# Reinforcement Learning and Optimal Control

# Tutorial 1 - $k$-armed bandit problems

Nadir Farhi

October 03, 2024

## Exercise 1

The objective of this exercise is to reproduce the illustrations seen in class on the 10-armed bandits problem.

- We consider the 10-armed bandit problems ($k = 10$).
- We simulate the learning process for 2000 steps.
- We use the $\varepsilon$-greedy policy with different values of $\varepsilon$: 0, 0.1, 0.01.
- We assume that the action values $q_*(a)$ follows Gaussian distributions $N(0,1)$ for every $a = 1,2,\ldots,10$.
- We assume that the received rewards after selecting action $A_t$ at time step $t$ follow the Gussian distributions $N(q_*(A_t),1)$.
- The initial estimates are assumed to be zero.

Questions:

1. Illustrate the curve of the average reward as a function of the number of steps.
2. Illustrate the curve of the optimal actions as a function of the number of steps.
3. Check if the illustrations seen in class are reproduced.

## Exercise 2

Same exercise, with:

- $k = 15$

- Number of steps: 3000, the actions values $q_*(a)$ follow $N(0,2)$
- The reward after selection of action $A_t$ follows $N\left(q_*(A_t), \frac{1}{2}\right)$
- Initial estimates: case 1: zero, case 2: 5
- The value of $\varepsilon$:
    - case 1: 0,
    - case 2: 0.1,
    - case 3: 0.01,
    - case 4: $\varepsilon\ (t)\ =\ 1/t$,
    - case 5: $\varepsilon(t) = 1/t^2$.

Questions:

1. Illustrate the curve of the average reward as a function of the number of steps.

2. Illustrate the curve of the optimal actions as a function of the number of steps.

3. Compare with the results obtained in Exercise 1.

# Exercise 3

We consider the 10-armed bandit problem.

- For the estimation of the value of an action, we use the simple average method with the incremental implementation.
- For the action selection we consider the following approaches:
    - The greedy method: $A(t)\ = \underset{a}{\operatorname{argmax}} Q_t(a)$ .
    - The $\varepsilon$-greedy method.
    - The UCB action selection:

$$A_t = arg\max_a \left( Q_t(a) + c\sqrt{\frac{\log t}{N_t(a)}} \right).$$

Question:

Simulate the 10-bandit problem, with different values of $\varepsilon$ and $c$, and compare the three proposed approaches.

# Exercise 4 - Gradient Bandit Algorithm

We consider the gradient bandit algorithm, with:

- $H_0(a) = 0, \forall a,$
- $\pi(a) = \frac{1}{k}, \forall a,$
- At $t+1$, after selecting action $A_t$, and receiving $R_t$, the action preferences $H$ are updated as seen in the class:

$$\begin{aligned} H_{t+1}(A_t) &\doteq H_t(A_t) + \alpha\big(R_t - \bar{R}_t\big)\big(1 - \pi_t(A_t)\big), && \text{and} \\ H_{t+1}(a) &\doteq H_t(a) - \alpha\big(R_t - \bar{R}_t\big)\pi_t(a), && \text{for all } a \neq A_t, \end{aligned}$$

- the action probabilities $\pi(a)$ are obtained according to the Soft-max distribution:

$$\Pr\{A_t = a\} \doteq \frac{e^{H_t(a)}}{\sum_{b=1}^{k} e^{H_t(b)}} \doteq \pi_t(a),$$

with $\alpha = 0.1, 0.2, 0.5, 0.7$, and $\overline{R_t}$ is obtained:

$t-1$

$$\overline{R_t} = \sum_{s=1}^{t-1} R_s \,/(t-1).$$

$s = 1$

Questions: For $k$ = 10, simulate the 10-armed bandit problem, and:

1. Illustrate the optimal action as a function of the number of steps.
2. Compare to the figures seen in class.
3. Resimulate for other value of $\alpha$, and compare.