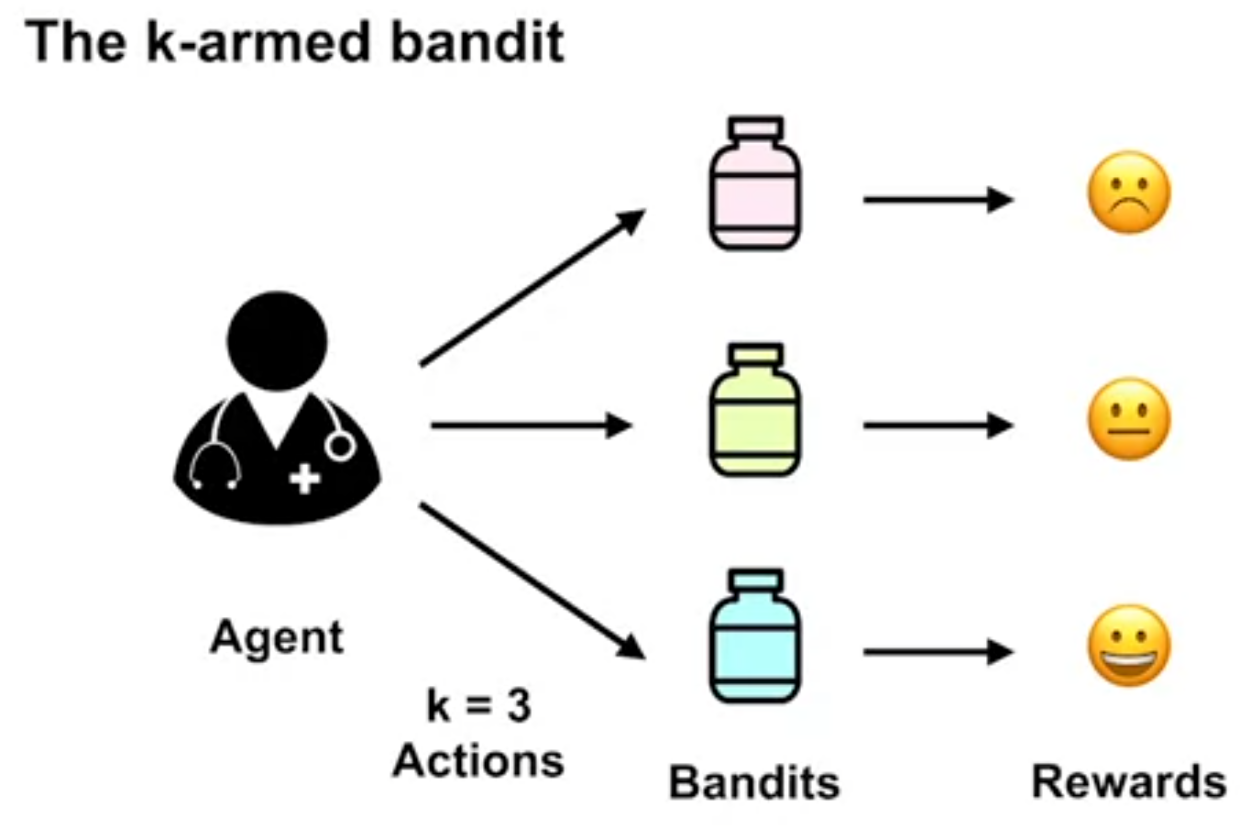
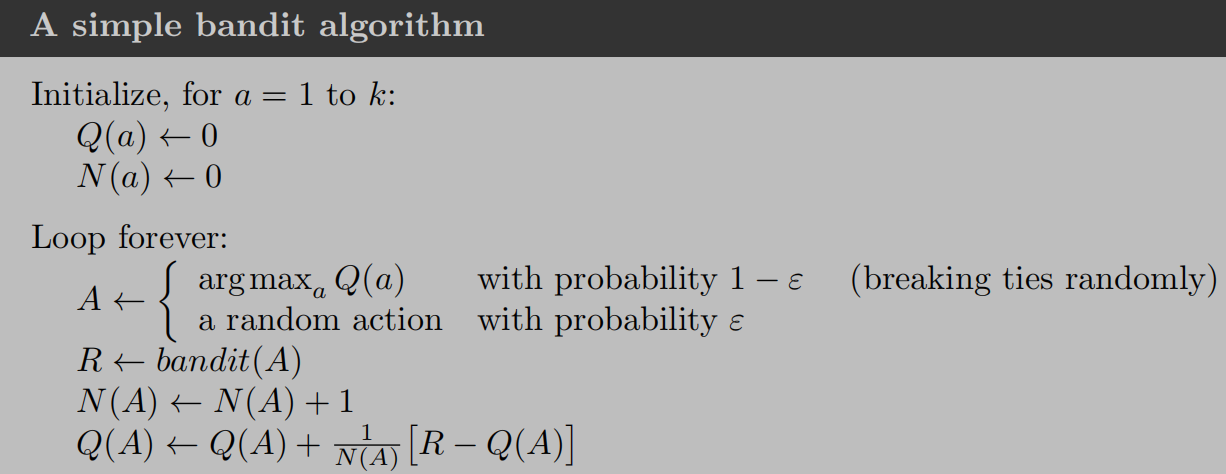
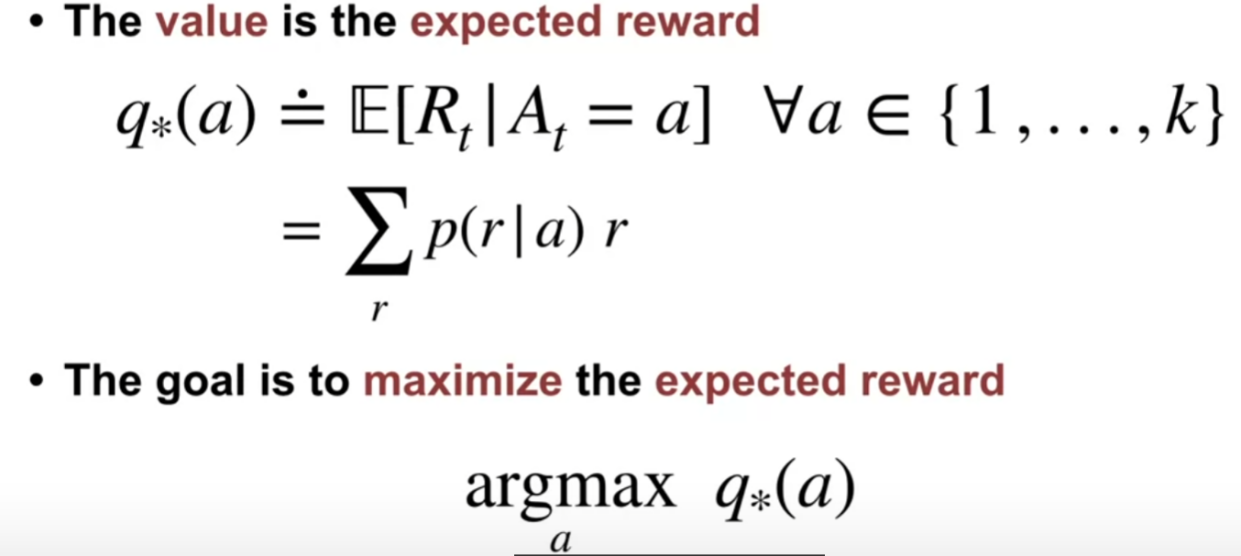
* K-armed bandit example





1/N(A) called the step size

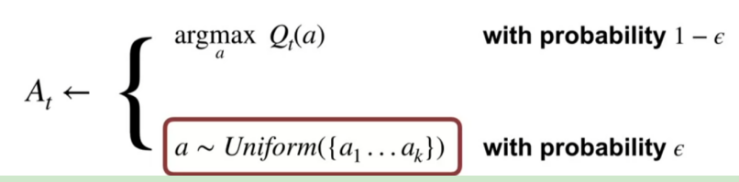
* Action values:



q star of a is defined as the expectation of reward received(R\_t), given we selected action a, for each possible action one through k. Inside the sum, we have multiplied the possible reward by the probability of observing that reward.

* Some methods of balancing exploiting vs exploration:

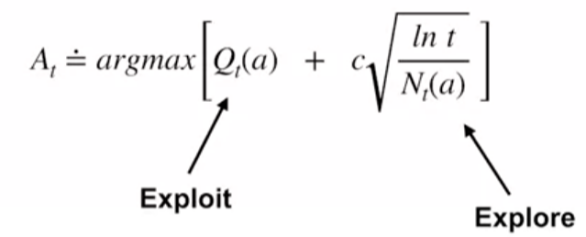
1. **Epsilon-greedy**: epsilon refers to the probability of choosing to explore



1. **Optimistic initial values:** optimistically initial Estimate value at the beginning, then the estimation value will decrease over the time.

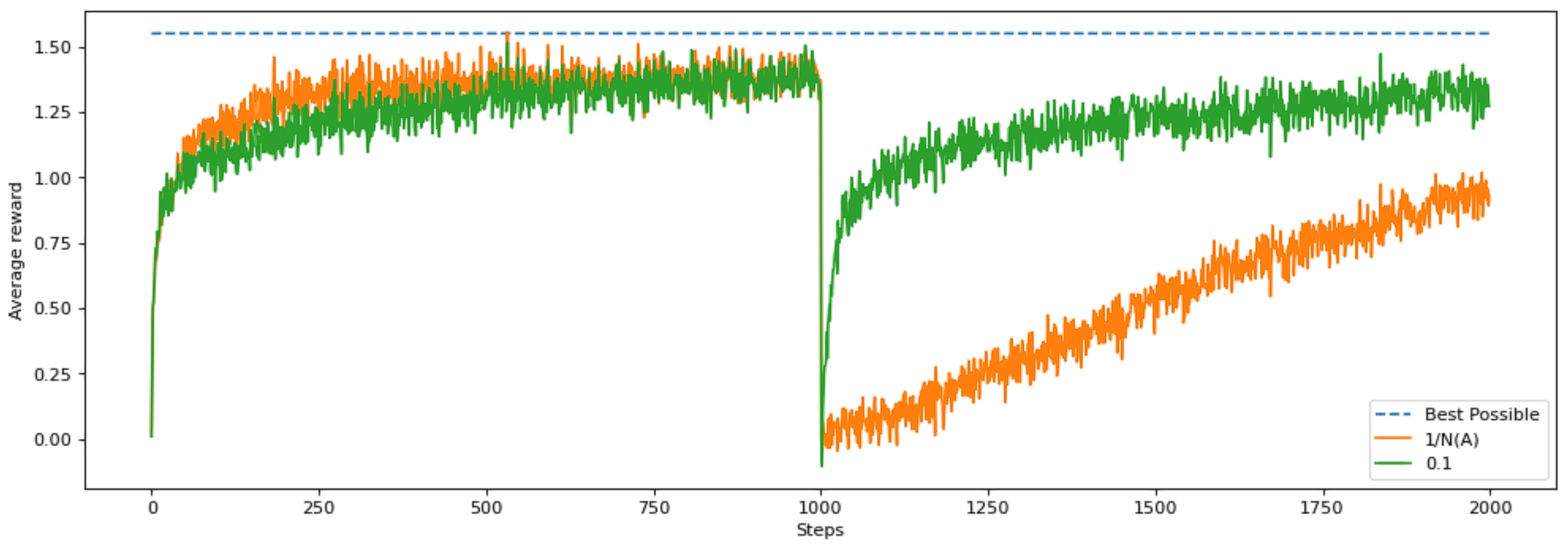
Cons: only drive exploration early in the learning, thus it is not suitable for non-stationary problem (An optimistic agent may have already settled on a particular action, and will not notice that a different action is better now)

1. **UCB (upper-confidence bound action selection):** instead of uniformly choosing an exploration action a(like epsilon-greedy), choose the action with higher upper bound of confident interval of Q\*(a)- expectation of reward received

****

* Non-stationary vs stationary multi-armed bandit problem: the reward distribution do not change over time

Ex: after 1000 steps we will randomly change the expected value of all of the arms



if the best action gets chosen 500 times. That means the step size for that action is 1/500 or 0.002. At each step when we update the value of the action and the value is going to move only 0.002 \* the error. That is a very tiny adjustment and it will take a long time for it to get to the true value.

The agent with step size 0.1, however, will always update in 1/10th of the direction of the error. This means that on average it will take ten steps for it to update its value to the sample mean.