Introduction to bandit theory

Victor Thuot[1]

[1]UMR MISTEA, INRAE Montpellier

semdoc IMAG – 4th of December 2024 – Montpellier

1. The multi-armed bandit model
   - Sequential and adaptive sampling
   - Regret minimization vs pure exploration
2. Algorithms for regret minimization
   - ETC
   - UCB
3. Conclusion

What is a multi-armed bandit ?

Main ideas

- **The bandit model** is a sequential game, where at each round, a learner chooses an action to make, and obtains a random reward depending on this action.

Main ideas

- **The bandit model** is a sequential game, where at each round, a learner chooses an action to make, and obtains a random reward depending on this action.

$\rightarrow$ Trade-off between exploitation and exploration

- exploit their current knowledge;
- explore unknown actions to gain knowledge for the future.

Exploration VS exploitation

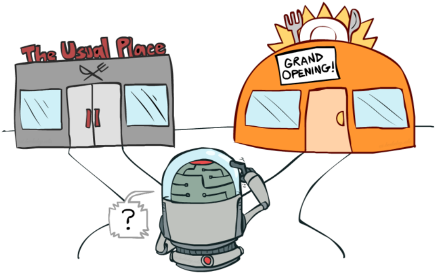
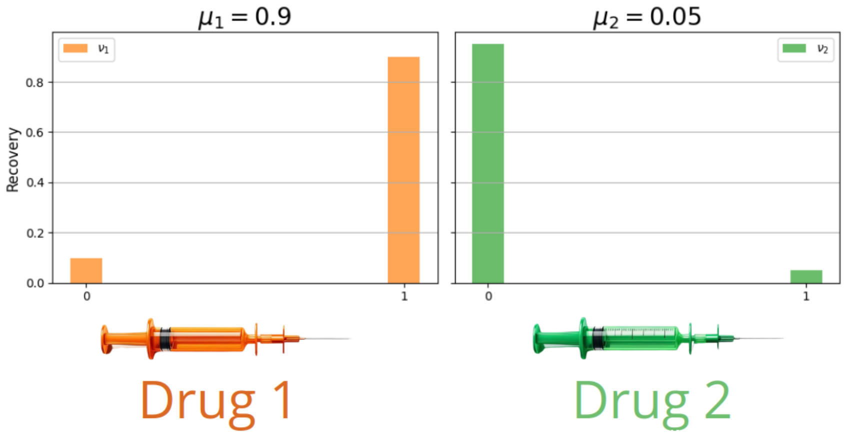


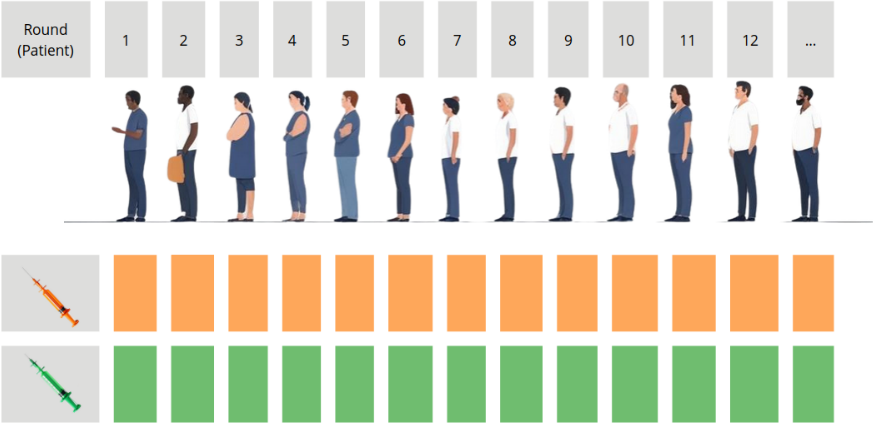
Figure: source: UC Berkeley AI course

- Two possible drugs 1 and 2
- Unknown probability of being cured μ_1 and μ_2
- At each round, choose drug 1 or 2, observe the response to the drug (binary)



Clinical-trial

- At each round, choose drug 1 or 2, observe the response to the chosen drug



Clinical-trial: randomized trial

- randomized trial: test half patients with 1 and half with 2

Round (Patient)	1	2	3	4	5	6	7	8	9	10	11	12	...
	1		1		1		0		1		1		...
		0		0		0		1		0		0	...

Clinical-trial: randomized trial

- randomized trial: test half patients with 1 and half with 2

Round (Patient)	1	2	3	4	5	6	7	8	9	10	11	12	...

	1		1		1		0		1		1		...
		0		0		0		1		0		0	...

- What is the problem ?

Clinical-trial: randomized trial

- randomized trial: test half patients with 1 and half with 2

Round (Patient)	1	2	3	4	5	6	7	8	9	10	11	12	...
	1		1		1		0		1		1		...
		0		0		0		1		0		0	...

- What is the problem ?
- Solution: adapt the treatment on the fly

Some leading examples

- **Clinical trial** [Chow and Chang, 2008, Thompson, 1933]

  When a patient arrives, the doctor chooses a treatment, and observes how the patient reacts to the treatment.

- **Ad placement** [Langford and Zhang, 2007]

  When a new user arrives, the website chooses one add to show, and observes if the user clicks on the add or not.

- **Dynamic pricing** [Den Boer, 2015]

  When a customer arrives, the store chooses a price offered to the customer, and observes if the customer buys or not the product.

Multi-armed-bandit model [Robbins, 1952]

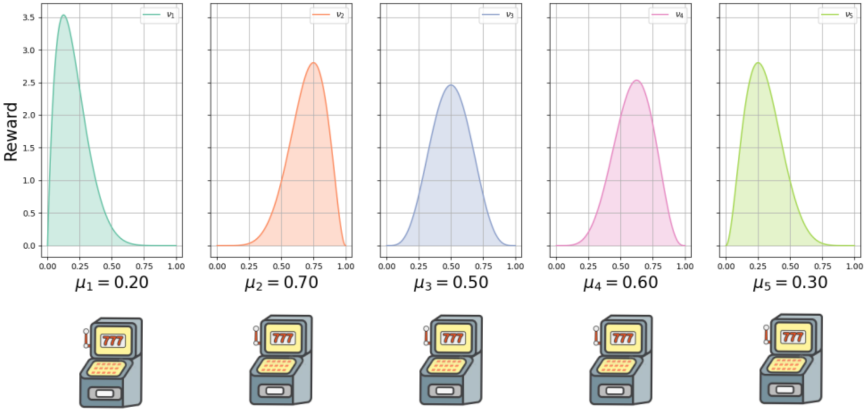


Figure: 5-armed bandit

Algorithm 1 Learning protocol

Input: K number of arms, T budget
for $t = 1, \dots, T$ **do**
  Choose one arm $A_t \in \{1, \dots, K\}$ based on the passed.
  Obtain a reward from the environment X_t
end for

Multi-armed-bandit model

Algorithm 2 Learning protocol

Input: K number of arms, T budget
for $t = 1, \dots, T$ **do**
  Choose one arm $A_t \in \{1, \dots, K\}$ based on the passed.
  Obtain a reward from the environment X_t
end for

- i.i.d reward: conditionally on $A_t = a$, $X_t \sim \nu_a$, where ν_a is a distribution which depends only on a

Multi-armed-bandit model

Algorithm 3 Learning protocol

Input: K number of arms, T budget
for $t = 1, \dots, T$ **do**
  Choose one arm $A_t \in \{1, \dots, K\}$ based on the passed.
  Obtain a reward from the environment X_t
end for

- i.i.d reward: conditionally on $A_t = a$, $X_t \sim \nu_a$, where ν_a is a distribution which depends only on a
- $(\nu_1, \dots, \nu_K)$ is called the environment
- $(\mu_1, \dots, \mu_K)$ denotes the associated means

Multi-armed-bandit model

Algorithm 4 Learning protocol

Input: K number of arms
for $t = 1, \dots, T$ **do**
  Choose one arm $A_t \in \{1, \dots, K\}$ based on the passed.
  Obtain a reward from the environment X_t
end for

- $N_a(t) := \sum_{s=1}^{t} \mathbb{1}_{A_s = a}$
- $\hat{\mu}_a(t) := \frac{1}{N_a(t)} \sum_{s=1}^{t} \mathbb{1}_{A_s = a} X_t$
- Denote as a^* the best choice such that $\mu_* = \max_a \mu_a$

Regret minimization vs pure exploration

Regret minimization :

- The reward X_t is in $\mathbb{R}$, it is seen as a reward.
- Cumulative Regret: $R_T = \sum_{t=1}^{T} \mathbb{E}[\mu_* - X_t]$
- Objective: minimize the cumulative regret

Pure exploration :

- The budget T is seen as a cost
- Simple Regret: $r_T = \mathbb{E}[\mu_* - X_T]$
- Objective:

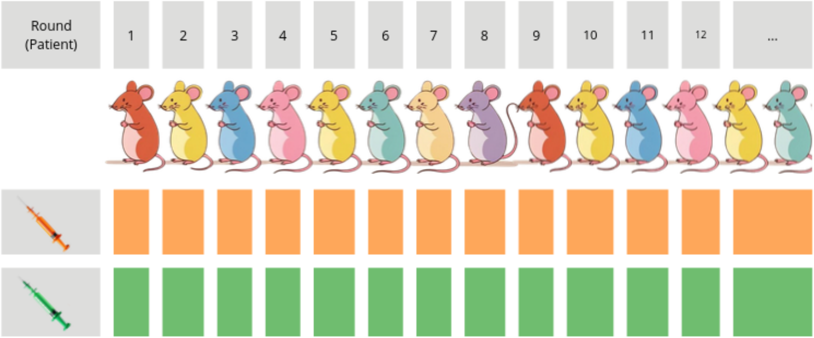
  minimize the simple regret
  minimize $\mathbb{P}(A_T \neq a_*)$

Regret minimization

- Objective: maximize the number of patient cured

Pure exploration

- Objective: identify the best treatment with the least probabiity of error



ETC: Explore ...

- exploration phase: choose each drug $m = 2$ and identify the best drug

Round (Patient)	1	2	3	4	5	6	7	8	9	10	11	12	...
ETC	exploration				exploitation								

	1		1										...
		0		1									...

ETC: ... Then Commit

- exploitation phase: commit to the best drug

Round (Patient)	1	2	3	4	5	6	7	8	9	10	11	12	...
ETC	exploration				exploitation								
	1		1		1	1	0	1	1	1	0	0	...
		0		1									...

ETC: Explore Then Commit

Algorithm 5 Explore-Then-Commit

Input: K number of arms, T budget, parameter $m \leqslant T/K$

for $t = 1, \ldots, mK$ **do**

  Choose $A_t = t \mod K$

end for

for $t = mK + 1, \ldots, T$ **do**

  Choose $A_t = \text{argmax}_a \, \hat{\mu}_a(Km)$

end for

Regret

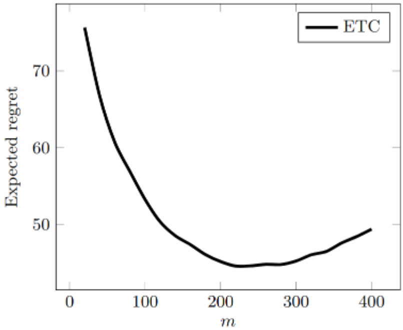


Figure: Expected regret for ETC over 10^5 trials on a Gaussian bandit with means $\mu_1 = 0, \mu_2 = 1/10$ [Lattimore and Szepesvári, 2020]

Exploration vs exploitation

Theorem

If $\nu_1, \ldots, \nu_K$ are 1-subGaussian,

$$R_T \leqslant m \sum_{i=1}^{K} \Delta_i + (T - Km) \sum_{i=1}^{K} \Delta_i \exp\left(-\frac{m\Delta_i^2}{4}\right)$$

- tuning m, exploration vs exploitation

Upper Confidence Bound Algorithm (UCB)

- Optimism in the face of uncertainty
- Confidence bound $UCB_a(t,\delta) = \begin{cases} +\infty & \text{if } T_a(t) = 0 \\ \hat{\mu}_a(t) + \sqrt{\frac{2\log(1/\delta)}{T_a(t)}} & \text{sinon.} \end{cases}$

Algorithm 6 Upper Confidence Bound

Input: K number of arms, tuning parameter δ
for $t = 1, \ldots, T$ **do**
  Choose $A_t = \text{argmax}_a\, UCB_a(t-1,\delta)$
  Update
end for

UCB

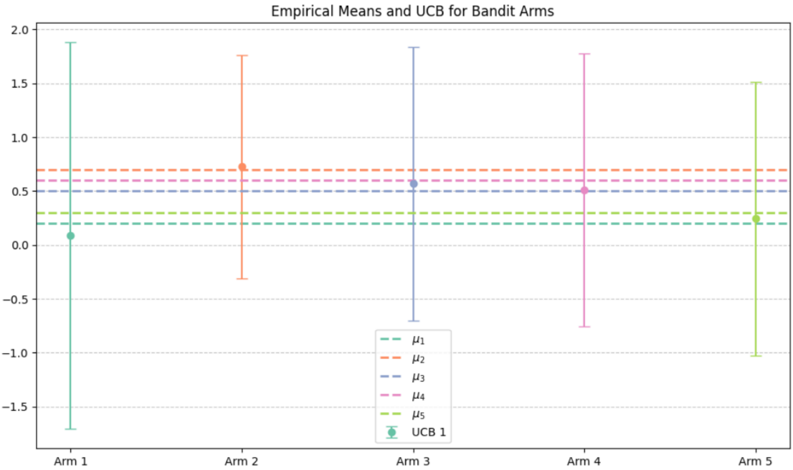


Figure: Upper confidence bounds after 10 rounds

UCB

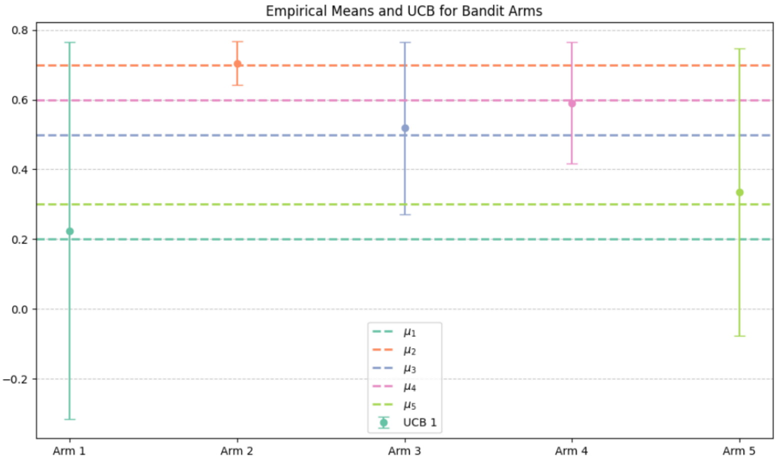


Figure: Upper confidence bounds after 1000 rounds

Many variations

- Non-stationary (automatic trading)
- Structured set of arms (dynamic pricing)
- Infinite or large set of arms
- Contextual : add a context C_t (dynamic pricing, recommendation system)
- Adversarial setting

Chow, S.-C. and Chang, M. (2008).
Adaptive design methods in clinical trials–a review.
Orphanet journal of rare diseases, 3:1–13.

Den Boer, A. V. (2015).
Dynamic pricing and learning: historical origins, current research, and new directions.
Surveys in operations research and management science, 20(1):1–18.

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

Langford, J. and Zhang, T. (2007).
The epoch-greedy algorithm for multi-armed bandits with side information.
Advances in neural information processing systems, 20.

Lattimore, T. and Szepesvári, C. (2020).
Bandit algorithms.
Cambridge University Press.

Robbins, H. (1952).
Some aspects of the sequential design of experiments.

Slivkins, A. et al. (2019).
Introduction to multi-armed bandits.
Foundations and Trends® in Machine Learning, 12(1-2):1–286.

Thompson, W. R. (1933).
On the likelihood that one unknown probability exceeds another in view of the evidence of two samples.
Biometrika, 25(3-4):285–294.

Take-Home Message

- The multi-armed bandit problem captures the fundamental trade-off between exploration and exploitation in sequential decision-making.
- Bandit methods are widely applicable, from optimizing treatments in clinical trials to dynamic pricing and recommendation systems.
- Many variations for each application.
- Bandit theory provides a rigorous and practical foundation for learning and decision-making under uncertainty.

4 Pure exploration

Best arm identification

- $\nu_1, \ldots, \nu_K$ environment of a K-armed bandit
- objective: identify the arm a_* with the best expected reward
- Fixed budget: budget T fixed, minimize $\mathbb{P}(A_T \neq a_*)$
- Fixed confidence: T is a stopping time chosen by the learner, objective: output A_T such that $\mathbb{P}(A_T \neq a_*) \leqslant \delta$

Sequential Halving Algorithm: Overview

Key Idea:

- Allocate budget iteratively across remaining arms.
- Eliminate the less promising arms in each round based on their empirical means.

Algorithm Steps:

1. Start with all arms $\{1, \ldots, K\}$ and divide the budget equally among them.
2. Compute the empirical mean reward for each arm.
3. Discard approximately half of the arms with the lowest means.
4. Repeat until only one arm remains.

Sequential Halving Algorithm

Algorithm 7 Upper Confidence Bound

Input: $S = \{1, \dots, K\}$ set of arms, budget T
$n = T / \lceil \log_2(K) \rceil$
for $s = 1, \dots, \lceil \log_2(K) \rceil$ **do**
  sample $n/|S|$ times each arm in S
  eliminate from S the half arms with the lowest expected mean
end for
return Remaining arm $\hat{a} \in S$

Joint framework

Let M be a $N \times d$ matrix.

- **learning protocol** – a learner observes sequentially and actively entries of the matrix with some sub-Gaussian noise
- **unknown structure** – there exists an unknown structure over the matrix that has to be recovered
- **objective** – the learner has to recover the unknown structure with a prescribed probability of error, while minimizing the budget spent

Active clustering problem

Problem

- **Observations** – one entire row (dimension d) at a time
- **Unknown structure** – there exists a partition of the rows G^*, so that, two rows μ_i and μ_j are in the same group, iff $\mu_i = \mu_j$.
- **Objective** – recover G^* with probability larger than $1 - \delta$

Active clustering problem through entries*

*with Maximilian Graf– PhD student in Potsdam

Problem

- **Observations** – one entry $I_j, J_t \in [N] \times [d]$ at a time
- **Unknown structure** – there exists a partition of the rows G^*, so that, two rows μ_i and μ_j are in the same group, iff $\mu_i = \mu_j$.
- **Objective** – recover G^* with probability larger than $1 - \delta$

Condorcet Winner Identification*

*work with El Mehdi Saad – Centrale Paris

- **Observations** – (I_t, J_t) a comparison between two experts
- **Unknown structure** – $N = d$, $M - \frac{1}{2}I$ antisymmetric, there exists a Condorcet Winner
- **Objective** – identify the **CW**