Multi-armed bandits

Mohit Pandey
mpandey@bu.edu

Chapter 2 of Sutton and Barto

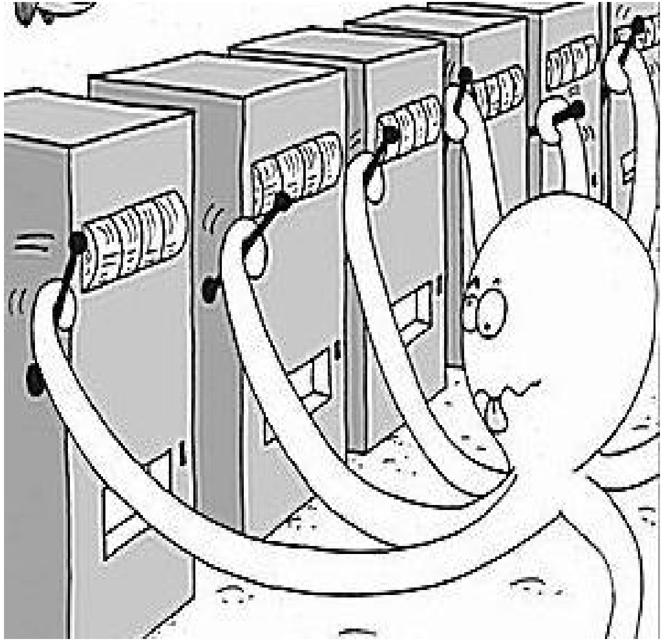
What is Multi-Armed Bandit?

One-Armed Bandit =
Slot Machine

source: walmart.com

What is Multi-Armed Bandit?

1. Multi-Armed Bandit = Multiple Slot Machine= more than one possible actions at each step
2. Objective: **maximize reward** in a casino



source: Microsoft Research

Problem Setting

- # of actions K , # of time steps T
- For each time step $t = 1, \ldots, T$
  - the reward vector $r_t = (r_{1,t}, \ldots, r_{K,t})$ is generated
  - the agent chooses an action $\mathbf{a}_t \in \{1, \ldots, K\}$
  - the agent receives the reward $r(a_t)$
- Remark: rewards of unchosen actions are not revealed

Problem Setting

- # of actions K , # of time steps T
- For each time step $t = 1, \ldots, T$
  - $\boxed{\text{the reward vector } r_t = (r_{1,t}, \ldots, r_{K,t}) \text{ is generated}}$
  - the agent chooses an action $\mathbf{a}_t \in \{1, \ldots, K\}$
  - the agent receives the reward $r(a_t)$
- Remark: rewards of unchosen actions are not revealed

Which Action to Choose?

- Value of an action: Expected reward *after* that action is chosen
- $q_*(a) = Exp(R_t | A_t = a)$
- We don't know $q_*(a)$
- Estimate of value of action $Q_t(a)$

$$Q_t(a) = \frac{\text{sum of rewards when } a \text{ was taken prior to t}}{\text{Number of times } a \text{ was taken prior to t}}$$

Which Action to Choose?

$$Q_t(a) = \frac{\text{sum of rewards when } a \text{ was taken prior to t}}{\text{Number of times } a \text{ was taken prior to t}}$$

For example, if an action has been taken n-1 times, then estimate of the value of that action will be:

$$Q_n(a) = \frac{R_1 + R_2 + \ldots + R_{n-1}}{n-1}$$

Example: K=4 bandit problem

Exercise 2.2

- Possible Actions: left (A=1), right (A=2), up (A=3), down (A=4)
- Let's suppose $Q_1(a)=0$ for all actions a
- Let's suppose this is how three steps are taken with their corresponding rewards:
    - $A_1=1, R_1=4$
    - $A_2=3$, $R_2=-1$
    - $A_3=1$, $R_3=2$
- Q function for each step:
    - $Q_1(a)=0$ for all actions a
    - $Q_2(A=1)=4$ and $Q_2(A=a)=0$ for all other actions
    - $Q_3(A=1)=4$, $Q_3(A=2)=3$ and $Q_3(A=a)=0$ for all other actions

Which Action to Choose?

- Let's try greedy action: $A_t = \arg\max Q_t(a)$
  - Maximize immediate reward by exploiting present knowledge
  - What if there is a better action which is unexplored?
- We need to also explore

Exploration-Exploitation Dilema

Fundamental question of RL: How to balance between exploration and exploitation?

source: RSM Discovery

Exploration-Exploitation Dilemma

- *Example*: finding the best restaurant in town:
  - Exploitation: keep going to your favorite restaurant
  - Exploration: taking the risk of trying a new one
- Two Naïve Algorithms
  - Random (= full exploration): choose action randomly
  - Greedy (= full exploitation): choose the best action according to your present knowledge

METHODS

- Greedy method and ε greedy method
- Optimistic Initial values
- Upper confidence bound action selection
- Gradient bandit algorithms

METHODS

- $\boxed{\text{Greedy method and } \varepsilon \text{ greedy method}}$
- Optimistic Initial values
- Upper confidence bound action selection
- Gradient bandit algorithms

Greedy & ε- greedy methods

- Greedy method
    - Always choose an action whose Q(a) is maximum
- ε-greedy method
    - Exploration: with probability $\mathcal{E}$, choose action randomly
    - Exploitation: with probability 1-$\mathcal{E}$, be greedy

Greedy & ε- greedy methods

A simple bandit algorithm

Initialize, for $a = 1$ to k:

$Q(a) \leftarrow 0$

$N(a) \leftarrow 0$

Loop forever:

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

$R \leftarrow bandit(A)$

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

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

Greedy & ε- greedy methods

A simple bandit algorithm

Initialize, for $a = 1$ to k:

$Q(a) \leftarrow 0$

$N(a) \leftarrow 0$

Loop forever:

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

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

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

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

How to determine rewards?

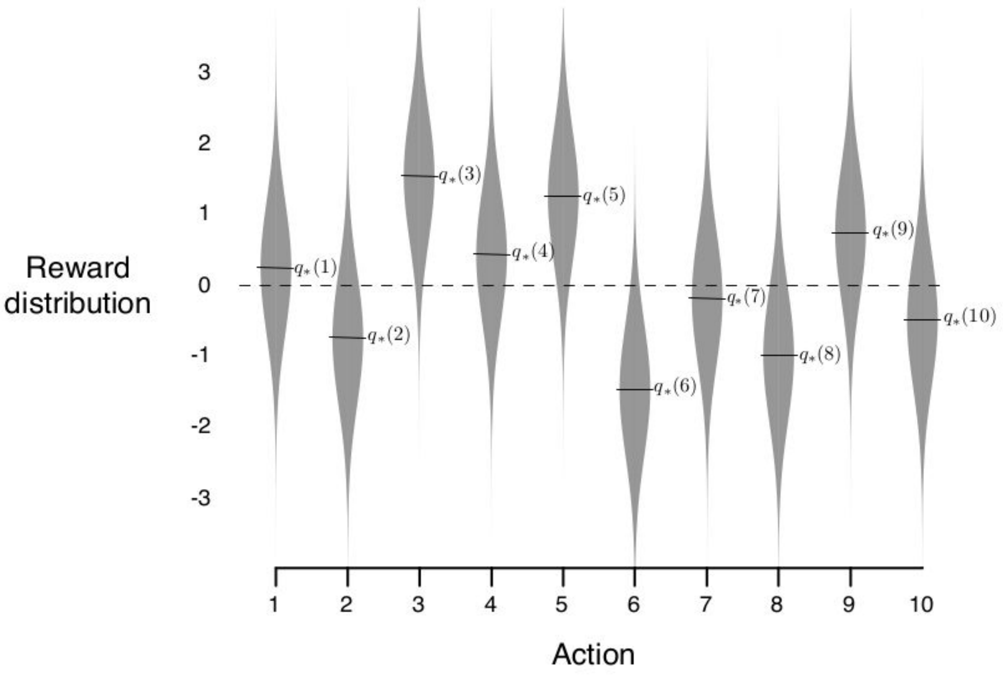
10-armed bandit

Choose a bandit problem → For each step, choose action so as to maximize the total expected reward $q_*(a)$

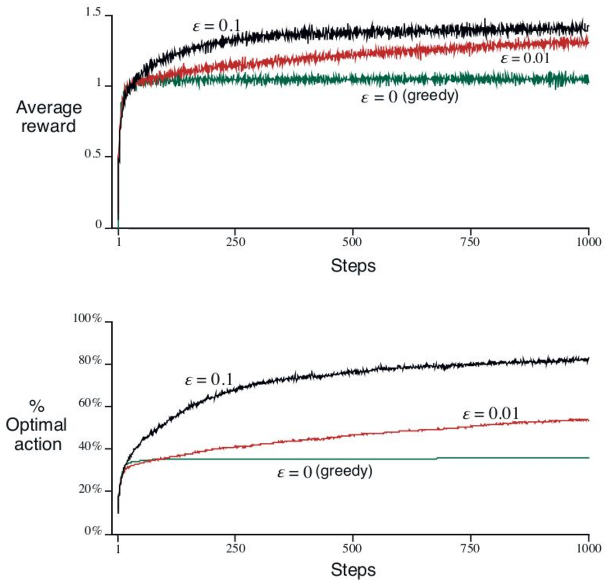
1. **Step 1:** For each bandit problem, action values $q_*(a)$ are chosen from normal distribution with mean=0 and variance=1
2. **Step 2:** Once the problem is determined, learning method chooses an action, whose R_t is selected from normal distribution with mean=$q_*(a)$ and variance=1

How to determine rewards?

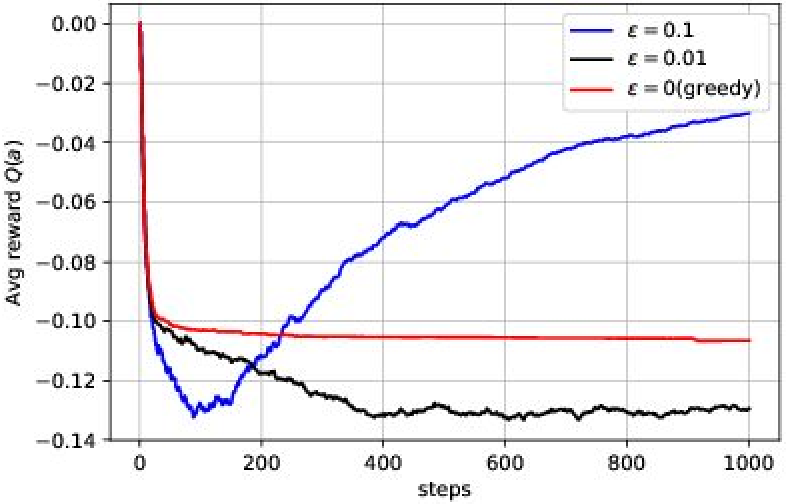
10-armed bandit



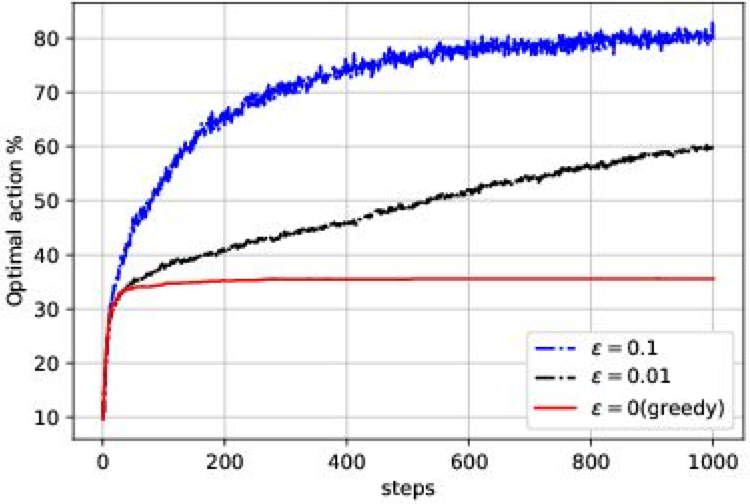
Results: Average rewards & Optimal action



My Results: Average rewards



My Results: Optimal action



Thank you!

mpandey@bu.edu

https://github.com/mohitpandey92/counterdiabatic-driving/blob/master/papers/machine%20learning/jupyter_code/sutton_book_ex/k_arm_bandit.ipynb