Tabular Solution Methods
Monte Carlo and Q-Learning

Alina Vereshchaka

University at Buffalo
avereshc@buffalo.edu

June 4, 2019
Overview

1. Exploration vs Exploitation

2. Monte Carlo (MC) Methods

3. Temporal-Difference

4. Q-Learning
Table of Contents

1 Exploration vs Exploitation

2 Monte Carlo (MC) Methods

3 Temporal-Difference

4 Q-Learning
Exploration vs Exploitation: Where to eat?
Online decision-making involves a fundamental choice:

- **Exploitation**: Make the best decision given current information (greedy)
- **Exploration**: Gather more information

The greedy algorithm selects action with highest value:

\[ a_t^* = \arg \max_a Q_t(s, a) \]
Exploration vs Exploitation

$\epsilon - greedy$ algorithm:

- With probability $\epsilon$ choose a random action $a$
- With probability $1 - \epsilon$ choose “greedy” action $a$ with the highest Q-value.
Table of Contents

1 Exploration vs Exploitation

2 Monte Carlo (MC) Methods

3 Temporal-Difference

4 Q-Learning
Monte Carlo Methods

In Model-free, we focus on figuring out the value functions directly from the interactions with the environment.

There are few approaches for solving these kind of problems:

- Monte Carlo
- Temporal-Difference approach (SARSA, Q-Learning)
Monte Carlo Methods

- Learns value functions directly from episodes of experience
- MC is model-free: no knowledge of MDP transitions / rewards
- MC learns from complete episodes: no bootstrapping
- Uses the simplest idea: Value = Mean Return
- All episodes must terminate
Algorithm 4 On-policy Monte Carlo control

1: Initialise $Q$ and $\pi$ arbitrarily
2: $\text{Returns}(s, a) \leftarrow$ empty list $\forall s \in S, a \in A$
3: repeat
4:   for $s \in S$ and $a \in A$ do
5:     Generate an episode using $\epsilon$-greedy $\pi$ starting with $s, a$
6:     for $s, a$ in the episode do
7:     ... $\text{Returns}(s, a) \leftarrow$ append return following $s, a$
8:     $Q(s, a) = \text{average}(\text{Returns}(s, a))$
9:   end for
10:  for $s$ in the episode do
11:   $\pi(s) = \arg \max_a Q(s, a)$
12: end for
13: end for
14: until convergence
1 Exploration vs Exploitation

2 Monte Carlo (MC) Methods

3 Temporal-Difference

4 Q-Learning
Temporal-Difference Learning is model-free and learns from episodes of experience.

- TD is model-free: no knowledge of MDP transition/rewards
- TD learning can learn from incomplete episodes, by bootstrapping.
- Updates targets with regard to existing estimates rather than exclusively relying on actual rewards and complete returns
Table of Contents

1 Exploration vs Exploitation

2 Monte Carlo (MC) Methods

3 Temporal-Difference

4 Q-Learning
Q-Learning Algorithm

1. Initialize $Q$
2. Choose action from $Q$
3. Perform action
4. Measure Reward
5. Update $Q$
Q-learning (off-policy TD control) for estimating $\pi \approx \pi^*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q(s, a)$, for all $s \in S^+, a \in A(s)$, arbitrarily except that $Q(terminal, \cdot) = 0$

Loop for each episode:
  Initialize $S$
  Loop for each step of episode:
    Choose $A$ from $S$ using policy derived from $Q$ (e.g., $\varepsilon$-greedy)
    Take action $A$, observe $R, S'$
    $Q(S, A) \leftarrow Q(S, A) + \alpha \left[ R + \gamma \max_a Q(S', a) - Q(S, A) \right]$
    $S \leftarrow S'$
  until $S$ is terminal