# A quick overview of Bandit, old and new

*Vianney Perchet*

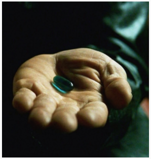
Journées MAS
August 2022

Crest, ENSAE & Criteo AI Lab

Session sponsored by the ANR BOLD

# Classical "Historical" Examples of Bandits Problems

- Size of data: $n$ patients with some proba of getting cured
- Choose one of two treatments to prescribe

 or 

- Patients **cured** or **dead**

1) **Inference:** Find the best treatment between the red and blue
2) **Cumul:** Save as many patientsvas possible

# Classical "Historical" Examples of Bandits Problems

- Size of data: $n$ banners with some proba of click
- Choose one of two ads to display

or

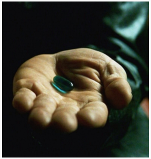
- Banner **clicked** or **ignored**

1) **Inference:** Find the best ad between the red and blue
2) **Cumul:** Get as many clicksvas possible

# Classical "Historical" Examples of Bandits Problems

- Size of data: $n$ patients with some proba of getting cured
- Choose one of two treatments to prescribe

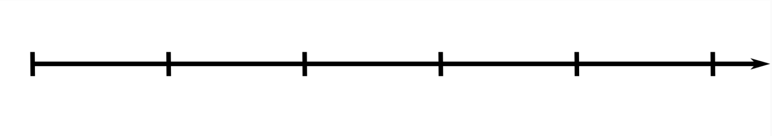
or



- Patients **cured** ♡ or **dead** ☠

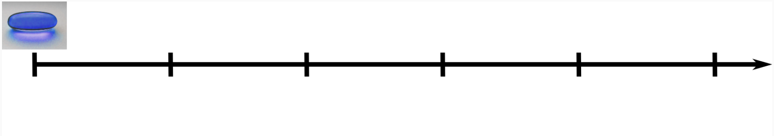
1) **Inference:** Find the best treatment between the red and blue
2) **Cumul:** Save as many patientsvas possible

# Two-Armed Bandit



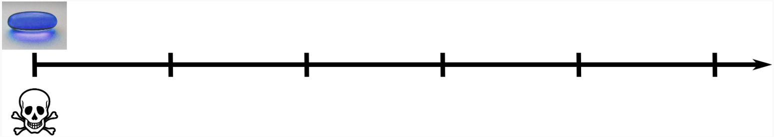
– Patients arrive and are treated sequentially.

# Two-Armed Bandit



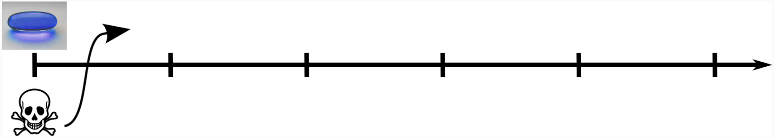
– Patients arrive and are treated sequentially.

# Two-Armed Bandit



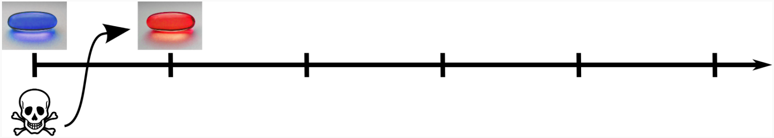
– Patients arrive and are treated sequentially.

# Two-Armed Bandit



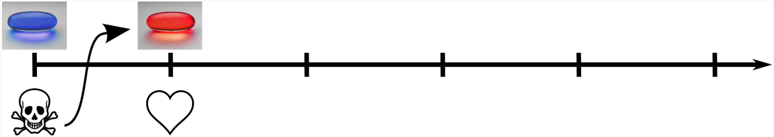
– Patients arrive and are treated sequentially.

# Two-Armed Bandit



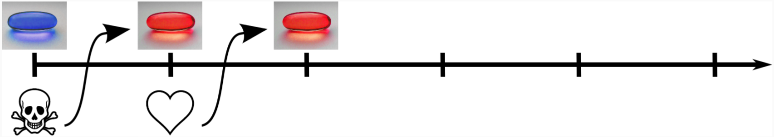
– Patients arrive and are treated sequentially.

# Two-Armed Bandit



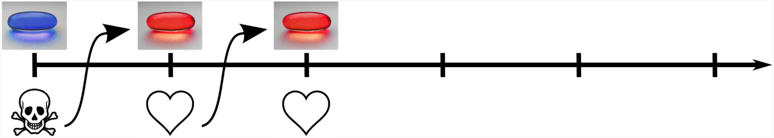
– Patients arrive and are treated sequentially.

# Two-Armed Bandit



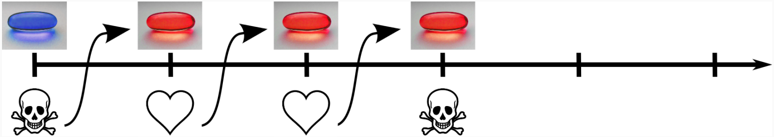
– Patients arrive and are treated sequentially.

# Two-Armed Bandit



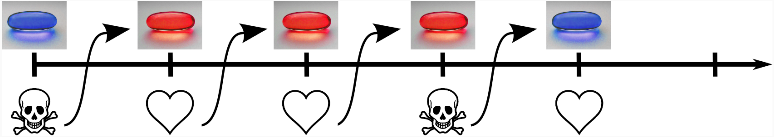
– Patients arrive and are treated sequentially.

# Two-Armed Bandit



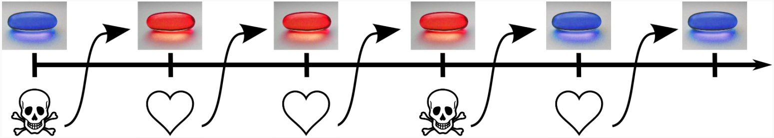
– Patients arrive and are treated sequentially.

# Two-Armed Bandit



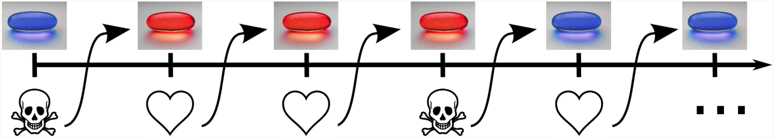
– Patients arrive and are treated sequentially.

# Two-Armed Bandit



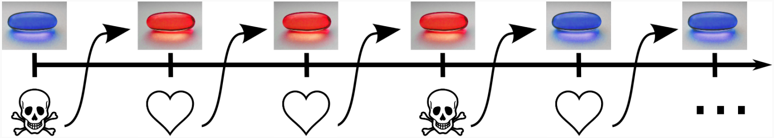
– Patients arrive and are treated sequentially.

# Two-Armed Bandit




– Patients arrive and are treated sequentially.

# Two-Armed Bandit



- Patients arrive and are treated sequentially.
- Save as many as possible.

# Outline

1. Bayesian environment
2. Non-Bayesian, yet stochastic environment
3. Some extensions & alternative models

# Bayesian Environment

# $K$-Armed Stochastic Bandit Problems

- $K$ arms $i \in \{1, \dots, K\}$, reward $X_t^i \in \mathbb{R}$ Gaussian/Bernoulli/...

$$X_1^i, X_2^i, \dots, \sim \mathcal{N}(\mu^i, 1) \quad \text{i.i.d.}$$

- **Prior** $\rho_i \in \mathcal{P}(\mathbb{R})$ (independent between arm)
- **Discount** $\gamma \in [0, 1]$ (termination proba.)
- Non-Anticipative Policy: $\pi_t\left(X_0^{\pi_0}, X_1^{\pi_1}, \dots, X_{t-1}^{\pi_{t-1}}\right) \in \{1, \dots, K\}$
- **Goal:** Maximize expected reward $\mathbb{E} \sum_{t=0}^{\infty} \gamma^t X_t^{\pi_t} = \mathbb{E} \sum_{t=0}^{\infty} \gamma^t \mu^{\pi_t}$

# Optimal Solution - Gittins Index

- Fully specified Bayesian pb. Optimal (non-tractable) strategies
- Simpler formulation (Gittins Index).
  Strategy optimal if it selects arm with the highest **Gittins index**
- **Gittins Index ??**
  - 2 arms. $X_t^1 \in [0,1]$ iid, $\mathbb{E}X_T^1 = \mu$ and $X_t^2 = \nu$ ($\mu$ unknown, $\nu$ known)
  - Optimal policy: selects 1 until $\tau$ then selects 2.
  - Gittins index (at $t$) = $\sup \{ \nu \mid \text{optimal policy selects 1 at time } t \}$

$$\nu_t^1 = \sup \left\{ \frac{\mathbb{E}\sum_{t=0}^{\tau} \gamma^t X_t^1}{\mathbb{E}\sum_{t=0}^{\tau} \gamma^t} \middle| \tau \text{ is a stopping time} \right\}$$

# Simple example of computations

- **Prior** on Arm 1: type $G$ with proba. $p$ and $B$ with proba. $1-p$
- Type $G$, reward $M$ a.s. ; type $B$, reward 0 a.s.
- Gittins index at 0 =

$$\frac{\mathbb{E}\sum_{t=0}^{\tau}\gamma^t X_t^1}{\mathbb{E}\sum_{t=0}^{\tau}\gamma^t} = \frac{p\frac{M}{1-\gamma}}{\frac{p}{1-\gamma}+(1-p)} = \frac{pM}{1-(1-p)\gamma}$$

- Gittens ind. **strictly larger** than expected reward ("exploration")
- Pros
  - **Reduction** from one $K$-arms problem to $K$ one-arm problem.
  - **Simple** decision policy (select highest index)
- Cons
  - Very **fragile**. All assumptions are necessary
  - **Computational burden** of indices.

# Simpler Bayesian Algorithm. Thompson Sampling

- **Prior** $\rho_i$ over the parametric family of distributions $\Theta_i$
- Repeat at each iteration
  - **Update** the prior w.r.t. the observation (**Bayesian** update)
  - **Pick** a parameter $\theta_i$ (for each arm) accordingly to the posterior
  - **Select** arm with the highest expectation given picked parameters
- Pros
  - **Simple** computations (Bayesian updates)

    Ex. Bernoulli + Beta prior: counting of success/failure
- Cons
  - **Sub-optimal** for the Bayesian pb
- Pros again
  - "almost" optimal, and **Optimal** for the **non-Bayesian** pb.

# Non-Bayesian Environment

# $K$-Armed Stochastic Bandit Problems

- $K$ actions $i \in \{1, \dots, K\}$, outcome $X_t^i \in \mathbb{R}$ Gaussian/Bernoulli

$$X_1^i, X_2^i, \dots, \sim \mathcal{N}(\mu^i, 1) \quad \textbf{i.i.d.}$$

- Non-Anticipative Policy: $\pi_t\left(X_1^{\pi_1}, X_2^{\pi_2}, \dots, X_{t-1}^{\pi_{t-1}}\right) \in \{1, \dots, K\}$
- Goal: Maximize expected reward $\quad \sum_{t=1}^T \mathbb{E} X_t^{\pi_t} = \sum_{t=1}^T \mu^{\pi_t}$
- Performance: Cumulative Regret

$$R_T = \max_{i \in \{1,\dots,K\}} \sum_{t=1}^T \mu^i - \sum_{t=1}^T \mu^{\pi_t} = \sum_i \Delta_i \sum_{t=1}^T \mathbb{1}\{\pi_t = i\}$$

with $\Delta_i = \mu^\star - \mu^i$, the “gap” or cost of error $i$.

# The pitfall of Reinforcement Learning : negative bias

- $\overline{X}_n^{(k)} = \frac{1}{n}\sum_{m=1}^{n} X_m^{(k)}$ **not** available, **only** $\widehat{X}_n^{(k)} = \dfrac{\sum_{m:k_m=k} X_m^{(k)}}{\sharp\{m : k_m = k\}}$
- with $k_n = \arg\max \widehat{X}_n^{(k)}$, $\mathbb{E}R_n = \Theta(n)$.

  because $\mathbb{E}[\widehat{X}_n^{(k)}] \leq \mu^{(k)}$ **negatively** biased
- **Positive** (vanishing) bias ? Tradeoff Exploitation / Exploration

**Hoeffding inequality: exponential decay**

$\left|\overline{X}_n^{(k)} - \mu^k\right| > \varepsilon$ with proba at most $2\exp\left(-2n\varepsilon^2\right)$.

Implies Finite number of $\varepsilon$-mistakes:

$$\mathbb{E}\sum_{n\in\mathbb{N}} \mathbb{1}\left\{\left|\overline{X}_n^{(k)} - \mu^k\right| > \varepsilon\right\} \leq \frac{1}{\varepsilon^2}$$

# Most Famous algorithm [Auer, Cesa-Bianchi, Fisher, '02]

- UCB - "Upper Confidence Bound"

$$\pi_{t+1} = \arg\max_i \left\{ \overline{X}_t^i + \sqrt{\frac{2\log(t)}{T^i(t)}} \right\},$$

where $T^i(t) = \sum_{t=1}^t \mathbb{1}\{\pi_t = i\}$ and $\overline{X}_t^i = \frac{1}{T_t^i} \sum_{s:\pi_s = i} X_s^i$.

**Regret:**

$$\mathbb{E}\, R_T \lesssim \sum_k \frac{\log(T)}{\Delta_k}$$

**Worst-Case:**

$$\mathbb{E}\, R_T \lesssim \sup_{\Delta} K \frac{\log(T)}{\Delta} \wedge T\Delta$$
$$\approx \sqrt{KT\log(T)}$$

## Ideas of proof $\pi_{t+1} = \arg\max_i \left\{ \overline{X}_t^i + \sqrt{\frac{2\log(t)}{T^i(t)}} \right\}$

- 2-lines proof:

$$\pi_{t+1} = i \neq \star \iff \overline{X}_t^\star + \sqrt{\frac{2\log(t)}{T^\star(t)}} \leq \overline{X}_t^i + \sqrt{\frac{2\log(t)}{T^i(t)}}$$

$$\text{“} \implies \text{”} \Delta_i \leq \sqrt{\frac{2\log(t)}{T^i(t)}} \implies T^i(t) \lesssim \frac{\log(t)}{\Delta_i^2}$$

- Number of mistakes grows as $\frac{\log(t)}{\Delta_i^2}$; each mistake costs $\Delta_i$.

$$\text{Regret at stage T} \lesssim \sum_i \frac{\log(T)}{\Delta_i^2} \times \Delta_i \approx \sum_i \frac{\log(T)}{\Delta_i}$$

- “ $\implies$ ” actually happens with overwhelming proba
- “optimal”: no algo with regret always smaller than $\sum_i \frac{\log(T)}{\Delta_i}$

# Optimality of Thompson Sampling and UCB ?

- Intuitions:
  - "Need" $\mathrm{KL}(\theta, \theta')$ samples to distinguish between $\theta$ and $\theta'$
  - Each sample cost $\mu - \mu'$ in regret (with $\mu = \mathbb{E}_{X \sim \theta} X$).
  - Regret should scale as $\sum_i \frac{\mu_* - \mu_i}{\mathrm{KL}(\theta_i, \theta^*)} \simeq \sum_i \frac{1}{\Delta_i}$
- Formally, Lai & Robbins'85. Any "**relevant**" algorithm satisfies

$$\liminf_{T \to \infty} \frac{R_T}{\log(T)} \geq \sum_i \frac{\mu_* - \mu_i}{\mathrm{KL}(\theta_i, \theta^*)}$$

  Relevant = expected regret always sub-polynomial.
- (variants of) UCB & Thompson Sampling "optimal"

  Remark: minimax regret $\Omega(\sqrt{KT})$

# Extensions

# Different Frameworks

- Best-Arm Identification
  - Do not minimize regret, identify $\star = \arg\max_k \mu^k$
    - **Fixed budget** of $T$ samples. Minimize proba. of mistake
    - **Fixed confidence** of $\delta \in [0,1]$. Minimize nb of samples
  - Algo: similar to regret minimization (UCB, successive elimination)
- Contextual Bandits
  - Reward depends on a covariate $Z_t$

    Observe $Z_t$, pick $k_t$, receive $\mu^{k_t}(X_t) + \text{noise}$
  - **Regularity** assumptions: $\mu^k(\cdot)$ linear, Lip., Holder, parametric....
  - Algo: combine Non-parametric regression with UCB
  - Typical regret in $T\left(\frac{K}{T}\right)^{\frac{\beta}{2\beta+d}}$ for $\beta$-Holder and $d$-dim. covariates
- Many more
  - Heavy-tail distribution, adversarial rewards (no assumption),...
  - Complete/Partial/Graph/Costly/Delayed observations...
  - Multi-player (with collisions, collusions, correlations...)