INTRODUCTION TO REINFORCEMENT LEARNING

Nathanaël FIJALKOW

$$RL \subseteq ML$$

ML:

$$\left.\begin{array}{l}\text{Data}\\ \text{Background knowledge}\end{array}\right\} \longrightarrow \text{Model} \longrightarrow \left\{\begin{array}{l}\text{Action}\\ \text{Prediction}\end{array}\right.$$

Supervised learning:

Training data is labelled

Objective: predict unlabelled data

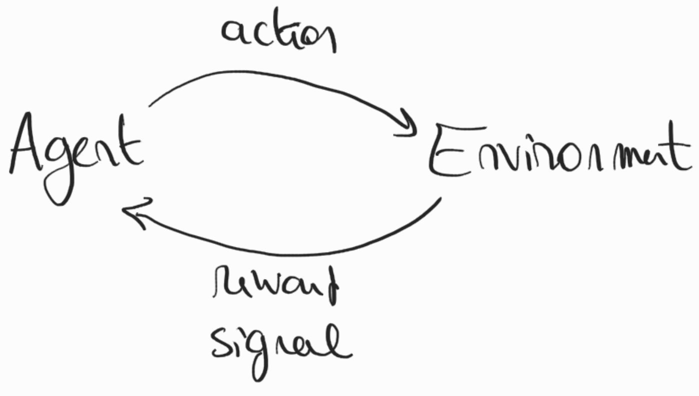
Unsupervised learning:

Training data is not labelled

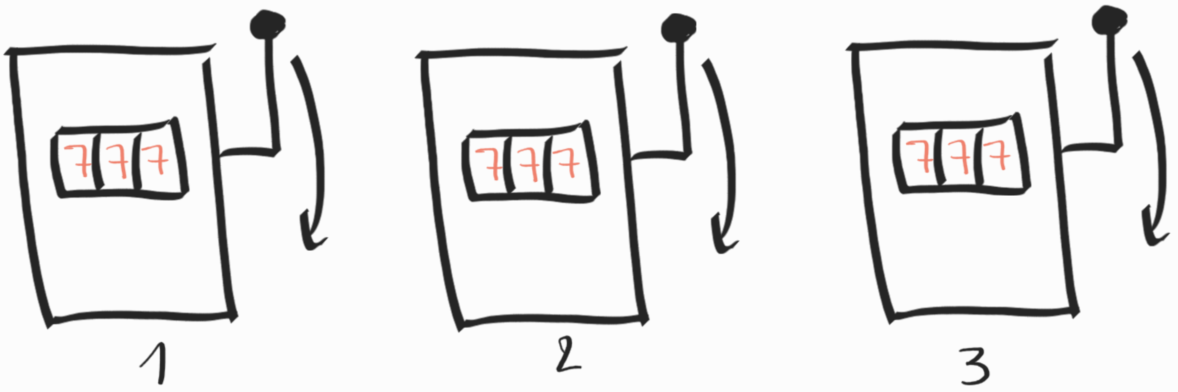
objective: finding structure / pattern

RL is neither supervised nor unsupervised !

RL:
- active learning
- reward-based



MULTI ARMED BANDITS



arm = machine = action

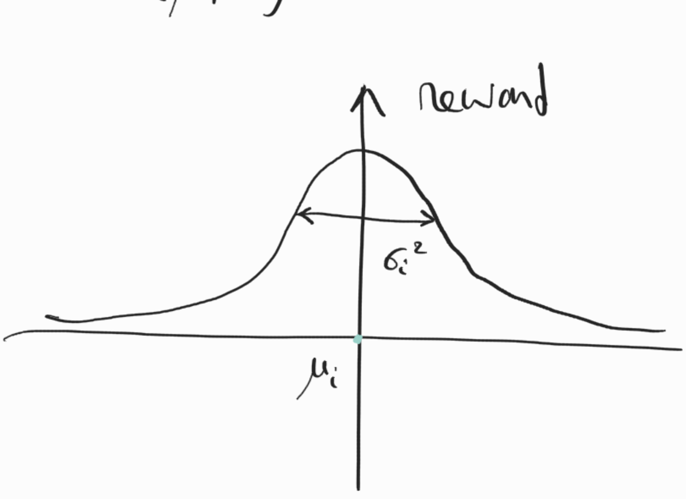
K machines : 1, ..., K

we assume that each machine i has a reward distribution δ_i

Bernoulli distribution $B(p)$

$$\begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

Normal distribution $\mathcal{N}(\mu, \sigma^2)$



Notation: $\mu = \mathbb{E}[\delta]$ $\sigma^2 = \mathrm{Var}(\delta)$

Scenario:

- we know the number of machines but NOT THE REWARD DISTRIBUTIONS

- at each time step t we pick an arm i and draw a reward $r(t) \in \mathbb{R}$ from δ_i

- Two (similar) goals:

  → (1) identify the best arm: $\arg\max_i \mu_i$

  → (2) maximise the total reward:

$$\sum_{t \geq 1} r(t) \leftarrow \text{reward at time } t$$

$K = 2$

$$\delta_1 = \begin{cases} 2 & \text{prob. } 1/2 \\ -1 & \text{prob } 1/2 \end{cases}$$

$$\delta_2 = \begin{cases} 100 & \text{prob. } 1/10 \\ 0 & \text{prob } 9/10 \end{cases}$$

time	machine	
1	1	$\longrightarrow r(1) = 2$
2	2	$\longrightarrow r(2) = 0$
3	2	$\longrightarrow r(3) = 0$
4	2	$\longrightarrow r(4) = 0$

Difficulty: we have to make choices based on incomplete statistics !

Trade off between:

- getting good information on all machines:
  EXPLORATION

- getting good rewards
  EXPLOITATION

after T steps

for each machine $i \in [1,K]$

we have a set of samples

↳ $\hat{\mu}_i(T)$ empirical expectation

Greedy strategy:

EXPLOITATION

$$\text{play} \quad \operatorname*{argmax}_{i\in[1,K]} \hat{\mu}_i(T)$$

ε-Greedy strategy:

ε fixed $\quad \varepsilon = 0.1$

$$\begin{array}{l} \text{exploration} \\ \\ \text{exploitation} \end{array} \begin{cases} \text{uniformly at random over all actions} & \text{w. probability } \varepsilon \\ \\ \text{greedy: } \operatorname*{argmax}_{i\in[1,K]} \hat{\mu}_i(T) & \text{w. probability } 1-\varepsilon \end{cases}$$

Notations

$$\underset{i \in [1,k]}{\arg\max} \; \mu_i \overset{\text{def}}{=} * \in [1,K] \text{ such that}$$

$$\mu_* = \max_{i \in [1,K]} \mu_i$$

Regret analysis

Regret(T) = R(T) = difference between

" best a posteriori "

and

" what we achieved "

$$R(T) = T \cdot \mu_k - \sum_{t=1}^{T} r(t)$$

⚠ μ_i is the actual expectation

$\hat{\mu}_i(T)$ is the empirical expectation at time T

$$\max \sum_{t \geq 1}^{T} r(t)$$

$$\iff$$

$$\min \quad R(T)$$

We use the regret for comparing different strategies

So far:

Greedy $\quad \underset{i}{\operatorname{argmax}} \ \hat{\mu}_i(T)$

ε-Greedy $\quad \begin{cases} \underset{i}{\operatorname{argmax}} \ \hat{\mu}_i(T) & \text{with probability } 1-\varepsilon \\ \text{uniform at random} & \text{with probability } \varepsilon \end{cases}$

Issues:
- exploration never stops : at least ε is lost
- exploration does not take existing info into account
- may take a long time to converge

UPPER CONFIDENCE BOUNDS (UCB)

$r(t)$: reward at time t

$$r(i,t) : \begin{cases} r(t) & \text{if } i \text{ was chosen at time } t \\ 0 & \text{o/w} \end{cases}$$

$$R(T) = T \cdot \mu_* - \sum_{t=1}^{T} r(t)$$

$$\hat{\mu}_i(T) = \frac{1}{n(i,T)} \sum_{t=1}^{T} r(i,t)$$

UCB

$$\underset{i}{\operatorname{argmax}} \; \hat{\mu}_i(T) + c(i,T)$$

$$c(i,T) = \sqrt{\frac{\log(T)}{n(i,T)}}$$

intuitions :

- if $n(i,T)$ is small (little information) then $c(i,T)$ is large : we need to explore
- if $n(i,T)$ is large then $c(i,T)$ is small, except when T grows, but exponentially less often

why $c(i,T) = \sqrt{\frac{\log(T)}{n(i,T)}}$?

Chernoff-Hoeffding bounds

let $y_1, \dots, y_n$ iid samples of Y

Law of large numbers: $\frac{1}{n}\sum_{i=1}^{n} y_i \longrightarrow \mathbb{E}[Y]$

Better: this happens fast!

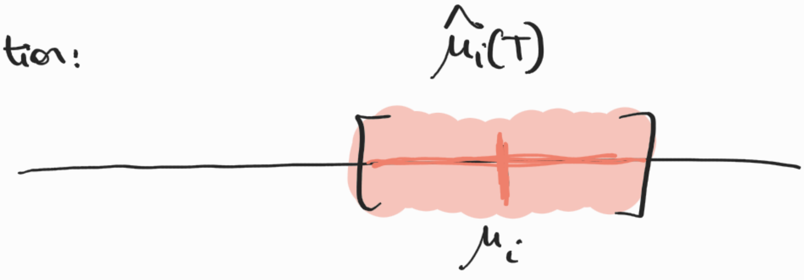
$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} y_i - \mathbb{E}[Y]\right| \geq c\right) \leqslant 2e^{-2c^2 n}$$

We apply it to our case:

$$\mathbb{P}\left(\left|\hat{\mu}_i(T) - \mu_i\right| \geq c(i,T)\right) \leqslant 2\,e^{-2\,c(i,T)^2 n(i,T)}$$

$c(i,T) = \sqrt{\frac{\log(T)}{n(i,T)}}$ gives $\frac{2}{T^2} \xrightarrow[T \to +\infty]{} 0$

Intuition:



with high probability

UPDATES

$$\hat{\mu}_i(T) = \frac{1}{n(i,T)} \sum_{t=1}^{T} r(i,t)$$

After choosing i :

$$\hat{\mu}_i(T+1) = \frac{1}{\underbrace{n(i,T+1)}_{n(i,T)+1}} \sum_{t=1}^{T+1} r(i,t)$$

$\downarrow$ Small calculations

$$\hat{\mu}_i(T+1) = \hat{\mu}_i(T) + \frac{1}{n(i,T)+1}\left[X(i,T+1) - \hat{\mu}_i(T)\right]$$

$$\text{NEW} = \text{OLD} + \alpha\left[\text{CURRENT} - \text{OLD}\right]$$

we call this step size

Two possible updates :

- empirical mean
- constant step size ($\alpha = 0.1$ for instance)

$$\hat{\mu}_i(T+1) = \hat{\mu}_i(T) + \alpha\left(\hat{\mu}_i(T) - X(i,T+1)\right)$$