Statistical Data Analysis

Jana de Wiljes

wiljes@uni-potsdam.de

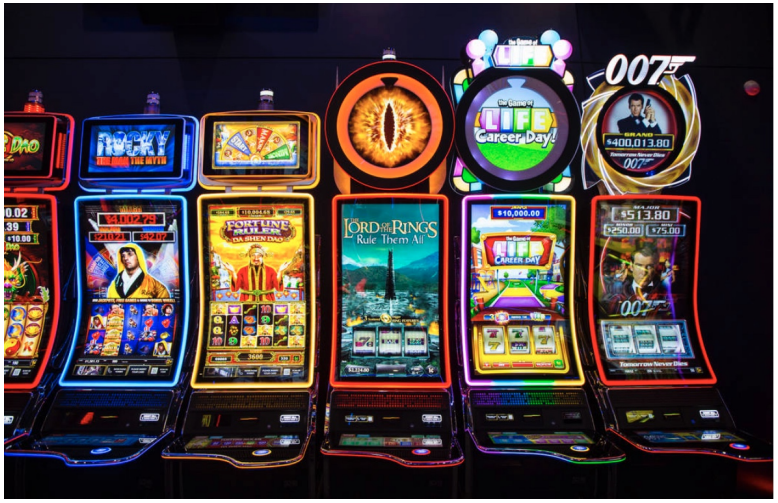
www.dewiljes-lab.com

January 4, 2023

Multi-Armed Stochastic Bandits

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

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

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

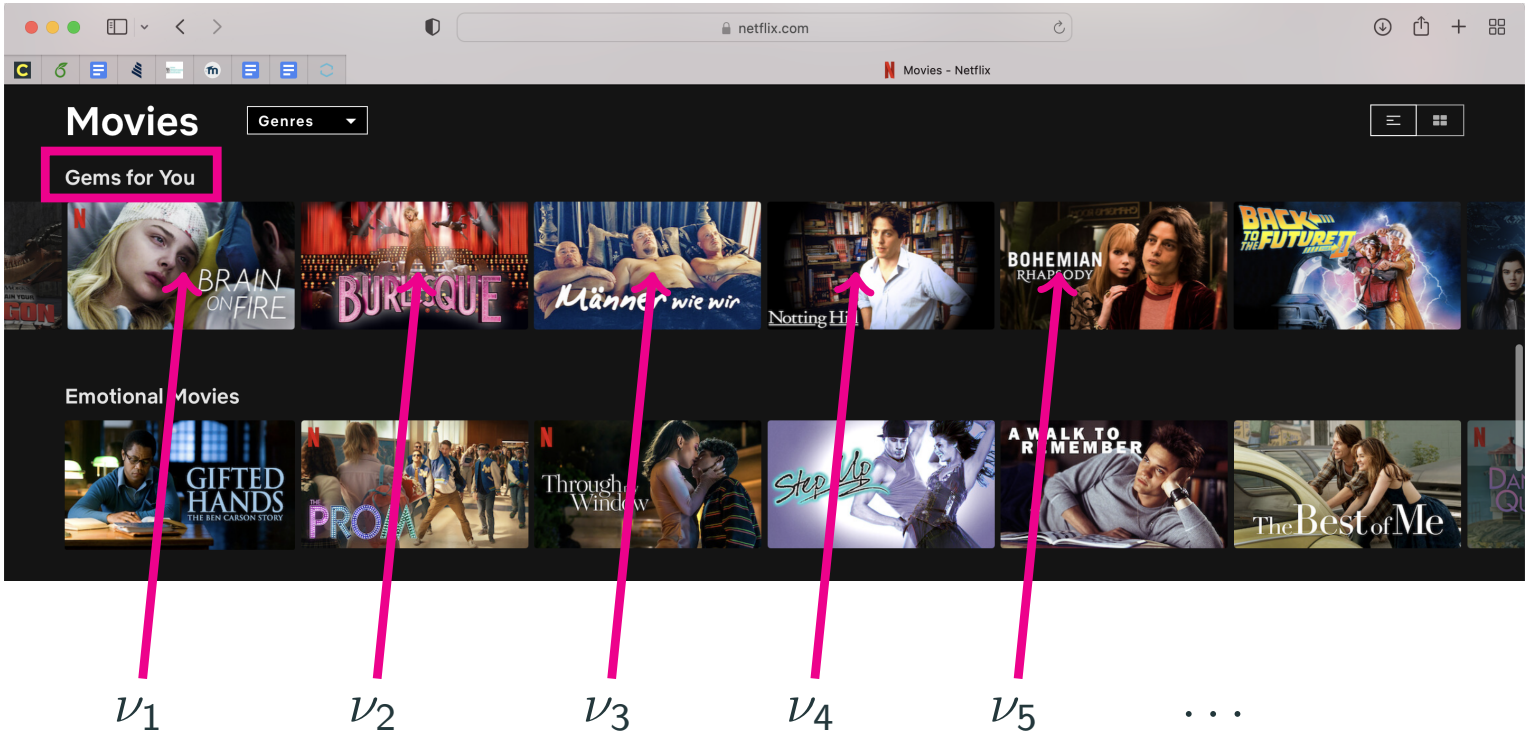
Bernoulli setting:

- $\{(\mathcal{B}(\mu_i))_i : \mu_i \in [0,1]\}$
- 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 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

$$\mathcal{R}_T(\nu, a_1, \ldots, a_T) = T\mu^*(\nu) - \mathbb{E}[\sum_{t=1}^{T} R_t] \tag{2}$$

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 $\mathcal{R}_T(\nu, \pi)$

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 $\mathcal{A}$ finite and horizon $T \in \mathbb{N}$, the regret $\mathcal{R}_T$ of policy π in ν satisfies

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

Naive strategies

Uniform Exploration:

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

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

Follow The Leader:

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

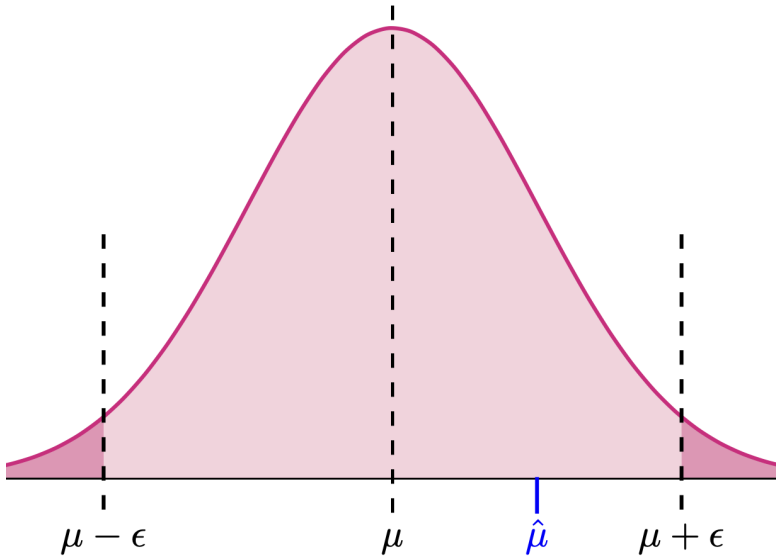
$$\hat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} R_{a,s} \mathbb{I}(A_s = a) \tag{6}$$

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

Goal: develop an algorithm that balances Exploration and Exploitation

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

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

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

Explore-Then-Commit

Algorithm 1 Explore-Then-Commit

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

Analysis for two arms

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

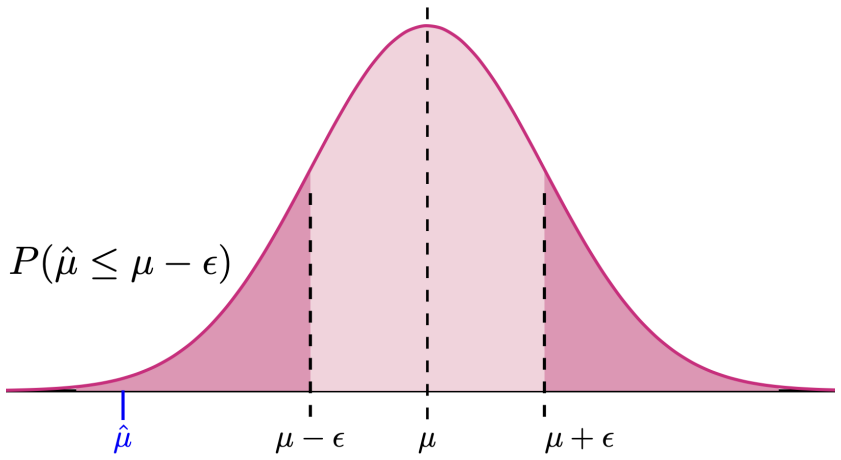
$$\begin{aligned}
\mathcal{R}_T(\nu, \pi_{ETC}) &= \Delta \mathbb{E}[N_2(T)] = \Delta m + (T - 2m)\mathbb{P}(A_{Km+1} = 2) \\
&\leq \Delta m + \Delta T \mathbb{P}[\hat{\mu}_2(Km) \geq \hat{\mu}_1(Km)] \\
&\leq \Delta m + \Delta T \mathbb{P}[\underbrace{\hat{\mu}_2(Km) - \mu_2 - (\hat{\mu}_1(Km) - \mu_1)}_{\sqrt{2/m}-\text{subgaussian}}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

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

$$\hat{\mu} - \underbrace{\sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}}_{\epsilon} \leq \mu \leq \hat{\mu} + \underbrace{\sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}}_{\epsilon} \tag{7}$$

with probability at least $1 - \delta$

UCB1

Algorithm 2 UCB1

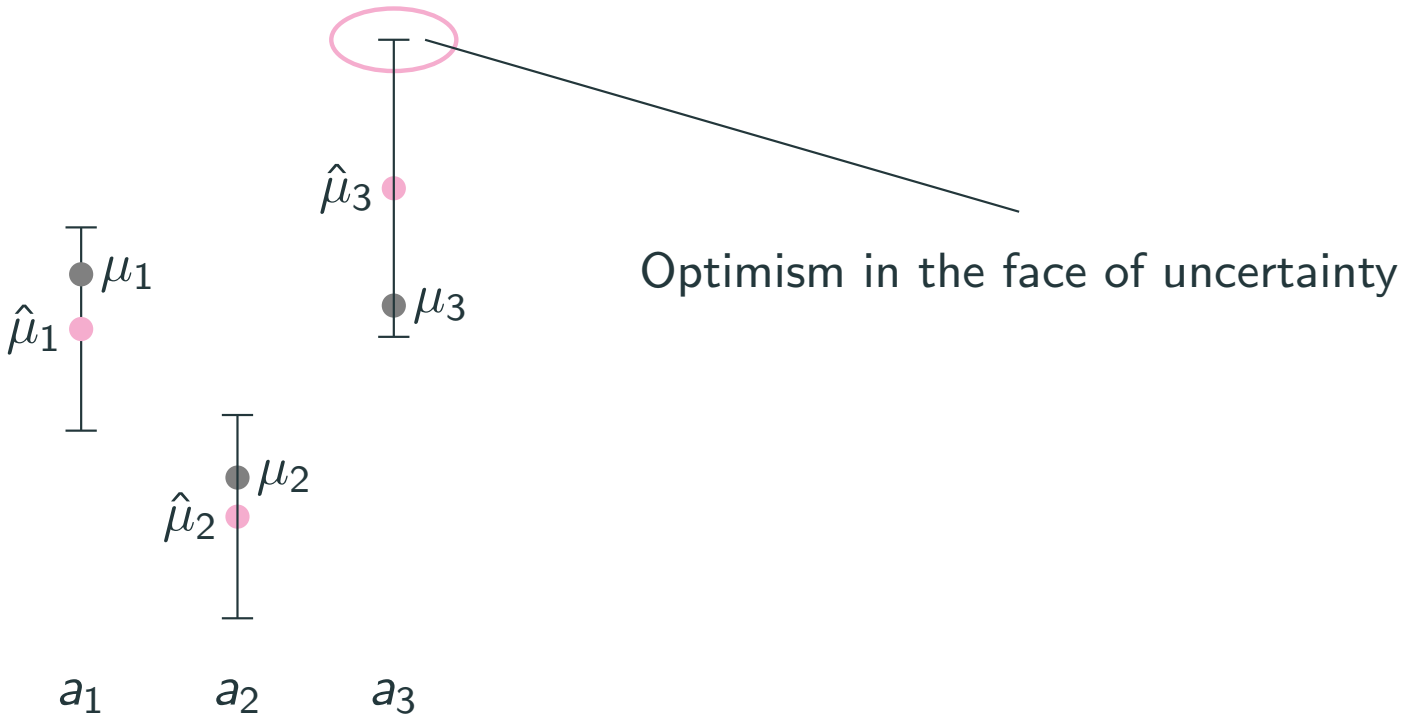
Initialization: Play each machine once;

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

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

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

end for



UCB1

Algorithm 3 UCB1

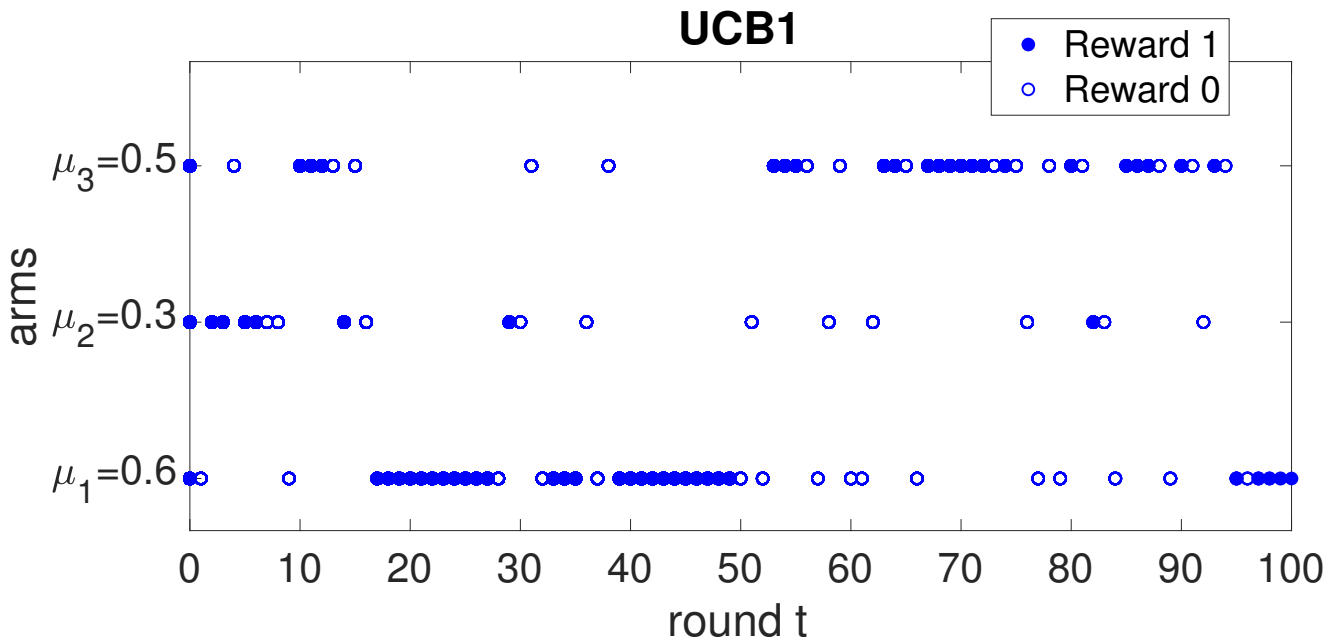
Initialization: Play each machine once;

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

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

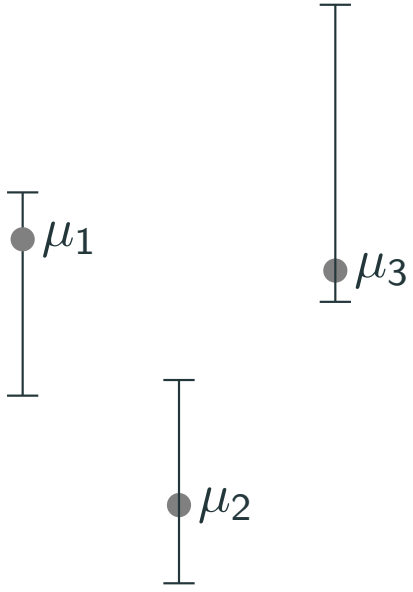
Update $\hat{\mu}_{a+1}(t+1)$ and $N_{t+1}(a+1)$

end for



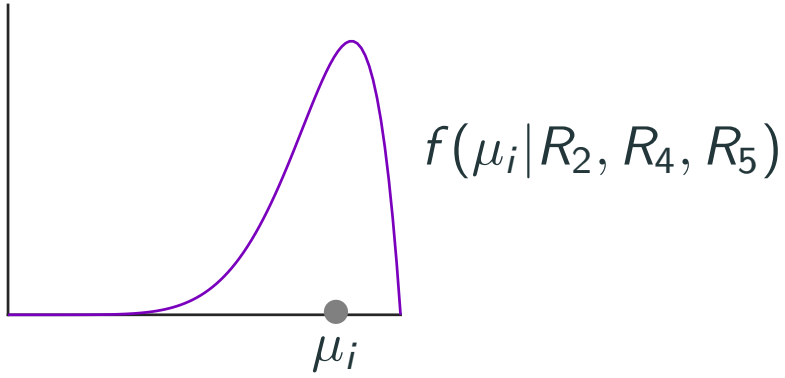
Bayesian approach

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



Bayesian approach

Use posterior densities to describe the uncertainty



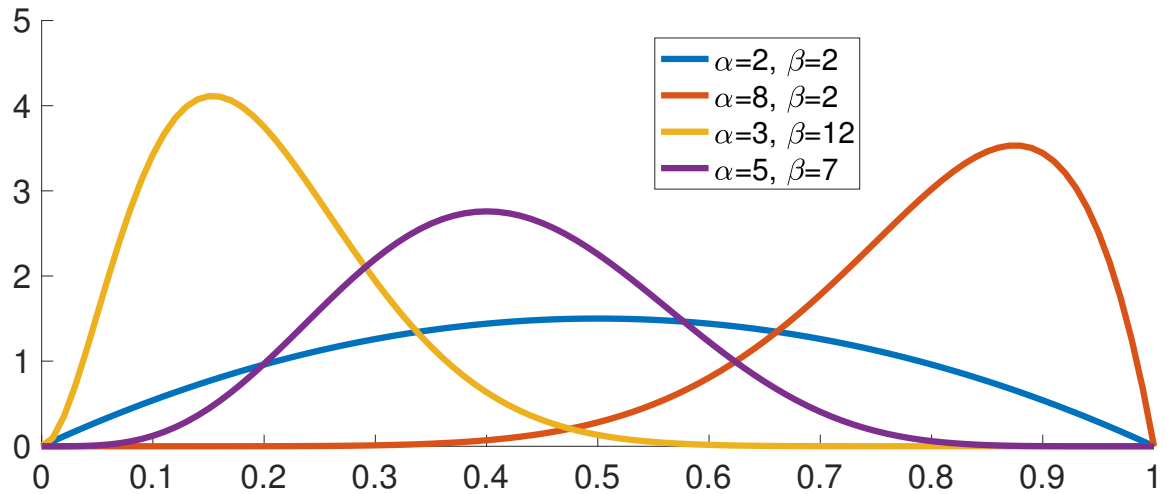
Prior distribution

Beta distribution

For α and β larger than zero the density of the beta distribution is given by

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

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



Thompson sampling

Algorithm 4 Thompson Sampling

Initialization: Play each machine once;

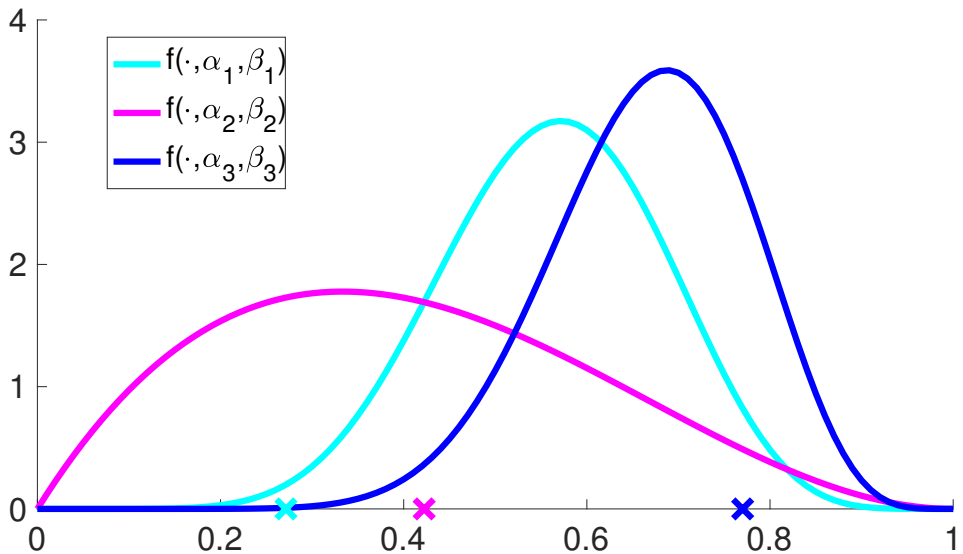
for $t = 1, 2, 3....$ **do**

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 5 Thompson Sampling

Initialization: Play each machine once;

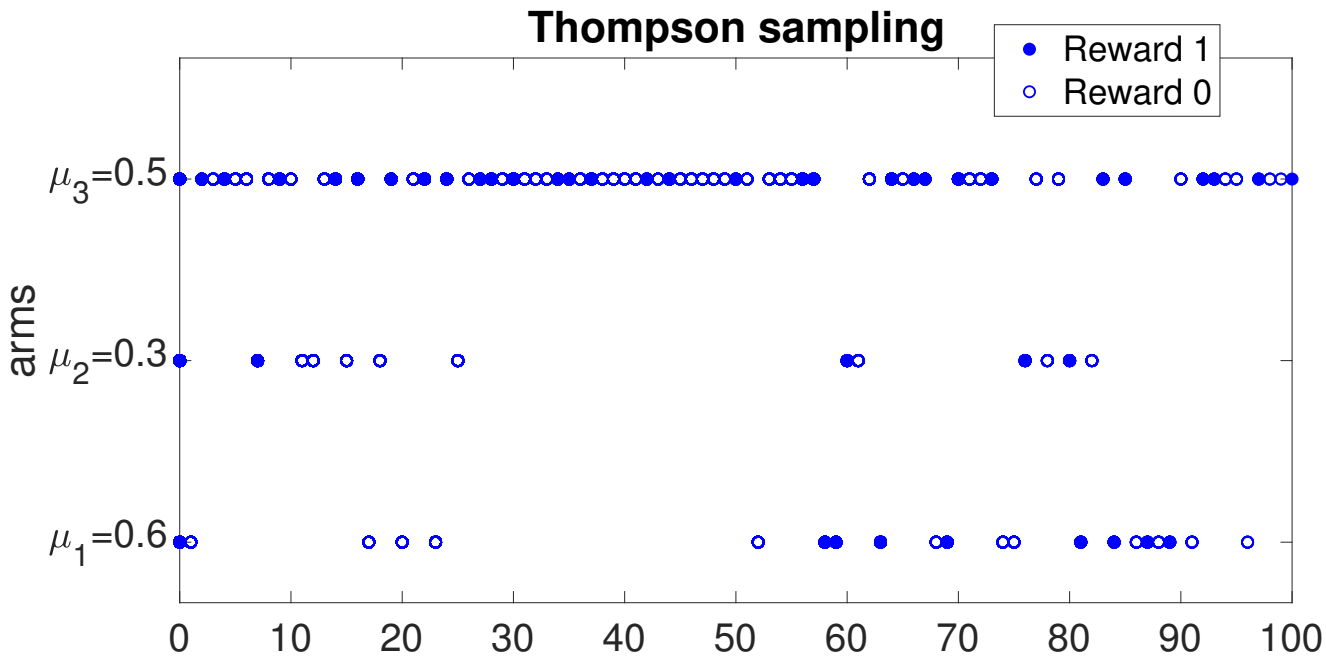
for $t = 1, 2, 3....$ **do**

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



References

- **Earliest reference:**
  1. W. Thompson (1933) *On the likelihood that one unknown probability exceeds another in view of the evidence of two samples*, Biometrika 25: 285-294
- **Introduction:**
  1. T. Lattimore and C. Szepesvari (2010) *Bandit Algorithms* Cambridge University Press
  2. L. Bottou, F. E. Curtis, J. Nocedal (2018) *Optimization Methods for Large-Scale Machine Learning*, SIAM Review
- **Key papers**
  1. P. Auer, N. Cesa-Bianchi and P. Fischer (2002) *Finite-time Analysis of the Multiarmed Bandit Problem.* Machine Learning, 47, 23556
  2. T. Lai and H. Robbins (1985) *Asymptotically efficient adaptive allocation rules.* Advances in Applied Mathematics, 6(1) :42.
- **Applications**
  1. L. Li, W. Chu, J. Langford and R. E. Schapire (2010) *A contextual-bandit approach to personalized news article recommendation.* IW3C2