C1M3: Worksheets

CMPUT 397
Fall 2020

Reminders: Sept 23, 2019

- Announcement sent out about Discussion Sessions; please fill out the Google Form
- I posted a resource (under Other Resources) with a simple proof for why the optimal state-value function is unique

A Few Questions from Slido

- I’ll go through some Slido questions
- Ask any additional questions in Zoom chat

Slido: Multiple Optimal Policies

- Multiple Optimal Policies: “If there are multiple optimal action at a state, is the probability of picking them always evenly distributed or can there be cases where it is not?”
  - Related: In the case of optimal policy, what will happen if the return of two actions are tie?
  - Related: Under what kind of situations can multiple optimal policies exist, if in the video it was said that you can simply combine policies that have higher values for a certain state into one policy that is higher than both original in those states?”

Slido: The Role of Gamma

- “In week 1 lectures, it made sense to diminish past rewards when calculating cumulative rewards, because we don't care about the past. Using the same logic, in week 2, shouldn't we discount rewards received at present rather than discounting future rewards (because future rewards should matter more)?”

Slido: Reward specification

- “I'm wondering about the rewarding intermediate steps, in the textbook they say it shouldn't be done as the agent could find a way to optimize this without achieving the goal. In the video, it mentioned providing an incentive for long stretch goals. What is best?”

Slido: Stochastic vs Deterministic Policies

- “Is a stochastic policy ever optimal? Or for a policy to be optimal, must it be deterministic?”
- “If it's deterministic, does it mean the policy is optimal?”
- Can you answer these questions for yourself?

Slido: Implementation

- “How are policies implemented in code? Can they be like a python list that gets updated and changed after each episode so that it gets closer to the optimal policy?”

- Alternative question: How do you represent conditional distributions in code?

Slido Misc

- “Curious if there have been any real-world applications of inverse-reinforcement learning.”

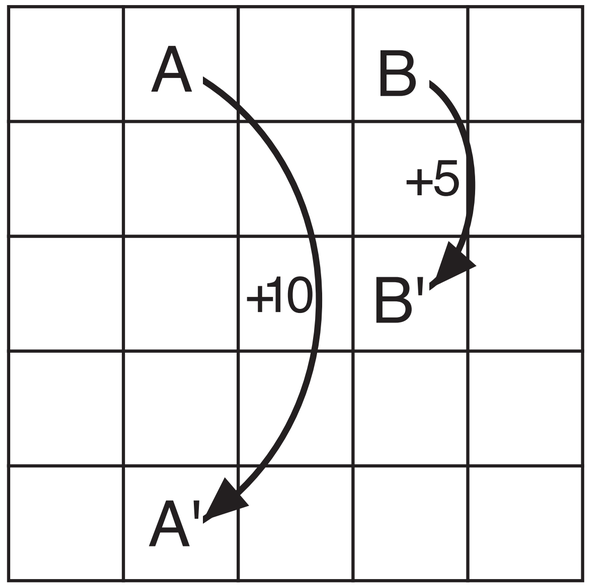
- “When we calculate the qstar(s,a), the defination tell us qstar is the maximum of expect value. But when we really compute it, we use the expect of maximum value. I think maximum of expect value is equal or less than expect of maximum value. So how can we avoid this?”

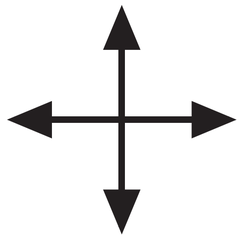
- “I am curious about the different ways the agent can avoid risk-aversive behaviour. Do descriptions of such behaviour have to be explicitly described and would we let the agent learn what the different types of risk-aversive behaviour are?”

Practice Question

The Bellman equation (3.10) must hold for each state for the value function v_\pi shown in Figure 3.2. As an example, show numerically that this equation holds for the center state, valued at +0.7, with respect to its four neighboring states, valued at +2.3, +0.4, -0.4, and +0.7. (These numbers are accurate only to one decimal place.). **Harder one:** verify the red state.

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r \,|\, s, a) \Big[r + \gamma v_\pi(s') \Big], \quad \text{for all } s \in \mathcal{S},$$





Actions

3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0

γ = 0.9
π = random
-1 reward on bump

Worksheet Question 1

Express the action-value function q_π in terms of v_π. The formula will also include p and π.