AV499 - AVD871 - Programming Assignment 7.

k-armed Bandits (multiarmed Bandits).

Not required to submit. Please work out on your own and compare with solution.

- Note that a k-armed bandit is the simplest version of a reinforcement learning (RL) problem where the system has a fixed state. In RL, a fundamental idea is that, although we do not know a model (F or p(a)) for the system or for the reward, we can interact with the system many times. That is, we can feed in control actions A_t into the system and observe the reward R_t and the next state S_{t+1} (which is fixed for a k-armed bandit). When learning about RL it is therefore required to build a simulation of the system - the simulated system will interact with the agent. This is your first task.

Suppose we have a k-armed bandit with $k = 3$ and rewards specified as follows for the 3 actions $\{1, 2, 3\}$

$R(s, 1) \sim$ Uniform $[0, 10]$.

$R(s, 2) \sim$ Exponential $[1/5]$

$R(s, 3) \sim$ Exponential $[2/5]$.

Setup a system simulator that will take actions A_t as input and produce R_t.

- Now implement an agent/controller which chooses actions according to a ϵ-greedy policy where $\epsilon = 0.1$. For this controller, plot the regret as a function of time. What do you observe? Do we need to do any "averaging" here to get a match with what we had learned in theory? Study the regret plot as a function of ϵ, the initial estimates $Q_0(a)$. Study the regret for a horizon of 500.