

Reinforcement Learning is a subfield under machine learning. In RL, the agent collects its own data by trial and error. Every time, an agent is asked to make a decision, it is given a few options. The agent gets a reward for selecting an option. The agent's job is to maximize the reward for its actions.

If the reward for actions was known, the problem of RL would be trivial. The agent can greedily select the actions with the highest reward. However, the reward most often is not known. What we know are the estimates of these rewards. The estimates are known as action values.

Theoretically, the expected reward is given as:-

$$q_j^*(a) = E[R_t | A_t = a] \quad \forall a \in \{1, \dots, k\}$$

i.e. the expected reward is the summation of expected rewards received if an action a is taken at times $1 - \dots - k$.

This can be found by

$$q_j^*(a) = p(x|a) \cdot r$$

Again the goal is to maximize this expected reward which can also be represented

as

$\arg\max q^*(a)$ ie the selection
of actions yielding a more expected reward.

In practical terms, we can find this by doing an experiment multiple times and plotting the results or using other statistical methods to find the kind of distribution. Then just find the expected value for that distribution.

↓
mean of the distribution

This problem is called the problem of bandits.

It is a smaller problem under RL and is a good intro to the field of RL.

Method of Estimating $q^*(a)$ [Sample Average Method]

One method of estimating $q^*(a)$ is called the sample average method. In the sample average method, we estimate the action value using previous values.

$$q^*(a)_n = \frac{a_1 + a_2 + \dots + a_{n-1}}{n-1}$$

$$= \frac{\sum_{i=1}^{n-1} a_i}{n-1}$$

It's $n-1$ because the value is based on previous

It's $n-1$ because the value is based on previous values.

This method has a problem that in the above form, the time and space complexity grows by $O(n)$. We can make this better using previous values as follows :-

$$\begin{aligned}
 q_v^*(a)_{n+1} &= \frac{\sum_{i=1}^n a_i}{n} \\
 &= \frac{1}{n} \left[a_n + \sum_{i=1}^{n-1} a_i \right] \\
 &= \frac{1}{n} \left[a_n + \frac{(n-1)}{(n-1)} \sum_{i=1}^{n-1} a_i \right] \rightarrow q_v^*(a)_n \\
 &= \frac{1}{n} \left[a_n + (n-1) q_v^*(a)_n \right] \\
 &= \frac{1}{n} \left[a_n + n q_v^*(a)_n - q_v^*(a)_n \right]
 \end{aligned}$$

$$q_v^*(a)_{n+1} = q_v^*(a)_n + \frac{1}{n} [a_n - q_v^*(a)_n]$$

This is a very common form in RL. It can also be written as

New Estimate = Old Estimate + StepSize [Target - Old Estimate]
or New Estimate = Old Estimate + StepSize × Error

The $\frac{1}{n}$ term is called the step size.

The choice of $\frac{1}{n}$ is varied according to the problem.

If we do not have a stationary RL problem, i.e. the distribution of action values varies with time, we use a constant step size b/w $(0, 1]$.

What this does is that it weighs the newer values more in the calculation of action value estimates.

This can be seen as follows:-

constant step size

$$q_v^*(a)_{n+1} = q_v^*(a)_n + \alpha [a_n - q_v^*(a)_n]$$

$$\begin{aligned}
 &= q_v^*(a)_n + \alpha a_n - \alpha q_v^*(a)_n \\
 &= \alpha a_n + (1-\alpha) q_v^*(a)_n \quad \text{--- } ① \\
 &= \alpha a_n + (1-\alpha) [\alpha a_{n-1} + (1-\alpha) q_v^*(a)_{n-1}] \quad [\text{using } ①] \\
 &= \alpha a_n + (1-\alpha) \alpha a_{n-1} + (1-\alpha)^2 q_v^*(a)_{n-1} \\
 &= \alpha a_n + (1-\alpha) \alpha a_{n-1} + (1-\alpha)^2 [\alpha a_{n-2} + (1-\alpha) q_v^*(a)_{n-2}] \\
 &= \alpha a_n + (1-\alpha) \alpha a_{n-1} + (1-\alpha)^2 \alpha a_{n-2} + (1-\alpha)^3 q_v^*(a)_{n-2} \\
 &\quad \vdots \\
 &= \alpha a_n + (1-\alpha) \alpha a_{n-1} + (1-\alpha)^2 \alpha a_{n-2} + \dots + \\
 &\quad \dots + (1-\alpha)^{n-2} \alpha a_2 + (1-\alpha)^{n-1} \alpha a_1 + (1-\alpha)^n q_v^*(a), \\
 &= \boxed{\{(1-\alpha)^n q_v^*(a) + \sum_{i=1}^n \alpha (1-\alpha)^{n-i} a_i\}}
 \end{aligned}$$

$$= \boxed{(1-\alpha)^n q^*(a) + \sum_{i=1}^n [\alpha (1-\alpha)^{n-i} a_i]}$$

From this $= n$, we can see that as we go closer to the current value, the weightage goes higher.

This is because $(1-\alpha) < 1$ and greater the value of i , smaller is $n-i$ and higher will be $(1-\alpha)^{n-i}$.

We also see that $q^*(a)$, ie the initial estimate influences the estimate hugely.

This can be used to influence exploration or exploitation as will be seen later.