Summary Chapter 2 RL Book [1] Submitted by: Dhawal Gupta at September 18, 2019

1 Problem Definition : K armed bandit

The problem can be defined by facing a slot machine consisting of k arms (/levers) which you can pull. Given a chance, you can select one arm to pull, given that choice of arm you receive a numerical reward (can be thought of as monetary payoff from slot machine) from and underlying unknown distribution, which may be different for each of the k arms. Our objective is to maximize the expected total reward that we can earn in say a given number of tries or time period.

1.1 Notation

We can define different aspects of the problem as follows:

a	The value of action selected
A_t	Random variable, action selected at time t
R_t	Reward collected at time t
$Q_t(a)$	Estimate value function of action a at time t
$q_*(a)$	Optimal value of action a
α	Constant learning rate
α_t	Learning rate at time t
$\doteq$	Defined as t
π_t	Policy followed at t

The optimal value function i.e. q_* for an action a can be defined as follows

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

1.2 Terms

Exploitation Policy : A policy π_t which tries to exploit the estimated value of an action at time t using Q_t and picks the action with the estimated highest return.
Exploration Policy : A policy π_t which tries to explore the different actions at time t and picks action using some criterion or randomly.
Stationary Problem : A k-armed bandit problem where the reward distribution of each arm stays unchanged over the duration of experiment.
Non Stationary Problem : A k-armed bandit problem in which the reward distribution of each arm (/action) may change over the period of time.

2 Action-value Methods

We can use sample averaging method to estimating the value of different actions

$$Q_t(a) \doteq \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}} \tag{2}$$

Either we can follow a greedy policy and select our action based on $argmax_a Q_t(a)$, or we can follow an ϵ-greedy policy, where we select a random action with probability ϵ otherwise we pick do exploitation

3 Incremental Implementation

Rather than keeping a track of all the previous rewards for a action a we will write it into a form of incremental form where we need to remember the previous estimate and number of times the action as been picked.

$$Q_{n+1} = Q_n + \frac{1}{n}[R_n - Q_n] \text{ for a single action} \tag{3}$$

4 Tracking a Non Stationary Problem

The problem with 3 is that the update error weight decays with the number of plays and the agent might not be able to adapt to changing reward distributions after a given amount of time. We can try to use a constant learning parameter α to counter this.

$$\begin{aligned} Q_{n+1} &\doteq Q_n + \alpha[R_n - Q_n] \\ &= (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i. \end{aligned} \tag{4}$$

5 Optimistic Initial Values

The book [1] shows that setting the initial action values estimates of all the actions to an optimistic values allows us to embed the process of exploration in the inherent process of doing a greedy search , but again that is plagued by the process of non stationary problems.

6 Upper-Confidence-Bound Action Selection

In this approach we try to reason on the uncertainty of different action and their potential of returning higher rewards than the current estimated maximum reward. One way to do it is via following:

$$A_t \doteq \underset{a}{argmax}[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}}] \tag{5}$$

Where $N_t(a)$ denotes the number of times action a has been selected prior to time t. c control the degree of exploration.

7 Gradient Bandit Algorithm

Now we try to learn a numerical preference for each action a denoted by $H_t(a)$. Larger preference implies more probability of taking the action, there is not interpretation in terms of reward. Only the relative preference matters and absolute values carry no significance.

$$PrA_t = a \doteq \frac{\exp^{H_t(a)}}{\sum_{b=1}^{k} exp^{H_t(b)}} \doteq \pi_t(a) \tag{6}$$

Update rules for the preference of action.

$$\begin{aligned} H_{t+1}(A_t) &\doteq H_t(A_t) + \alpha(R_t - \hat{R}_t)(1 - \pi_t(A_t)), \text{and} \\ H_{t+1}(a) &\doteq H_t(a) - \alpha(R_t - \hat{R}_t)\pi_t(a), \text{ for all } a \neq A_t \end{aligned} \tag{7}$$

Here $\hat{R}_t$ is the average of all rewards up to the time t, including it.

7.1 Contextual Bandits

The contextual bandits problem extends the current simple problem into the case where we have n different k armed slot machine, and given the id of a machine we need to learn to pull the best arm of the machine. This is only short of the full reinforcement learning problem , if taking the current action decided the future state (or machine) we might have to take an action on (play on).

References

[1] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. 2018.