The k-bandit Problem

In the k-bandit problem, the decision maker (or agent) chooses between k actions and receives rewards based on the action they choose. The k-bandit problem involves decision making under the circumstance where the consequence for every action is unknown.

Every action returns a value which can be called an action value.

The value of selecting an action $q_*(a)$ is defined as an expected reward R_t when taking the action A_t for all actions a in the set of k actions.

$$\begin{aligned} q_*(a) &\doteq \mathbb{E}[R_t | A_t = a] \; \forall a \in \{1, 2, \dots, k\} \\ &= \sum_r p(r|a)\, r \end{aligned}$$

The objective of the equation is to find the highest expected action value

Learning Action Values

1. **Sample-average method**
   This method estimates the action value by taking the sum of rewards when taking an action divided for the number of time that action is taken

$$\begin{aligned} q_t(a) &\doteq \frac{sum\ of\ rewards\ when\ a\ taken\ prior\ to\ t}{total\ of\ times\ a\ taken\ prior\ to\ t} \\ &= \frac{\sum_{i=1}^{t-1} R_i}{t-1} \end{aligned}$$

2. **Incremental Update Rule**
   This instance is used under the circumstance where the number of data points is not constant. It can help update the action value without storing the data of previous records.

$$\begin{aligned} Q_{n+1} &= \frac{1}{n} \sum_{i=1}^{n} R_i \\ &= Q_n + \frac{1}{n}(R_n - Q_n) \end{aligned}$$

Exploration vs Exploitation

1. **Tradeoff**

   Exploration requires trials on various actions to gain more knowledge about the task.
   Exploitation applies the action which yields the best result for the next iterations.
   You cannot choose to do both jobs at the same time.

2. **Epsilon-Greedy Action selection**

   We can choose to exploit with a small chance of exploration. The following method is called Epsilon-Greedy action selection, with the epsilon ε stands for the probability of choosing exploration.

$$A_t \leftarrow \begin{cases} \underset{a}{argmax}\, Q_t(a) \ \ with\ the\ probability\ of\ 1-\varepsilon \\ a \sim Uniform(\{a_1 \ldots a_k\})\ with\ the\ probability\ of\ 1-\varepsilon \end{cases}$$

3. **Optimistic Initial Values**

   By setting high initial values for all actions equally, the agent can be encouraged to explore all the options in early learning and provide adjustments on the action values later.

   However, the drawbacks of this method are:

   - It only encourages early exploration. This means the agent will stop exploring after a period which is not well-suited for non-stationary problems.
   - There is no specific guideline to set optimistic initial values.

4. **Upper-Confidence Bound Action Selection**

   a. **Uncertainty in Estimates**

      Uncertainty in estimate refers to the degree of doubt or variability associated with a predicted value or outcome. In simple terms, it's an acknowledgment that no estimate is 100% accurate, and the actual result may differ from what's predicted.

      It can be represented by an estimate value point and the two confidence intervals: the upper bound and the lower bound. The larger the distance between the confidence intervals, the more uncertain the estimate value will be.

   b. **Optimism in the Face of Uncertainty**

      If we are uncertain about something, we should optimistically assume it is a good thing. In a set of actions, the action with the highest upper bound will always be picked up first regardless of the distance of the interval of that action.

We can use Upper-confidence Bound Action Selection by selecting the action with the highest action value plus our upper-confidence bound exploration term.

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