Multi-Armed Bandits: an introduction

Emilie Kaufmann

EURO PhD school, July 2022
Why bandits?

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

an agent facing arms in a Multi-Armed Bandit

How to sequentially choose which arm to pull in order to maximize our profit?
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?
Dynamic allocation of computational resource

Numerical experiments:

- where to evaluate a costly function in order to find its maximum?

Artificial intelligence for games:

- how to choose the next game to simulate in order to find the best move to play next?
1 The multi-armed bandit problem

2 Fixing the greedy strategy

3 Upper Confidence Bound (UCB) algorithms

4 Towards optimal algorithms
The Multi-Armed Bandit Setting

\[ K \text{ arms} \leftrightarrow K \text{ rewards streams } (X_{a,t})_{t \in \mathbb{N}} \]

At round \( t \), an agent:

- chooses an arm \( A_t \)
- receives a reward \( R_t = X_{A_t,t} \)

Sequential sampling strategy (bandit algorithm):

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

Goal: Maximize \( \sum_{t=1}^{T} R_t \).
The **Stochastic** Multi-Armed Bandit Setting

\( K \) arms \( \leftrightarrow \) \( K \) probability distributions : \( \nu_a \) has mean \( \mu_a \)

At round \( t \), an agent:

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

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

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

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

\( \Rightarrow \) a particular reinforcement learning problem
Clinical trials

**Historical motivation** [Thompson, 1933]

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)

For the $t$-th visitor of a website,

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

Goal: maximize the sum of ratings
Regret of a bandit algorithm

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

\[
\mu_* = \max_{a \in \{1, \ldots, K\}} \mu_a \quad \text{and} \quad a_* = \arg\max_{a \in \{1, \ldots, K\}} \mu_a.
\]

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

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

What regret rate can we achieve?

- consistency: \( \frac{R_\nu(A, T)}{T} \rightarrow 0 \)
- can we be more precise?
Regret decomposition

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

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

Proof.

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

\[ = \mathbb{E} \left[ \sum_{t=1}^{T} (\mu_\star - \mu_{A_{t}}) \right] \]

\[ = \sum_{a=1}^{K} \frac{\mu_\star - \mu_a}{\Delta_a} \mathbb{E} \left[ \sum_{t=1}^{T} 1(A_{t} = a) \right]. \]
Regret decomposition

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

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

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 \( \Delta_a \)'s

\( \Rightarrow \) Exploration / Exploitation trade-off
The greedy strategy

Select each arm once and, for \( t \geq K \), exploit the current knowledge:

\[
A_{t+1} = \arg\max_{a \in [K]} \hat{\mu}_a(t)
\]

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 and, for \( t \geq K \), exploit the current knowledge:

\[
A_{t+1} = \arg\max_{a \in [K]} \hat{\mu}_a(t)
\]

where

\[
N_a(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = a) \text{ is the number of selections of arm } a
\]

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

Properties:

👍 a simple (non-parametric) algorithm

👎 suffers linear regret

e.g. in a two armed Bernoulli bandit with means \( \mu_1 > \mu_2 \)

\[
R_\nu(T) \geq (1 - \mu_1)\mu_2(\mu_1 - \mu_2) \times (T - 1)
\]
Outline

1. The multi-armed bandit problem
2. Fixing the greedy strategy
3. Upper Confidence Bound (UCB) algorithms
4. Towards optimal algorithms
Explore-Then-Commit

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

- draw each arm $m$ times
- compute the empirical best arm $\hat{a} = \arg\max_a \hat{\mu}_a(Km)$
- 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 \( m \in \{1, \ldots, T/K\} \),

- draw each arm \( m \) times
- compute the empirical best arm \( \hat{a} = \arg\max_a \hat{\mu}_a(Km) \)
- 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. \( \mu_1 > \mu_2, \Delta := \mu_1 - \mu_2 \).

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

\( \hat{\mu}_{a,m} \): empirical mean of the first \( m \) observations from arm \( a \)
Explore-Then-Commit

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

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

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

$\Rightarrow$ EXPLORATION followed by EXPLOITATION

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

$$R_\nu(\text{ETC}, T) = \Delta \mathbb{E}[N_2(T)]$$

$$= \Delta \mathbb{E}[m + (T - 2m)1(\hat{a} = 2)]$$

$$\leq \Delta m + (\Delta T) \times \mathbb{P}(\hat{\mu}_{2,m} \geq \hat{\mu}_{1,m})$$

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

$\rightarrow$ requires a concentration inequality
Technical tool: Concentration Inequalities

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

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

- $\nu_a$ bounded in $[0, 1]$: $1/4$ sub-Gaussian
- $\nu_a = \mathcal{N}(\mu_a, \sigma^2)$: $\sigma^2$ sub-Gaussian

Hoeffding inequality

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

$$\mathbb{P}\left(\frac{Z_1 + \cdots + Z_s}{s} \geq \mu + \chi\right) \leq e^{-\frac{s\chi^2}{2\sigma^2}}$$
Technical tool : Concentration Inequalities

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

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

- $\nu_a$ bounded in $[0, 1]$ : $1/4$ sub-Gaussian
- $\nu_a = \mathcal{N}(\mu_a, \sigma^2)$ : $\sigma^2$ sub-Gaussian

Hoeffding inequality

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

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

Emilie Kaufmann | CRISTAL
Explore-Then-Commit

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

- draw each arm \( m \) times
- compute the empirical best arm \( \hat{a} = \arg\max_a \mu_a(Km) \)
- keep playing this arm until round \( T \)
\[ A_{t+1} = \hat{a} \text{ for } t \geq Km \]

\[ \Rightarrow \text{EXPLORATION followed by EXPLOITATION} \]

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

Assumption: \( \nu_1, \nu_2 \) are bounded in \([0, 1]\).

\[
R_\nu(T) = \Delta \mathbb{E}[N_2(T)] \\
= \Delta \mathbb{E}[m + (T - 2m)1(\hat{a} = 2)] \\
\leq \Delta m + (\Delta T) \times P(\hat{\mu}_{2,m} \geq \hat{\mu}_{1,m})
\]

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

\( \rightarrow \) Hoeffding’s inequality
Explore-Then-Commit

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

- draw each arm $m$ times
- compute the empirical best arm $\hat{a} = \arg\max_a \hat{\mu}_a(Km)$
- 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: $\nu_1, \nu_2$ are bounded in $[0, 1]$.

$$R_\nu(T) = \Delta \mathbb{E}[N_2(T)] = \Delta \mathbb{E}[m + (T - 2m)1(\hat{a} = 2)] \leq \Delta m + (\Delta T) \times \exp(-m\Delta^2/2)$$

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

$\rightarrow$ Hoeffding’s inequality
Explore-Then-Commit

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

- draw each arm \( m \) times
- compute the empirical best arm \( \hat{a} = \arg\max_a \hat{\mu}_a(Km) \)
- 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:** \( \nu_1, \nu_2 \) are bounded in \([0, 1]\).

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

\[ R_\nu(ETC, T) \leq \frac{2}{\Delta} \left[ \log \left( \frac{T\Delta^2}{2} \right) + 1 \right]. \]
Explore-Then-Commit

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

- draw each arm \( m \) times
- compute the empirical best arm \( \hat{a} = \arg\max_a \hat{\mu}_a(Km) \)
- 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: \( \nu_1, \nu_2 \) are bounded in \([0, 1]\).

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

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

+ logarithmic regret!
- requires the knowledge of \( T \) and \( \Delta \)
Sequential Explore-Then-Commit

★ explore uniformly until a random time of the form

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

★ \( \hat{a}_\tau = \arg\max_a \hat{\mu}_a(\tau) \) and \( A_{t+1} = \hat{a}_\tau \) for \( t \in \{\tau + 1, \ldots, T\} \)

⇒ [Garivier et al., 2016] for two Gaussian arms, for \( c = 8 \), same regret as ETC, without the knowledge of \( \Delta \)

⇒ ... but larger regret as that of the best fully sequential strategy
Another possible fix: $\epsilon$-greedy

The $\epsilon$-greedy rule [Sutton and Barto, 1998] is a simple randomized way to alternate exploration and exploitation.

**$\epsilon$-greedy strategy**

At round $t$,

- with probability $\epsilon$
  
  \[ A_t \sim \mathcal{U}\left(\{1, \ldots, K\}\right) \]

- with probability $1 - \epsilon$
  
  \[ A_t = \arg\max_{a=1,\ldots,K} \hat{\mu}_a(t). \]

$\rightarrow$ Linear regret: $R_{\nu}$ ($\epsilon$-greedy, $T$) $\geq \epsilon \frac{K-1}{K} \Delta_{\min} T$

\[ \Delta_{\min} = \min_{a: \mu_a < \mu_*} \Delta_a \]
Another possible fix: $\epsilon$-greedy

$\epsilon_t$-greedy strategy

At round $t$,
- with probability $\epsilon_t := \min\left(1, \frac{K}{d^2t}\right)$
  \[ A_t \sim \mathcal{U}\left(\{1, \ldots, K\}\right) \]
- with probability $1 - \epsilon_t$
  \[ A_t = \arg\max_{a=1,\ldots,K} \hat{\mu}_a(t-1). \]

Theorem [Auer et al., 2002]

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

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

1. The multi-armed bandit problem
2. Fixing the greedy strategy
3. Upper Confidence Bound (UCB) algorithms
4. Towards optimal algorithms
The optimism principle

**Step 1**: construct a set of statistically plausible models

- For each arm $a$, build a confidence interval on the mean $\mu_a$:

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

- \text{LCB} = \text{Lower Confidence Bound}
- \text{UCB} = \text{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

$$A_{t+1} = \arg\max_{a=1,...,K} \text{UCB}_a(t).$$
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 $\sigma^2$ sub-Gaussian

Hoeffding inequality, reloaded

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

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

We need $UCB_a(t)$ such that

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

→ tool: concentration inequalities

Example: rewards are $\sigma^2$ sub-Gaussian

Hoeffding inequality, reloaded

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

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

⚠️ Cannot be used directly in a bandit model as the number of observations from each arm is random!
How to build confidence intervals?

- \( N_a(t) = \sum_{s=1}^{t} 1(A_s = a) \) 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 \( \mu_a \) 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?

$N_a(t) = \sum_{s=1}^{t} 1(A_s=a)$ 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 $\mu_a$ 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

$UCB(\alpha)$ selects $A_{t+1} = \arg\max_a \ UCB_a(t)$ where

$$UCB_a(t) = \hat{\mu}_a(t) + \sqrt{\frac{\alpha \log(t)}{N_a(t)}}.$$ 

- exploitation term
- exploration bonus

- popularized by [Auer et al., 2002] for bounded rewards: UCB1, for $\alpha = 2$

- the analysis of $UCB(\alpha)$ was further refined to hold for $\alpha > 1/2$ in that case [Bubeck, 2010, Cappé et al., 2013]
A UCB algorithm in action
Regret of UCB($\alpha$)

**Theorem**

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

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

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

**Proof:**
A worse-case regret bound

Corollary

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

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 \( \Delta > 0 \),

\[ \mathcal{R}_v(T) = \sum_{a: \Delta_a \leq \Delta} \Delta_a \mathbb{E}[N_a(T)] + \sum_{a: \Delta_a \geq \Delta} \Delta_a \mathbb{E}[N_a(T)] \]

\[ \leq \Delta T + \sum_{a: \Delta_a \geq \Delta} \left(C \frac{\log(T)}{\Delta_a} + D \Delta_a\right) \]

\[ \leq \Delta T + \frac{CK \log(T)}{\Delta} + D \left(\sum_{a=1}^{K} \Delta_a\right) \]

\[ = 2\sqrt{CKT \log(T)} + D \left(\sum_{a=1}^{K} \Delta_a\right) \quad \text{for} \quad \Delta = \sqrt{\frac{CK \log(T)}{T}} \]
An improved problem-dependent result

Context: $\sigma^2$ sub-Gaussian rewards

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

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

Theorem [Cappé et al.’13]

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

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

For UCB(\(\alpha\)) applied to \(\sigma^2\)-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. Fixing the greedy strategy
3. Upper Confidence Bound (UCB) algorithms
4. Towards optimal algorithms
The Lai and Robbins lower bound

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

\[ \nu \leftrightarrow \mu = (\mu_1, \ldots, \mu_K) \]

**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,

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

[Lai and Robbins, 1985]
The Lai and Robbins lower bound

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

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

Key tool: Kullback-Leibler divergence.

Kullback-Leibler divergence

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

Theorem

For uniformly good algorithm,

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

[Lai and Robbins, 1985]
The Lai and Robbins lower bound

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

\[
\nu \leftrightarrow \mu = (\mu_1, \ldots, \mu_K)
\]

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

[Lai and Robbins, 1985]
UCB compared to the lower bound

Gaussian distributions with variance $\sigma^2$

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

- **Upper bound**: for UCB($\alpha$) 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 $\sigma^2$

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

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

$\rightarrow$ 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($\alpha$) with $\alpha = 1/2$

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

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

$\rightarrow$ UCB is *not* asymptotically optimal for Bernoulli rewards...
The \textit{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 Bernoulli rewards,

\[ \mathbb{P}(\text{UCB}_a(t) > \mu_a) \gtrsim 1 - \frac{1}{t \log(t)}. \]
An asymptotically optimal algorithm

kl-UCB selects \( A_{t+1} = \arg\max_a UCB_a(t) \) with

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

**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

\[
\mathcal{R}_\mu(\text{kl-UCB}, T) \simeq \left( \sum_{a: \mu_a < \mu_*} \frac{\Delta_a}{\text{kl}(\mu_a, \mu_*)} \right) \log(T).
\]
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 $\nu$ with rewards supported in $[0, 1]$ such that

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

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. (kl)-UCB only achieves $O(\sqrt{KT \log(T)})$
Going further

We saw different type of frequentist algorithms:

- either based on comparing (MLE) estimates of the mean rewards (ETC, $\varepsilon$-greedy)
- or using confidence intervals (UCB, $kl$-UCB)

Next lecture: Bayesian bandits
Going further

Perspectives:

- algorithms which are asymptotically optimal and minimax optimal
  [Garivier et al., 2018]

- algorithms which are asymptotically optimal for different families of distributions (e.g. one algorithm for Gaussian and Bernoulli bandits)
  [Baudry et al., 2020]

- algorithms which are robust to adversarial rewards (Best Of Both worlds)
  [Zimmert and Seldin, 2021]

- algorithms which are robust to non-stationary rewards
  [Garivier and Moulines, 2011, Suk and Kpotufe, 2022]
References

The Bandit Book

by [Lattimore and Szepesvari, 2019]


On explore-then-commit strategies. 

On Upper-Confidence Bound Policies For Switching Bandit Problems. 
In Algorithmic Learning Theory (ALT), pages 174–188. PMLR.

Asymptotically efficient adaptive allocation rules. 
Advances in Applied Mathematics, 6(1) :4–22.

Bandit Algorithms. 
Cambridge University Press.

A contextual-bandit approach to personalized news article recommendation. 
In WWW.

Some aspects of the sequential design of experiments. 

Tracking most significant arm switches in bandits.
In Conference On Learning Theory COLT.

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.

Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits.