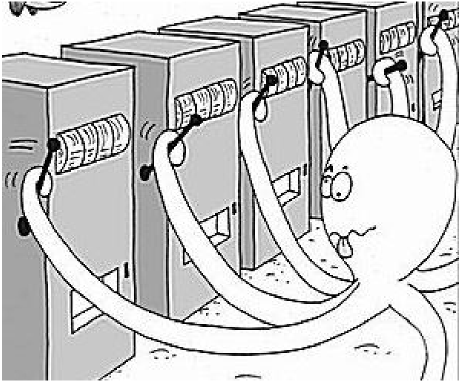
Stochastic Multi-Armed Bandits

Emilie Kaufmann

RLSS Summer School
Barcelona, June 2023

Why bandits ?

- one-armed bandit = old name for a slot machine



an agent facing arms in a Multi-Armed Bandit

Sequential resource allocation

Clinical trials

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



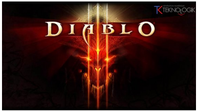



- Which treatment should be allocated to the next patient based on responses observed on previous patients ?

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 ?

Outline

1. The multi-armed bandit problem
2. Simple fixes of the greedy strategy
3. Optimistic Exploration
   - A simple UCB algorithm
   - Towards optimal algorithms
4. Randomized Exploration
   - Thompson Sampling
   - Non Parametric Approaches

Outline

1. The multi-armed bandit problem
2. Simple fixes of the greedy strategy
3. Optimistic Exploration
   - A simple UCB algorithm
   - Towards optimal algorithms
4. Randomized Exploration
   - Thompson Sampling
   - Non Parametric Approaches

The Multi-Armed Bandit Setting

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**) :

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

Goal : Maximize $\sum_{t=1}^{T} R_t$.

The Stochastic Multi-Armed Bandit Setting

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**) :

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

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

➜ solving a one-state MDP for the finite-horizon critirion

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

Regret of a bandit algorithm

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

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

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.

$$\begin{aligned}
\mathcal{R}_\nu(\mathcal{A}, T) &= \mu_\star T - \mathbb{E}\left[\sum_{t=1}^{T} X_{A_t,t}\right] = \mu_\star T - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{A_t}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} (\mu_\star - \mu_{A_t})\right] \\
&= \sum_{a=1}^{K} \underbrace{\mu_\star - \mu_a}_{\Delta_a} \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(A_t = a)\right]}_{N_a(T)}.
\end{aligned}$$

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

The greedy strategy

Select each arm once, then exploit the current knowledge :

$$A_{t+1} = \underset{a \in [K]}{\mathrm{argmax}}\ \hat{\mu}_a(t)$$

where

- $N_a(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = a)$ is the number of selections of arm a
- $\hat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} X_s \mathbb{1}(A_s = a)$ is the empirical mean of the rewards collected from arm a

The greedy strategy

Select each arm once, then exploit the current knowledge :

$$A_{t+1} = \underset{a\in[K]}{\mathrm{argmax}}\ \hat{\mu}_a(t)$$

where

- $N_a(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = a)$ is the number of selections of arm a
- $\hat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} X_s \mathbb{1}(A_s = a)$ is the empirical mean of the rewards collected from arm a

Thre greedy strategy can fail ! $\nu_1 = \mathcal{B}(\mu_1), \nu_2 = \mathcal{B}(\mu_2)$, $\mu_1 > \mu_2$

$$\mathbb{E}[N_2(T)] \geq (1-\mu_1)\mu_2 \times (T-1)$$

➜ **Exploitation** is not enough, we need to add some **exploration**

Outline

1 The multi-armed bandit problem

2 Simple fixes of the greedy strategy

3 Optimistic Exploration
- A simple UCB algorithm
- Towards optimal algorithms

4 Randomized Exploration
- Thompson Sampling
- Non Parametric Approaches

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

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

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

$\rightarrow$ requires a concentration inequality

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

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

$\Rightarrow$ 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

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. $\mu_1 > \mu_2$, $\Delta := \mu_1 - \mu_2$.

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

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. $\mu_1 > \mu_2$, $\Delta := \mu_1 - \mu_2$.

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

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

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

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. $\mu_1 > \mu_2$, $\Delta := \mu_1 - \mu_2$.

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

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

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

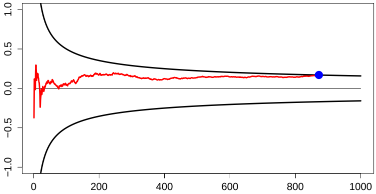
+ logarithmic regret !
− requires the knowledge of T and Δ

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

... but larger regret as that of the best fully sequential strategy

Another possible fix : ϵ-greedy

The ϵ-greedy rule [Sutton and Barto, 1998] is a simple randomized 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 $1-\epsilon$
$$A_t = \underset{a=1,\dots,K}{\mathrm{argmax}} \ \hat{\mu}_a(t).$$

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

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

Another possible fix : ϵ-greedy

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

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

Outline

1 The multi-armed bandit problem

2 Simple fixes of the greedy strategy

3 Optimistic Exploration
- A simple UCB algorithm
- Towards optimal algorithms

4 Randomized Exploration
- Thompson Sampling
- Non Parametric Approaches

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

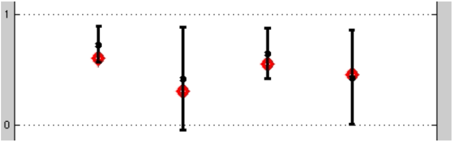


FIGURE – Confidence intervals on the means after t rounds

The optimism principle

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

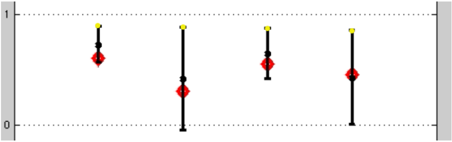


FIGURE – Confidence intervals on the means after t rounds

- That is, select

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

Outline

1 The multi-armed bandit problem

2 Simple fixes of the greedy strategy

3 Optimistic Exploration
- A simple UCB algorithm
- Towards optimal algorithms

4 Randomized Exploration
- Thompson Sampling
- Non Parametric Approaches

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

Reminder : Hoeffding inequality

Z_i i.i.d. with mean μ s.t. $\mathbb{E}[e^{\lambda(Z_1-\mu)}] \leq e^{\frac{\lambda^2\sigma^2}{2}}$. For all $s \geq 1$

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

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

Reminder : Hoeffding inequality

Z_i i.i.d. with mean μ s.t. $\mathbb{E}[e^{\lambda(Z_1-\mu)}] \leq e^{\frac{\lambda^2\sigma^2}{2}}$. 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

$$\mathbb{P}\left(\mu_a \leq \hat{\mu}_a(t) + \sqrt{\frac{6\sigma^2 \log(t)}{N_a(t)}}\right) \geq 1 - \frac{1}{t^2}$$

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

$$\mathbb{P}\left(\mu_a \leq \hat{\mu}_a(t) + \sqrt{\frac{6\sigma^2 \log(t)}{N_a(t)}}\right) \geq 1 - \frac{1}{t^2}$$

Proof.

$$\mathbb{P}\left(\mu_a > \hat{\mu}_a(t) + \sqrt{\frac{6\sigma^2 \log(t)}{N_a(t)}}\right) \leq \mathbb{P}\left(\exists s \leq t : \mu_a > \hat{\mu}_{a,s} + \sqrt{\frac{6\sigma^2 \log(t)}{s}}\right)$$

$$\leq \sum_{s=1}^{t} \mathbb{P}\left(\hat{\mu}_{a,s} < \mu_a - \sqrt{\frac{6\sigma^2 \log(t)}{s}}\right) \leq \sum_{s=1}^{t} \frac{1}{t^3} = \frac{1}{t^2}.$$

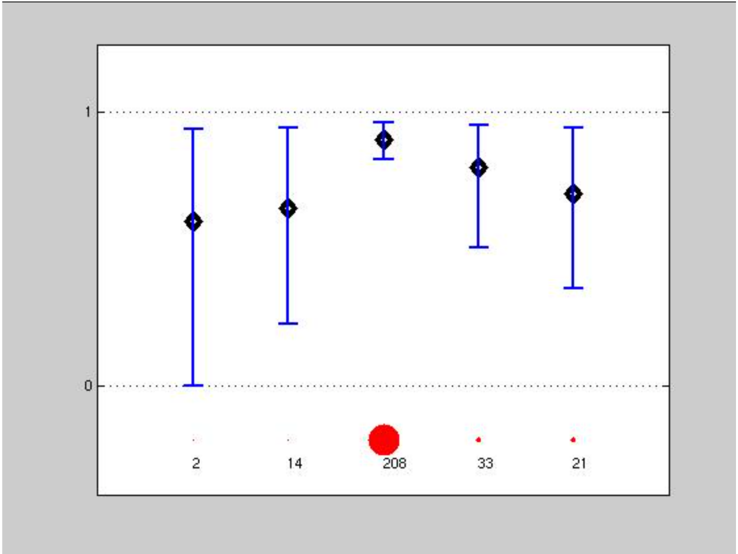
A first UCB algorithm

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

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

- this form of UCB was first proposed for Gaussian rewards [Katehakis and Robbins, 1995]
- 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



A regret bound for UCB(α)

Theorem

For σ^2-subGaussian rewards, the UCB algorithm with parameter $\alpha = 6\sigma^2$ satisfies, for any sub-optimal arm a,

$$\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] \leq \frac{24\sigma^2}{\Delta_a^2}\log(T) + 1 + \frac{\pi^2}{3}$$

where $\Delta_a = \mu_\star - \mu_a$.

Consequence :

$$\mathcal{R}_\nu(\mathrm{UCB}(6\sigma^2), T) \leq \left(\sum_{a:\mu_a<\mu_\star} \frac{24\sigma^2}{\Delta_a}\right)\log(T) + \left(1 + \frac{\pi^2}{3}\right)\sum_{a=1}^{K}\Delta_a$$

Proof (1/2)

For each arm $i \in \{1, a\}$, define the two ends of the confidence interval :

$$\begin{aligned} \mathrm{UCB}_i(t) &= \hat{\mu}_i(t) + \sqrt{\frac{6\sigma^2 \log(t)}{N_i(t)}} \\ \mathrm{LCB}_i(t) &= \hat{\mu}_i(t) - \sqrt{\frac{6\sigma^2 \log(t)}{N_i(t)}} \end{aligned}$$

and the *good event*

$$\mathcal{E}_t = (\mu_1 < \mathrm{UCB}_1(t)) \cap (\mu_a > \mathrm{LCB}_a(t))$$

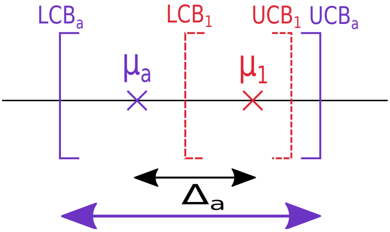
- **Step 1** : Hoeffding inequality + union bound :

$$\mathbb{P}\left(\mathcal{E}_t^c\right) \leq \mathbb{P}\left(\mu_1 > \hat{\mu}_1(t) + \sqrt{\frac{6\sigma^2 \log(t)}{N_1(t)}}\right) + \mathbb{P}\left(\mu_a < \hat{\mu}_a(t) - \sqrt{\frac{6\sigma^2 \log(t)}{N_a(t)}}\right) \leq \frac{2}{t^2}$$

Proof (2/2)

- **Step 2 :** What happens on the good event ?

$$(A_{t+1} = a) \cap (\mu_1 < \mathrm{UCB}_1(t)) \cap (\mu_a > \mathrm{LCB}_a(t))$$

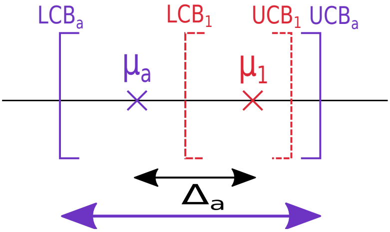


$$\Rightarrow \quad N_a(t) \leq \frac{24\sigma^2 \log(t)}{\Delta_a^2}$$

Proof (2/2)

- **Step 2 :** What happens on the good event ?

$$(A_{t+1} = a) \cap (\mu_1 < \mathrm{UCB}_1(t)) \cap (\mu_a > \mathrm{LCB}_a(t))$$



$$\Rightarrow \quad N_a(t) \leq \frac{24\sigma^2 \log(t)}{\Delta_a^2}$$

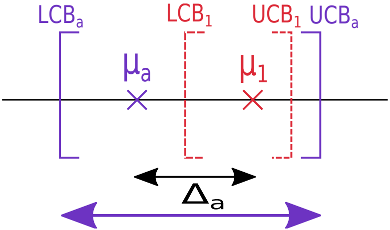
- **Step 3 :** Putting everything together

$$\begin{aligned} \mathbb{E}[N_a(T)] &\leq 1 + \sum_{t=K}^{T-1} \mathbb{P}(\mathcal{E}_t^c) + \sum_{t=K}^{T-1} \mathbb{P}(A_{t+1} = a, \mathcal{E}_t) \\ &\leq 1 + \frac{\pi^2}{3} + \sum_{t=K}^{T-1} \mathbb{P}\left(A_{t+1} = a, N_a(t) \leq \frac{24\sigma^2 \log(T)}{\Delta_a^2}\right) \end{aligned}$$

Proof (2/2)

- **Step 2 :** What happens on the good event ?

$$(A_{t+1} = a) \cap (\mu_1 < \mathrm{UCB}_1(t)) \cap (\mu_a > \mathrm{LCB}_a(t))$$



$$\Rightarrow \quad N_a(t) \leq \frac{24\sigma^2 \log(t)}{\Delta_a^2}$$

- **Step 3 :** Putting everything together

$$\begin{aligned} \mathbb{E}[N_a(T)] &\leq 1 + \sum_{t=K}^{T-1} \mathbb{P}\left(\mathcal{E}_t^c\right) + \sum_{t=K}^{T-1} \mathbb{P}\left(A_{t+1} = a, \mathcal{E}_t\right) \\ &\leq 1 + \frac{\pi^2}{3} + \frac{24\sigma^2 \log(T)}{\Delta_a^2} \end{aligned}$$

A worse-case regret bound

Corollary

$$\mathcal{R}_\nu(\mathrm{UCB}(6\sigma^2), T) \leq 10\sqrt{KT\log(T)} + \left(1 + \frac{\pi^2}{3}\right)\left(\sum_{a=1}^{K} \Delta_a\right)$$

Proof. For any algorithm satisfying $\mathbb{E}[N_a(T)] \leq C\frac{\log(T)}{\Delta_a} + D$ for all sub-optimal arm a, for any $\Delta > 0$,

$$\begin{aligned}
\mathcal{R}_\nu(T) &= \sum_{a:\Delta_a \leq \Delta} \Delta_a \mathbb{E}[N_a(T)] + \sum_{a:\Delta_a \geq \Delta} \Delta_a \mathbb{E}[N_a(T)] \\
&\leq \Delta T + \sum_{a:\Delta_a \geq \Delta} \left(C\frac{\log(T)}{\Delta_a} + D\Delta_a\right) \\
&\leq \Delta T + \frac{CK\log(T)}{\Delta} + D\left(\sum_{a=1}^{K} \Delta_a\right) \\
&= 2\sqrt{CKT\log(T)} + D\left(\sum_{a=1}^{K} \Delta_a\right) \quad \text{for } \Delta = \sqrt{\frac{CK\log(T)}{T}}
\end{aligned}$$

Best known problem-dependent bound

Context : σ^2 sub-Gaussian rewards

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

($c = 0$ corresponds to $\mathrm{UCB}(\alpha)$ *with* $\alpha = 2\sigma^2$*)*

Theorem [Cappé et al.'13]

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

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

Summary

For UCB(α) applied to σ^2-subGaussian reward, setting $\alpha = 2\sigma^2$ yields

- a problem-dependent regret bound of

$$\left(\sum_{a=1}^{K} \frac{2\sigma^2}{\Delta_a}\right) \log(T) + o(\log(T))$$

- a worse-case regret of order

$$O\left(\sqrt{KT \log(T)}\right)$$

➜ how good are these regret rates ?

Outline

1 The multi-armed bandit problem

2 Simple fixes of the greedy strategy

3 Optimistic Exploration
- A simple UCB algorithm
- Towards optimal algorithms

4 Randomized Exploration
- Thompson Sampling
- Non Parametric Approaches

A worse-case lower bound

Theorem [Cesa-Bianchi and Lugosi, 2006]

Fix $T \in \mathbb{N}$. For every bandit algorithm $\mathcal{A}$, there exists a stochastic bandit model ν with rewards supported in $[0,1]$ such that

$$\mathcal{R}_\nu(\mathcal{A}, T) \geq \frac{1}{20}\sqrt{KT}$$

- worse-case model :

$$\left\{\begin{array}{lll} \nu_a & = & \mathcal{B}(1/2) \text{ for all } a \neq i \\ \nu_i & = & \mathcal{B}(1/2 + \Delta) \end{array}\right.$$

with $\Delta \simeq \sqrt{K/T}$.

Remark. UCB achieves $O(\sqrt{KT\log(T)})$ (near-optimal)
There exists worse-case optimal algorithms, e.g., MOSS or Tsallis-Inf [Audibert and Bubeck, 2010, Zimmert and Seldin, 2021]

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 $\nu = (\nu_{\mu_1}, \dots, \nu_{\mu_K})$, $\mu_a \in \mathcal{I}$.

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

Key tool : Kullback-Leibler divergence.

Kullback-Leibler divergence

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

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]

UCB compared to the lower bound

Gaussian distributions with variance σ^2

- **Lower bound** : $\mathbb{E}[N_a(T)] \gtrsim \frac{2\sigma^2}{(\mu_\star - \mu_a)^2} \log(T)$
- **Upper bound** : for UCB(α) with $\alpha = 2\sigma^2$

$$\mathbb{E}[N_a(T)] \lesssim \frac{2\sigma^2}{(\mu_\star - \mu_a)^2} \log(T)$$

➜ UCB is asymptotically optimal for Gaussian rewards !

UCB compared to the lower bound

Gaussian distributions with variance σ^2

- **Lower bound** : $\mathbb{E}[N_a(T)] \gtrsim \frac{2\sigma^2}{(\mu_\star - \mu_a)^2} \log(T)$
- **Upper bound** : for UCB(α) with $\alpha = 2\sigma^2$

$$\mathbb{E}[N_a(T)] \lesssim \frac{2\sigma^2}{(\mu_\star - \mu_a)^2} \log(T)$$

➜ UCB is asymptotically optimal for Gaussian rewards !

Bernoulli distributions (bounded, $\sigma^2 = 1/4$)

- **Lower bound** : $\mathbb{E}[N_a(T)] \gtrsim \frac{1}{\mathrm{kl}(\mu_a, \mu_\star)} \log(T)$
- **Upper bound** : for UCB(α) with $\alpha = 1/2$

$$\mathbb{E}[N_a(T)] \lesssim \frac{1}{2(\mu_\star - \mu_a)^2} \log(T)$$

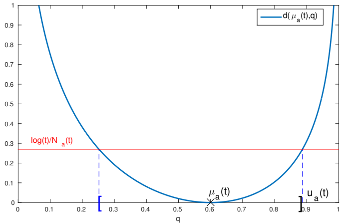
Pinsker's inequality : $\mathrm{kl}(\mu_a, \mu_\star) > 2(\mu_\star - \mu_a)^2$

➜ UCB is *not* asymptotically optimal for Bernoulli rewards...

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 in a one-dimensional exponential family [a],

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

a. e.g., Bernoulli, Gaussian with known variances, Poisson, Exponential

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 Bernoulli rewards (and one-dimenionsal exponential families) :

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

A variant : the IMED algorithm

An interesting alternative proposed by [Honda and Takemura, 2015], that slightly departs from an *index policy*.[1]

Indexed Minimum Emprical Divergence

Compute

$$\hat{\mu}_\star(t) = \max_{a \in [K]} \hat{\mu}_a(t)$$

and select

$$A_{t+1} = \underset{a \in [K]}{\mathrm{argmin}} \left[N_a(t) \mathrm{kl}\left(\hat{\mu}_a(t), \hat{\mu}_\star(t)\right) + \log(N_a(t)) \right]$$

➜ IMED is also asymptotically optimal for exponential families (and beyond)

1. in an index policy, the index computed for each arm depends on the history of this arm only, whereas $\hat{\mu}_\star(t)$ depends on all arms

Outline

1 The multi-armed bandit problem

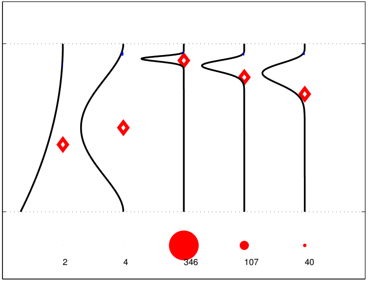
2 Simple fixes of the greedy strategy

3 Optimistic Exploration
- A simple UCB algorithm
- Towards optimal algorithms

4 Randomized Exploration
- Thompson Sampling
- Non Parametric Approaches

A Bayesian algorithm

$\pi_a(0)$: prior distribution on μ_a
$\pi_a(t) = \mathcal{L}(\mu_a | Y_{a,1}, \dots, Y_{a,N_a(t)})$: posterior distribution on μ_a



Two equivalent interpretations :

- [Thompson, 1933] : "randomize the arms according to their posterior probability being optimal"
- modern view : "draw a possible bandit model from the posterior distribution and act optimally in this sampled model"

Russo et al. 2018, *A Tutorial on Thompson Sampling*

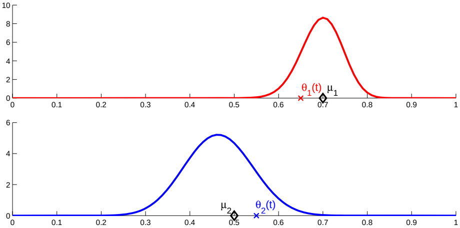
A Bayesian algorithm : Thompson Sampling

Input : a prior distribution $\pi(0)$

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

Thompson Sampling for Bernoulli distributions $\quad \nu_a = \mathcal{B}(\mu_a)$

- $\pi_a(0) = \mathcal{U}([0,1])$
- $\pi_a(t) = \mathrm{Beta}\left(S_a(t)+1; N_a(t)-S_a(t)+1\right)$



A Bayesian algorithm : Thompson Sampling

Input : a prior distribution $\pi(0)$

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

Thompson Sampling for Bernoulli distributions $\nu_a = \mathcal{B}(\mu_a)$

- $\pi_a(0) = \mathcal{U}([0,1])$
- $\pi_a(t) = \mathrm{Beta}\left(S_a(t) + 1; N_a(t) - S_a(t) + 1\right)$

Thompson Sampling for Gaussian distributions $\nu_a = \mathcal{N}(\mu_a, \sigma^2)$

- $\pi_a(0) \propto 1$
- $\pi_a(t) = \mathcal{N}\left(\hat{\mu}_a(t); \frac{\sigma^2}{N_a(t)}\right)$

Regret bounds

Upper bound on sub-optimal selections

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

where $\mathrm{kl}(\mu_a, \mu_\star)$ is the KL divergence between ν_a and $\nu_{a_\star}$

- proved for Bernoulli bandits, with a uniform prior [Kaufmann et al., 2012, Agrawal and Goyal, 2013]
- for 1-dimensional exponential families, with a conjuguate prior [Agrawal and Goyal, 2017, Korda et al., 2013]

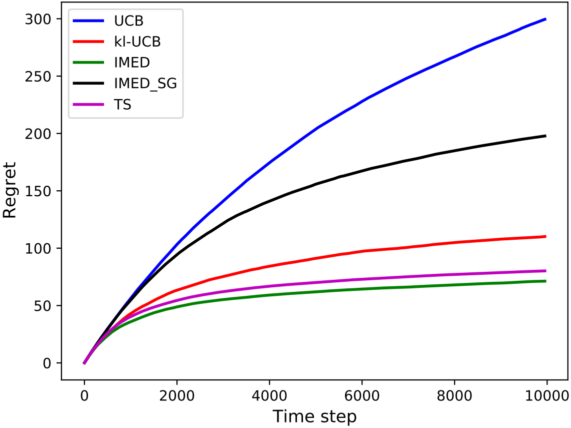
➜ Thompson Sampling is **asymptotically optimal** in these cases

- beyond 1-parameter models, the prior has to be well chosen... [Honda and Takemura, 2014]

Practical performance

Bernoulli arms

$$\boldsymbol{\mu} = [0.1 \ 0.05 \ 0.05 \ 0.05 \ 0.02 \ 0.02 \ 0.02 \ 0.01 \ 0.01 \ 0.01]$$



Outline

1 The multi-armed bandit problem

2 Simple fixes of the greedy strategy

3 Optimistic Exploration
- A simple UCB algorithm
- Towards optimal algorithms

4 Randomized Exploration
- Thompson Sampling
- Non Parametric Approaches

Non parametric algorithms

Thompson Sampling relies on a parametric assumption to maintain a posterior distribution

- Gaussian rewards with known variance : TS with Gaussian prior
- Bernoulli rewards* : TS with Beta prior

Idea : replace the posterior sampling step by a non-parametric history-resampling method

*A binarization trick can be used to handle more general bounded rewards

Perturbed History Exploration

First idea : Non-parameteric Bootstrap

- $\mathcal{H}_{a,t} = (Y_{a,1}, \dots, Y_{a,N_a(t)})$: history of collected rewards from arm a
- sample $N_a(t)$ rewards from $\mathcal{H}_{a,t}$ with replacement, and average them to define an index $B_a(t)$
- $A_{t+1} = \mathrm{argmax}_a \, B_a(t)$

[Kveton et al., 2019b] : linear regret even for two Bernoulli arms

➜ possible fix : Perturbing the history

Perturbed History Exploration (PHE)

$B_a(t)$ is the empirical means of the rewards in $\mathcal{H}_{a,t}$ and $\alpha \times N_a(t)$ fake rewards drawn iid from $\mathcal{B}(1/2)$

➜ $\alpha > 2$: logarithmic regret for bounded rewards in $[0, 1]$ [Kveton et al., 2019a]

Non Parametric Thompson Sampling

Context : rewards bounded in $[0, B]$
Idea : random re-weighting of the augmented history

[Riou and Honda, 2020]

Index of arm a after t rounds

- $\mathcal{H}_{a,t} = (Y_{a,1}, \dots, Y_{a,N_a(t)}, B)$: history of collected rewards from arm a augmented by the upper bound B on the support
- $w_{a,t} \sim \mathrm{Dir}(\underbrace{1, \dots, 1}_{N_a(t)+1})$ a random probability vector

$$B_a(t) = \sum_{s=1}^{N_a(t)} w_{a,t}(s) Y_{a,s} + B w_{a,t}(N_a(t)+1)$$

Non Parameteric Thompson Sampling

Let $\mathcal{B}$ be the set of distributions that are supported on $[0, B]$.

Theorem [Riou and Honda, 2020]

On an instance $\nu = (\nu_1, \dots, \nu_K)$ such that $\nu_a \in \mathcal{B}$ for all a.

$$\mathcal{R}_\nu(\text{NPTS}, T) \leq \sum_{a:\mu_a<\mu_\star} \frac{\Delta_a \log T}{\mathcal{K}_{\text{inf}}(\nu_a, \mu_\star)} + o(\log T) \; .$$

where $\mathcal{K}_{\text{inf}}(\nu, \mu) = \inf\left\{\mathrm{KL}\left(\nu, \nu'\right) : \nu' \in \mathcal{B} : \mathbb{E}_{X\sim\nu'}[X] \geq \mu\right\}$.

➜ matching the lower bound of [Burnetas and Katehakis, 1996] for general (possibly non-parametric) reward distributions

More on Non-Parametric Algorithms

- Extending the idea of Non Parameteric Thompson Sampling beyond bounded distributions

➜ Dirichlet Sampling [Baudry et al., 2021]

- Sub-sampling algorithms : fair pairwise comparison between arms based on sub-sampling the most selected one

➜ BESA[Baransi et al., 2014], SSMC [Chan, 2020]
SDA algorithms [Baudry et al., 2020]

Conclusion

We saw several principles to solve the exploration/exploitation trade-off in a simple bandit model, with strong guarantees on their regret, e.g.,

- the use of confidence intervals
- posterior sampling or randomized mechanisms

They can be extended to more challenging tasks such that contextual bandits or regret minimization in reinforcement learning
(see tomorrow's classes)

Bandit strategies such as UCB have also served as an inspiration for some Monte-Carlo Tree Search strategies
(see this afternoon's class)

References

The Bandit Book

by [Lattimore and Szepesvari, 2019]

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.

Audibert, J.-Y. and Bubeck, S. (2010).
Regret Bounds and Minimax Policies under Partial Monitoring.
Journal of Machine Learning Research.

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

Baransi, A., Maillard, O., and Mannor, S. (2014).
Sub-sampling for multi-armed bandits.
In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML / PKDD*.

Baudry, D., Kaufmann, E., and Maillard, O.-A. (2020).
Sub-sampling for Efficient Non-Parametric Bandit Exploration.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

Baudry, D., Saux, P., and Maillard, O. (2021).
From optimality to robustness : Dirichlet sampling strategies in stochastic bandits.

In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Burnetas, A. and Katehakis, M. (1996).
Optimal adaptive policies for sequential allocation problems.
Advances in Applied Mathematics, 17(2) :122–142.

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.

Cesa-Bianchi, N. and Lugosi, G. (2006).
Prediction, Learning and Games.
Cambridge University Press.

Chan, H. P. (2020).
The multi-armed bandit problem : An efficient nonparametric solution.
The Annals of Statistics, 48(1).

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)*.

Honda, J. and Takemura, A. (2014).
Optimality of Thompson Sampling for Gaussian Bandits depends on priors.
In *Proceedings of the 17th conference on Artificial Intelligence and Statistics*.

Honda, J. and Takemura, A. (2015).
Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards.
Journal of Machine Learning Research, 16 :3721–3756.

Katehakis, M. and Robbins, H. (1995).
Sequential choice from several populations.
Proceedings of the National Academy of Science, 92 :8584–8585.

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*.

Kveton, B., Szepesvári, C., Ghavamzadeh, M., and Boutilier, C. (2019a).
Perturbed-history exploration in stochastic multi-armed bandits.
In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*.

Kveton, B., Szepesvári, C., Vaswani, S., Wen, Z., Lattimore, T., and Ghavamzadeh, M. (2019b).
Garbage in, reward out : Bootstrapping exploration in multi-armed bandits.
In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

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*.

Riou, C. and Honda, J. (2020).
Bandit algorithms based on thompson sampling for bounded reward distributions.
In *Algorithmic Learning Theory (ALT)*.

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.

Zimmert, J. and Seldin, Y. (2021).
Tsallis-inf : An optimal algorithm for stochastic and adversarial bandits.
Journal of Machine Learning Research, 22 :28 :1–28 :49.