Chapter 2 : Multi-Armed Bandits

December 12, 2019

> The most 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.

- Evaluative feedback indicates how good the action taken was, but not which action is better (depends entirely on the action)
- Instructive feedback indicates which action to take, independently of the taken action. (intependent of the action taken)

This chapter : **RL where we learn to act in only one situation** (nonassociative setting). Towards the end of the chapter, we tackle the associative problem (which action to take in more than one situation)

1 A k-armed Bandit Problem

- One choice among k different options (or actions)
- After each choice, we receive a numerical reward chosen from a stationary probability distribution depending on the action selected.
- Goal is to maximize the expected total reward over some time period (*time steps*)

In this problem, each of the k actions have an expected or mean reward (the *value* of the action). Action selected at time t is $A(t)$, and corresponding reward is $R(t)$. Then, the expected reward $q_*(a)$ is given by :

$$q_*(a) \doteq \mathbb{E}\left[R_t | A_t = a\right] \tag{1}$$

This is a trivial problem if we know the value of each action. We assume we don't know them, but we may have estimates ($Q_t(a)$). We would like $Q_t(a)$ to be close to $q_*(a)$.

If we maintain the estimates at each time step, we have one action with largest estimated reward, which we call the *greedy action*. When selecting them, we are **exploiting** the current knowledge. If we select another one, we are **exploring** the system.

Ways to balance between exploration and exploitation might be complicated. Here, we focus only on simple approaches.

2 Action-value Methods

Ways to evaluate values of actions. One natural way to etimate it is average of rewards actually received :

$$Q_t(a) \doteq \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t} = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}} \tag{2}$$

Note that if the denominator is 0, then we define a default value. As the denominator goes to infinity, the LLN says that $Q_t(a)$ converges to $q_*(a)$. This is called the **sample-average method**. (Maybe not the best method, but simple enough for now).

Simplest action selection rule is to select one of the highest estimated values (greedy).

$$A_t \doteq \arg\max_a Q_t(a) \tag{3}$$

This focusses on exploitation and not exploration. Simple approach to also do exploration would be to sample greedily most of the time, but sometimes (e.g. with probability ϵ) select an action at random (called ϵ-*greedy* methods)

> In ϵ-*greedy* action selection, for the case of two actions and $\epsilon = 0.5$, what is the probability that the greedy action is selected?
>
> Answer : 0.75 (half of the time, we select the greedy, the other half, we select one action at random, so 50% of these times (25% of all times), we select the greedy also)