Multi-armed Bandits

Saúl Díaz Infante Velasco

Multi-armed Bandits

A verry important feature distinguishing reinforcement learning from other types of learning is that it uses training information that evaluates the actions taken rather than instructs by giving correct actions.

A k-armed Bandit Problem

We consier the following setup:

- You are faced repeatedly with a choice among k diferent options, or actions.
- After a choice you recive a numerical reward chosen from a stationary probability distribution that depends on the action you selected
- Your objective is to maximize the expected total reward over some time period, for example, over 1000 action selections, or time steps.

The problem is named by analogy to a slot machine, or `one-armed bandit`, except that it has k levers instead of one.

Each of the k actions has an expected or mean reward given that that action is selected; let us call this the value of that action.

We denote the action selected on time step t as A_t, and the corresponding reward as R_t . The value then of an arbitrary action a, denoted $q_*(a)$, is the expected reward given that a is selected:

$$q_*(a) = E[R_t | A_t = a].$$