_1 = \mu_a + \Delta_a$, and using results for subgaussian densities

$$\mathbb{P}\Big(\hat{\mu}_{au_a} + \sqrt{\frac{2\log(1/\delta)}{u_a}} \geq \mu_1\Big) = \mathbb{P}\Big(\hat{\mu}_{au_a} - \mu_a \geq \Delta_a - \sqrt{\frac{2\log(1/\delta)}{u_a}}\Big) \tag{51}$$

$$\leq \mathbb{P}\Big(\hat{\mu}_{au_a} - \mu_a \geq c\Delta_a\Big) \leq \exp\Big(-\frac{u_a c^2 \Delta_a^2}{2}\Big) \tag{52}$$

Combining the bounds yields:

$$\mathbb{P}(G_a^c) \leq T\delta + \exp\Big(-\frac{u_a c^2 \Delta_a^2}{2}\Big) \tag{53}$$

and inserting in (42)

$$\mathbb{E}[N_a(T)] \leq u_a + \Big(T\delta + \exp\Big(-\frac{u_a c^2 \Delta_a^2}{2}\Big)\Big)T \tag{54}$$

It remains to choose $u_a \in [T]$ satisfying (50). A natural choice is the smallest integer for which the eq. holds, which is

$$u_a = \left\lceil \frac{2\log(1/\delta)}{(1-c)^2\Delta_a^2} \right\rceil \tag{55}$$

Regret bound UCB

Note that for this choice of u_a can be larger than T, yet in this case the (54) holds immediately as $N_a(T) \leq T$. Inserting the choice of u_a and $\delta = 1/T^2$ in (54) yields

$$\mathbb{E}[N_a(T)] \leq \left\lceil \frac{2\log(1/\delta)}{(1-c)^2\Delta_a^2} \right\rceil + \left(T\delta + \exp\left(-\frac{u_a c^2 \Delta_a^2}{2} \right) \right) T \tag{56}$$

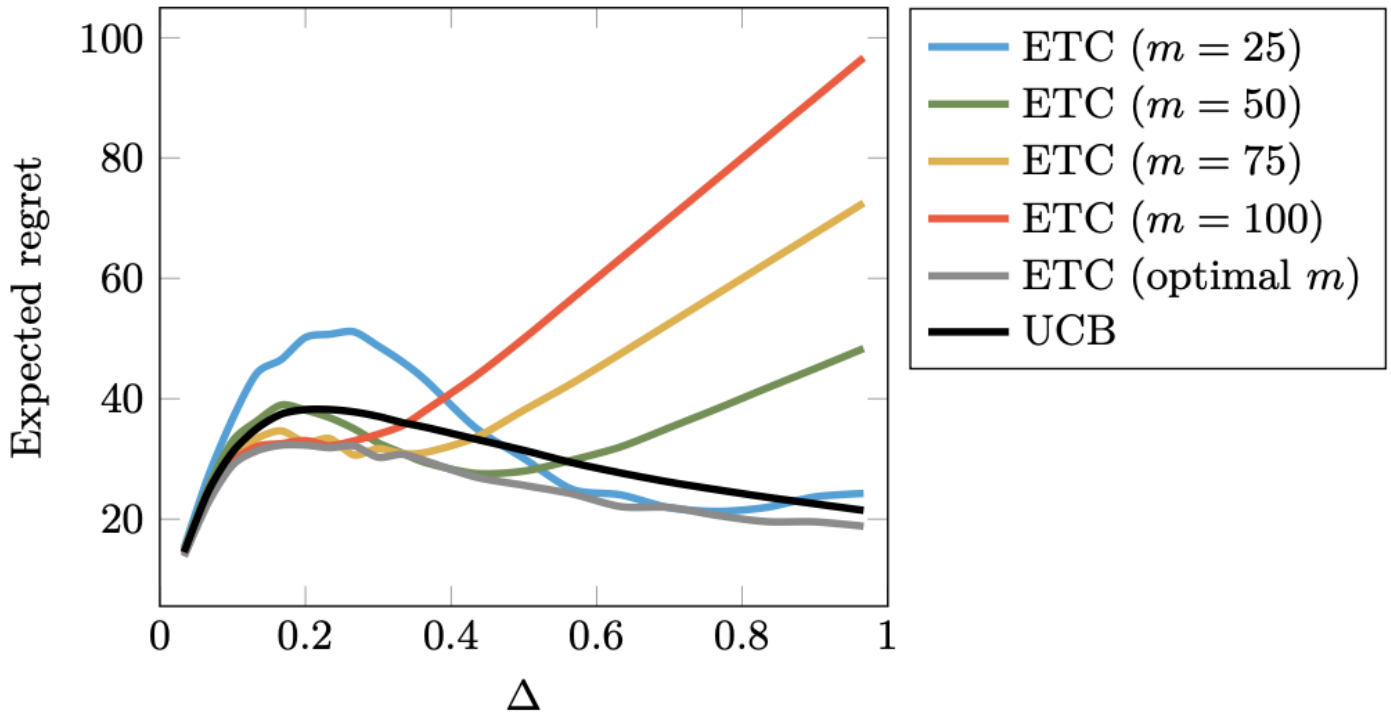
$$= \left\lceil \frac{2\log(T^2)}{(1-c)^2\Delta_a^2} \right\rceil + 1 + T^{1-2c^2/(1-c)^2} \tag{57}$$

All that remains is to choose $c \in (0,1)$. The second term will contribute a polynomial dependence on T unless $2c^2/(1-c)^2 \geq 1$. However, if c is chosen too close to 1, then the first term blows up. Somewhat arbitrarily we choose $c = 1/2$, which leads to

$$\mathbb{E}[N_a(T)] \leq 3 + \frac{16\log(T)}{\Delta_a^2} \tag{58}$$

□

Regret bound UCB



Example with $T = 1000$ and $k = 2$ and unit variance Gaussian rewards with means 0 and Δ respectively.

Lower bounds

Ideas:

- For any policy you give me, I will give you an instance of a bandit problem ν on which the regret is at least L
- If you give me a *reasonable* policy, then its regret on any instance ν is at least $L(\nu)$

Definition: For a a policy π on a set of stochastic bandit environments $\mathcal{E}$ the worst-case regret is defined via

$$\mathcal{R}_T(\pi, \mathcal{E}) = \sup_{\nu \in \mathcal{E}} \mathcal{R}(\pi, \nu) \tag{59}$$

Definition: The minimax regret is

$$\mathcal{R}_T^*(\pi, \mathcal{E}) = \inf_{\pi \in \Pi} \mathcal{R}_T(\pi, \mathcal{E}) = \inf_{\pi \in \Pi} \sup_{\nu \in \mathcal{E}} \mathcal{R}(\pi, \nu) \tag{60}$$

where Π is the set of all policies.

Lower bounds

Kullback-Leibler divergence: for different bandit models ν_μ and $\nu_{\mu'}$ parameterized through μ and μ' we define

$$\text{div}_{KL}(\nu_\mu, \nu_{\mu'}) = \mathbb{E}\Big[\log \frac{d\nu_\mu}{d\nu_{\mu'}}(X)\Big] \tag{61}$$

Examples:

- for Bernoulli bandits:

$$div_{KL}(\mu, \mu') := \mu \log\left(\frac{\mu}{\mu'}\right) + (1-\mu)\log\left(\frac{(1-\mu)}{(1-\mu')}\right) \tag{62}$$

- for Gaussian bandits (different means but the same variance σ^2)

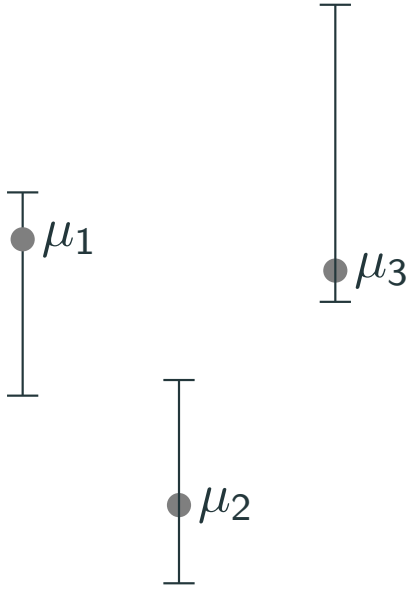
$$\text{div}_{KL}(\mu, \mu') := \frac{(\mu-\mu')^2}{2\sigma^2} \tag{63}$$

Lai and Robbins lower bound: For uniformly good algorithm

$$\mu_a < \mu^* \implies \lim_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log(T)} \geq \frac{1}{\text{div}_{KL}(\mu_a, \mu^*)} \tag{64}$$

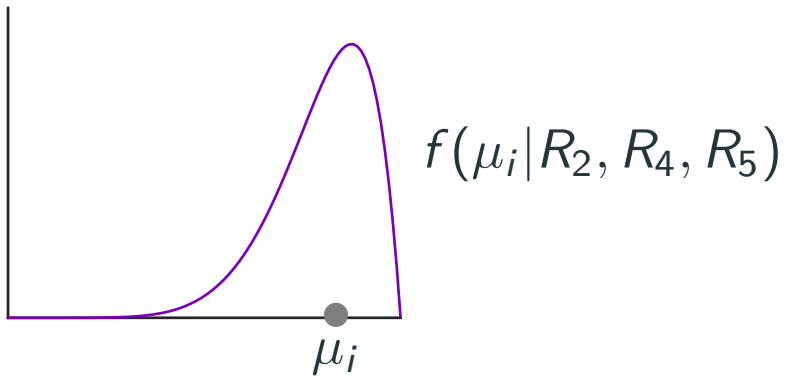
Bayesian approach

So far: Frequentist approximation of the statistics such as the mean via MLE estimators



Bayesian approach

Use posterior densities to describe the uncertainty



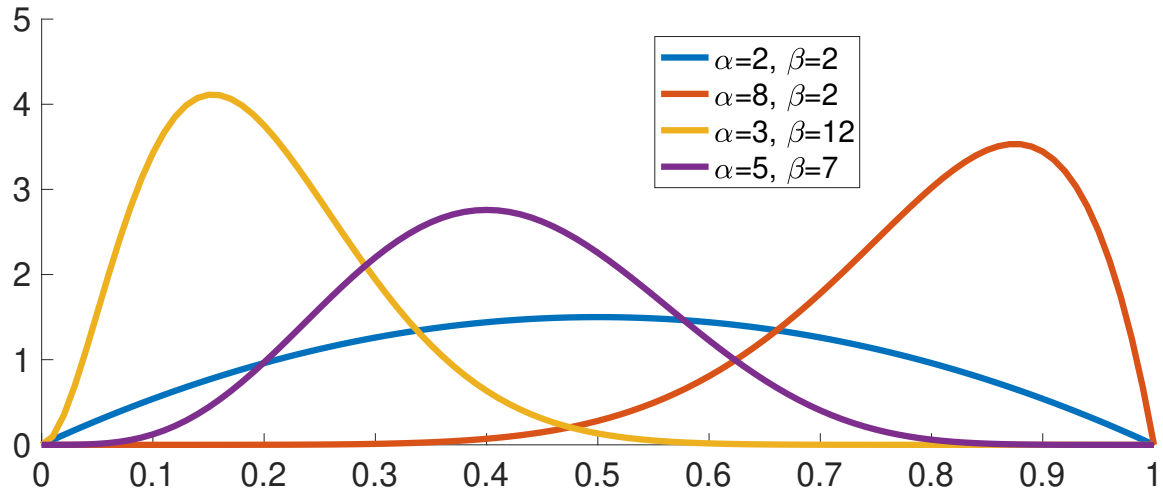
Prior distribution

Beta distribution

For α and β larger than zero and

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

with $\Gamma(n) = (n-1)!$ being the gamma function.



Thompson Sampling

Algorithm 5 Thompson Sampling

Initialization: Play each machine once;

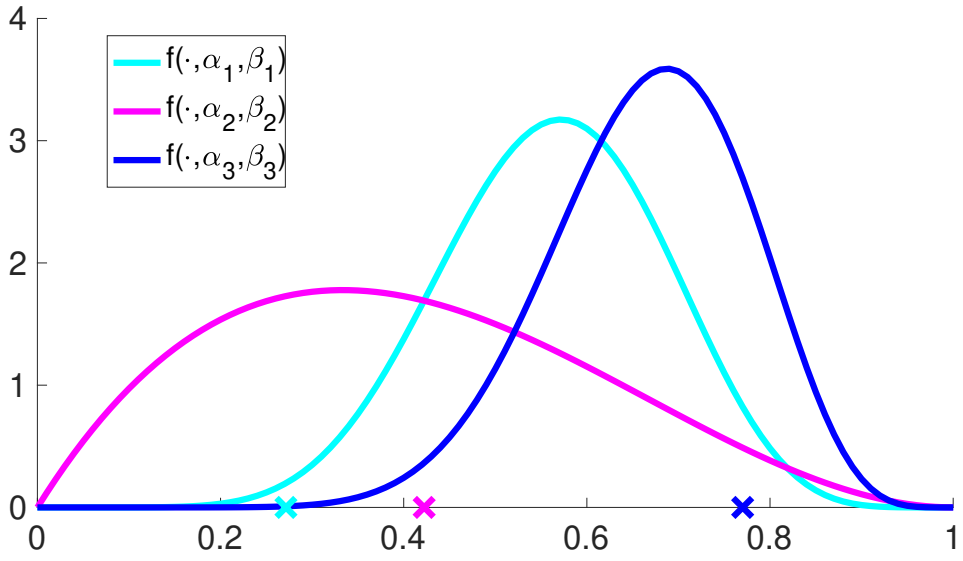
for $t = 1, 2, 3....$ **do**

Set $\alpha_a = \sum_{s=1}^{t} R_{s,a}\mathbb{I}(A_s = a) + 1$ and $\beta_a = N_a(t) - \alpha_a + 1$

Draw $x_t(a) \sim f(\cdot, \alpha_a, \beta_a)\ \forall a$

Choose the action $a_t = \arg\max_{a'} x_t(a')$;

end for



Thompson Sampling

Algorithm 6 Thompson Sampling

Initialization: Play each machine once;

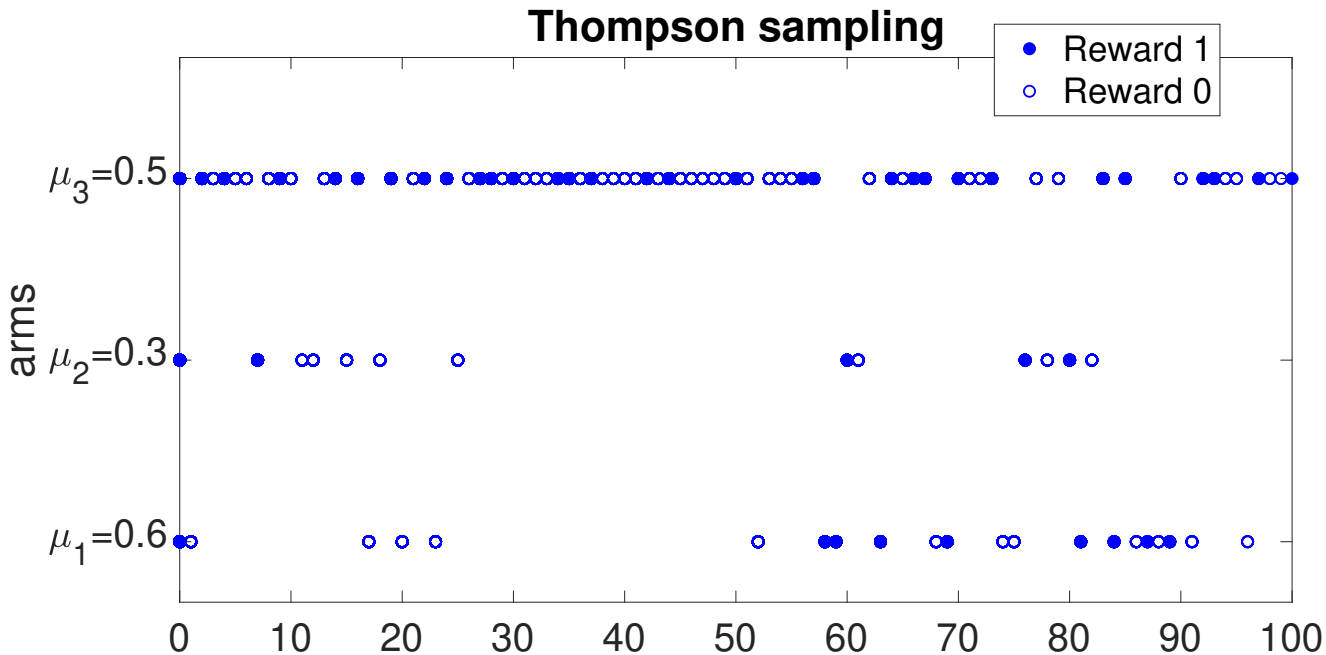
for $t = 1, 2, 3....$ **do**

Set $\alpha_a = \sum_{s=1}^{t} R_{s,a} \mathbb{I}(A_s = a) + 1$ and $\beta_a = N_a(t) - \alpha_a + 1$

Draw $x_t(a) \sim f(\cdot, \alpha_a, \beta_a)\ \forall a$

Choose the action $a_t = \arg\max_{a'} x_t(a')$;

end for



Regret bounds for Thompson Sampling

Problem-dependent regret: For all $\epsilon > 0$

$$\mathbb{E}_\mu[N_a(T)] \leq (1+\epsilon)\frac{1}{\mathrm{div}_{KL}(\mu, \mu^*)}\log(T) + o(\log(T)) \tag{65}$$

This results holds :

- for Bernoulli bandits, with a uniform prior [Kaufmann et al., 2012, Agrawal and Goyal, 2013]
- for Gaussian bandits, with Gaussian prior [Agrawal and Goyal, 2017]
- for exponential family bandits, with Jeffrey's prior [Korda et al., 2013]

Problem-independent regret: [Agrawal and Goyal, 2017] For Bernoulli and Gaussian bandits, Thompson Sampling satisfies

$$\mathcal{R}_T \leq \mathcal{O}\left(\sqrt{KT\log(T)}\right) \tag{66}$$