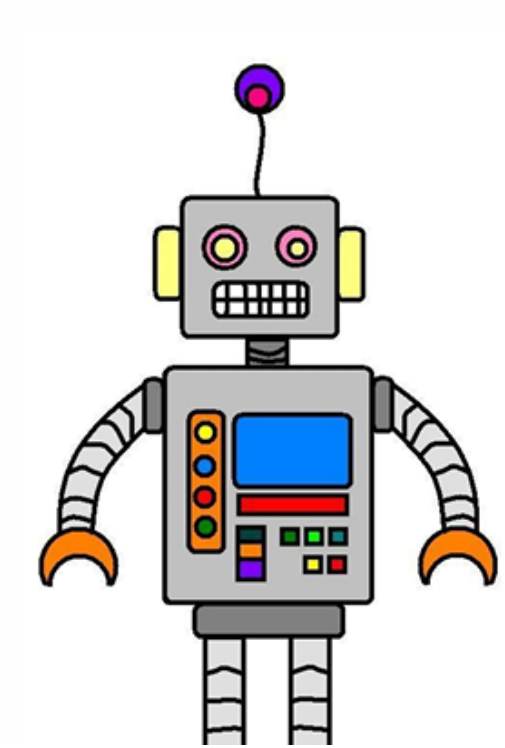
Reinforcement Learning

Fateme Taroodi

Exploration & Exploitation:

- **Exploration:** Exploration involves trying out different actions that may not necessarily lead to the best-known rewards but could provide valuable information for the agent's learning process.
  Goal: The goal of exploration is to gather more knowledge about the environment, especially about areas the agent hasn't visited much yet, to improve its long-term decision-making.
  Behavior: The agent takes actions that may not seem optimal based on past experiences to explore unknown parts of the state-action space.

- **Exploitation:** Exploitation involves using the knowledge the agent has gained so far to make the best decision in order to maximize rewards.
  Goal: The goal of exploitation is to take the action that is expected to yield the highest reward based on the agent's current knowledge, which is typically the action that maximizes the expected Q-value.
  Behavior: The agent sticks to actions that have already proven to be successful in the past.

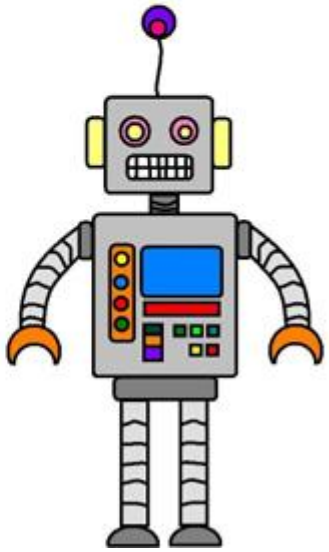


k-armed Bandit Problem

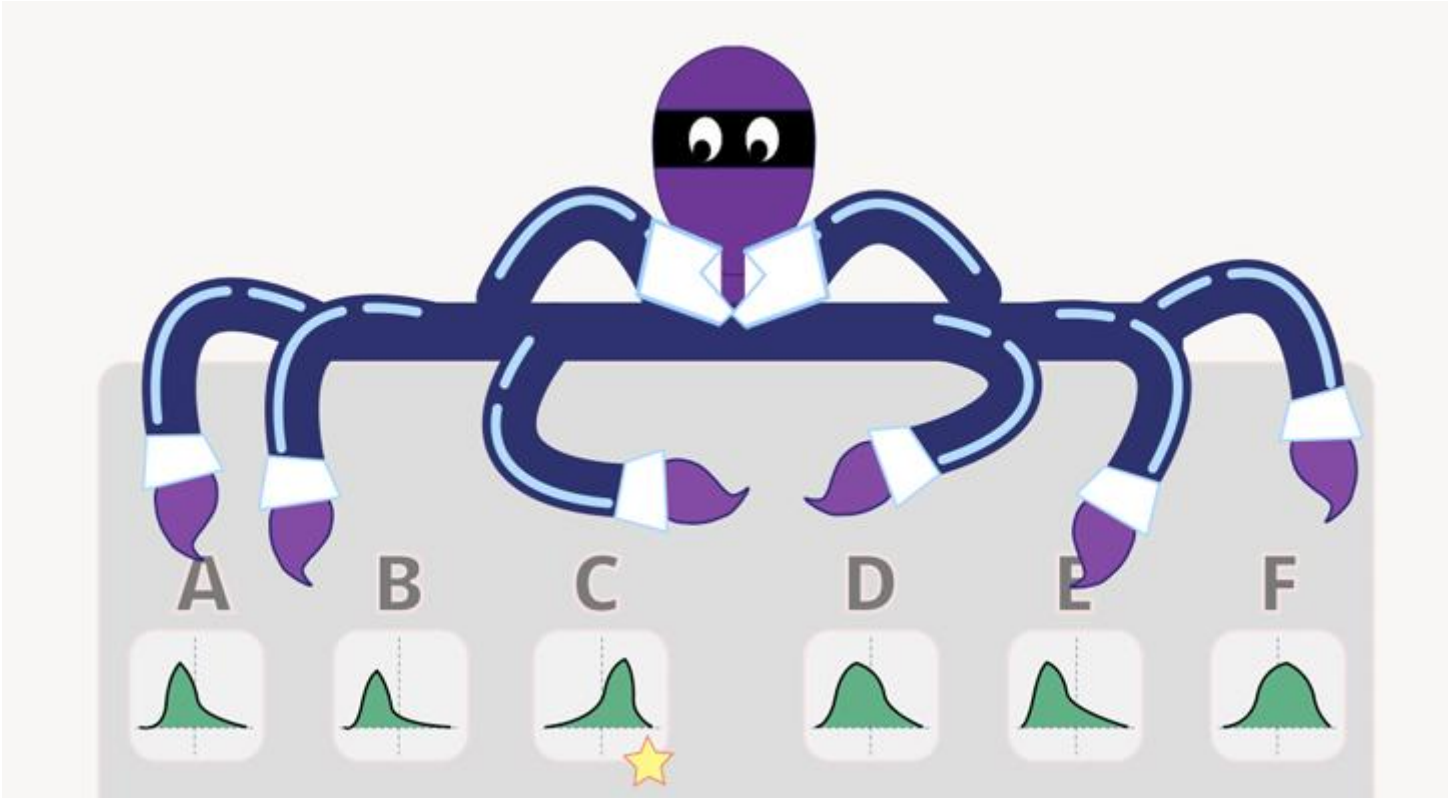
Fateme Taroodi

6 Machines

1000 tries



Fateme Taroodi

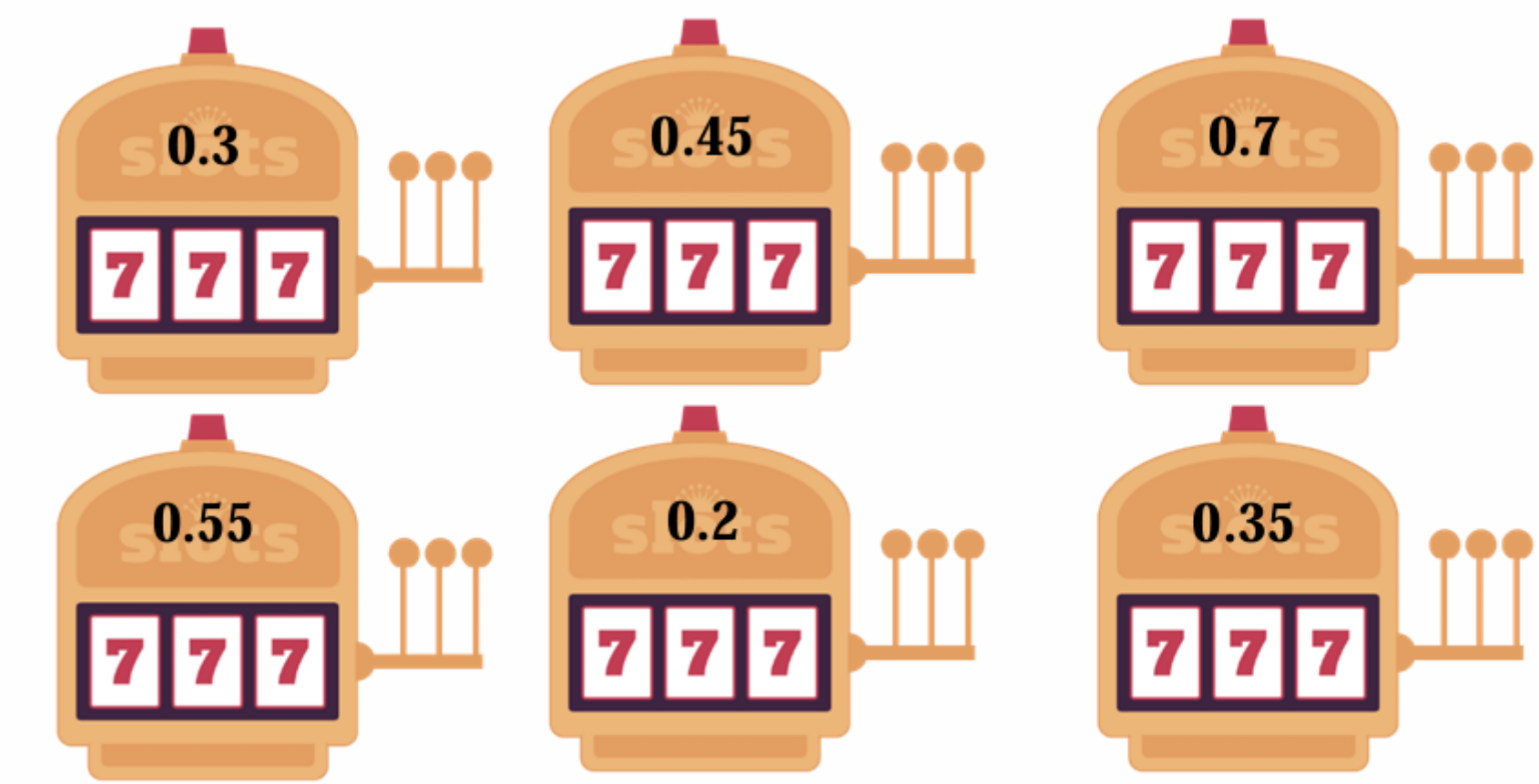


Fateme Taroodi

sample-average

$$Q_t(a) \doteq \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t} = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

$$q_*(a) \doteq \mathbb{E}[R_t \mid A_t = a]$$



Fateme Taroodi

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

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

$$
\begin{aligned}
Q_{n+1} &= \frac{1}{n}\Big(R_n + (n-1)Q_n\Big) \\
&= \frac{1}{n}\Big(R_n + nQ_n - Q_n\Big) \\
&= Q_n + \frac{1}{n}\Big[R_n - Q_n\Big],
\end{aligned}
$$

$$
NewEstimate \leftarrow OldEstimate + StepSize \Big[Target - OldEstimate\Big]
$$

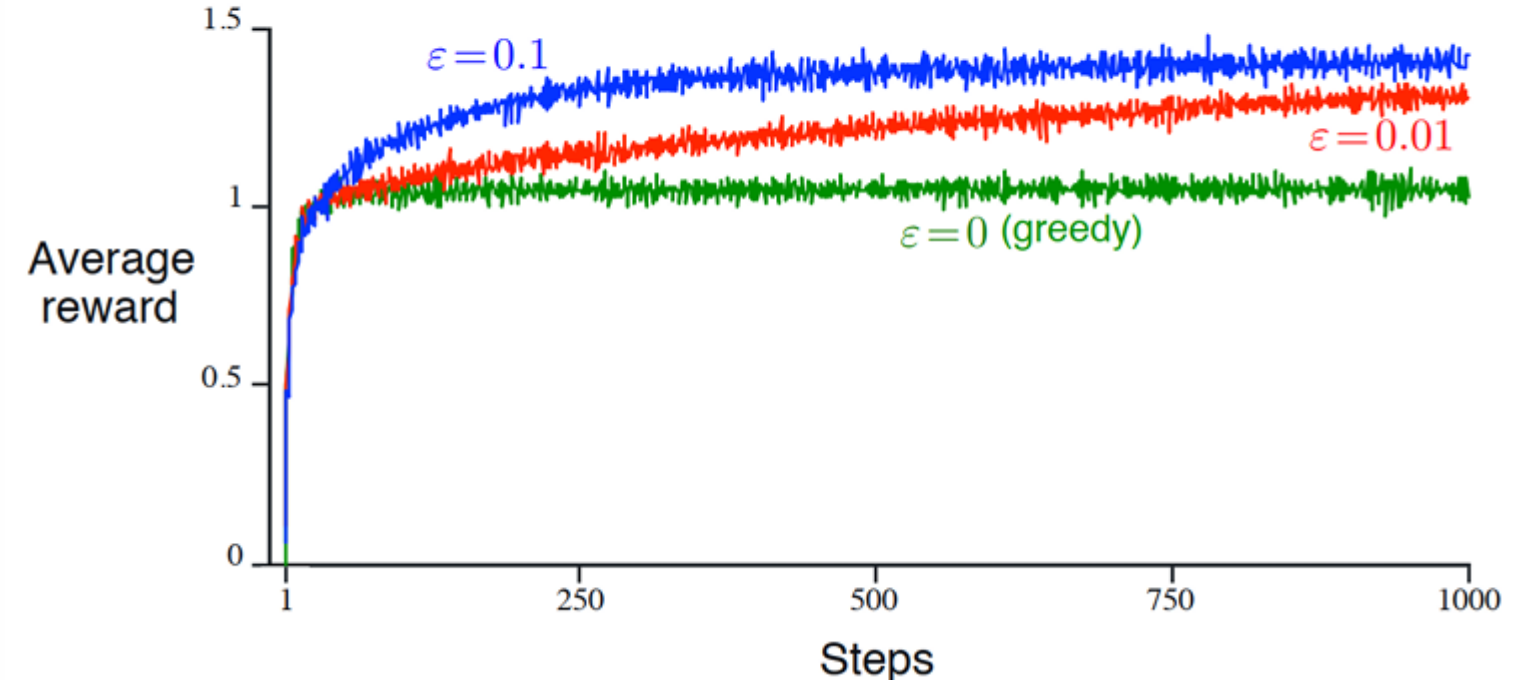
?

Strategies Incorporating Exploration and Exploitation:

- **Epsilon-Greedy Strategy**

$$\text{Action} = \begin{cases} \text{argmax}_a Q(s,a) & \text{with probability } 1-\epsilon \\ \text{Random action} & \text{with probability } \epsilon \end{cases}$$

ε-greedy



A simple bandit algorithm

Initialize, for $a = 1$ to k:

$Q(a) \leftarrow 0$

$N(a) \leftarrow 0$

Loop forever:

$$A \leftarrow \begin{cases} \operatorname{argmax}_a Q(a) & \text{with probability } 1-\varepsilon \quad \text{(breaking ties randomly)} \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$$

$R \leftarrow bandit(A)$

$N(A) \leftarrow N(A) + 1$

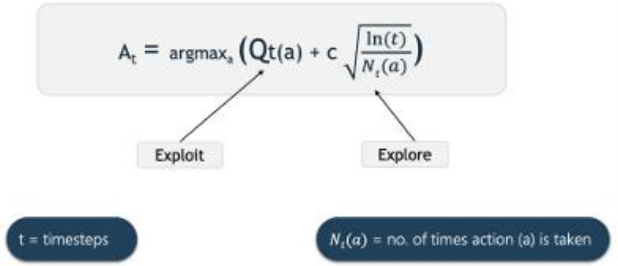
$Q(A) \leftarrow Q(A) + \frac{1}{N(A)}\left[R - Q(A)\right]$

Upper confidence bound

Fateme Taroodi

- **Upper Confidence Bound (UCB):**

The **Upper Confidence Bound (UCB)** method is a strategy used for balancing **exploration** and **exploitation** in decision-making problems, especially in contexts like the **multi-armed bandit problem** and other RL settings.

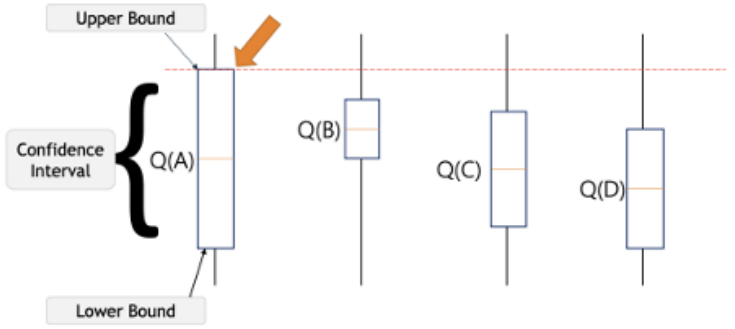


• N(s_t,a): The number of times action a has been selected in state s_t (or the count of how many times the agent has explored this action) before time t.

The formula consists of two terms:

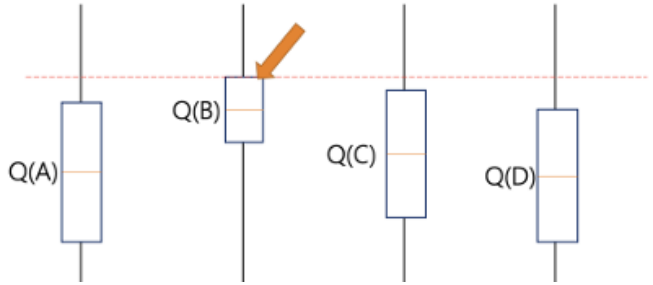
1.Exploitative term: Q_t(a), which represents the estimated Q-value for an action. This term encourages the agent to select actions that have already shown higher Q-values based on previous observations. This helps the agent to exploit what it already knows is working well.

2.Explorative term: $c\sqrt{\frac{\ln(t)}{N_t(a)}}$, which represents the **confidence interval** around the estimated Q-value, increasing as the action a is explored less frequently. This term is an **exploration bonus**. The more an action has been tried (i.e., the larger $N_t(a)$ is), the smaller this whole term will be. Conversely, if an action has been tried fewer times, the bonus increases, encouraging the agent to explore this action more



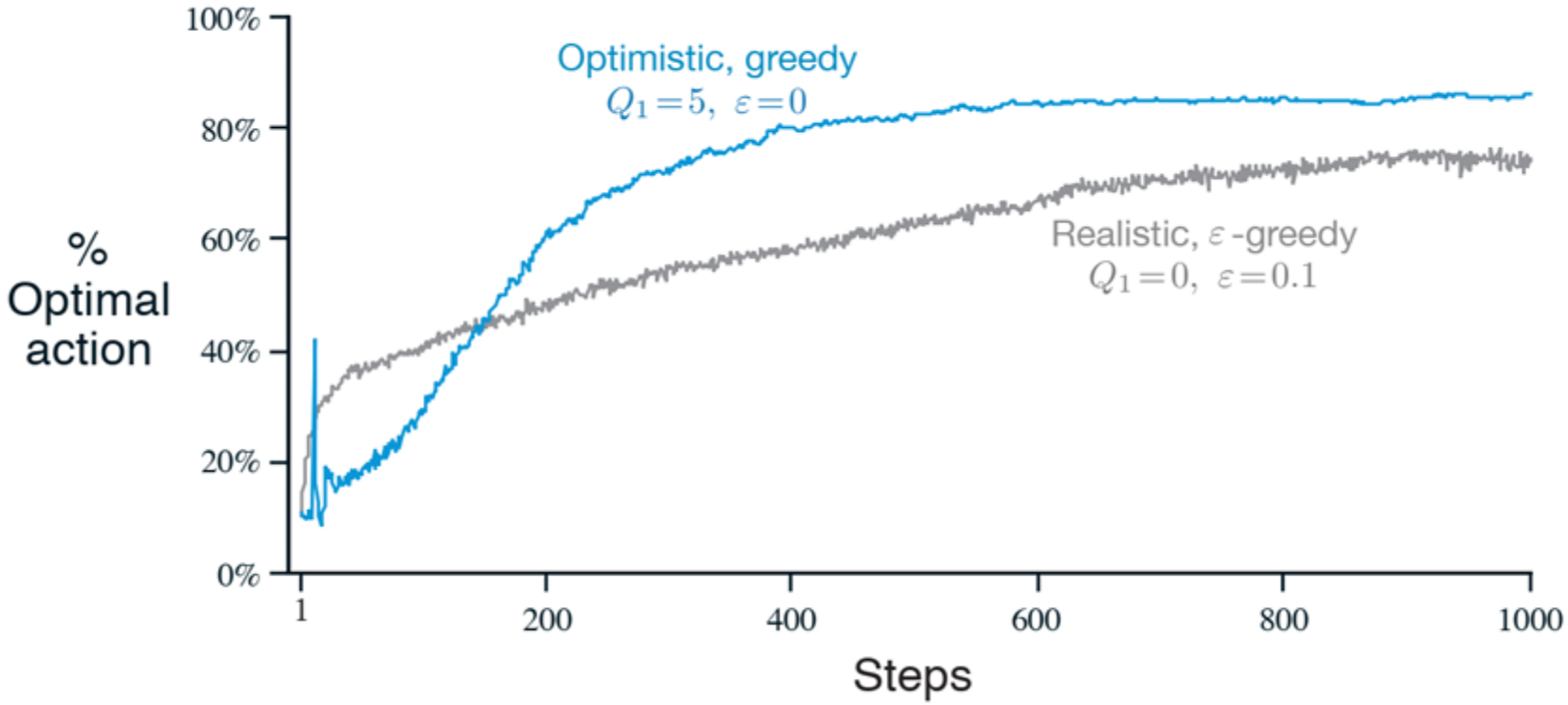
Q(A) in the above picture represents the current action-value estimate for action *A*. The brackets represent a confidence interval around *Q(A)* which says that we are confident that the actual action-value of action *A* lies somewhere in this region.
According to the UCB algorithm, it will optimistically pick the action that has the highest upper bound i.e. *A*. By doing this either it will have the highest value and get the highest reward, or by taking that we will get to learn about an action we know least about.

After each round the confidence bound will be shorten for the machine which is selected thereby avoiding biasness towards one machine thus giving chance to others.



Optimistic, greedy

Fateme Taroodi


100%
80%
60%
40%
20%
0%
% Optimal action
Optimistic, greedy
$Q_1 = 5,\ \varepsilon = 0$
Realistic, ε-greedy
$Q_1 = 0,\ \varepsilon = 0.1$
1
200
400
600
800
1000
Steps

A simple bandit algorithm

Initialize, for $a = 1$ to k:

$Q(a) \leftarrow 0$ ✖ 100

$N(a) \leftarrow 0$

Loop forever:

$$A \leftarrow \begin{cases} \arg\max_a Q(a) & \text{with probability } 1-\varepsilon \quad \text{(breaking ties randomly)} \\ \text{a random action} & \text{with probability } \varepsilon = 0 \end{cases}$$

$R \leftarrow bandit(A)$

$N(A) \leftarrow N(A) + 1$

$Q(A) \leftarrow Q(A) + \frac{1}{N(A)}\left[R - Q(A)\right]$