Sequential Decision Making

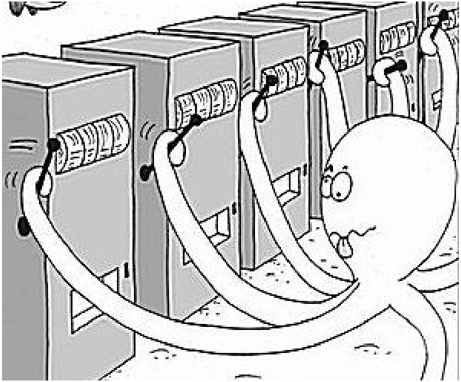
Lecture 2 : Stochastic bandits

Emilie Kaufmann

M2 Data Science, 2022/2023

Why bandits ?

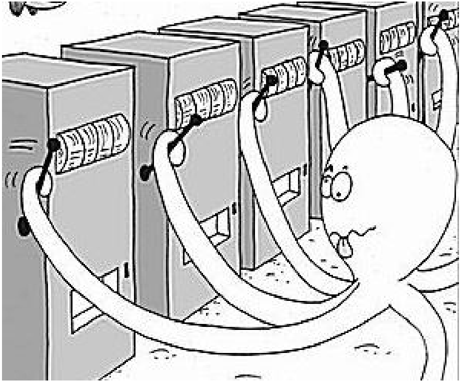
- Make money in a casino ? (one-armed bandit = slot machine)



an agent facing arms in a Multi-Armed Bandit

Why bandits ?

- Make money in a casino ? (one-armed bandit = slot machine)



an agent facing arms in a Multi-Armed Bandit

NO !

Sequential resource allocation

Clinical trials

- K treatment for a given symptom (with unknown effect)

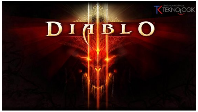





- What treatment should be allocated to the next patient based on responses observed on previous patients ?

Online advertisement

- K adds that can be displayed



- Which add should be displayed for a user, based on the previous clicks of previous (similar) users ?

Useful reference

The Bandit Book

by [Lattimore and Szepesvari, 2019]

The Multi-Armed Bandit Setup

K **arms** $\leftrightarrow$ K rewards streams $(X_{a,t})_{t\in\mathbb{N}}$

At round t, an agent :

- chooses an arm A_t
- receives a reward $R_t = X_{A_t,t}$

Sequential sampling strategy (**bandit algorithm**) :

$$A_{t+1} = F_t(A_1, R_1, \ldots, A_t, R_t).$$

Goal (for now !) : Maximize $\sum_{t=1}^{T} R_t$.

The Stochastic Multi-Armed Bandit Setup

K **arms** $\leftrightarrow$ K probability distributions : ν_a has mean μ_a

ν_1

ν_2

ν_3

ν_4

ν_5

At round t, an agent :

- chooses an arm A_t
- receives a reward $R_t = X_{A_t,t} \sim \nu_{A_t}$

Sequential sampling strategy (**bandit algorithm**) :

$$A_{t+1} = F_t(A_1, R_1, \ldots, A_t, R_t).$$

Goal (for now !) : Maximize $\mathbb{E}\left[\sum_{t=1}^{T} R_t\right]$

➜ a particular reinforcement learning problem

Clinical trials

Historical motivation [Thompson, 1933]





$\mathcal{B}(\mu_1)$ $\mathcal{B}(\mu_2)$ $\mathcal{B}(\mu_3)$ $\mathcal{B}(\mu_4)$ $\mathcal{B}(\mu_5)$

For the t-th patient in a clinical study,

- chooses a treatment A_t
- observes a response $R_t \in \{0, 1\} : \mathbb{P}(R_t = 1 | A_t = a) = \mu_a$

Goal : maximize the expected number of patients healed

Online content optimization

Modern motivation (\$\$) [Li et al., 2010]
(recommender systems, online advertisement)

ν_1

ν_2

ν_3

ν_4

ν_5

For the t-th visitor of a website,

- recommend a movie A_t
- observe a rating $R_t \sim \nu_{A_t}$ (e.g. $R_t \in \{1, \dots, 5\}$)

Goal : maximize the sum of ratings

Outline

1. Performance measure and first strategies

2. Best achievable regret

3. Mixing Exploration and Exploitation
   - Upper Confidence Bound algorithms

4. Bayesian algorithms
   - Thompson Sampling

Regret of a bandit algorithm

Bandit instance : $\nu = (\nu_1, \nu_2, \dots, \nu_K)$, mean of arm a : $\mu_a = \mathbb{E}_{X\sim\nu_a}[X]$.

$$\mu_\star = \max_{a\in\{1,\dots,K\}} \mu_a \qquad a_\star = \underset{a\in\{1,\dots,K\}}{\operatorname{argmax}} \mu_a.$$

Maximizing rewards $\leftrightarrow$ selecting $a_\star$ as much as possible
$\leftrightarrow$ minimizing the regret [Robbins, 1952]

$$\mathcal{R}_\nu(\mathcal{A}, T) := \underbrace{T\mu_\star}_{\substack{\text{sum of rewards of}\\ \text{an oracle strategy}\\ \text{always selecting } a_\star}} - \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} R_t\right]}_{\substack{\text{sum of rewards of}\\ \text{the strategy } \mathcal{A}}}$$

What regret rate can we achieve ?

- consistency : $\frac{\mathcal{R}_\nu(\mathcal{A},T)}{T} \to 0$
- can we be more precise ?

Regret decomposition

$N_a(t)$: number of selections of arm a in the first t rounds
$\Delta_a := \mu_\star - \mu_a$: sub-optimality gap of arm a

Regret decomposition

$$\mathcal{R}_\nu(\mathcal{A}, T) = \sum_{a=1}^{K} \Delta_a \mathbb{E}\left[N_a(T)\right].$$

Proof.

Regret decomposition

$N_a(t)$: number of selections of arm a in the first t rounds
$\Delta_a := \mu_\star - \mu_a$: sub-optimality gap of arm a

Regret decomposition

$$\mathcal{R}_\nu(\mathcal{A}, T) = \sum_{a=1}^{K} \Delta_a \mathbb{E}\left[N_a(T)\right].$$

A strategy with small regret should :

- select not too often arms for which $\Delta_a > 0$
- ... which requires to try all arms to estimate the values of the Δ_a's

$\Rightarrow$ Exploration / Exploitation trade-off

Two naive strategies

- **Idea 1 :** Uniform Exploration

Draw each arm T/K times

$\Rightarrow$ EXPLORATION
$$\mathcal{R}_\nu(\mathcal{A}, T) = \left(\frac{1}{K}\sum_{a:\mu_a>\mu_\star} \Delta_a\right) T$$

Two naive strategies

- **Idea 1 :** Uniform Exploration

Draw each arm T/K times

$\Rightarrow$ EXPLORATION $\mathcal{R}_\nu(\mathcal{A}, T) = \left(\frac{1}{K}\sum_{a:\mu_a > \mu_\star} \Delta_a\right) T$

- **Idea 2 :** Follow The Leader

$$A_{t+1} = \underset{a \in \{1,\dots,K\}}{\text{argmax}} \ \hat{\mu}_a(t)$$

where

$$\hat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} X_{a,s} \mathbb{1}_{(A_s = a)}$$

is an estimate of the unknown mean μ_a.

$\Rightarrow$ EXPLOITATION $\mathcal{R}_\nu(\mathcal{A}, T) \geq (1-\mu_1) \times \mu_2 \times (\mu_1 - \mu_2) T$

(Bernoulli arms)

A better idea : Explore-Then-Commit

Given $m \in \{1, \ldots, T/K\}$,

- draw each arm m times
- compute the empirical best arm $\hat{a} = \text{argmax}_a \, \hat{\mu}_a(Km)$
- keep playing this arm until round T

$$A_{t+1} = \hat{a} \text{ for } t \geq Km$$

$\Rightarrow$ EXPLORATION followed by EXPLOITATION

A better idea : Explore-Then-Commit

Given $m \in \{1, \dots, T/K\}$,

- draw each arm m times
- compute the empirical best arm $\hat{a} = \text{argmax}_a \ \hat{\mu}_a(Km)$
- keep playing this arm until round T

$$A_{t+1} = \hat{a} \ \text{ for } t \geq Km$$

$\Rightarrow$ EXPLORATION followed by EXPLOITATION

Analysis for two arms. $\mu_1 > \mu_2$, $\Delta := \mu_1 - \mu_2$.

$$\begin{aligned} \mathcal{R}_\nu(\texttt{ETC}, T) &= \Delta \mathbb{E}[N_2(T)] \\ &= \Delta \mathbb{E}\left[m + (T-2m)\mathbb{1}\left(\hat{a} = 2\right)\right] \\ &\leq \Delta m + (\Delta T) \times \mathbb{P}\left(\hat{\mu}_{2,m} \geq \hat{\mu}_{1,m}\right) \end{aligned}$$

$\hat{\mu}_{a,m}$: empirical mean of the first m observations from arm a

A better idea : Explore-Then-Commit

Given $m \in \{1, \ldots, T/K\}$,

- draw each arm m times
- compute the empirical best arm $\hat{a} = \text{argmax}_a \ \hat{\mu}_a(Km)$
- keep playing this arm until round T

$$A_{t+1} = \hat{a} \ \text{ for } t \geq Km$$

$\Rightarrow$ EXPLORATION followed by EXPLOITATION

Analysis for two arms. $\mu_1 > \mu_2$, $\Delta := \mu_1 - \mu_2$.

$$\begin{aligned}
\mathcal{R}_\nu(\texttt{ETC}, T) &= \Delta \mathbb{E}[N_2(T)] \\
&= \Delta \mathbb{E}\left[m + (T - 2m)\mathbb{1}\left(\hat{a} = 2\right)\right] \\
&\leq \Delta m + (\Delta T) \times \mathbb{P}\left(\hat{\mu}_{2,m} \geq \hat{\mu}_{1,m}\right)
\end{aligned}$$

$\hat{\mu}_{a,m}$: empirical mean of the first m observations from arm a

$\rightarrow$ requires a concentration inequality

Intermezzo : Concentration Inequalities

Sub-Gaussian random variables : $Z - \mu$ is σ^2-subGaussian if

$$\mathbb{E}[Z] = \mu \quad \text{and} \quad \mathbb{E}\left[e^{\lambda(Z-\mu)}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}}. \tag{1}$$

Hoeffding inequality

Z_i i.i.d. satisfying (1). For all $s \geq 1$

$$\mathbb{P}\left(\frac{Z_1 + \dots + Z_s}{s} \geq \mu + x\right) \leq e^{-\frac{sx^2}{2\sigma^2}}$$

Proof : Cramér-Chernoff method

- ν_a bounded in $[a, b]$: $(b-a)^2/4$ sub-Gaussian (Hoeffding's lemma)
- $\nu_a = \mathcal{N}(\mu_a, \sigma^2)$: σ^2 sub-Gaussian

Intermezzo : Concentration Inequalities

Sub-Gaussian random variables : $Z - \mu$ is σ^2-subGaussian if

$$\mathbb{E}[Z] = \mu \quad \text{and} \quad \mathbb{E}\left[e^{\lambda(Z-\mu)}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}}. \tag{1}$$

Hoeffding inequality

Z_i i.i.d. satisfying (1). For all $s \geq 1$

$$\mathbb{P}\left(\frac{Z_1 + \dots + Z_s}{s} \leq \mu - x\right) \leq e^{-\frac{sx^2}{2\sigma^2}}$$

Proof : Cramér-Chernoff method

- ν_a bounded in $[a, b]$: $(b-a)^2/4$ sub-Gaussian (Hoeffding's lemma)
- $\nu_a = \mathcal{N}(\mu_a, \sigma^2)$: σ^2 sub-Gaussian

A better idea : Explore-Then-Commit

Given $m \in \{1, \ldots, T/K\}$,

- draw each arm m times
- compute the empirical best arm $\hat{a} = \text{argmax}_a\ \hat{\mu}_a(Km)$
- keep playing this arm until round T

$$A_{t+1} = \hat{a} \ \text{ for } t \geq Km$$

$\Rightarrow$ EXPLORATION followed by EXPLOITATION

Analysis for two arms. $\mu_1 > \mu_2$, $\Delta := \mu_1 - \mu_2$.

Assumption : ν_1, ν_2 are bounded in $[0, 1]$.

$$\begin{aligned}
\mathcal{R}_\nu(T) &= \Delta \mathbb{E}[N_2(T)] \\
&= \Delta \mathbb{E}\left[m + (T - 2m)\mathbb{1}\left(\hat{a} = 2\right)\right] \\
&\leq \Delta m + (\Delta T) \times \mathbb{P}\left(\hat{\mu}_{2,m} \geq \hat{\mu}_{1,m}\right)
\end{aligned}$$

$\hat{\mu}_{a,m}$: empirical mean of the first m observations from arm a

$\rightarrow$ Hoeffding's inequality

A better idea : Explore-Then-Commit

Given $m \in \{1, \ldots, T/K\}$,

- draw each arm m times
- compute the empirical best arm $\hat{a} = \text{argmax}_a \ \hat{\mu}_a(Km)$
- keep playing this arm until round T
$$A_{t+1} = \hat{a} \ \text{ for } t \geq Km$$

$\Rightarrow$ EXPLORATION followed by EXPLOITATION

Analysis for two arms. $\mu_1 > \mu_2$, $\Delta := \mu_1 - \mu_2$.

Assumption : ν_1, ν_2 are bounded in $[0, 1]$.

$$\begin{aligned} \mathcal{R}_\nu(T) &= \Delta \mathbb{E}[N_2(T)] \\ &= \Delta \mathbb{E}\left[m + (T - 2m)\mathbb{1}\left(\hat{a} = 2\right)\right] \\ &\leq \Delta m + (\Delta T) \times \exp(-m\Delta^2/2) \end{aligned}$$

$\hat{\mu}_{a,m}$: empirical mean of the first m observations from arm a

$\rightarrow$ Hoeffding's inequality

A better idea : Explore-Then-Commit

Given $m \in \{1, \ldots, T/K\}$,

- draw each arm m times
- compute the empirical best arm $\hat{a} = \text{argmax}_a \ \hat{\mu}_a(Km)$
- keep playing this arm until round T
$$A_{t+1} = \hat{a} \ \text{ for } t \geq Km$$

$\Rightarrow$ EXPLORATION followed by EXPLOITATION

Analysis for two arms. $\mu_1 > \mu_2$, $\Delta := \mu_1 - \mu_2$.

Assumption : ν_1, ν_2 are bounded in $[0, 1]$.

For $m = \frac{2}{\Delta^2} \log\left(\frac{T\Delta^2}{2}\right)$,
$$\mathcal{R}_\nu(\texttt{ETC}, T) \leq \frac{2}{\Delta}\left[\log\left(\frac{T\Delta^2}{2}\right) + 1\right].$$

A better idea : Explore-Then-Commit

Given $m \in \{1, \dots, T/K\}$,

- draw each arm m times
- compute the empirical best arm $\hat{a} = \mathrm{argmax}_a \ \hat{\mu}_a(Km)$
- keep playing this arm until round T

$$A_{t+1} = \hat{a} \ \text{ for } t \geq Km$$

$\Rightarrow$ EXPLORATION followed by EXPLOITATION

Analysis for two arms. $\mu_1 > \mu_2$, $\Delta := \mu_1 - \mu_2$.

Assumption : ν_1, ν_2 are bounded in $[0, 1]$.

For $m = \frac{2}{\Delta^2} \log\left(\frac{T\Delta^2}{2}\right)$,

$$\mathcal{R}_\nu(\texttt{ETC}, T) \leq \frac{2}{\Delta}\left[\log\left(\frac{T\Delta^2}{2}\right) + 1\right].$$

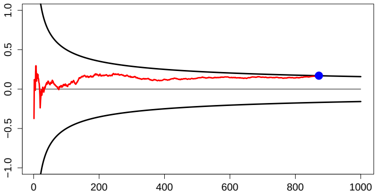
\+ logarithmic regret !

— requires the knowledge of T and Δ

Sequential Explore-Then-Commit

- explore uniformly until a random time of the form

$$\tau = \inf\left\{t \in \mathbb{N} : |\hat{\mu}_1(t) - \hat{\mu}_2(t)| > \sqrt{\frac{c\log(T/t)}{t}}\right\}$$

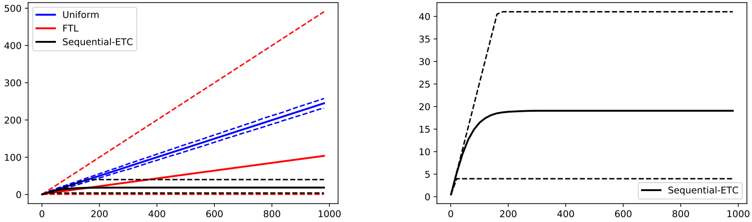


- $\hat{a}_\tau = \mathrm{argmax}_a\ \hat{\mu}_a(\tau)$ and $(A_{t+1} = \hat{a}_\tau)$ for $t \in \{\tau+1, \dots, T\}$

➜ [Garivier et al., 2016] for two Gaussian arms, for $c = 8$, same regret as ETC, without the knowledge of Δ

Numerical illustration

$$\nu_1 = \mathcal{N}(1,1) \qquad \nu_2 = \mathcal{N}(1.5,1)$$



Expected regret estimated over $N = 500$ runs for Sequential-ETC versus two naive baselines.

(dashed lines : empirical 0.05% and 0.95% quantiles of the regret)

Outline

Examples of regret rates

For two-armed bandits with bounded rewards, $\Delta = |\mu_1 - \mu_2|$

$$\mathcal{R}_\nu(\texttt{ETC}, T) \lesssim \frac{2}{\Delta} \log\left(T\Delta^2\right).$$

➜ problem-dependent logarithmic regret bound

Remark : blows up when Δ tends to zero...

$$\begin{aligned}
\mathcal{R}_\nu(\texttt{ETC}, T) & \lesssim & \min\left[\frac{2}{\Delta}\log\left(T\Delta^2\right), \Delta T\right] \\
& \leq & \sqrt{T}\max_{u>0}\left(\min\left[\frac{2}{u}\log(u^2); u\right]\right) \\
& \leq & C\sqrt{T}.
\end{aligned}$$

➜ problem-independent square-root regret bound

The Lai and Robbins lower bound

Context : a parametric bandit model where each arm is parameterized by its mean $\nu = (\nu_{\mu_1}, \dots, \nu_{\mu_K})$, $\mu_a \in \mathcal{I}$.

$$\nu \quad \leftrightarrow \quad \boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$$

Key tool : Kullback-Leibler divergence.

Kullback-Leibler divergence

$$\mathrm{kl}(\mu, \mu') := \mathrm{KL}\left(\nu_\mu, \nu_{\mu'}\right) = \mathbb{E}_{X \sim \nu_\mu}\left[\log \frac{d\nu_\mu}{d\nu_{\mu'}}(X)\right]$$

Theorem

For *uniformly good* algorithm,

$$\mu_a < \mu_\star \Rightarrow \liminf_{T \to \infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]}{\log T} \geq \frac{1}{\mathrm{kl}(\mu_a, \mu_\star)}$$

[Lai and Robbins, 1985]

The Lai and Robbins lower bound

Context : a parametric bandit model where each arm is parameterized by its mean $\nu = (\nu_{\mu_1}, \ldots, \nu_{\mu_K})$, $\mu_a \in \mathcal{I}$.

$$\nu \quad \leftrightarrow \quad \boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$$

Key tool : Kullback-Leibler divergence.

Kullback-Leibler divergence

$$\mathrm{kl}(\mu, \mu') := \frac{(\mu - \mu')^2}{2\sigma^2} \quad \text{(Gaussian bandits)}$$

Theorem

For *uniformly good* algorithm,

$$\mu_a < \mu_\star \Rightarrow \liminf_{T \to \infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]}{\log T} \geq \frac{1}{\mathrm{kl}(\mu_a, \mu_\star)}$$

[Lai and Robbins, 1985]

The Lai and Robbins lower bound

Context : a parametric bandit model where each arm is parameterized by its mean $\nu = (\nu_{\mu_1}, \dots, \nu_{\mu_K})$, $\mu_a \in \mathcal{I}$.

$$\nu \quad \leftrightarrow \quad \boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$$

Key tool : Kullback-Leibler divergence.

Kullback-Leibler divergence

$$\mathrm{kl}(\mu, \mu') := \mu \log\left(\frac{\mu}{\mu'}\right) + (1-\mu)\log\left(\frac{1-\mu}{1-\mu'}\right) \quad \text{(Bernoulli bandits)}$$

Theorem

For *uniformly good* algorithm,

$$\mu_a < \mu_\star \Rightarrow \liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]}{\log T} \geq \frac{1}{\mathrm{kl}(\mu_a, \mu_\star)}$$

[Lai and Robbins, 1985]

Some room for better algorithms !

A particular case of parameteric and bounded distributions :

$$\nu_1 = \mathcal{B}(\mu_1) \qquad \nu_2 = \mathcal{B}(\mu_2)$$

- **Regret of ETC :** $\mathcal{R}_\nu(\mathrm{ETC}, T) \lesssim \frac{2}{\Delta} \log\left(T\Delta^2\right)$
- **Lower bound :** $\mathcal{R}_\nu(\mathcal{A}, T) \gtrsim \frac{\Delta}{\mathrm{kl}(\mu_2,\mu_1)} \log\left(T\Delta^2\right)$

Pinsker's inequality : $\mathrm{kl}(\mu_2, \mu_1) \geq 2(\mu_1 - \mu_2)^2$.

➜ Explore-Then-Commit does not match the lower bound...

Outline

A simple strategy : ϵ-greedy

The ϵ-greedy rule [Sutton and Barto, 1998] is the simplest way to alternate exploration and exploitation.

ϵ-greedy strategy

At round t,

- with probability ϵ

$$A_t \sim \mathcal{U}(\{1, \dots, K\})$$

- with probability $1 - \epsilon$

$$A_t = \underset{a=1,\dots,K}{\text{argmax}} \ \hat{\mu}_a(t).$$

→ Linear regret : $\mathcal{R}_\nu\left(\epsilon\texttt{-greedy}, T\right) \geq \epsilon \frac{K-1}{K} \Delta_{\min} T$.

$$\Delta_{\min} = \min_{a:\mu_a < \mu_\star} \Delta_a$$

A simple strategy : ϵ-greedy

A simple fix :

ϵ_t-greedy strategy

At round t,

- with probability $\epsilon_t := \min\left(1, \frac{K}{d^2 t}\right)$
$$A_t \sim \mathcal{U}(\{1, \dots, K\})$$
- with probability $1 - \epsilon_t$
$$A_t = \underset{a=1,\dots,K}{\operatorname{argmax}} \ \hat{\mu}_a(t-1).$$

Theorem [Auer et al., 2002]

If $0 < d \leq \Delta_{\min}$, $\mathcal{R}_\nu\left(\epsilon_t\text{-greedy}, T\right) = O\left(\frac{K \log(T)}{d^2}\right)$.

➜ requires the knowledge of a lower bound on $\Delta_{\min}$...

Outline

The optimism principle

Step 1 : construct a set of statistically plausible models

- For each arm a, build a confidence interval on the mean μ_a :

$$\mathcal{I}_a(t) = [\mathrm{LCB}_a(t), \mathrm{UCB}_a(t)]$$

LCB = Lower Confidence Bound
UCB = Upper Confidence Bound

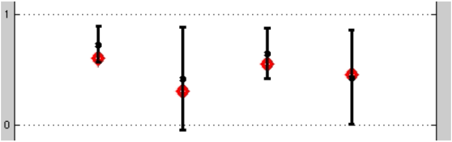


FIGURE – Confidence intervals on the means after t rounds

The optimism principle

Step 2 : act as if the best possible model were the true model
(optimism in face of uncertainty)

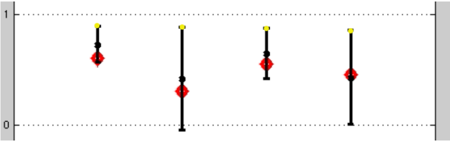


FIGURE – Confidence intervals on the means after t rounds

$$\text{Optimistic bandit model} = \underset{\boldsymbol{\mu} \in \mathcal{C}(t)}{\operatorname{argmax}} \max_{a=1,\dots,K} \mu_a$$

- That is, select

$$A_{t+1} = \underset{a=1,\dots,K}{\operatorname{argmax}} \, \mathrm{UCB}_a(t).$$

How to build confidence intervals ?

We need $\mathrm{UCB}_a(t)$ such that

$$\mathbb{P}\left(\mu_a \leq \mathrm{UCB}_a(t)\right) \gtrsim 1 - t^{-1}.$$

➜ tool : concentration inequalities

Example : rewards are σ^2 sub-Gaussian

Hoeffding inequality, reloaded

Z_i i.i.d. satisfying (1). For all $s \geq 1$

$$\mathbb{P}\left(\frac{Z_1 + \cdots + Z_s}{s} < \mu - x\right) \leq e^{-\frac{sx^2}{2\sigma^2}}$$

How to build confidence intervals ?

We need $\mathrm{UCB}_a(t)$ such that

$$\mathbb{P}\left(\mu_a \leq \mathrm{UCB}_a(t)\right) \gtrsim 1 - t^{-1}.$$

➜ tool : concentration inequalities

Example : rewards are σ^2 sub-Gaussian

Hoeffding inequality, reloaded

Z_i i.i.d. satisfying (1). For all $s \geq 1$

$$\mathbb{P}\left(\frac{Z_1 + \dots + Z_s}{s} < \mu - x\right) \leq e^{-\frac{sx^2}{2\sigma^2}}$$

⚠ Cannot be used directly in a bandit model as the number of observations from each arm is random !

How to build confidence intervals ?

- $N_a(t) = \sum_{s=1}^{t} \mathbb{1}_{(A_s=a)}$ number of selections of a after t rounds
- $\hat{\mu}_{a,s} = \frac{1}{s}\sum_{k=1}^{s} Y_{a,k}$ average of the first s observations from arm a
- $\hat{\mu}_a(t) = \hat{\mu}_{a,N_a(t)}$ empirical estimate of μ_a after t rounds

Hoeffding inequality + union bound

$$\mathbb{P}\left(\mu_a \leq \hat{\mu}_a(t) + \sigma\sqrt{\frac{\beta \log(t)}{N_a(t)}}\right) \geq 1 - \frac{1}{t^{\frac{\beta}{2}-1}}$$

How to build confidence intervals ?

- $N_a(t) = \sum_{s=1}^{t} \mathbb{1}_{(A_s = a)}$ number of selections of a after t rounds
- $\hat{\mu}_{a,s} = \frac{1}{s}\sum_{k=1}^{s} Y_{a,k}$ average of the first s observations from arm a
- $\hat{\mu}_a(t) = \hat{\mu}_{a,N_a(t)}$ empirical estimate of μ_a after t rounds

Hoeffding inequality + union bound

$$\mathbb{P}\left(\mu_a \leq \hat{\mu}_a(t) + \sigma\sqrt{\frac{\beta \log(t)}{N_a(t)}}\right) \geq 1 - \frac{1}{t^{\frac{\beta}{2}-1}}$$

Proof.

$$\mathbb{P}\left(\mu_a > \hat{\mu}_a(t) + \sigma\sqrt{\frac{\beta \log(t)}{N_a(t)}}\right) \leq \mathbb{P}\left(\exists s \leq t : \mu_a > \hat{\mu}_{a,s} + \sigma\sqrt{\frac{\beta \log(t)}{s}}\right)$$

$$\leq \sum_{s=1}^{t} \mathbb{P}\left(\hat{\mu}_{a,s} < \mu_a - \sigma\sqrt{\frac{\beta \log(t)}{s}}\right) \leq \sum_{s=1}^{t} \frac{1}{t^{\beta/2}} = \frac{1}{t^{\beta/2-1}}.$$

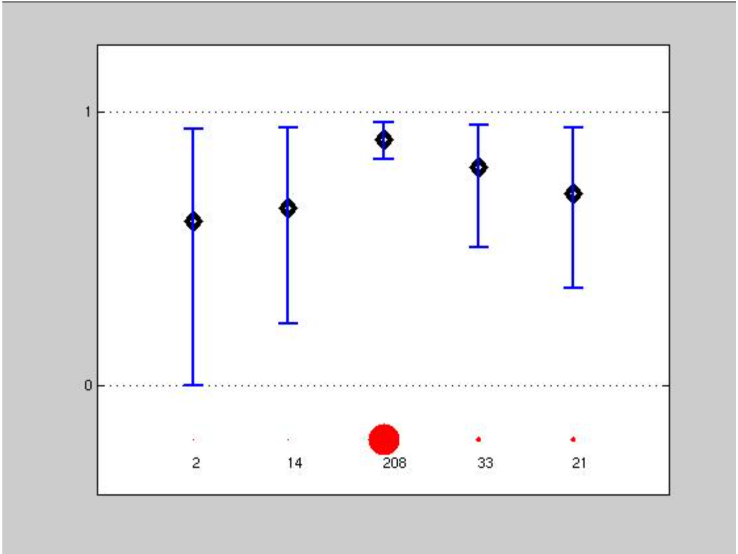
A first UCB algorithm

UCB(α) selects $A_{t+1} = \mathrm{argmax}_a \ \mathrm{UCB}_a(t)$ where

$$\mathrm{UCB}_a(t) = \underbrace{\hat{\mu}_a(t)}_{\text{exploitation term}} + \underbrace{\sqrt{\frac{\alpha \log(t)}{N_a(t)}}}_{\text{exploration bonus}}.$$

- popularized by [Auer et al., 2002] for bounded rewards : UCB1, for $\alpha = 2$
- the analysis of UCB(α) was further refined to hold for $\alpha > 1/2$ in that case [Bubeck, 2010, Cappé et al., 2013]

A UCB algorithm in action



Regret of UCB(α) for bounded rewards

Theorem

For every $\alpha > 1$ and every sub-optimal arm a, there exists a constant $C_\alpha > 0$ such that

$$\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] \leq \frac{4\alpha}{(\mu_\star - \mu_a)^2} \log(T) + C_\alpha.$$

Proof :

An improved result

Context : σ^2 sub-Gaussian rewards

$$\mathrm{UCB}_a(t) = \hat{\mu}_a(t) + \sqrt{\frac{2\sigma^2(\log(t) + c \log\log(t))}{N_a(t)}}$$

Theorem [Cappé et al.'13]

For $c \geq 3$, the UCB algorithm associated to the above index satisfy

$$\mathbb{E}[N_a(T)] \leq \frac{2\sigma^2}{(\mu_\star - \mu_a)^2} \log(T) + C_{\boldsymbol{\mu}} \sqrt{\log(T)}.$$

An improved result

Context : σ^2 sub-Gaussian rewards

$$\mathrm{UCB}_a(t) = \hat{\mu}_a(t) + \sqrt{\frac{2\sigma^2(\log(t) + c\log\log(t))}{N_a(t)}}$$

Theorem [Cappé et al.'13]

For $c \geq 3$, the UCB algorithm associated to the above index satisfy

$$\mathbb{E}[N_a(T)] \leq \frac{2\sigma^2}{(\mu_\star - \mu_a)^2}\log(T) + C_{\boldsymbol{\mu}}\sqrt{\log(T)}.$$

- Gaussian rewards :

$$\mathcal{R}_\nu(\mathrm{UCB}, T) \lesssim \left(\sum_{a:\mu_a<\mu_\star}\frac{2\sigma^2}{\Delta_a}\right)\log(T).$$

➜ matching the Lai and Robbins lower bound ! asymptotically optimal

An improved result

Context : σ^2 sub-Gaussian rewards

$$\mathrm{UCB}_a(t) = \hat{\mu}_a(t) + \sqrt{\frac{2\sigma^2(\log(t) + c\log\log(t))}{N_a(t)}}$$

Theorem [Cappé et al.'13]

For $c \geq 3$, the UCB algorithm associated to the above index satisfy

$$\mathbb{E}[N_a(T)] \leq \frac{2\sigma^2}{(\mu_\star - \mu_a)^2}\log(T) + C_{\boldsymbol{\mu}}\sqrt{\log(T)}.$$

- Bernoulli rewards :

$$\mathcal{R}_\nu(\mathrm{UCB}, T) \lesssim \left(\sum_{a:\mu_a<\mu_\star} \frac{1}{2\Delta_a}\right)\log(T)$$

➜ optimal ?

An improved result

Context : σ^2 sub-Gaussian rewards

$$\mathrm{UCB}_a(t) = \hat{\mu}_a(t) + \sqrt{\frac{2\sigma^2(\log(t) + c\log\log(t))}{N_a(t)}}$$

Theorem [Cappé et al.'13]

For $c \geq 3$, the UCB algorithm associated to the above index satisfy

$$\mathbb{E}[N_a(T)] \leq \frac{2\sigma^2}{(\mu_\star - \mu_a)^2}\log(T) + C_{\boldsymbol{\mu}}\sqrt{\log(T)}.$$

- Bernoulli rewards :

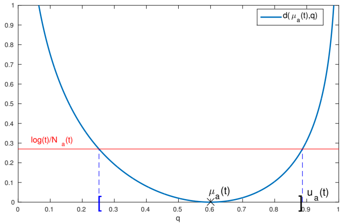
$$\mathcal{R}_\nu(\mathrm{UCB}, T) \neq \left(\sum_{a:\mu_a<\mu_\star} \frac{\Delta_a}{\mathrm{kl}(\mu_a, \mu_\star)}\right)\log(T)$$

➜ not matching the Lai and Robbins lower bound

The kl-UCB algorithm

Exploits the KL-divergence in the lower bound !

$$\mathrm{UCB}_a(t) = \max\left\{ q \in [0,1] : \mathrm{kl}\left(\hat{\mu}_a(t), q\right) \leq \frac{\log(t)}{N_a(t)} \right\}.$$



A tighter concentration inequality [Garivier and Cappé, 2011]

For rewards that belong to a 1-d exponential family (e.g. Bernoulli)

$$\mathbb{P}(\mathrm{UCB}_a(t) > \mu_a) \gtrsim 1 - \frac{1}{t\log(t)}.$$

An asymptotically optimal algorithm

kl-UCB selects $A_{t+1} = \mathrm{argmax}_a \, \mathrm{UCB}_a(t)$ with

$$\mathrm{UCB}_a(t) = \max\left\{ q \in [0,1] : \mathrm{kl}\left(\hat{\mu}_a(t), q\right) \leq \frac{\log(t) + c \log\log(t)}{N_a(t)} \right\}.$$

Theorem [Cappé et al., 2013]

If $c \geq 3$, for every arm such that $\mu_a < \mu_\star$,

$$\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] \leq \frac{1}{\mathrm{kl}(\mu_a, \mu_\star)} \log(T) + C_{\boldsymbol{\mu}} \sqrt{\log(T)}.$$

- asymptotically optimal for rewards in a 1-d exponential family :

$$\mathcal{R}_{\boldsymbol{\mu}}(\text{kl-UCB}, T) \simeq \left(\sum_{a : \mu_a < \mu_\star} \frac{\Delta_a}{\mathrm{kl}(\mu_a, \mu_\star)} \right) \log(T).$$

Outline

1. Performance measure and first strategies
2. Best achievable regret
3. Mixing Exploration and Exploitation
   - Upper Confidence Bound algorithms
4. Bayesian algorithms
   - Thompson Sampling

Frequentist versus Bayesian bandit

$\nu_{\boldsymbol{\mu}} = (\nu^{\mu_1}, \dots, \nu^{\mu_K}) \in (\mathcal{P})^K.$

- Two probabilistic models

Frequentist model	**Bayesian model**
$\mu_1, \dots, \mu_K$ unknown parameters	$\mu_1, \dots, \mu_K$ drawn from a prior distribution : $\mu_a \sim \pi_a$
arm a : $(Y_{a,s})_s \overset{\text{i.i.d.}}{\sim} \nu^{\mu_a}$	arm a : $(Y_{a,s})_s \vert \boldsymbol{\mu} \overset{\text{i.i.d.}}{\sim} \nu^{\mu_a}$

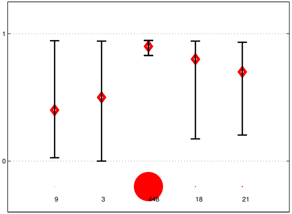
- The regret can be computed in each case

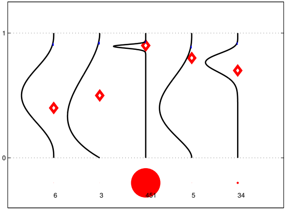
Frequentist regret (regret)	Bayesian regret (Bayes risk)
$\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}, T) = \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=1}^{T} (\mu_\star - \mu_{A_t})\right]$	$\mathrm{R}^{\pi}(\mathcal{A}, T) = \mathbb{E}_{\boldsymbol{\mu}\sim\pi}\left[\sum_{t=1}^{T} (\mu_\star - \mu_{A_t})\right]$ $= \int \mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}, T) d\pi(\boldsymbol{\mu})$

Frequentist and Bayesian algorithms

- Two types of tools to build bandit algorithms :

Frequentist tools	Bayesian tools
MLE estimators of the means Confidence Intervals	Posterior distributions $\pi_a^t = \mathcal{L}(\mu_a \| Y_{a,1}, \dots, Y_{a,N_a(t)})$





Example : Bernoulli bandits

Bernoulli bandit model $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$

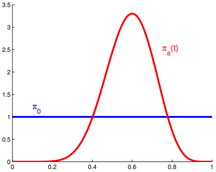
- **Bayesian view** : $\mu_1, \dots, \mu_K$ are random variables

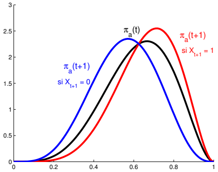
  prior distribution : $\mu_a \sim \mathcal{U}([0,1])$

➜ posterior distribution :

$$\begin{aligned} \pi_a(t) &= \mathcal{L}(\mu_a | R_1, \dots, R_t) \\ &= \mathrm{Beta}\Big(\underbrace{S_a(t)}_{\#ones} + 1, \underbrace{N_a(t) - S_a(t)}_{\#zeros} + 1\Big) \end{aligned}$$

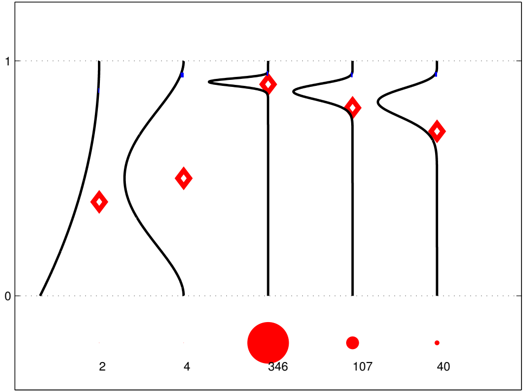
$S_a(t) = \sum_{s=1}^{t} R_s \mathbb{1}_{(A_s = a)}$ sum of the rewards.





Bayesian algorithm

A Bayesian bandit algorithm exploits the posterior distributions of the means to decide which arm to select.



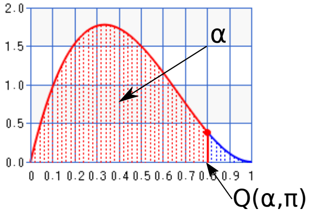
First example : Bayes-UCB

- $\Pi_0 = (\pi_1(0), \dots, \pi_K(0))$ be a prior distribution over $(\mu_1, ..., \mu_K)$
- $\Pi_t = (\pi_1(t), \dots, \pi_K(t))$ be the posterior distribution over the means $(\mu_1, ..., \mu_K)$ after t observations

Bayes-UCB selects at time $t+1$

$$A_{t+1} = \underset{a=1,\dots,K}{\mathrm{argmax}} \; Q\left(1 - \frac{1}{t(\log t)^c}, \pi_a(t)\right)$$

where $Q(\alpha, \pi)$ is the quantile of order α of the distribution π.



First example : Bayes-UCB

- $\Pi_0 = (\pi_1(0), \dots, \pi_K(0))$ be a prior distribution over $(\mu_1, ..., \mu_K)$
- $\Pi_t = (\pi_1(t), \dots, \pi_K(t))$ be the posterior distribution over the means $(\mu_1, ..., \mu_K)$ after t observations

Bayes-UCB selects at time $t+1$

$$A_{t+1} = \underset{a=1,\dots,K}{\text{argmax}} \; Q\left(1 - \frac{1}{t(\log t)^c}, \pi_a(t)\right)$$

where $Q(\alpha, \pi)$ is the quantile of order α of the distribution π.

Bernoulli reward with uniform prior :

- $\pi_a(0) \overset{i.i.d}{\sim} \mathcal{U}([0,1]) = \text{Beta}(1,1)$
- $\pi_a(t) = \text{Beta}(S_a(t) + 1, N_a(t) - S_a(t) + 1)$

First example : Bayes-UCB

- $\Pi_0 = (\pi_1(0), \dots, \pi_K(0))$ be a prior distribution over $(\mu_1, ..., \mu_K)$
- $\Pi_t = (\pi_1(t), \dots, \pi_K(t))$ be the posterior distribution over the means $(\mu_1, ..., \mu_K)$ after t observations

Bayes-UCB selects at time $t+1$

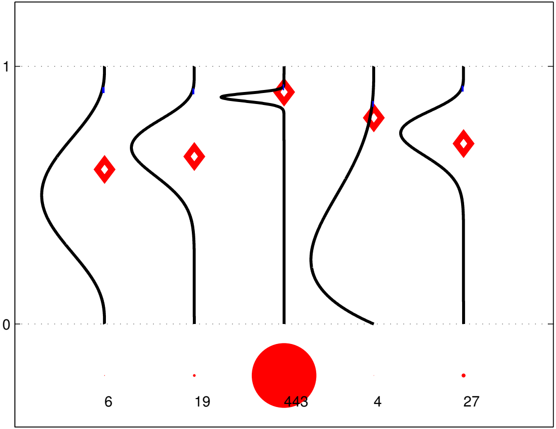
$$A_{t+1} = \underset{a=1,...,K}{\operatorname{argmax}} \; Q\left(1 - \frac{1}{t(\log t)^c}, \pi_a(t)\right)$$

where $Q(\alpha, \pi)$ is the quantile of order α of the distribution π.

Gaussian rewards with Gaussian prior :

- $\pi_a(0) \overset{i.i.d}{\sim} \mathcal{N}(0, \kappa^2)$
- $\pi_a(t) = \mathcal{N}\left(\frac{S_a(t)}{N_a(t)+\sigma^2/\kappa^2}, \frac{\sigma^2}{N_a(t)+\sigma^2/\kappa^2}\right)$

Bayes UCB in action



- Bayes-UCB is also asymptotically optimal for Bernoulli distribution

Outline

1. Performance measure and first strategies

2. Best achievable regret

3. Mixing Exploration and Exploitation
   - Upper Confidence Bound algorithms

4. Bayesian algorithms
   - Thompson Sampling

Thompson Sampling

A very old idea : [Thompson, 1933].

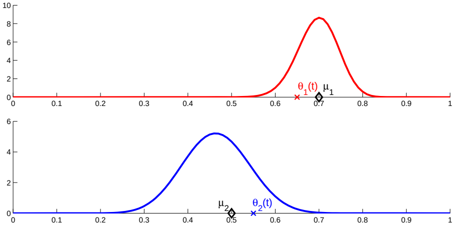
Two equivalent interpretations :

- "select an arm at random according to its probability of being the best"
- "draw a possible bandit model from the posterior distribution and act optimally in this sampled model"

$\neq$ optimistic

Thompson Sampling : a randomized Bayesian algorithm

$$\left\{\begin{array}{l} \forall a \in \{1..K\}, \quad \theta_a(t) \sim \pi_a(t) \\ A_{t+1} = \underset{a=1\dots K}{\text{argmax}}\ \theta_a(t). \end{array}\right.$$



Thompson Sampling is asymptotically optimal

Problem-dependent regret

$$\forall \epsilon > 0, \quad \mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] \leq (1+\epsilon)\frac{1}{\mathrm{kl}(\mu_a, \mu_\star)}\log(T) + o_{\mu,\epsilon}(\log(T)).$$

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]

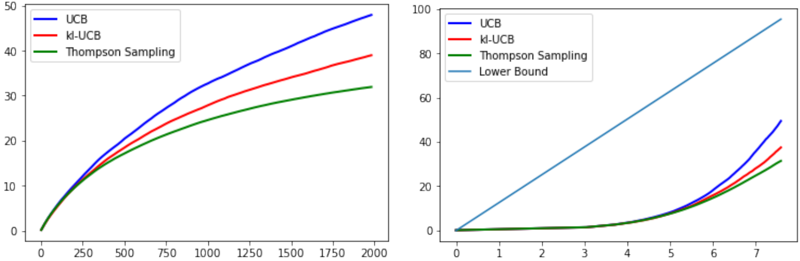
Problem-independent regret [Agrawal and Goyal, 2017]

For Bernoulli and Gaussian bandits, Thompson Sampling satisfies

$$\mathcal{R}_{\boldsymbol{\mu}}(\mathtt{TS}, T) = O\left(\sqrt{KT\log(T)}\right).$$

Bayesian versus Frequentist algorithms

- Regret up to $T = 2000$ (average over $N = 200$ runs) as a function of T (resp. $\log(T)$)



$$\mu = [0.1\ 0.15\ 0.2\ 0.25]$$

Summary

Several ways to solve the exploration/exploitation trade-off, mostly

- the optimism-in-face-of-uncertainty principle (UCB)
- posterior sampling (Thompson Sampling)

What do they need ?

- UCB : the hability to build a confidence region for the unknown model parameters and compute the best possible model
- Thompson Sampling : the ability to define a prior distribution and sample from the corresponding posterior distribution

➜ these principles can be extended to more challenging bandit problems (and to reinforcement learning !)

Agrawal, S. and Goyal, N. (2013).
Further Optimal Regret Bounds for Thompson Sampling.
In *Proceedings of the 16th Conference on Artificial Intelligence and Statistics*.

Agrawal, S. and Goyal, N. (2017).
Near-optimal regret bounds for thompson sampling.
J. ACM, 64(5) :30 :1–30 :24.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002).
Finite-time analysis of the multiarmed bandit problem.
Machine Learning, 47(2) :235–256.

Bubeck, S. (2010).
Jeux de bandits et fondation du clustering.
PhD thesis, Université de Lille 1.

Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., and Stoltz, G. (2013).
Kullback-Leibler upper confidence bounds for optimal sequential allocation.
Annals of Statistics, 41(3) :1516–1541.

Garivier, A. and Cappé, O. (2011).
The KL-UCB algorithm for bounded stochastic bandits and beyond.
In *Proceedings of the 24th Conference on Learning Theory*.

Garivier, A., Kaufmann, E., and Lattimore, T. (2016).
On explore-then-commit strategies.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kaufmann, E., Korda, N., and Munos, R. (2012).
Thompson Sampling : an Asymptotically Optimal Finite-Time Analysis.
In *Proceedings of the 23rd conference on Algorithmic Learning Theory*.

Korda, N., Kaufmann, E., and Munos, R. (2013).
Thompson Sampling for 1-dimensional Exponential family bandits.
In *Advances in Neural Information Processing Systems*.

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

Lattimore, T. and Szepesvari, C. (2019).
Bandit Algorithms.
Cambridge University Press.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010).
A contextual-bandit approach to personalized news article recommendation.
In *WWW*.

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

Sutton, R. and Barto, A. (1998).
Reinforcement Learning : an Introduction.
MIT press.

Thompson, W. (1933).
On the likelihood that one unknown probability exceeds another in view of the evidence of two samples.
Biometrika, 25 :285–294.