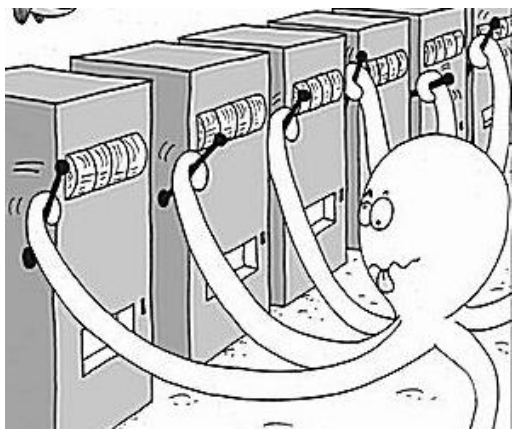
# Bandits
# (Machine Learning, SIR)

Matthieu Geist (CentraleSupélec)
matthieu.geist@centralesupelec.fr

- (multi-armed) bandit, the most basic example of a sequential decision problem with a trade-off between exploration and exploitation:
  - a gambler is facing a number of options;
  - at each time step, he chooses an option and receives a reward;
  - the goal is to maximize the total sum of rewards obtained in a sequence of allocations;
  - the player must balance the exploitation of actions that did well in the past and the exploration of actions that could give higher reward in the future.



Some applications:

- clinical trials;
- online services (*e.g.*, ad placement);
- cognitive radio;
- *etc.*
- less directly, playing Go, black-box optimization, *etc.*

- Formalism:
  - each arm $i \in \{1, \dots, K\}$ is associated to an *unknown* probability measure $\nu_i$;
  - typically, rewards are assumed to be bounded (in $[0, 1]$, without loss of generality);
  - at each time step $t = 1, 2, \dots$, the player chooses an arm $I_t \in \{1, \dots, K\}$ (based on past choices and observations) and receives a reward $X_{I_t,t}$ drawn from $\nu_{I_t}$, independently from the past;
  - expectation of the $i^{\text{th}}$ arm: $\mu_i = \mathrm{E}[X_{i,t}] = \int x d\nu_i(x)$;
  - highest expectation and corresponding arm:

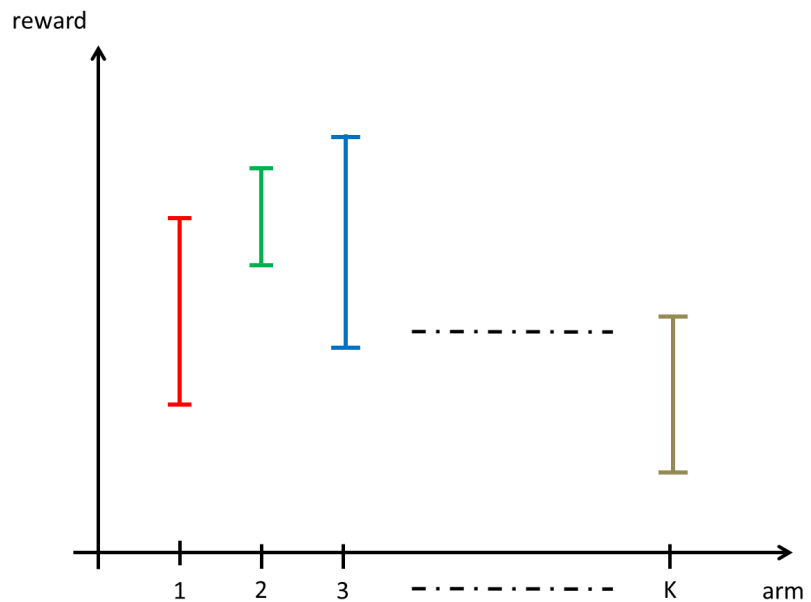
$$\mu_* = \max_{1 \le i \le K} \mu_i \text{ and } i_* \in \operatorname*{argmax}_{1 \le i \le K} \mu_i$$

- Ideal (but unreachable) strategy: $I_t = i_*$.
- Regret, quantifying the quality of a strategy:

$$\mathrm{R}_n = n\mu_* - \mathrm{E}\left[\sum_{t=1}^{n} \mu_{I_t}\right].$$

- Define:
  - the number of times the player selected arm $i$ during the first $s$ rounds,
$$T_i(s) = \sum_{t=1}^{s} \mathbb{I}_{\{I_t = i\}};$$
  - the suboptimality gap of arm $i$,
$$\Delta_i = \mu_* - \mu_i.$$
- One can easily check that
$$\sum_{i=1}^{K} T_i(n) = n \text{ and } \sum_{t=1}^{n} \mu_{I_t} = \sum_{i=1}^{K} \mu_i T_i(n)$$
- The regret can be rewritten as
$$\mathrm{R}_n = n\mu_* - \mathrm{E}\left[\sum_{t=1}^{n} \mu_{I_t}\right] = \sum_{i=1}^{K} \Delta_i \mathrm{E}\left[T_i(n)\right].$$
- A good strategy should control the (expected) number of time a suboptimal arm is played.

- Possible strategies (from empirical expectations of each arm):
  - greedy strategy
  - $\epsilon$-greedy strategy
  - softmax strategy
  - ...
- Here, we consider optimism in the face of uncertainty
  - act greedily resp. to the most "favorable" case
  - here, highest upper-bound of arm's confidence interval



Let $X_1, \ldots, X_n$ be i.i.d. (independent and identically distributed) random variables. Write $\mu = \mathrm{E}[X_1]$ their common expectation and $\mu_n$ the related empirical mean: $\mu_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

**Theorem (Hoeffding's inequality)**

*Assume that there exists a convex function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ such that*

$$\forall \lambda \geq 0, \quad \ln \mathrm{E}\left[e^{\lambda(X_1-\mu)}\right] \leq \psi(\lambda) \text{ and } \ln \mathrm{E}\left[e^{\lambda(\mu-X_1)}\right] \leq \psi(\lambda).$$

*Define the Legendre-Fenchel transform of $\psi$ as*

$$\psi_*(\epsilon) = \sup_{\lambda \geq 0} \left(\lambda\epsilon - \psi(\lambda)\right).$$

*Then:*

$$\mathrm{P}\left(\mu_n - \mu \geq \epsilon\right) \leq e^{-n\psi_*(\epsilon)} \text{ and } \mathrm{P}\left(\mu - \mu_n \geq \epsilon\right) \leq e^{-n\psi_*(\epsilon)}.$$

PAC (Probably Approximately Correct) equivalent formulation: with probability at least $1-\delta$, we have

$$\mu \leq \mu_n + \psi_*^{-1}\left(\frac{1}{n}\ln\frac{1}{\delta}\right).$$

- What about bounded rewards?

**Lemma (Hoeffding)**

*Let $Y$ be a random variable such that $\mathrm{E}[Y] = 0$ and $c \leq Y \leq d$ almost surely. Then, for any $s \geq 0$, we have*

$$\mathrm{E}\left[e^{sY}\right] \leq e^{s^2 \frac{(d-c)^2}{8}}.$$

**Corollary (Hoeffding)**

*Assume that $0 \leq X_1 \leq 1$ almost surely. Then we have*

$$\mathrm{P}\left(\mu_n - \mu \geq \epsilon\right) \leq e^{-2n\epsilon^2} \text{ and } \mathrm{P}\left(\mu - \mu_n \geq \epsilon\right) \leq e^{-2n\epsilon^2}.$$

PAC form: with probability at least $1 - \delta$, we have

$$\mu \leq \mu_n + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}.$$

- The UCB (Upper Confidence Bound) strategy applies the principle of optimism in the face of uncertainty:
  - sample mean of rewards obtained by pulling arm $i$ for $s$ times:
$$\mu_{i,s} = \frac{1}{s}\sum_{t=1}^{s} X_{i,t};$$
  - from Hoeffding, w.p. at least $1-\delta$,
$$\mu_i < \mu_{i,s} + \psi_*^{-1}\left(\frac{1}{n}\ln\frac{1}{\delta}\right);$$
  - recall arm $i$ has been played $s = T_i(t-1)$ before round $t$;
  - let choose $\delta = \frac{1}{t^\alpha}$ where $\alpha > 0$ is a free parameter
- $(\alpha, \psi)$-UCB strategy
$$I_t \in \operatorname*{argmax}_{1\le i\le K}\left(\mu_{i,T_i(t-1)} + \psi_*^{-1}\left(\frac{\alpha\ln t}{T_i(t-1)}\right)\right).$$
- for rewards bounded in $[0,1]$, use $\psi_*(\epsilon) = 2\epsilon^2 \Leftrightarrow \psi_*^{-1}(u) = \sqrt{\frac{u}{2}}$:
$$I_t \in \operatorname*{argmax}_{1\le i\le K}\left(\mu_{i,T_i(t-1)} + \sqrt{\frac{\alpha\ln t}{2T_i(t-1)}}\right).$$

**Theorem (Regret analysis)**

*Assume that the rewards distributions satisfy the assumption of the Hoeffding's theorem. Then the $(\alpha, \psi)$-UCB strategy with $\alpha > 2$ satisfies*

$$\mathrm{R}_n \leq \sum_{i:\Delta_i>0} \left( \frac{\alpha \Delta_i}{\psi_* \left( \frac{\Delta_i}{2} \right)} \ln n + \frac{\alpha}{\alpha - 2} \right).$$

*If rewards are bounded in $[0, 1]$, the bound on the regret simplifies as (using the fact that $\psi_*(\epsilon) = 2\epsilon^2$):*

$$\mathrm{R}_n \leq \sum_{i:\Delta_i>0} \left( \frac{2\alpha}{\Delta_i} \ln n + \frac{\alpha}{\alpha - 2} \right).$$

## More on bandits

- Other topics with stochastic bandits:
  - identify the best arm with a fixed budget or a fixed confidence (without paying attention to gathered rewards);
  - stochastic bandits from a Bayesian viewpoint;
  - etc.
- Other kinds of bandits:
  - adverserial bandits;
  - contextual bandits;
  - Markovian bandits;
  - linear bandits;
  - etc.