HSLU Lucerne University of Applied Sciences and Arts

Multiarmed Bandits

Reinforcement Learning
September 22, 2022

Learning Objectives

- Define multi-armed bandits as an RL problem
- Understand the meaning of value and policy functions
- Understand exploration and why RL algorithms need it
- Know epsilon and epsilon greedy policies
- Define the update of the value function from experience

Multi-Armed Bandit Problem

- k Slot Machines with reward distributed by (different) probability functions
- 1000 coins to play
- Goal: Maximise the total reward
- What is the best **policy** to play?

Multi-Armed Bandit

How would **you** play?

We want to teach an agent to play a multi-armed bandit:

- What are the possible actions?
- How does the value function look like?

Actions in Reinforcement Learning

- An agent **evaluates** actions in RL
- An agent **does not instruct** the correct actions
- An agent employs **active exploration** to search for good (or the best) behavior

(this refers to training, a trained agent will follow the learned (optimal) policy)

Formulation of the problem

- The actual value of an action a is the expected reward

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

(which is not known)

- The estimated value of an action a is called the (action-) value function

$$Q_t(a)$$

which we would like to be close to the true value

Action-Value Methods

A simple method to estimate the action values is to average the rewards whenever the action was taken

$$Q_n \doteq \frac{R_1 + R_2 + \cdots + R_{n-1}}{n-1}$$

(however, we need to keep all the rewards)

Incremental calculation

We would prefer an incremental calculation instead

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

General Update Formula

$$Q_{n+1} = Q_n + \frac{1}{n}(R_n - Q_n)$$

The last line in the previous equation can be written as

$$\boxed{\mathrm{NewEstimate} \leftarrow \mathrm{OldEstimate} + \mathrm{StepSize}\ [\mathrm{Target} - \mathrm{OldEstimate}]}$$

The term [Target – OldEstimate] is an *error* that we want to reduce by taking a step towards the Target

Many RL algorithm use this formula with different values of the error and the StepSize

Exploitation and Exploration

Exploitation:

- Exploit current knowledge by taking the action with the maximal estimated value

- Greedy action

Exploration:

- Explore the value of other actions to get better estimates

- Non greedy actions

Epsilon Greedy Methods

Exploitation:

With probability 1-ε:

- Take action with maximal $Q_t(a)$ (greedy action)

Exploration:

With probability ε:

- Take any valid action with equal probability

Implementation:

- Draw random variable in [0..1]
- Compare with threshold ε

Simple Multi Armed Bandit Agent

A simple bandit algorithm

Initialize, for $a = 1$ to k:

$Q(a) \leftarrow 0$

$N(a) \leftarrow 0$

Loop forever:

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

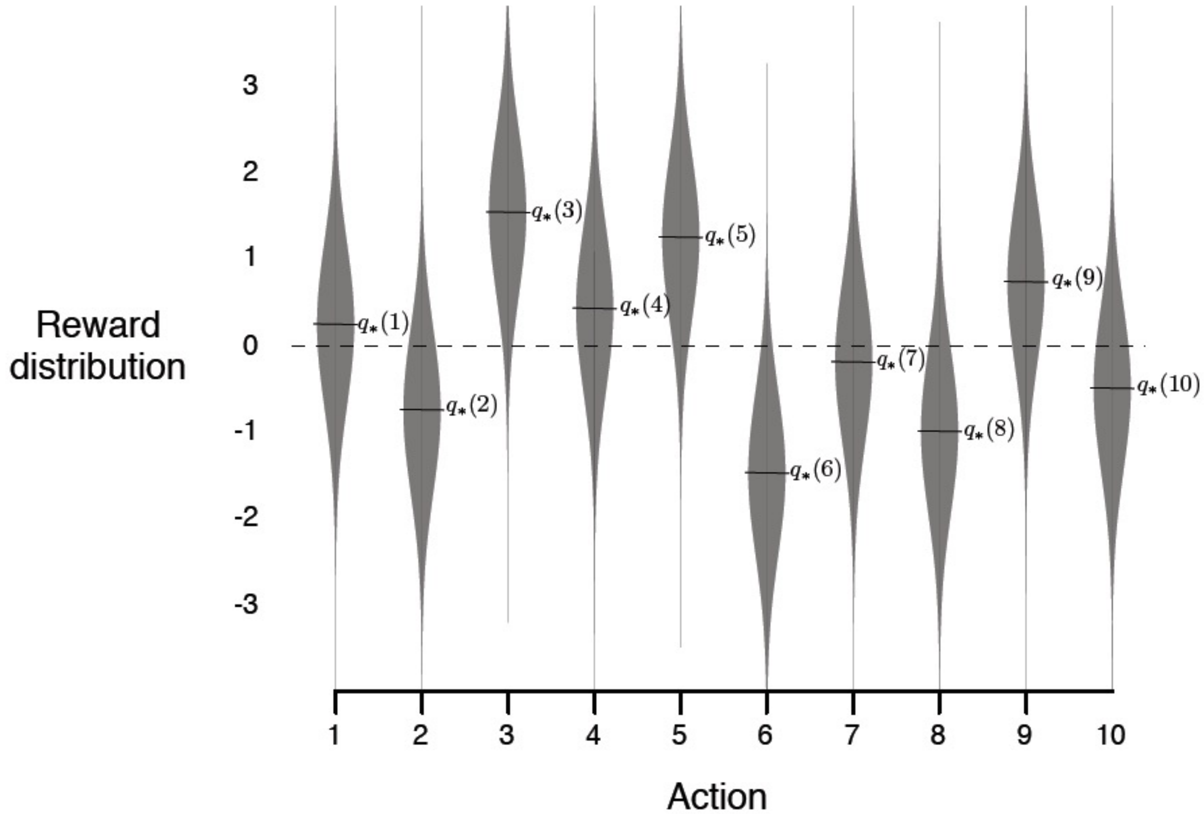
$R \leftarrow \text{bandit}(A)$

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

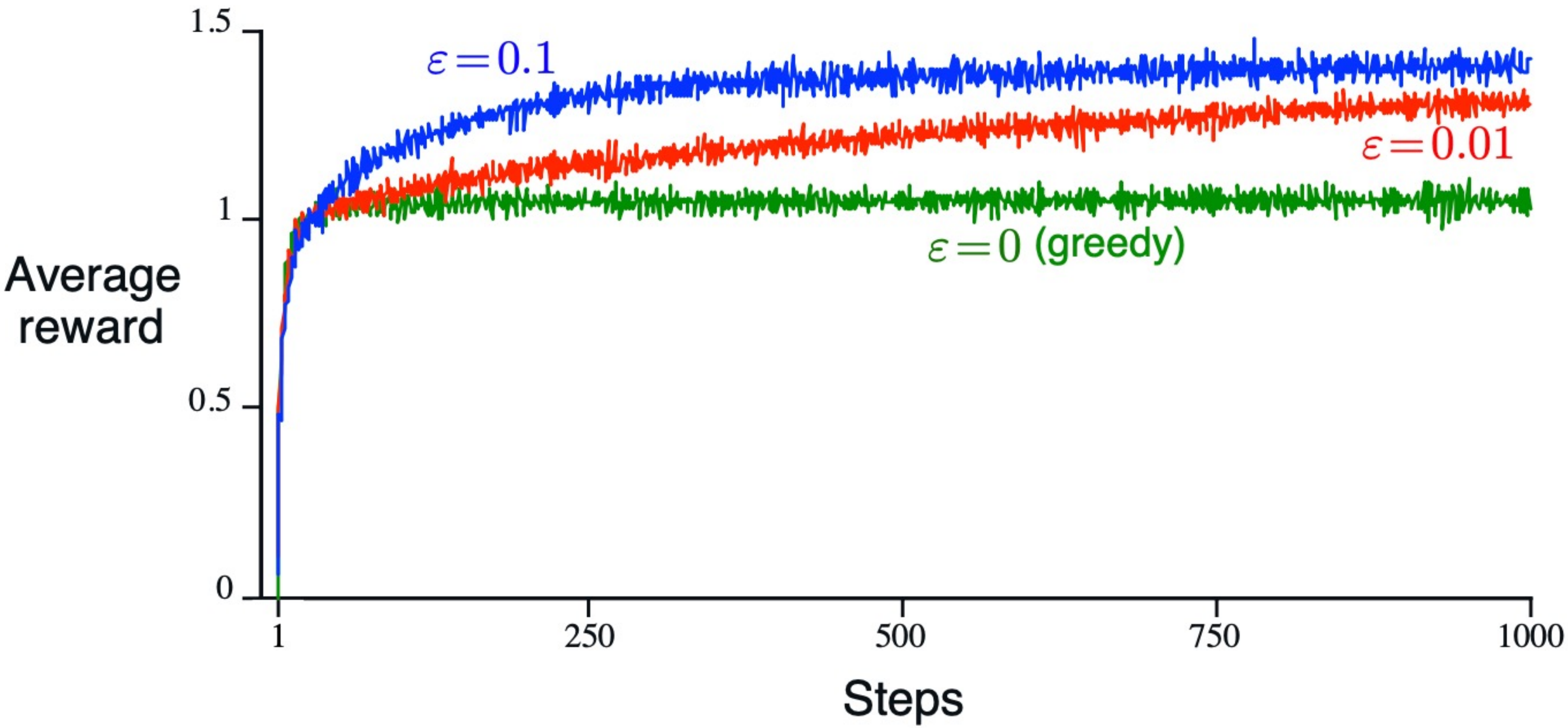
$Q(A) \leftarrow Q(A) + \frac{1}{N(A)}\left[R - Q(A)\right]$

Testbed 10-armed bandits

- 10 bandits with different mean reward (drawn from Normal probability distribution with mean 0)
- Return reward with Normal distribution (sigma=1.0) around mean value



Comparison of exploration



Upper Confidence Bound

- Epsilon-greedy methods are not selecting the most promising methods during exploration, and
- Epsilon-greedy methods are not efficient once the best method has been found
- A better method is the upper confidence bound (UCB) algorithm that includes a term to measure the uncertainty in the estimate

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

Number of times that this action has been selected previously

Upper Confidence Bound