Reinforcement Learning and Decision Making Under Uncertainty
IN-STK 5100/9100

Spring 2022
A. George, UiO
8th February, 2022

Mandatory Assignment 2

Submission Deadline: 22nd February, 2022

Exercise 5. Optimal Strategy

Consider the finite horizon k-armed bandit problem with stationary reward distributions. Assuming the reward distributions are known, consider a strategy that always pulls an arm with maximal value $q^{max} = \max_{a \in [k]} q^*(a)$.

a) Show that this strategy achieves maximal expected total reward.

b) Show that this strategy achieves a regret of 0.

Exercise 6. Optimistic Initial Values

Consider the k-armed bandit problem with stationary reward distributions over rewards $\{0, 1\}$. Let ϵ-Greedy with $\epsilon = 0$ be implemented with a fixed step size $\alpha = 1/4$ for the updates on action-value estimations, i.e., $Q^{t+1}(A_t) = Q^t(A_t) + \alpha \cdot [R_t - Q^t(A_t)]$. To which value can the initial estimates of action-values $Q^0(a)$ be set so that we are guaranteed to pull every arm at least 2 times?

Exercise 7. Weighted Averages

Consider the k-armed bandit problem. To update the estimated action-values in ϵ-Greedy, we can consider the weighted average with step size $\alpha_t(a) \in (0, 1]$ such that $Q^{t+1}(a) = Q^t(a) + \alpha_t(a)(R_t - Q^t(a))$ if $a = A_t$.

a) Show that the weights of the rewards in the weighted average with constant step size indeed sum up to 1, i.e.,

$$(1-\alpha)^{N^{t+1}(a)} + \sum_{i=1,\dots,t} \alpha(1-\alpha)^{N^{t+1}(a) - N^{i+1}(a)} \cdot \mathbb{I}_{A_i = a} = 1 \text{ for } \alpha \in (0, 1],$$

where $N^{j+1}(a)$ is the number of pulls after the j-th round.
Hint: Consider $x := N^{t+1}(a)$ to replace all $N^{\cdots}(a)$.

b) Let's say we implement ϵ-Greedy ($\epsilon > 0$) with weighted average action-value estimations that use step size $\alpha_t(a) = \frac{1}{t+1}$. Assume the reward distributions are stationary. Can we guarantee convergence of $Q^t(a) \to q^*(a)$ for $T \to \infty$ for all actions a? Why?