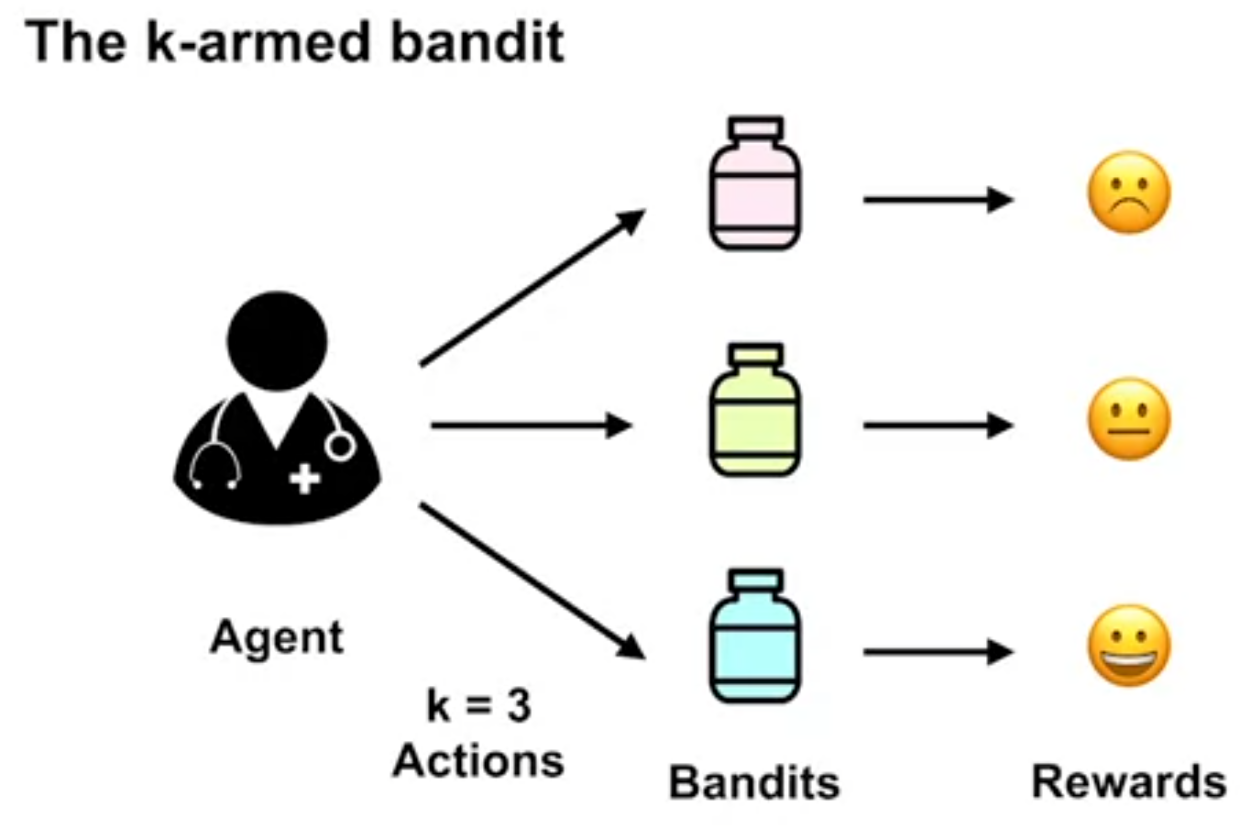
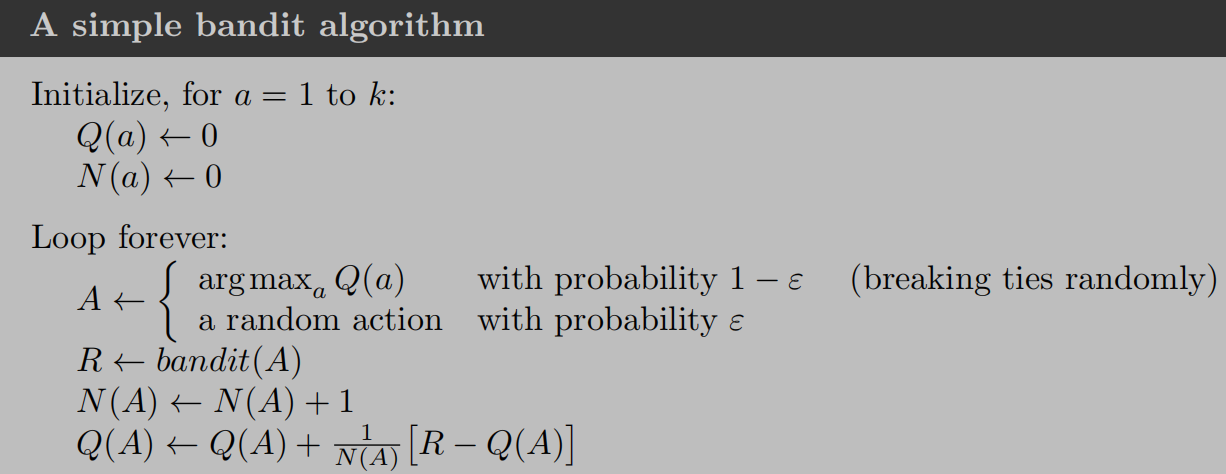
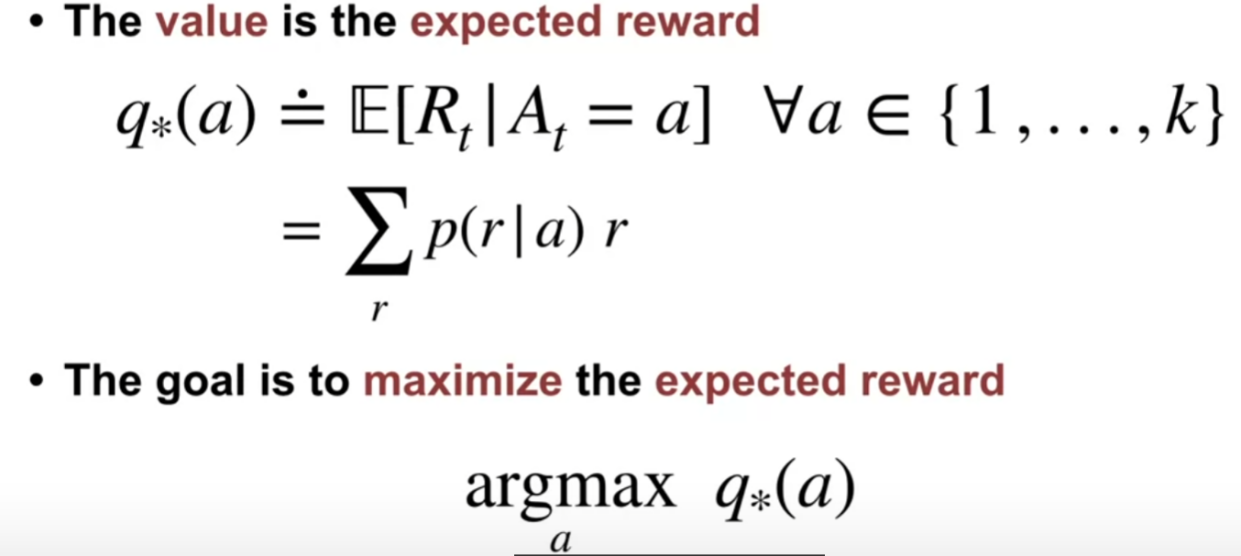
* K-armed bandit example





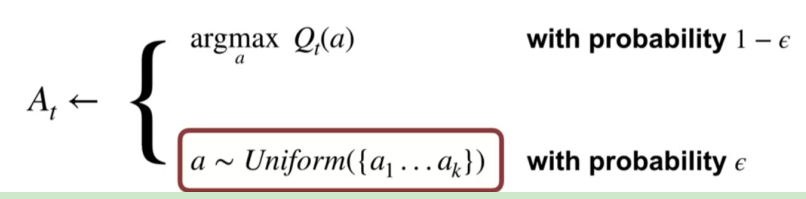
* Action values:



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:

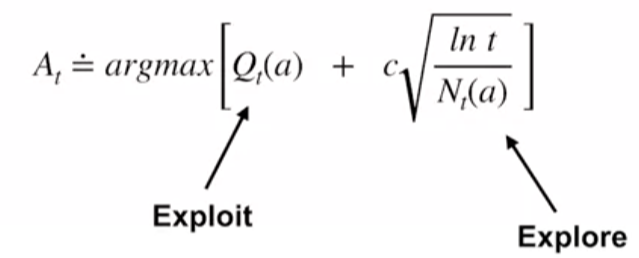
1. **Epsilon-greedy**: epsilon refers to the probability of choosing to explore



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 probabilities do not change over time