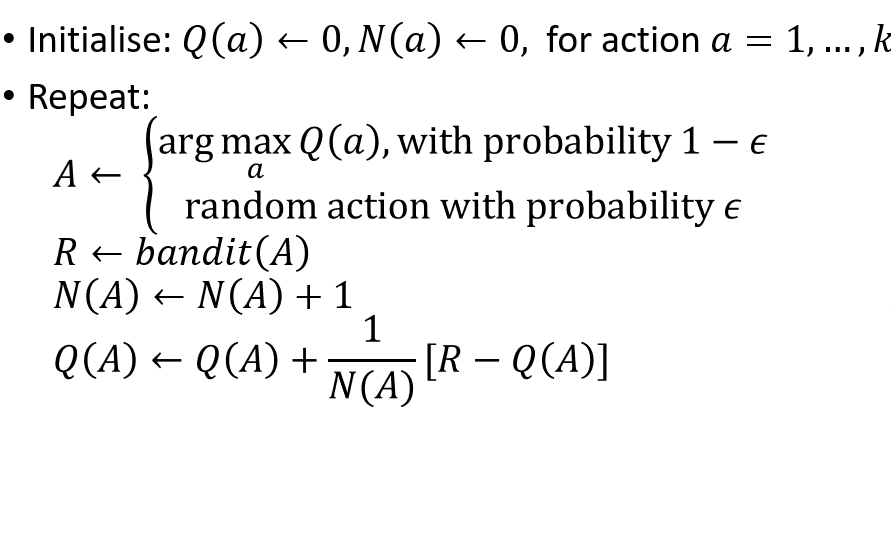
# Multi armed bandit problem

Read Chapter 2 of Introduction to Reinforcement Learning.

The algorithm for the n-bandit problem is the following (it exploits the incremental implementation provided in the book):



Given that the incremental implementation is used in the algorithm then we only need to store the sample average of each action a so far and its number of occurrences so far. The sample average of an action a so far is denoted as Q(a) while its number of occurrences so far is denoted as N(a).

In the initialisation step, the sample average of each action a and its number of occurrences is set to 0.

Subsequently, the core of the algorithm starts which is an infinite loop. At each iteration of this loop we may use a greedy action value approach or a ε-greedy action value approach. In the case of the pseudocode above, an ε-greedy action value approach is used as at each iteration we select the action with the highest sample average so far with probability (1-ε) and a random action with probability ε.

We perform the action resulting from the above steps and record the reward that we have obtained with that action. We increment the number of occurrences of the occurred action and subsequently we update its sample average.