Exploration vs. Exploitation

Objectives

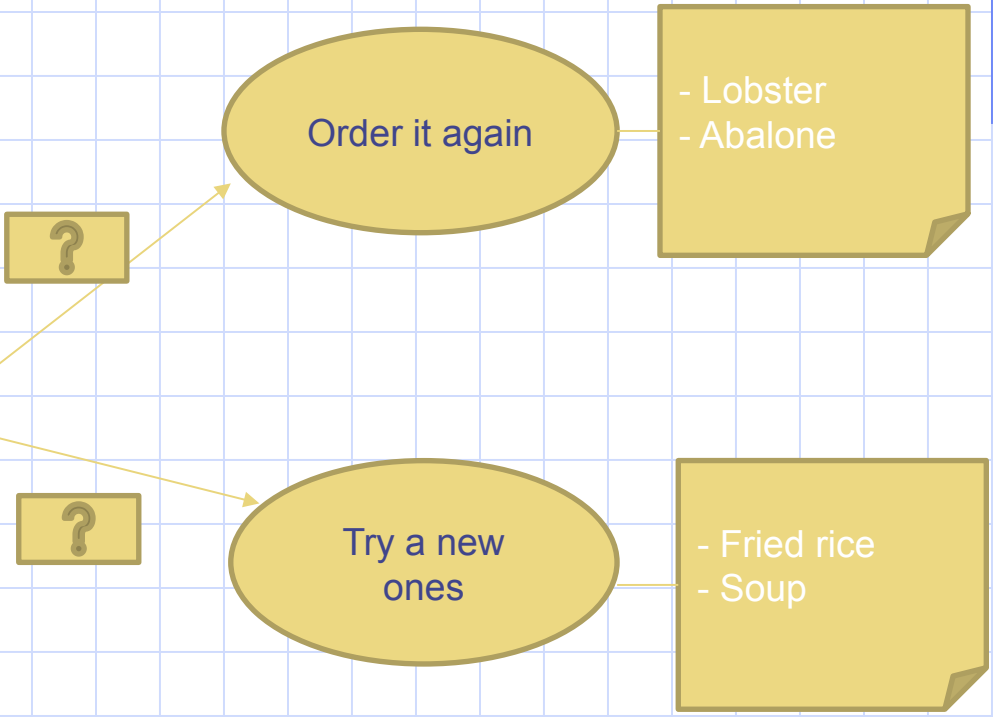
- Understand exploration and exploitation tradeoff
- Understand the method for balance the exploration and exploitation

Exploration and Exploitation

- Discuss example:
  - Mr. Manh is going out to eat with his lecturer Mr HoaDNT. When he get to the restaurant, he will have to decide what to order. He has been there a few times before and he always **order the same thing**. So he know that you will **be quite happy if he order it again**. Many of the **other items though look really tasty**.
  - How does he decide when to order the same good meal again, or try something new?

Exploration and Exploitation

- Discuss example:



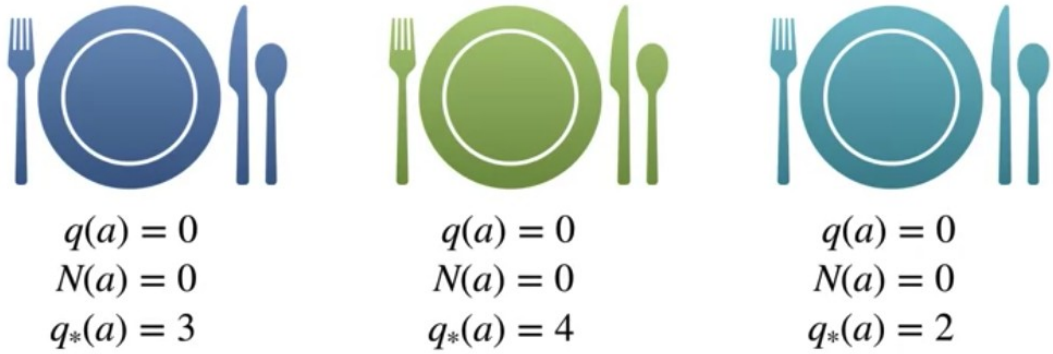
- What is Exploration and Exploitation? → discuss

Exploration and Exploitation

- Exploration:
  - Exploration involves trying out new or less familiar options to gain more information about the environment or task.
  - This is necessary to discover potentially better strategies or actions that may lead to higher rewards in the long run.
  - Exploration is crucial for learning and discovering optimal solutions.
  - Improve knowledge for long-term benefit

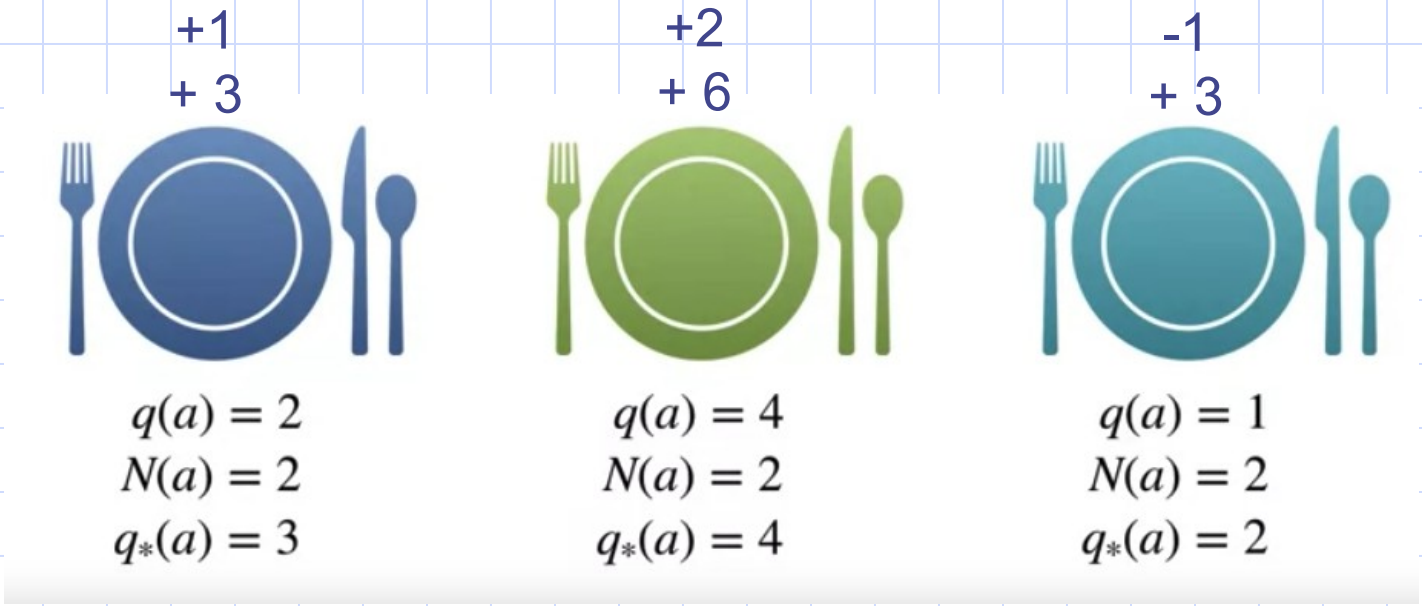
Exploration and Exploitation

- Example of exploratory behavior
  - Each of these plates represents a meal at your favorite restaurant, and you're trying to choose which meal to order.
  - Q of a is the estimated value for picking that meal.
  - N of a is the number of times you have picked that meal,
    q s



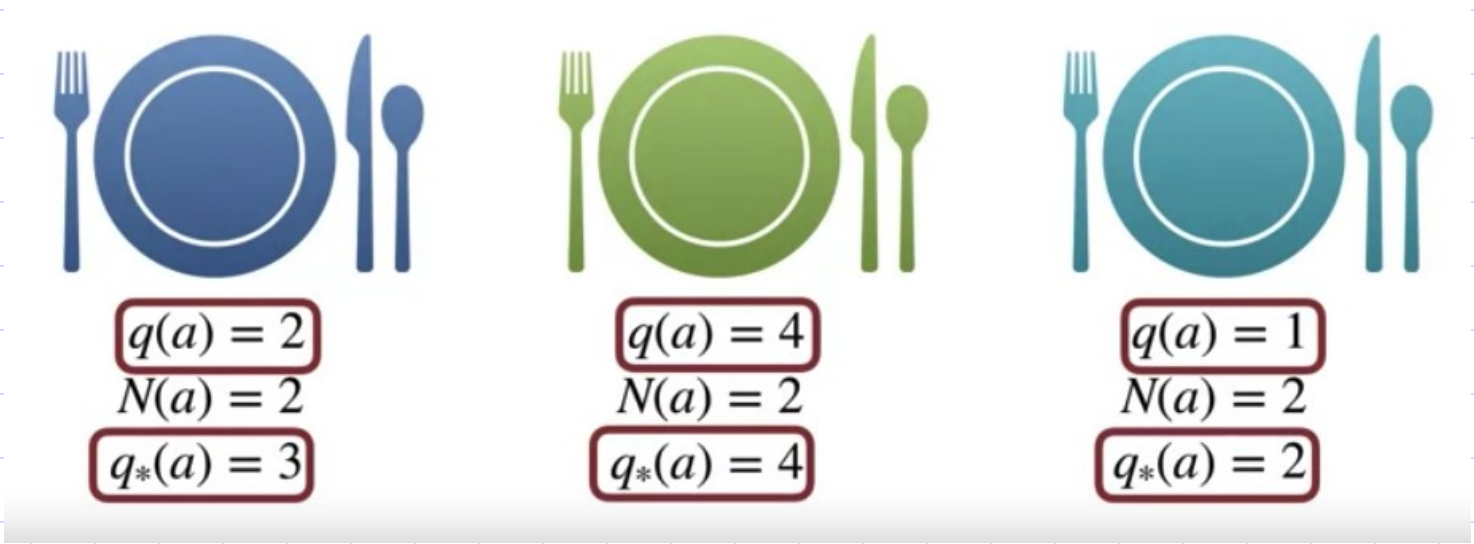
Exploration and Exploitation

- Example of exploratory behavior
  - Each time you visit this restaurant, you follow a strict regimen and choose each meal in a Round Robin fashion.



Exploration and Exploitation

- Example of exploratory behavior
  - After some time, you'll find the best meal to order



Exploration and Exploitation

- Exploitation:
  - Exploitation involves maximizing immediate rewards by choosing actions that have been found to be effective based on past experiences or knowledge.
  - Exploitation exploits the current knowledge to make decisions that are likely to yield the highest immediate rewards.
  - Exploitation may lead to suboptimal decisions if there are better but undiscovered options.
  - Exploit knowledge for short- term benefit

Exploration and Exploitation

- Example of exploitation behavior

$q(a) = 0$
$N(a) = 0$
$q_*(a) = 3$

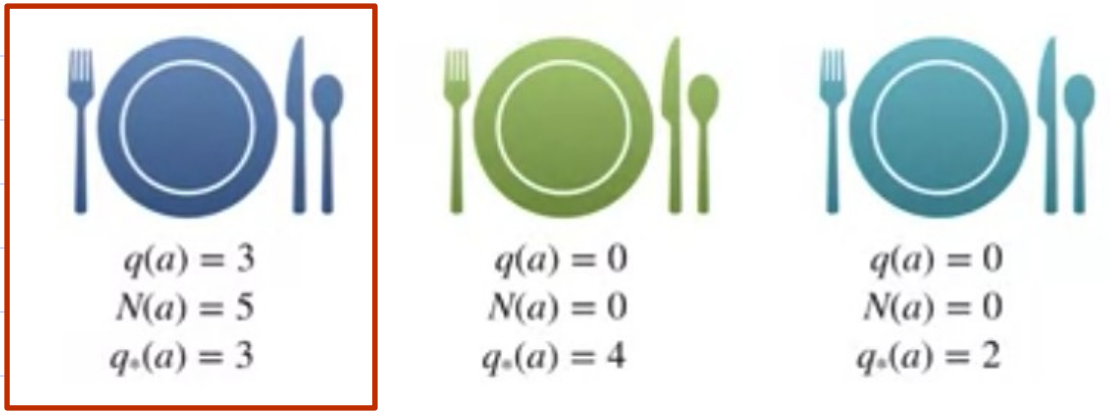
$q(a) = 0$
$N(a) = 0$
$q_*(a) = 4$

$q(a) = 0$
$N(a) = 0$
$q_*(a) = 2$

Exploration and Exploitation

- Example of exploitation behavior
  - The agent received a positive reward making the value for that meal higher.
  - The estimated values for the other actions are zero.
  - The greedy action is always the same, to pick the first meal.

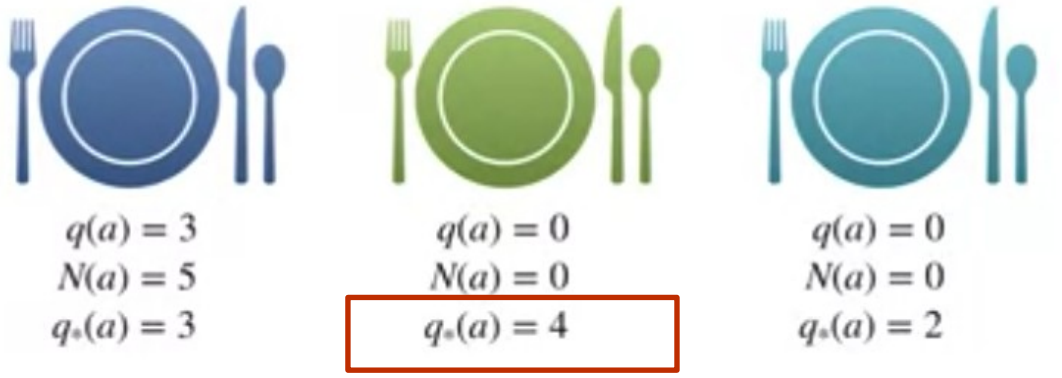
+1
+ 3
+ 5
+ 1
+5



Exploration and Exploitation

- Example of exploitation behavior
  - The agent never saw any samples for the other meals.
  - The estimated values for the other two actions remain far from the true values, which means the agent neve

+1
+ 3
+ 5
+ 1
+5



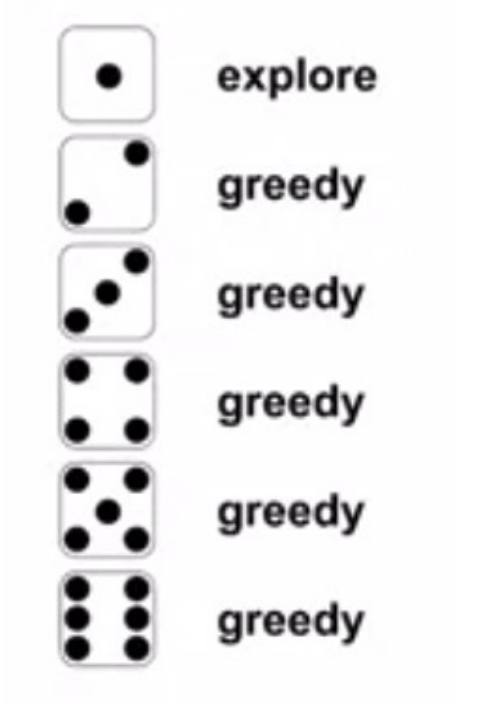
FPT Education
FPT UNIVERSITY

Exploration and Exploitation

- How do we choose explore or exploit?
  - When we explore:
    - Improve knowledge for long-term benefit
    - We get more accurate estimates of our values.
  - When we exploit,
    - Exploit knowledge for short-term benefit
    - We might get more reward.
  - We cannot however choose to do both simultaneously.

Exploration and Exploitation

- Choose randomly- Epsilon Greedy method: We could choose to exploit most of the time with a small chance of exploring
  - We could roll a dice. If it lands on one, then we'll explore.
  - Otherwise, we'll choose the greedy action.
  - Where epsilon refers to the probability of choosing to explore. In this case, epsilon will be equal to one over six.



Exploration and Exploitation

- Epsilon Greedy method:
  - The action that we select on time-step t is:
    - The greedy action with probability one minus epsilon
    - or a random action with probability epsilon.

$$A_t \leftarrow \begin{cases} \underset{a}{\text{argmax}}\ Q_t(a) & \textbf{with probability } 1-\epsilon \\ a \sim Uniform(\{a_1 \ldots a_k\}) & \textbf{with probability } \epsilon \end{cases}$$

Exploration and Exploitation

- Example:
  - Python code implementing epsilon-greedy exploration and exploitation. This code creates an agent that interacts with a multi-armed bandit environment and uses epsilon-greedy strategy to balance exploration and exploitation.

Exploration and Exploitation

- Example

```python
# 1.3 Example for Exploration and Exploitation- HoaDNt@fe.edu.vn
import numpy as np

class EpsilonGreedyAgent:
    def __init__(self, num_actions, epsilon=0.1):
        self.num_actions = num_actions
        self.epsilon = epsilon
        self.action_values = np.zeros(num_actions)
        self.action_counts = np.zeros(num_actions)

    def select_action(self):
        if np.random.rand() < self.epsilon:
            # Randomly choose an action for exploration
            action = np.random.randint(self.num_actions)
        else:
            # Choose the greedy action for exploitation
            action = np.argmax(self.action_values)
        return action

    def update_value(self, action, reward):
        self.action_counts[action] += 1
        # Update action-value estimate using incremental update rule
        self.action_values[action] += (1 / self.action_counts[action]) * (reward - self.action_values[action])

# Create a simple multi-armed bandit environment
class MultiArmedBandit:
    def __init__(self, num_arms):
        self.num_arms = num_arms
        self.true_action_values = np.random.normal(0, 1, num_arms)

    def get_reward(self, action):
        # Reward is sampled from a normal distribution with mean true action value and unit variance
        return np.random.normal(self.true_action_values[action], 1)
```

Exploration and Exploitation

- Example

```
34
35 # Initialize the environment and agent
36 num_arms = 10
37 num_steps = 1000
38 agent = EpsilonGreedyAgent(num_arms)
39
40 # Interaction loop
41 bandit = MultiArmedBandit(num_arms)
42 total_rewards = 0
43 for step in range(num_steps):
44     action = agent.select_action()
45     reward = bandit.get_reward(action)
46     agent.update_value(action, reward)
47     total_rewards += reward
48
49 print("Total rewards obtained:", total_rewards)
50 print("Estimated action values:", agent.action_values)
```

```
Total rewards obtained: 1133.394282414202
Estimated action values: [ 0.66969484  0.75865086  0.77213978  0.76244285 -1.14527773  0.97416503
 -2.01990784  0.86064529 -0.50302905  1.25977875]
```

Summary

- Understand exploration and exploitation tradeoff
- Understand the method for balance the exploration and exploitation

Q & A