

Quiz Time

Check out Slack

We'll begin the class at 9h15

Reinforcement Learning

Some Insights from the Bandit Literature

Emilie Kaufmann & Claire Vernade



M2 MVA, 2024/2025

RL : Taking a step back

RL \leftrightarrow Learn a good policy in an unknown Markov Decision Process

Good policy : according to some notion of **value**

$$V^\pi(s) = \mathbb{E}^\pi \left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \middle| s_1 = s \right] \text{ or } V^\pi(s) = \mathbb{E}^\pi \left[\sum_{t=1}^H r_t \middle| s_1 = s \right]$$

RL : Taking a step back

RL \leftrightarrow Learn a **good policy** in an unknown Markov Decision Process

Good policy : according to some notion of **value**

$$V^\pi(s) = \mathbb{E}^\pi \left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \middle| s_1 = s \right] \text{ or } V^\pi(s) = \mathbb{E}^\pi \left[\sum_{t=1}^H r_t \middle| s_1 = s \right]$$

In previous lectures, you learnt about **optimize policies in MDPs** and about **approximate Value Iteration** algorithms like Q-learning.

RL : Taking a step back

RL \leftrightarrow Learn a **good policy** in an unknown Markov Decision Process

Good policy : according to some notion of **value**

$$V^\pi(s) = \mathbb{E}^\pi \left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \middle| s_1 = s \right] \text{ or } V^\pi(s) = \mathbb{E}^\pi \left[\sum_{t=1}^H r_t \middle| s_1 = s \right]$$

In previous lectures, you learnt about **optimize policies in MDPs** and about **approximate Value Iteration** algorithms like Q-learning.
So far, **convergence**, when proved, is asymptotic.

Main questions

- ▶ What is the convergence rate for RL algorithms ?
- ▶ How is it influenced by the exploration strategy ?
- ▶ How about stochastic environments ?

RL : Taking a step back

RL \leftrightarrow Learn a good policy in an unknown Markov Decision Process

Good policy : according to some notion of value

$$V^\pi(s) = \mathbb{E}^\pi \left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \middle| s_1 = s \right] \text{ or } V^\pi(s) = \mathbb{E}^\pi \left[\sum_{t=1}^H r_t \middle| s_1 = s \right]$$

In previous lectures, you learnt about optimize policies in MDPs and about approximate Value Iteration algorithms like Q-learning.
So far, convergence, when proved, is asymptotic.

Main questions

- ▶ What is the convergence rate for RL algorithms ?
- ▶ How is it influenced by the exploration strategy ?
- ▶ How about stochastic environments ?

Objectives of this lecture

- ▶ In-depth study of the simplest MDP : the multi-armed bandit
 - Stochastic bandit algorithms
 - Towards a more realistic model : contextual bandits
 - Regret or Sample complexity ?
- ▶ By the end of this lecture, you will :
 - Understand the notion of Regret
 - Know how to bound it for simple algorithms : ε -greedy, UCB
 - Know the Optimism Principle and its consequences on algorithm design

Reinforcement Learning

Lecture 7 : Multi-armed bandits

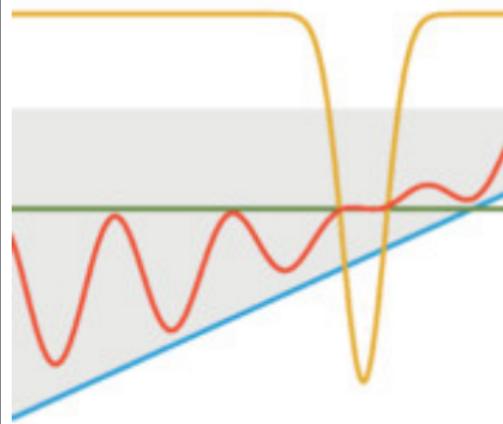
Emilie Kaufmann & Claire Vernade



M2 MVA, 2024/2025

Bandit Algorithms

TOR LATTIMORE
CSABA SZEPESVÁRI

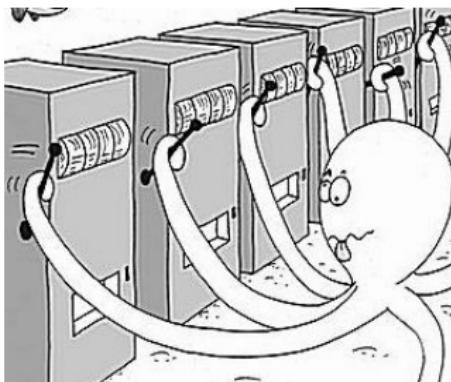


'The Bandit Book'
by [Lattimore and Szepesvari, 2019]

Stochastic bandit : a simple MDP

A stochastic multi-armed bandit model is an MDP with a single state s_0

- ▶ unknown reward distribution $\nu_{s_0, a}$ with mean $r(s_0, a)$
- ▶ transition $p(s_0 | s_0, a) = 1$
- ▶ the agent repeatedly chooses between the same set of actions



an agent facing arms in a Multi-Armed Bandit

Typical applications

Clinical trials

- ▶ K treatments 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, movie recommendation

- ▶ Catalog of K movies



- ▶ Which movie should be displayed for a user, based on the feedback from similar users ?

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**) :

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

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 $R_t = X_{A_t, t} \sim \nu_{A_t}$

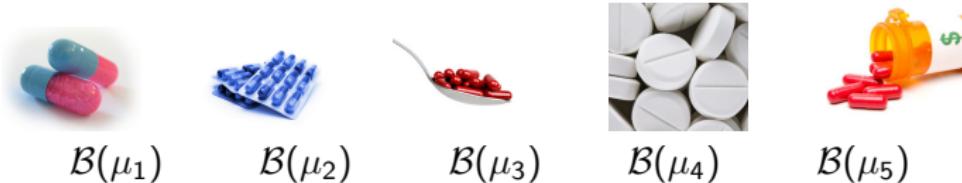
Sequential sampling strategy (**bandit algorithm**) :

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

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

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

Regret of a bandit algorithm

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

$$\mu_* = \max_{a \in \{1, \dots, K\}} \mu_a \quad a_* = \operatorname{argmax}_{a \in \{1, \dots, K\}} \mu_a.$$

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

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

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

Can we think of a better (equivalent) definition of the Regret ?

$$\mathcal{R}_\nu(\mathcal{A}, T) := T\mu_\star - \mathbb{E} \left[\sum_{t=1}^T R_t \right]$$

Regret decomposition

Regret decomposition

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

Proof.

Regret decomposition

Regret decomposition

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

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

⇒ Exploration / Exploitation trade-off

The greedy strategy

Select each arm once, then **exploit** the current knowledge :

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

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 :

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

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 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 Fixing the greedy strategy
- 2 Optimistic Exploration
 - A simple UCB algorithm
 - Towards optimal algorithms
- 3 Randomized Exploration : Thompson Sampling
- 4 Contextual Bandits
 - Lin-UCB
 - Linear Thompson Sampling
- 5 Bandits beyond Regret

Explore-Then-Commit

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

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

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

⇒ EXPLORATION followed by EXPLOITATION

Explore-Then-Commit

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

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

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

⇒ EXPLORATION followed by EXPLOITATION

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

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

$\hat{\mu}_{a,m}$: 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 $\hat{a} = \operatorname{argmax}_a \hat{\mu}_a(Km)$
- ▶ keep playing this arm until round T

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

⇒ EXPLORATION followed by EXPLOITATION

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

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

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

A Concentration Inequality

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

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

Hoeffding inequality

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

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

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

A Concentration Inequality

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

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

Hoeffding inequality

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

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

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

Explore-Then-Commit

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

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

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

⇒ 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 $[0, 1]$.

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

$\hat{\mu}_{a,m}$: empirical mean of the first m observations from arm a
→ Hoeffding's inequality

Explore-Then-Commit

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

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

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

⇒ 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 $[0, 1]$.

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

$\hat{\mu}_{a,m}$: empirical mean of the first m observations from arm a
→ Hoeffding's inequality

Explore-Then-Commit

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

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

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

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

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

Explore-Then-Commit

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

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

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

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

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

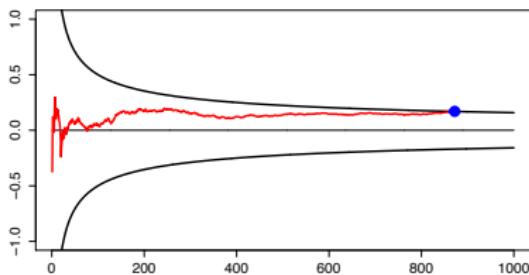
- + logarithmic regret !
- requires the knowledge of T and Δ

Sequential Explore-Then-Commit

- ▶ explore uniformly until a random time of the form

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

- ▶ $\hat{a}_\tau = \operatorname{argmax}_a \hat{\mu}_a(\tau)$ 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 $1 - \epsilon$

$$A_t = \operatorname*{argmax}_{a=1, \dots, K} \hat{\mu}_a(t).$$

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 $1 - \epsilon$

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

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

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

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 $1 - \epsilon_t$

$$A_t = \operatorname{argmax}_{a=1, \dots, K} \hat{\mu}_a(t-1).$$

Theorem [Auer et al., 2002]

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

→ requires the knowledge of a lower bound on Δ_{\min}

Outline

- 1 Fixing the greedy strategy
- 2 Optimistic Exploration
 - A simple UCB algorithm
 - Towards optimal algorithms
- 3 Randomized Exploration : Thompson Sampling
- 4 Contextual Bandits
 - Lin-UCB
 - Linear Thompson Sampling
- 5 Bandits beyond Regret

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 :

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

LCB = Lower Confidence Bound

UCB = Upper Confidence Bound

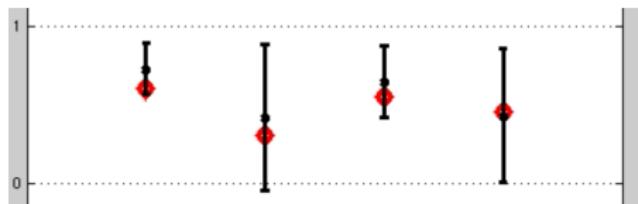


FIGURE – Confidence intervals on the means after t rounds

The optimism principle

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

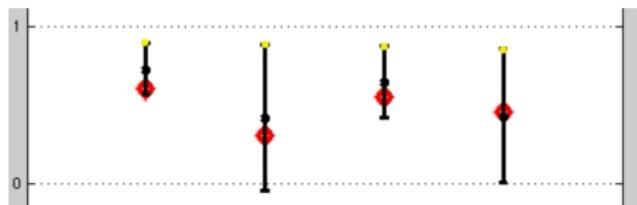


FIGURE – Confidence intervals on the means after t rounds

- ▶ That is, select

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

Outline

- 1 Fixing the greedy strategy
- 2 Optimistic Exploration
 - A simple UCB algorithm
 - Towards optimal algorithms
- 3 Randomized Exploration : Thompson Sampling
- 4 Contextual Bandits
 - Lin-UCB
 - Linear Thompson Sampling
- 5 Bandits beyond Regret

How to build confidence intervals ?

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

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

→ 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 $\text{UCB}_a(t)$ such that

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

→ 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.

$$\begin{aligned} \mathbb{P}\left(\mu_a > \hat{\mu}_a(t) + \sqrt{\frac{6\sigma^2 \log(t)}{N_a(t)}}\right) &\leq \mathbb{P}\left(\exists s \leq t : \mu_a > \hat{\mu}_{a,s} + \sqrt{\frac{6\sigma^2 \log(t)}{s}}\right) \\ &\leq \sum_{s=1}^t \mathbb{P}\left(\hat{\mu}_{a,s} < \mu_a - \sqrt{\frac{6\sigma^2 \log(t)}{s}}\right) \leq \sum_{s=1}^t \frac{1}{t^3} = \frac{1}{t^2}. \end{aligned}$$

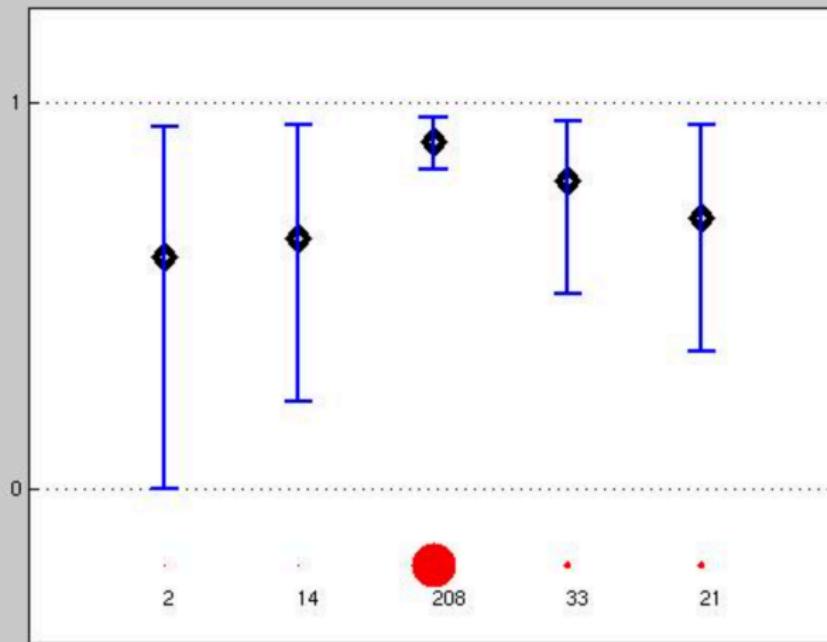
A first UCB algorithm

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

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

- ▶ 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]

UCB in action



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 ,

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

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

Consequence :

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

Proof (1/2)

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

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

and the *good event*

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

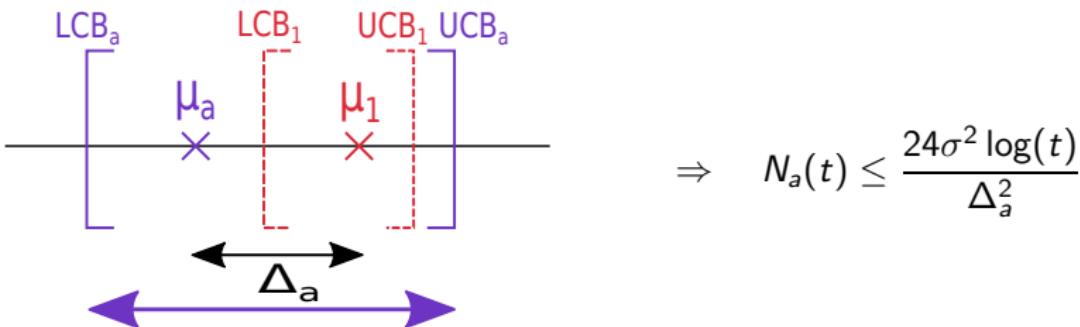
► **Step 1** : Hoeffding inequality + union bound :

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

Proof (2/2)

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

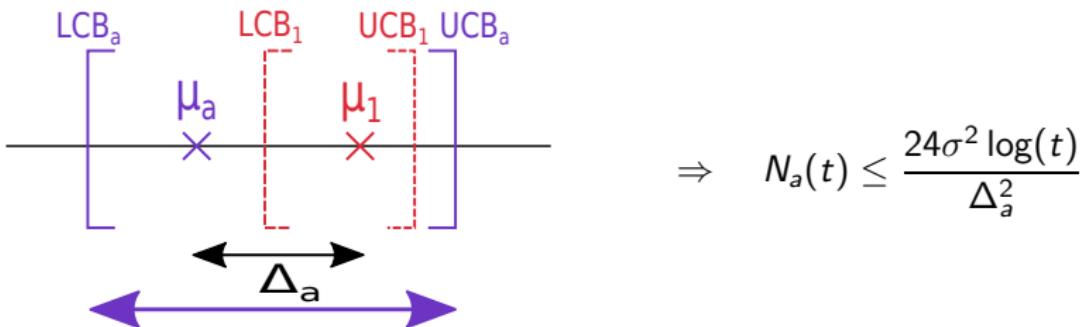
$$(A_{t+1} = a) \cap (\mu_1 < \text{UCB}_1(t)) \cap (\mu_a > \text{LCB}_a(t))$$



Proof (2/2)

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

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



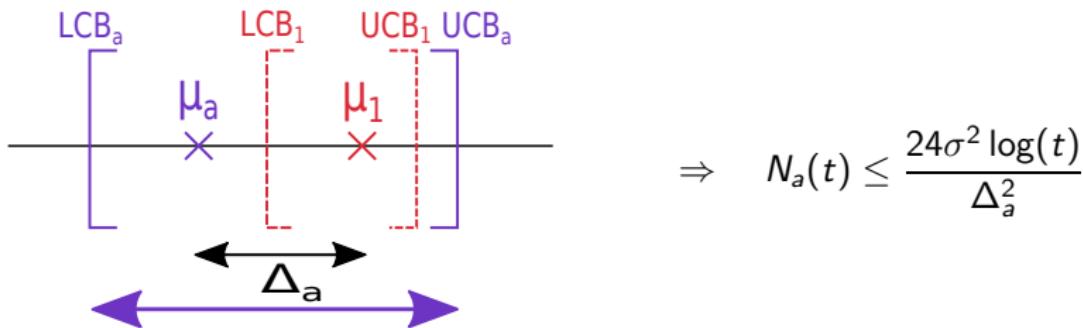
- ▶ Step 3 : Putting everything together

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

Proof (2/2)

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

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



- **Step 3 :** Putting everything together

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

A worst-case regret bound

Corollary

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

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 $\Delta > 0$,

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

Best known problem-dependent bound

Context : σ^2 sub-Gaussian rewards

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

($c = 0$ corresponds to UCB(α) with $\alpha = 2\sigma^2$)

Theorem [Cappé et al.'13]

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

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

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 Fixing the greedy strategy
- 2 Optimistic Exploration
 - A simple UCB algorithm
 - Towards optimal algorithms
- 3 Randomized Exploration : Thompson Sampling
- 4 Contextual Bandits
 - Lin-UCB
 - Linear Thompson Sampling
- 5 Bandits beyond Regret

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 :

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

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

Remark. UCB achieves $\mathcal{O}(\sqrt{KT \log(T)})$ (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 $\nu = (\nu_{\mu_1}, \dots, \nu_{\mu_K})$, $\mu_a \in \mathcal{I}$.

$$\nu \leftrightarrow \mu = (\mu_1, \dots, \mu_K)$$

Key tool : Kullback-Leibler divergence.

Kullback-Leibler divergence

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

Theorem

For uniformly good algorithm,

$$\mu_a < \mu_* \Rightarrow \liminf_{T \rightarrow \infty} \frac{\mathbb{E}_\mu[N_a(T)]}{\log T} \geq \frac{1}{\text{kl}(\mu_a, \mu_*)}$$

[Lai and Robbins, 1985]

The Lai and Robbins lower bound

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

$$\nu \leftrightarrow \mu = (\mu_1, \dots, \mu_K)$$

Key tool : Kullback-Leibler divergence.

Kullback-Leibler divergence

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

Theorem

For uniformly good algorithm,

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

[Lai and Robbins, 1985]

The Lai and Robbins lower bound

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

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

Key tool : Kullback-Leibler divergence.

Kullback-Leibler divergence

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

Theorem

For *uniformly good* algorithm,

$$\mu_a < \mu_* \Rightarrow \liminf_{T \rightarrow \infty} \frac{\mathbb{E}_\mu[N_a(T)]}{\log T} \geq \frac{1}{\text{kl}(\mu_a, \mu_*)}$$

[Lai and Robbins, 1985]

UCB compared to the lower bound

Gaussian distributions with variance σ^2

► **Lower bound** : $\mathbb{E}[N_a(T)] \gtrsim \frac{2\sigma^2}{(\mu_* - \mu_a)^2} \log(T)$

► **Upper bound** : for UCB(α) with $\alpha = 2\sigma^2$

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

→ UCB is asymptotically optimal for Gaussian rewards !

UCB compared to the lower bound

Gaussian distributions with variance σ^2

► **Lower bound** : $\mathbb{E}[N_a(T)] \gtrsim \frac{2\sigma^2}{(\mu_* - \mu_a)^2} \log(T)$

► **Upper bound** : for UCB(α) with $\alpha = 2\sigma^2$

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

→ UCB is asymptotically optimal for Gaussian rewards !

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

► **Lower bound** : $\mathbb{E}[N_a(T)] \gtrsim \frac{1}{\text{kl}(\mu_a, \mu_*)} \log(T)$

► **Upper bound** : for UCB(α) with $\alpha = 1/2$

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

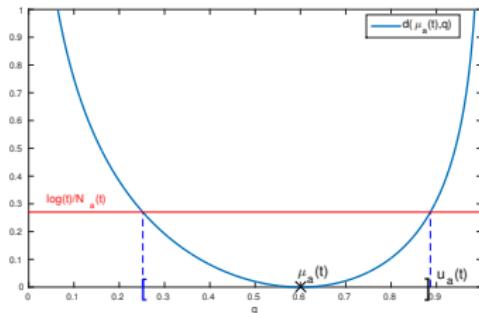
Pinsker's inequality : $\text{kl}(\mu_a, \mu_*) > 2(\mu_* - \mu_a)^2$

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

The kl-UCB algorithm

Exploits the KL-divergence in the lower bound !

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



A tighter concentration inequality [Garivier and Cappé, 2011]

For rewards in a one-dimensional exponential family^a,

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

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

An asymptotically optimal algorithm

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

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

Theorem [Cappé et al., 2013]

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

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

- ▶ **asymptotically optimal** for Bernoulli rewards (and one-dimensional exponential families) :

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

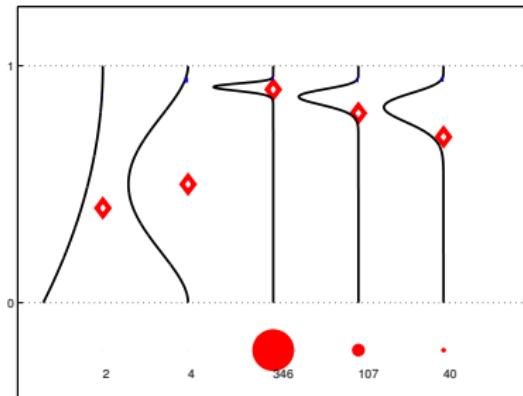
Outline

- 1 Fixing the greedy strategy
- 2 Optimistic Exploration
 - A simple UCB algorithm
 - Towards optimal algorithms
- 3 Randomized Exploration : Thompson Sampling
- 4 Contextual Bandits
 - Lin-UCB
 - Linear Thompson Sampling
- 5 Bandits beyond Regret

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”

Thompson Sampling

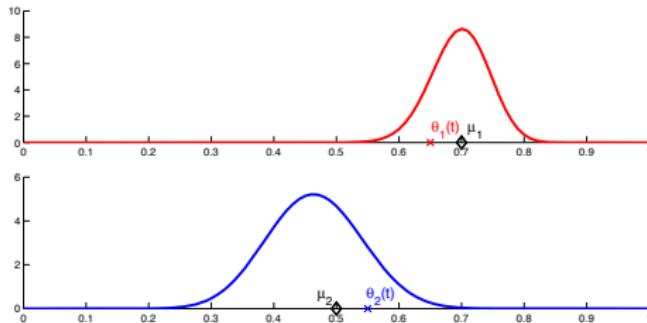
Input : a prior distribution $\pi(0)$

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

Thompson Sampling for Bernoulli distributions

$$\nu_a = \mathcal{B}(\mu_a)$$

- ▶ $\pi_a(0) = \mathcal{U}(0, 1)$
- ▶ $\pi_a(t) = \text{Beta}(S_a(t) + 1; N_a(t) - S_a(t) + 1)$



Thompson Sampling

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

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

Thompson Sampling for Bernoulli distributions

$$\nu_a = \mathcal{B}(\mu_a)$$

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

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

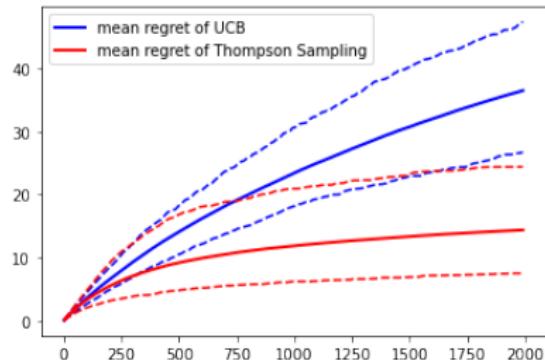
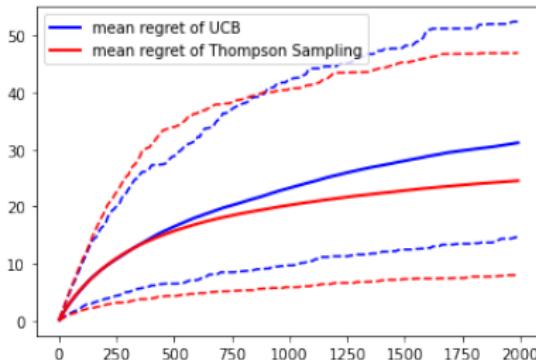
$$\forall a \neq a_*, \quad \mathbb{E}_\mu[N_a(T)] \leq \frac{\log(T)}{\text{kl}(\mu_a, \mu_*)} + o_\mu(\log(T)).$$

where $\text{kl}(\mu_a, \mu_*)$ is the KL divergence between ν_a and ν_{a_*}

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

Regret curves for UCB ($\alpha = 1/2$) and Thompson Sampling on two Bernoulli bandit problems, averaged over 500 runs.



Who is who ? Try it out !

$$\mu_A = [0.45 \ 0.5 \ 0.6]$$

$$\mu_B = [0.1 \ 0.05 \ 0.02 \ 0.01]$$

Summary so far

Several important ideas to tackle the exploration/exploitation challenge in a simple multi-armed bandit model with independent arms :

- ▶ Explore then Commit
- ▶ ε -greedy
- ▶ Optimistic algorithms : Upper Confidence Bounds strategies
- ▶ Randomized (Bayesian) exploration : Thompson Sampling

Can these ideas be extended to more **structured** models that are better suited for applications ?

Outline

- 1 Fixing the greedy strategy
- 2 Optimistic Exploration
 - A simple UCB algorithm
 - Towards optimal algorithms
- 3 Randomized Exploration : Thompson Sampling
- 4 Contextual Bandits
 - Lin-UCB
 - Linear Thompson Sampling
- 5 Bandits beyond Regret

Motivation



Which movie should Netflix recommend to a particular user, given the ratings provided by previous users ?

- to make good recommendation, we should take into account the characteristics of the movies / users

Arm in $\{1, 2, \dots, K\} \leftrightarrow$ Context vector in some space \mathcal{X}

A **contextual bandit model** incorporates two components :

- ▶ a sequential interaction protocol :
pick an arm, receive a reward
- ▶ a regression model for the dependency between context and reward

Generic Contextual Bandit Model

In each round t , the agent

- ▶ is given a set of *arms* $\mathcal{X}_t \subseteq \mathcal{X}$ *(can be different in each round)*
- ▶ selects an *arm* $x_t \in \mathcal{X}_t$
- ▶ receives a reward

$$r_t = f_*(x_t) + \varepsilon_t$$

where

- $f_* : \mathcal{X} \rightarrow \mathbb{R}$ is an unknown regression function
- ε_t is a centered noise, independent from previous data

Generic Contextual Bandit Model

In each round t , the agent

- ▶ is given a set of *arms* $\mathcal{X}_t \subseteq \mathcal{X}$ (*can be different in each round*)
- ▶ selects an *arm* $x_t \in \mathcal{X}_t$
- ▶ receives a reward

$$r_t = f_*(x_t) + \varepsilon_t$$

where

- $f_* : \mathcal{X} \rightarrow \mathbb{R}$ is an unknown regression function
- ε_t is a centered noise, independent from previous data

Example

- user t : descriptor $c_t \in \mathbb{R}^p$
- item a : descriptor $x_a \in \mathbb{R}^{p'}$
- build a user-item feature vector : $\psi : (x_a, c_t) \mapsto x_{t,a} = \psi(x_a, c_t) \in \mathbb{R}^d$

$$\mathcal{X}_t = \{x_{t,a}, a \in \mathcal{K}_t\}$$

Contextual linear bandits

In each round t , the agent

- ▶ receives a (finite) set of arms $\mathcal{X}_t \subseteq \mathbb{R}^d$
- ▶ chooses an arm $x_t \in \mathcal{X}_t$
- ▶ gets a reward $r_t = \theta_\star^\top x_t + \varepsilon_t$

where

- $\theta_\star \in \mathbb{R}^d$ is an unknown regression vector
- ε_t is a centered noise, independent from past data

Assumption : σ^2 - sub-Gaussian noise

$$\forall \lambda \in \mathbb{R}, \mathbb{E} [e^{\lambda \varepsilon_t} | \mathcal{F}_{t-1}, x_t] \leq e^{\frac{\lambda^2 \sigma^2}{2}}$$

e.g., Gaussian noise, bounded noise.

Contextual linear bandits

In each round t , the agent

- ▶ receives a (finite) set of arms $\mathcal{X}_t \subseteq \mathbb{R}^d$
- ▶ chooses an arm $x_t \in \mathcal{X}_t$
- ▶ gets a reward $r_t = \theta_\star^\top x_t + \varepsilon_t$

where

- $\theta_\star \in \mathbb{R}^d$ is an unknown regression vector
- ε_t is a centered noise, independent from past data

(Pseudo)-regret for contextual bandit

maximizing expected total reward \leftrightarrow minimizing the (expectation of)

$$R_T(\mathcal{A}) = \sum_{t=1}^T \left(\max_{x \in \mathcal{X}_t} \theta_\star^\top x - \theta_\star^\top x_t \right)$$

- in each round, comparison to a possibly different optimal action !

Tools

Algorithms will rely on estimates / confidence regions / posterior distributions for $\theta_* \in \mathbb{R}^d$.

- ▶ design matrix (with regularization parameter $\lambda > 0$)

$$B_t^\lambda = \lambda I_d + \sum_{s=1}^t x_s x_s^\top$$

- ▶ regularized least-square estimate

$$\hat{\theta}_t^\lambda = (B_t^\lambda)^{-1} \left(\sum_{s=1}^t r_s x_s \right)$$

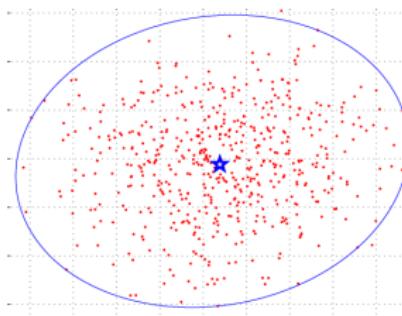
- ▶ estimate of the expected reward of an arm $x \in \mathbb{R}^d$: $x^\top \hat{\theta}_t^\lambda$
- sufficient for ε -greedy or ETC, but not for smarter algorithms...

Outline

- 1 Fixing the greedy strategy
- 2 Optimistic Exploration
 - A simple UCB algorithm
 - Towards optimal algorithms
- 3 Randomized Exploration : Thompson Sampling
- 4 Contextual Bandits
 - Lin-UCB
 - Linear Thompson Sampling
- 5 Bandits beyond Regret

How to build (tight) confidence interval on the mean rewards ?

Idea : rely on a confidence ellipsoid around $\hat{\theta}_t^\lambda$



$$\theta_* \in \left\{ \theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_t^\lambda\|_A \leq \beta_t \right\}$$

Why? For all invertible positive semi-definite matrix A ,

$$\forall x \in \mathbb{R}^d, \quad \left| x^\top \theta_* - x^\top \hat{\theta}_t^\lambda \right| \leq \|x\|_{A^{-1}} \left\| \theta_* - \hat{\theta}_t^\lambda \right\|_A$$

$$\|x\|_A = \sqrt{x^\top Ax}$$

How to build (tight) confidence interval on the mean rewards ?

Wanted : $\theta_\star \in \left\{ \theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_t^\lambda\|_A \leq \beta_t \right\}$

Example of threshold [Abbasi-Yadkori et al., 2011]

Assuming that the noise ε_t is σ^2 -sub-Gaussian, and that for all t and $x \in \mathcal{X}_t$, $\|x\| \leq L$, we have

$$\mathbb{P} \left(\exists t \in \mathbb{N}^* : \|\theta_\star - \hat{\theta}_t^\lambda\|_{B_t^\lambda} > \beta(t, \delta) \right) \leq \delta$$

with $\beta(t, \delta) = \sigma \sqrt{2 \log(1/\delta) + d \log(1 + t \frac{L}{d\lambda})} + \sqrt{\lambda} \|\theta_\star\|$.

→ Letting

$$C_t(\delta) = \left\{ \theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_t^\lambda\|_{B_t^\lambda} \leq \beta(t, \delta) \right\},$$

one has $\mathbb{P}(\forall t \in \mathbb{N}, \theta_\star \in C_t(\delta)) \geq 1 - \delta$.

A Lin-UCB algorithm

Consequence :

$$\mathbb{P}\left(\forall t \in \mathbb{N}^*, \forall x \in \mathcal{X}_{t+1}, \underbrace{x^\top \theta_\star}_{\text{unknown mean of arm } x} \leq \underbrace{x^\top \hat{\theta}_t^\lambda + \|x\|_{(B_t^\lambda)^{-1}} \beta(t, \delta)}_{\text{Upper Confidence Bound}}\right) \geq 1 - \delta.$$

One can assign to each arm $x \in \mathcal{X}_{t+1}$

$$\text{UCB}_x(t) = \underbrace{x^\top \hat{\theta}_t^\lambda}_{\text{empirical mean (exploitation term)}} + \underbrace{\|x\|_{(B_t^\lambda)^{-1}} \beta(t, \delta)}_{\text{exploration bonus}}$$

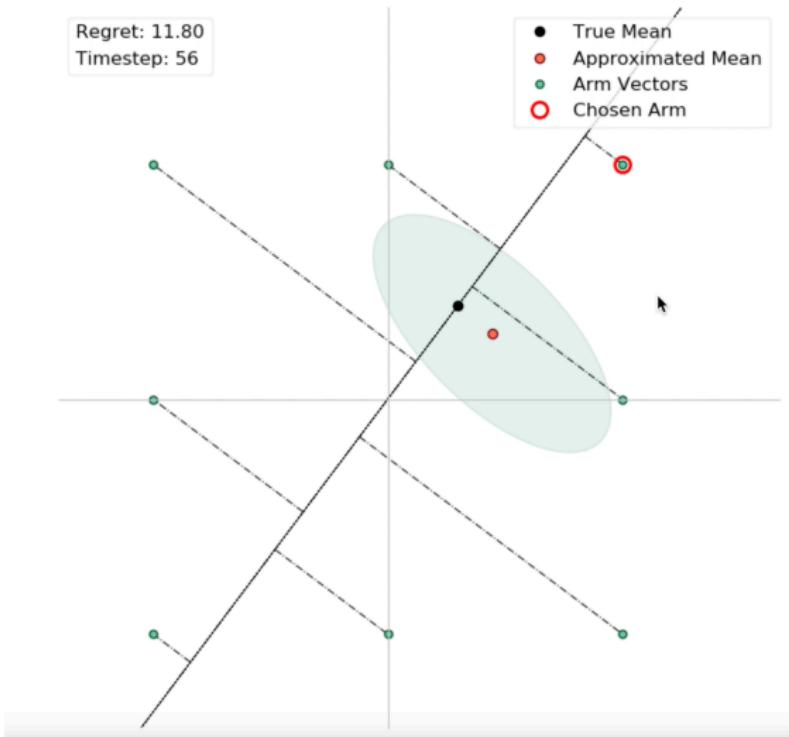
Lin-UCB

In each round $t + 1$, the algorithm selects

$$x_{t+1} = \operatorname{argmax}_{x \in \mathcal{X}_{t+1}} \left[x^\top \hat{\theta}_t^\lambda + \|x\|_{(B_t^\lambda)^{-1}} \beta(t, \delta) \right]$$

(many algorithms of this style, with different choices of $\beta(t, \delta)$)

LinUCB in action



Theoretical guarantees

We want to bound the **pseudo-regret**

$$R_T(\text{Lin-UCB}) = \sum_{t=1}^T \left(\max_{x \in \mathcal{X}_t} \theta_\star^\top x - \theta_\star^\top x_t \right)$$

or its expectation, the **regret** $\mathcal{R}_T(\text{Lin-UCB}) = \mathbb{E}[R_T(\text{Lin-UCB})]$.

Lemma

One can prove that, with probability larger than $1 - \delta$,

$$\forall T \in \mathbb{N}^*, R_T(\text{Lin-UCB}) \leq C\beta(T, \delta) \sqrt{dT \log(T)}$$

- ▶ with the choice of $\beta(t, \delta)$ presented before, with high probability

$$R_T(\text{Lin-UCB}) = \mathcal{O}(d\sqrt{T} \log(T) + \sqrt{dT \log(T) \log(1/\delta)})$$

- ▶ choosing $\delta = 1/T$, $\mathcal{R}_T(\text{Lin-UCB}) = \mathcal{O}(d\sqrt{T} \log(T))$

Outline

- 1 Fixing the greedy strategy
- 2 Optimistic Exploration
 - A simple UCB algorithm
 - Towards optimal algorithms
- 3 Randomized Exploration : Thompson Sampling
- 4 Contextual Bandits
 - Lin-UCB
 - Linear Thompson Sampling
- 5 Bandits beyond Regret

A Bayesian view on Linear Regression

Bayesian model :

- ▶ likelihood : $r_t = \theta_\star^\top x_t + \varepsilon_t$
- ▶ prior : $\theta_\star \sim \mathcal{N}(0, \kappa^2 I_d)$

Assuming further that the noise is Gaussian : $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$, the posterior distribution of θ_\star has a closed form :

$$\theta_\star | x_1, r_1, \dots, x_t, r_t \sim \mathcal{N}\left(\hat{\theta}_t^\lambda, \sigma^2 (B_t^\lambda)^{-1}\right)$$

with

- $B_t^\lambda = \lambda I_d + \sum_{s=1}^t x_s x_s^\top$
- $\hat{\theta}_t^\lambda = (B_t^\lambda)^{-1} (\sum_{s=1}^t r_s x_s)$ is the regularized least square estimate

with a regularization parameter $\lambda = \frac{\sigma^2}{\kappa^2}$.

Thompson Sampling for Linear Bandits

Recall the Thompson Sampling principle :

“draw a possible model from the posterior distribution and act optimally in this sampled model”

Thompson Sampling in linear bandits

In each round $t + 1$,

$$\begin{aligned}\tilde{\theta}_t &\sim \mathcal{N}(\hat{\theta}_t^\lambda, \sigma^2 (B_t^\lambda)^{-1}) \\ x_{t+1} &= \underset{x \in \mathcal{X}_{t+1}}{\operatorname{argmax}} x^\top \tilde{\theta}_t\end{aligned}$$

Numerical complexity : one need to draw a sample from a multivariate Gaussian distribution, e.g.

$$\tilde{\theta}_t = \hat{\theta}_t^\lambda + \sigma (B_t^\lambda)^{-1/2} X$$

where X is a vector with d independent $\mathcal{N}(0, 1)$ entries.

Theoretical guarantees

[Agrawal and Goyal, 2013b] analyze a *variant* of Thompson Sampling using some “posterior inflation” :

$$\begin{aligned}\tilde{\theta}_t &\sim \mathcal{N}\left(\hat{\theta}_t^1, v^2 (B_t^1)^{-1}\right) \\ x_{t+1} &= \underset{x \in \mathcal{X}_{t+1}}{\operatorname{argmax}} x^\top \tilde{\theta}_t\end{aligned}$$

where $v = \sigma \sqrt{9d \ln(T/\delta)}$.

Theorem

If the noise is σ^2 -sub-Gaussian, the above algorithm satisfies

$$\mathbb{P}\left(R_T(\text{TS}) = \mathcal{O}\left(d^{3/2} \sqrt{T} \left[\ln(T) + \sqrt{\ln(T) \ln(1/\delta)}\right]\right)\right) \geq 1 - \delta.$$

- ▶ slightly worse than Lin-UCB... in theory
- ▶ do we need the posterior inflation ?

Beyond linear bandits

Depending on the application, other parametric models may be better suited than the simple linear model, for example the **logistic model**.

$$\begin{aligned}\mathbb{P}(r_t = 1|x_t) &= \frac{1}{1 + e^{-\theta_*^\top x_t}} \\ \mathbb{P}(r_t = 0|x_t) &= \frac{e^{-\theta_*^\top x_t}}{1 + e^{-\theta_*^\top x_t}}\end{aligned}$$

e.g., clic / no-clic on an add depending on a user/add feature $x_t \in \mathbb{R}^d$

- ▶ [Filippi et al., 2010] : first UCB style algorithm for Generalized Linear Bandit models
- ▶ Thompson Sampling for logistic bandits [Dumitrascu et al., 2018]
- ▶ going further : UCB/TS for neural bandits !

Outline

- 1 Fixing the greedy strategy
- 2 Optimistic Exploration
 - A simple UCB algorithm
 - Towards optimal algorithms
- 3 Randomized Exploration : Thompson Sampling
- 4 Contextual Bandits
 - Lin-UCB
 - Linear Thompson Sampling
- 5 Bandits beyond Regret

Bandits without rewards ?



$$\mathcal{B}(\mu_1)$$

$$\mathcal{B}(\mu_2)$$

$$\mathcal{B}(\mu_3)$$

$$\mathcal{B}(\mu_4)$$

$$\mathcal{B}(\mu_5)$$

For the t -th patient in a clinical study,

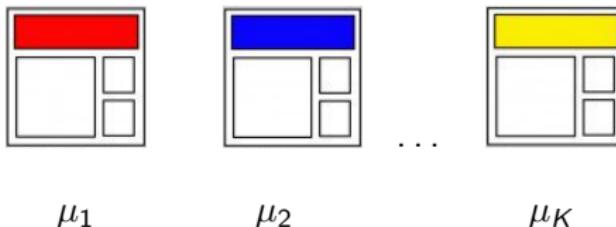
- ▶ chooses a treatment A_t
- ▶ observes a response $X_t \in \{0, 1\} : \mathbb{P}(X_t = 1) = \mu_{A_t}$

Maximize rewards \leftrightarrow cure as many patients as possible

Alternative goal : identify as quickly as possible the best treatment
(without trying to cure patients during the study)

Bandits without rewards ?

Probability that some version of a website generates a conversion :



Best version : $a_* = \underset{a=1, \dots, K}{\operatorname{argmax}} \mu_a$

Sequential protocol : for the t -th visitor :

- ▶ display version A_t
- ▶ observe conversion indicator $X_t \sim \mathcal{B}(\mu_{A_t})$.

Maximize rewards \leftrightarrow maximize the number of conversions

Alternative goal : identify the best version
(without trying to maximize conversions during the test)

A Pure Exploration Problem

Goal : identify an arm with mean close to μ_* as quickly and accurately as possible \simeq identify

$$a_* = \operatorname{argmax}_{a=1,\dots,K} \mu_a.$$

Algorithm : made of three components :

- sampling rule : A_t (arm to explore)
- recommendation rule : B_t (current guess for the best arm)
- stopping rule τ (when do we stop exploring ?)

Probability of error

The probability of error after T rounds is

$$p_\nu(T) = \mathbb{P}_\nu(B_T \neq a_*).$$

A Pure Exploration Problem

Goal : identify an arm with mean close to μ_* as quickly and accurately as possible \simeq identify

$$a_* = \operatorname{argmax}_{a=1,\dots,K} \mu_a.$$

Algorithm : made of three components :

- sampling rule : A_t (arm to explore)
- recommendation rule : B_t (current guess for the best arm)
- stopping rule τ (when do we stop exploring ?)

Simple regret [Bubeck et al., 2011]

The simple regret after n rounds is

$$r_\nu(n) = \mu_* - \mu_{B_n}.$$

A Pure Exploration Problem

Goal : identify an arm with mean close to μ_* as quickly and accurately as possible \simeq identify

$$a_* = \operatorname{argmax}_{a=1,\dots,K} \mu_a.$$

Algorithm : made of three components :

- sampling rule : A_t (arm to explore)
- recommendation rule : B_t (current guess for the best arm)
- stopping rule τ (when do we stop exploring ?)

Simple regret [Bubeck et al., 2011]

The simple regret after n rounds is

$$r_\nu(n) = \mu_* - \mu_{B_n}.$$

$$\Delta_{\min} p_\nu(T) \leq \mathbb{E}_\nu[r_\nu(T)] \leq \Delta_{\max} p_\nu(T)$$

Several objectives

Algorithm : made of three components :

- sampling rule : A_t (arm to explore)
- recommendation rule : B_t (current guess for the best arm)
- stopping rule τ (when do we stop exploring ?)

► Objectives studied in the literature :

Fixed-budget setting	Fixed-confidence setting
<u>input</u> : budget T	<u>input</u> : risk parameter δ (tolerance parameter ϵ)
$\tau = T$ $\text{minimize } \mathbb{P}(B_T \neq a_*)$ or $\mathbb{E}[r_T(\nu)]$	$\text{minimize } \mathbb{E}[\tau]$ $\mathbb{P}(B_\tau \neq a_*) \leq \delta$ or $\mathbb{P}(r_\nu(\tau) > \epsilon) \leq \delta$
[Bubeck et al., 2011] [Audibert et al., 2010]	[Even-Dar et al., 2006]

Can we use UCB ?

Context : bounded rewards (ν_a supported in $[0, 1]$)

We know good algorithms to maximize rewards, for example $UCB(\alpha)$

$$A_{t+1} = \operatorname{argmax}_{a=1, \dots, K} \hat{\mu}_a(t) + \sqrt{\frac{\alpha \ln(t)}{N_a(t)}}$$

- ▶ How good is it for best arm identification ?

Can we use UCB ?

Context : bounded rewards (ν_a supported in $[0, 1]$)

We know good algorithms to maximize rewards, for example $UCB(\alpha)$

$$A_{t+1} = \operatorname{argmax}_{a=1,\dots,K} \hat{\mu}_a(t) + \sqrt{\frac{\alpha \ln(t)}{N_a(t)}}$$

- ▶ How good is it for best arm identification ?

Possible recommendation rules :

Empirical Best Arm (EBA)	$B_t = \operatorname{argmax}_a \hat{\mu}_a(t)$
Most Played Arm (MPA)	$B_t = \operatorname{argmax}_a N_a(t)$
Empirical Distribution of Plays (EDP)	$B_t \sim p_t, \text{ where } p_t = \left(\frac{N_1(t)}{t}, \dots, \frac{N_K(t)}{t} \right)$

[Bubeck et al., 2011]

Can we use UCB ?

Context : bounded rewards (ν_a supported in $[0, 1]$)

We know good algorithms to maximize rewards, for example $UCB(\alpha)$

$$A_{t+1} = \operatorname{argmax}_{a=1,\dots,K} \hat{\mu}_a(t) + \sqrt{\frac{\alpha \ln(t)}{N_a(t)}}$$

- ▶ How good is it for best arm identification ?

Possible recommendation rules :

Empirical Best Arm (EBA)	$B_t = \operatorname{argmax}_a \hat{\mu}_a(t)$
Most Played Arm (MPA)	$B_t = \operatorname{argmax}_a N_a(t)$
Empirical Distribution of Plays (EDP)	$B_t \sim p_t$, where $p_t = \left(\frac{N_1(t)}{t}, \dots, \frac{N_K(t)}{t} \right)$

[Bubeck et al., 2011]

Can we use UCB ?

► UCB + Empirical Distribution of Plays

$$\begin{aligned}\mathbb{E}[r_\nu(T)] &= \mathbb{E}[\mu_* - \mu_{B_T}] = \mathbb{E}\left[\sum_{b=1}^K(\mu_* - \mu_b)\mathbb{1}_{(B_T=b)}\right] \\ &= \mathbb{E}\left[\sum_{b=1}^K(\mu_* - \mu_b)\mathbb{P}(B_T = b|\mathcal{F}_T)\right] \\ &= \mathbb{E}\left[\sum_{b=1}^K(\mu_* - \mu_b)\frac{N_b(T)}{T}\right] \\ &= \frac{1}{T}\sum_{b=1}^K(\mu_* - \mu_b)\mathbb{E}[N_b(T)] \\ &= \frac{\mathcal{R}_\nu(T)}{T}.\end{aligned}$$

→ a conversion from cumulative regret to simple regret !

Can we use UCB ?

► UCB + Empirical Distribution of Plays

$$\mathbb{E}[r_\nu(\text{UCB}(\alpha), T)] \leq \frac{\mathcal{R}_\nu(\text{UCB}(\alpha), T)}{T} \leq \frac{C(\nu) \ln(T)}{T}$$

Can we use UCB ?

► UCB + Empirical Distribution of Plays

$$\mathbb{E} [r_\nu (\text{UCB}(\alpha), T)] \leq \frac{\mathcal{R}_\nu(\text{UCB}(\alpha), T)}{T} \leq C \sqrt{\frac{K \ln(T)}{T}}$$

Can we use UCB ?

- ▶ UCB + Empirical Distribution of Plays

$$\mathbb{E}[r_\nu(\text{UCB}(\alpha), T)] \leq \frac{\mathcal{R}_\nu(\text{UCB}(\alpha), T)}{T} \leq C \sqrt{\frac{K \ln(T)}{T}}$$

- ▶ Almost optimal in the **worse case**

Lower bound [Bubeck et al., 2011]

For every algorithm \mathcal{A} , there exists a bandit instance ν in which

$$\mathbb{E}[r_\nu(\mathcal{A}, T)] \geq \frac{1}{20} \sqrt{\frac{K}{T}}$$

Can we use UCB ?

- ▶ UCB + Empirical Distribution of Plays

$$\mathbb{E} [r_\nu (\text{UCB}(\alpha), T)] \leq \frac{\mathcal{R}_\nu(\text{UCB}(\alpha), T)}{T} \leq C \sqrt{\frac{K \ln(T)}{T}}$$

- ▶ ... but potentially bad in the **problem-dependent** regime

The simple regret or the **uniform sampling** strategy decays exponentially :

$$\mathbb{E}_\nu [r_\nu (\text{Unif}, T)] \leq (K - 1) \Delta_{\max} \exp \left(-\frac{1}{2} \frac{T}{K} \Delta_{\min}^2 \right)$$

- UCB does not always provably outperform **uniform sampling**...

(Problem-dependent) sample complexity

With Uniform Sampling, the number of sample needed to get an error probability smaller than δ is of order

$$T \simeq \frac{K}{\Delta_{\min}^2} \log \left(\frac{1}{\delta} \right)$$

(assuming, e.g. rewards in $[0,1]$)

- ▶ Can be improved for smarter algorithms to

$$T \simeq \mathcal{O} \left(H(\nu) \log \left(\frac{1}{\delta} \right) \right)$$

where

$$H(\nu) = \sum_{a=1}^K \frac{1}{\Delta_a^2} \quad \text{with} \quad \Delta_{a_*} = \min_{a \neq a_*} \Delta_a .$$

(and more precise complexity measures for parametric distributions
[Garivier and Kaufmann, 2016])

Fixed Budget : Sequential Halving

Input : total number of plays T

Idea : split the budget in $\log_2(K)$ phases of equal length, eliminate the worst half of the remaining arms after each phase.

Initialisation : $S_0 = \{1, \dots, K\}$;

For $r = 0$ **to** $\lceil \ln_2(K) \rceil - 1$, **do**

sample each arm $a \in S_r$ $t_r = \left\lfloor \frac{T}{|S_r| \lceil \log_2(K) \rceil} \right\rfloor$ times ;

let $\hat{\mu}_a^r$ be the empirical mean of arm a ;

let S_{r+1} be the set of $\lceil |S_r|/2 \rceil$ arms with largest $\hat{\mu}_a^r$

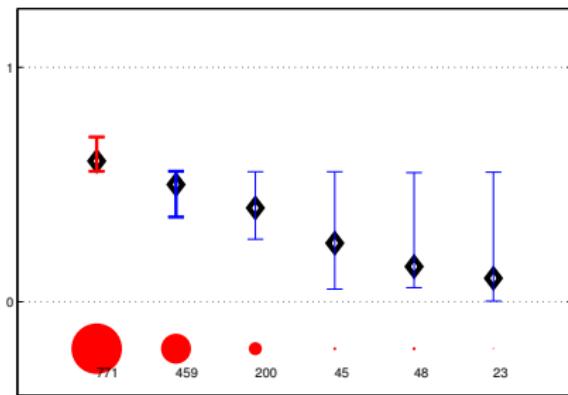
Output : B_T the unique arm in $S_{\lceil \log_2(K) \rceil}$

Theorem [Karnin et al., 2013]

$$\mathbb{P}_\nu(B_T \neq a_*) \leq 3 \log_2(K) \exp\left(-\frac{T}{8 \log_2(K) H(\nu)}\right).$$

Fixed Confidence : LUCB

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



► At round t , draw

$$B_t = \underset{b}{\operatorname{argmax}} \hat{\mu}_b(t)$$

$$C_t = \underset{c \neq B_t}{\operatorname{argmax}} \text{UCB}_c(t)$$

► Stop at round t if

$$\text{LCB}_{B_t}(t) > \text{UCB}_{C_t}(t) - \epsilon$$

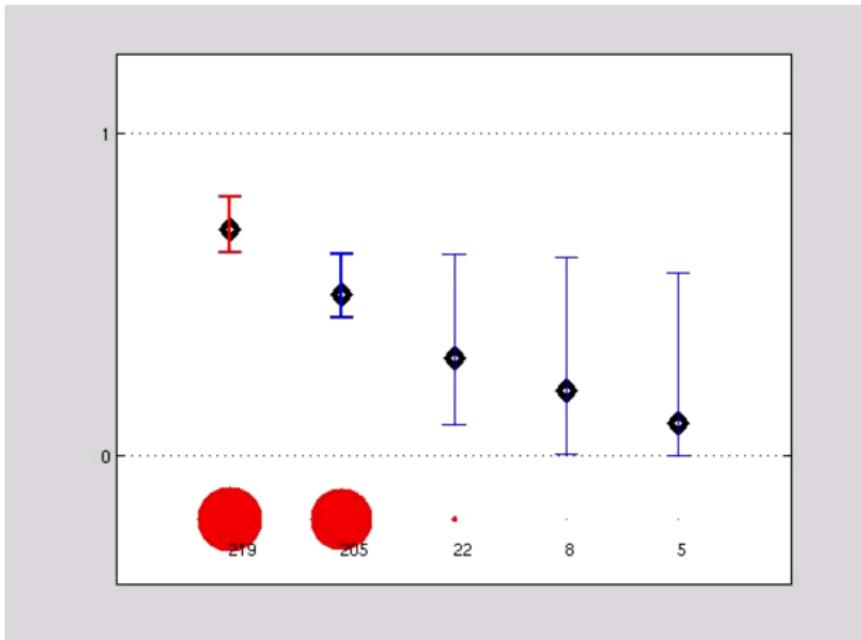
Theorem [Kalyanakrishnan et al., 2012]

For well-chosen confidence intervals, $\mathbb{P}_\nu(\mu_{B_\tau} > \mu_\star - \epsilon) \geq 1 - \delta$ and

$$\mathbb{E}[\tau_\delta] = \mathcal{O}\left(\left[\sum_{a=1}^K \frac{1}{\Delta_a^2 \vee \epsilon^2}\right] \ln\left(\frac{1}{\delta}\right)\right)$$

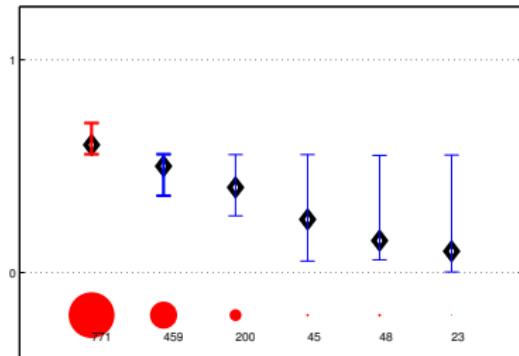
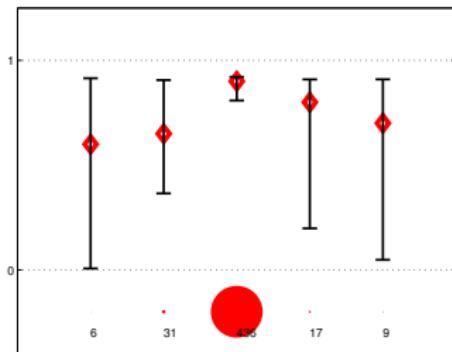
(kl)-LUCB in action

$$\begin{aligned} \text{UCB}_a(t) &= \max \left\{ q \in [0, 1] : N_a(t) \text{kl}(\hat{\mu}_a(t), q) \leq \log(Ct^2/\delta) \right\} \\ \text{LCB}_a(t) &= \min \left\{ q \in [0, 1] : N_a(t) \text{kl}(\hat{\mu}_a(t), q) \leq \log(Ct^2/\delta) \right\} \end{aligned}$$



A comparison with UCB

Regret minimizing algorithms and Best Arm Identification algorithms behave quite differently



Number of selections and confidence intervals for KL-UCB (left) and KL-LUCB (right)

Conclusion

In bandits, ε -greedy can be replaced by smarter algorithms

- ▶ both for learning while maximizing rewards *(regret)*
- ▶ and for fast identification of the best action *(sample complexity)*

Two important tools :

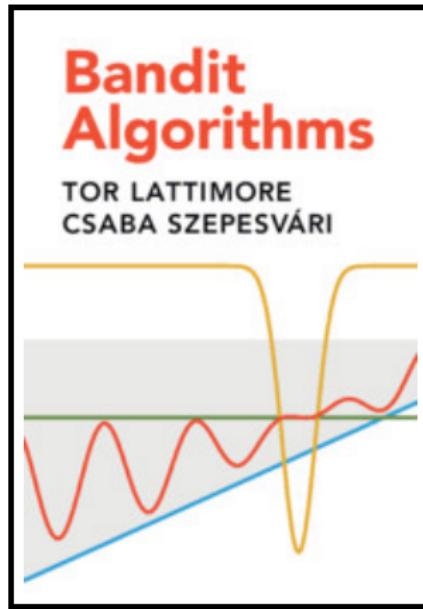
- ▶ confidence intervals
- ▶ posterior distributions

to better take into account the uncertainty and perform more efficient (“directed”) exploration.

Those tools can also be used in contextual bandit models.

How about general [Markov Decision Processes](#) ?

References



The Bandit Book

by [Lattimore and Szepesvari, 2019]

-  Abbasi-Yadkori, Y., D.Pál, and C.Szepesvári (2011).
Improved Algorithms for Linear Stochastic Bandits.
In *Advances in Neural Information Processing Systems*.
-  Agrawal, S. and Goyal, N. (2013a).
Further Optimal Regret Bounds for Thompson Sampling.
In *Proceedings of the 16th Conference on Artificial Intelligence and Statistics*.
-  Agrawal, S. and Goyal, N. (2013b).
Thompson Sampling for Contextual Bandits with Linear Payoffs.
In *International Conference on Machine Learning (ICML)*.
-  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.
-  Audibert, J.-Y., Bubeck, S., and Munos, R. (2010).
Best Arm Identification in Multi-armed Bandits.
In *Proceedings of the 23rd Conference on Learning Theory*.
-  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.
-  Bubeck, S., Munos, R., and Stoltz, G. (2011).
Pure Exploration in Finitely Armed and Continuous Armed Bandits.
Theoretical Computer Science 412, 1832-1852, 412 :1832–1852.
-  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.
-  Dumitrascu, B., Feng, K., and Engelhardt, B. E. (2018).
PG-TS : improved thompson sampling for logistic contextual bandits.
In *Advances in Neural Information Processing Systems (NeurIPS)*.
-  Even-Dar, E., Mannor, S., and Mansour, Y. (2006).
Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems.
Journal of Machine Learning Research, 7 :1079–1105.
-  Filippi, S., Cappé, O., Garivier, A., and Szepesvári, C. (2010).
Parametric Bandits : The Generalized Linear case.

In *Advances in Neural Information Processing Systems*.

-  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. and Kaufmann, E. (2016).
Optimal best arm identification with fixed confidence.
In *Proceedings of the 29th 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*.
-  Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. (2012).
PAC subset selection in stochastic multi-armed bandits.
In *International Conference on Machine Learning (ICML)*.
-  Karnin, Z., Koren, T., and Somekh, O. (2013).
Almost optimal Exploration in multi-armed bandits.
In *International Conference on Machine Learning (ICML)*.
-  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*.
-  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.
-  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.