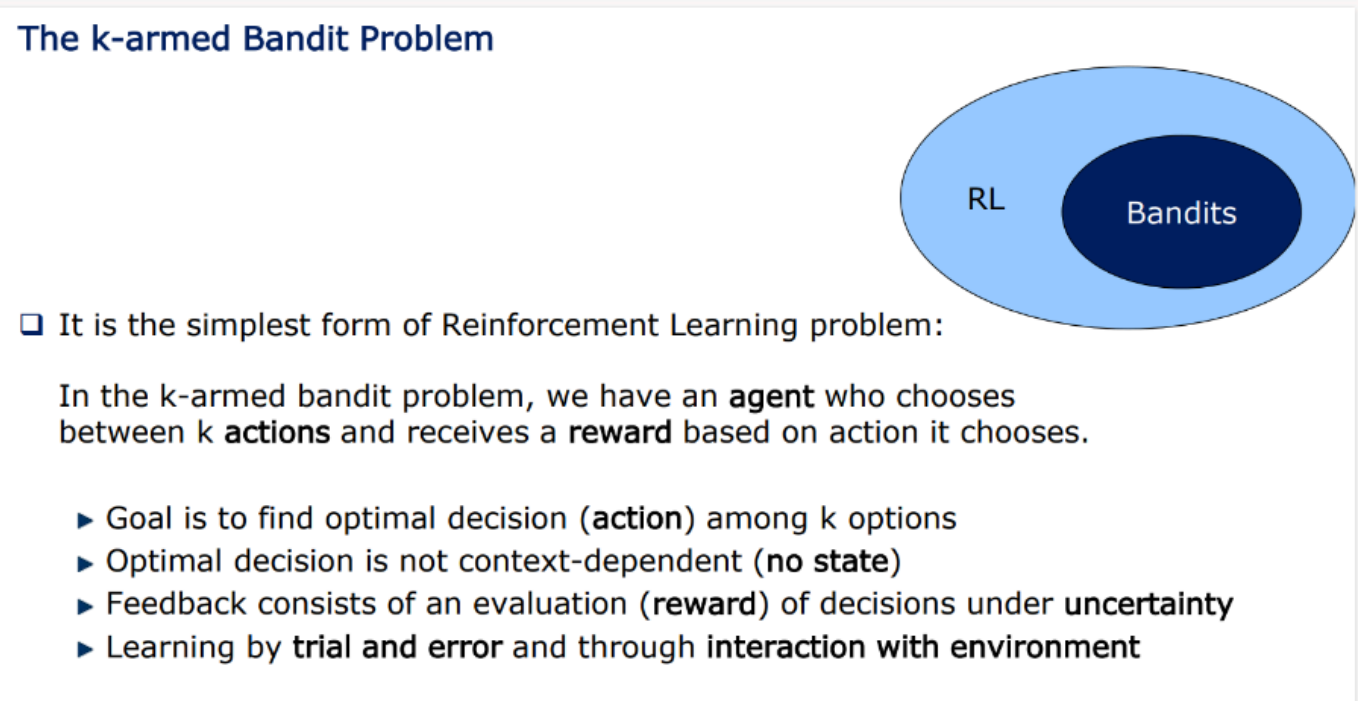
Multi-Armed Bandits

These are algorithm used to solve a problem where I have to take a decision in an uncertain scenario(Ex: select/advice a movie).



It is like a Markov Decision Problem with a single state and with not knowing the result of the choice. The name derive from the fact that the actions taken are named bandits.

Action Values

The value of each action is defined as the expected reward:
$q^*(a) \doteq \mathbb{E}[A_t = a] = \sum p(r|a)r, \forall a \in 1, .., k$
The goal of the agent is to maximize the expected reward: $argmax_a q^*(a)$
Indicator function: a function that is equal to 1 if the value is a specified one 0 otherwise
The more sample I go through the more the empirical average is going to be close to the real value.

Estimate of q*(a)

- As $p(r|a)$ is not known, we estimate $q^*(a)$ from experience:

sum of rewards when a chosen before step t

$$Q_t(a) \doteq \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_t=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_t=a}}$$

times a chosen before step t

Incremental update and non-stationary problems

Incremental update of action-values

- Let's consider the update of a single action:

$$Q_n = \frac{1}{n-1}\sum_{i=1}^{n-1} R_i$$

$$\begin{aligned} Q_{n+1} &= \frac{1}{n}\sum_{i=1}^{n} R_i = \frac{1}{n}\left(R_n + (n-1)\frac{1}{n-1}\sum_{i=1}^{n-1} R_i\right) \\ &= \frac{1}{n}\left(R_n + (n-1)Q_n\right) \\ &= Q_n + \frac{1}{n}\left(R_n - Q_n\right) \end{aligned}$$

Incremental update of action-values

new estimate

step size

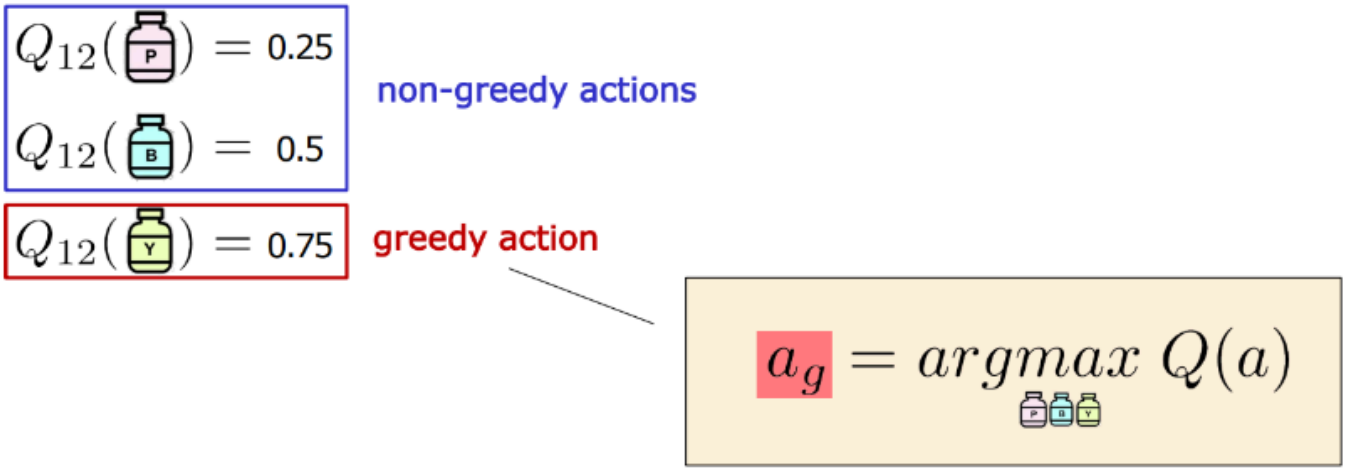
$$Q_{n+1} = Q_n + \frac{1}{n}\left(R_n - Q_n\right)$$

old estimate

target – old estimate

I can also comprehend updates in the parameter of my evaluated actions that I am looking at with the addition of a parameter that will indicates how much I have to look at the past samples.

Epsilon-Greedy Action Selection



I create a budget for exploration/exploitation and see if I am in a case where I am exploring or exploiting. It is called greedy as I am taking a choice on the base of what I think is the optimal choice in base of the data I have without looking at an actual optimization process.

$$A_t = \begin{cases} argmax_a\ Q_t(a) & \text{with probability } 1-\epsilon \\ Uniform(\{a_1, ..., a_k\}) & \text{with probability } \epsilon \end{cases}$$

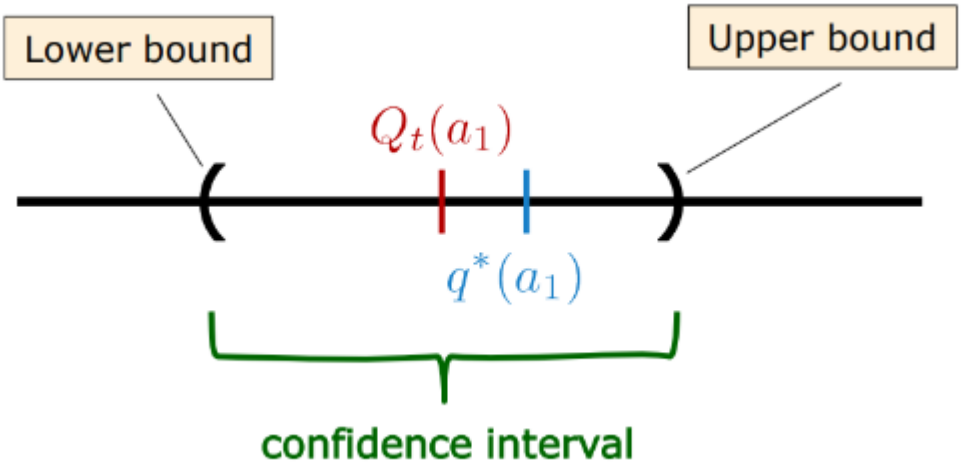
Optmistic Initial Values

I tweak my initial value in the function.
Limitations of optimistic initial values

- Optimistic initial values only drive early exploration
- They are not well-suited for non-stationary problems
- We may not know what othe optimistic initial value should be

UCB Action Selection

Can I behave in a smart way and select uniformly. Assume I get a fair evaluation so it is reasonable that I should explore some actions that I am really sure for my action in several time.



I want to narrow down the confidential interval only when I am certain it will be the best move(Hit the one with the greater upperbound).