“All have their worth,” said Yavanna,

“and each contributes to the worth of the others”.

J.R.R. Tolkien, *The Silmarillion*

CMPUT 365
Introduction to RL

Marlos C. Machado

Plan

- Exercises and Answer Questions

Reminder

You **should be enrolled in the private session** we created in Coursera for CMPUT 365.

I **cannot** use marks from the public repository for your course marks.

You **need** to **check**, **every time**, if you are in the private session and if you are submitting quizzes and assignments to the private section.

The deadlines in the public session **do not align** with the deadlines in Coursera.

If you have any questions or concerns, **talk with the TAs** or email us `cmput365@ualberta.ca`.

Reminders & Notes

- The marks for Coursera's first module are now available on eClass.
- Midterm is this Friday.
  - There will be an attendance sheet.
  - You need to have your OneCard with you and show it to us.
  - No calculator, no cheat sheet.
- If you are not enrolled in Coursera's private session, Sample-based Learning Methods, please enroll!

Please, interrupt me at any time!

https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

6. What is the exploration/exploitation tradeoff?

- ○ The agent wants to maximize the amount of reward it receives over its lifetime. To do so it needs to avoid the action it believes is worst to exploit what it knows about the environment. However to discover which arm is truly worst it needs to explore different actions which potentially will lead it to take the worst action at times.
- ○ The agent wants to explore the environment to learn as much about it as possible about the various actions. That way once it knows every arm's true value it can choose the best one for the rest of the time.
- ○ The agent wants to explore to get more accurate estimates of its values. The agent also wants to exploit to get more reward. The agent cannot, however, choose to do both simultaneously.

10. Imagine, an agent is in a maze-like gridworld. You would like the agent to find the goal, as quickly as possible. You give the agent a reward of +1 when it reaches the goal and the discount rate is 1.0, because this is an episodic task. When you run the agent its finds the goal, but does not seem to care how long it takes to complete each episode. How could you fix this? (**Select all that apply**)

- ☐ Set a discount rate less than 1 and greater than 0, like 0.9.
- ☐ Give the agent -1 at each time step.
- ☐ Give the agent a reward of 0 at every time step so it wants to leave.
- ☐ Give the agent a reward of +1 at every time step.

Exercise 3.10 Prove the second equality in (3.10).

$$G_t = \sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}.$$

6. Suppose $\gamma = 0.8$ and the reward sequence is $R_1 = 5$ followed by an infinite sequence of 10s. What is G_0?

- 55
- 45
- 15

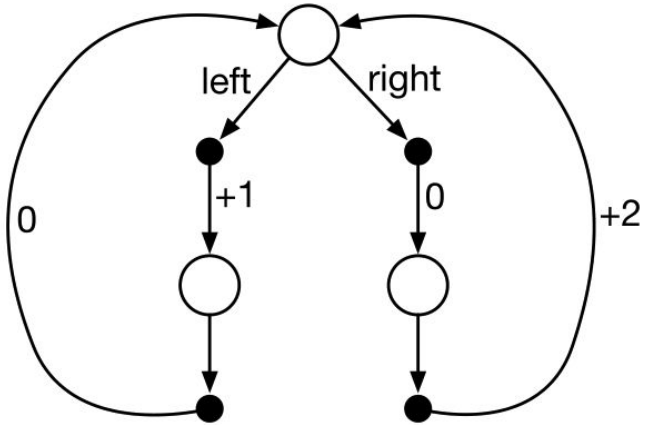
Suppose $\gamma = 0.8$ and we observe the following sequence of rewards: $R_1 = -3, R_2 = 5, R_3 = 2, R_4 = 7$, and $R_5 = 1$, with $T = 5$. What is G_0? Hint: Work Backwards and recall that $G_t = R_{t+1} + \gamma G_{t+1}$.

- 11.592
- 12
- 6.2736
- 8.24
- -3

10. Give an equation for some π_* in terms of v_* and the four-argument p.

- $\pi_*(a|s) = \max_{a'} \sum_{s',r} p(s', r|s, a')[r + \gamma v_*(s')]$
- $\pi_*(a|s) = 1 \text{ if } v_*(s) = \max_{a'} \sum_{s',r} p(s', r|s, a')[r + \gamma v_*(s')], \text{ else } 0$
- $\pi_*(a|s) = 1 \text{ if } v_*(s) = \sum_{s',r} p(s', r|s, a)[r + \gamma v_*(s')], \text{ else } 0$
- $\pi_*(a|s) = \sum_{s',r} p(s', r|s, a)[r + \gamma v_*(s')]$

2. Consider the continuing Markov decision process shown below. The only decision to be made is in the top state, where two actions are available, left and right. The numbers show the rewards that are received deterministically after each action. There are exactly two deterministic policies, π_{left} and π_{right}. Indicate the optimal policies if $\gamma = 0$? If $\gamma = 0.9$? If $\gamma = 0.5$? [Select all that apply]



- ☐ For $\gamma = 0$, π_{left}
- ☐ For $\gamma = 0.5$, π_{left}
- ☐ For $\gamma = 0.5$, π_{right}
- ☐ For $\gamma = 0$, π_{right}
- ☐ For $\gamma = 0.9$, π_{right}
- ☐ For $\gamma = 0.9$, π_{left}

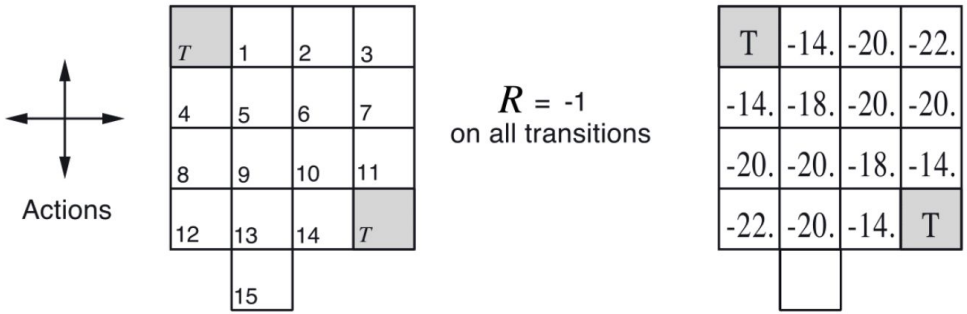
5. The word synchronous means "at the same time". The word asynchronous means "not at the same time". A dynamic programming algorithm is: [Select all that apply]

- ☐ Synchronous, if it systematically sweeps the entire state space at each iteration.
- ☐ Asynchronous, if it updates some states more than others.
- ☐ Asynchronous, if it does not update all states at each iteration.

8. Why are dynamic programming algorithms considered planning methods? [Select all that apply]

- ☐ They learn from trial and error interaction.
- ☐ They use a model to improve the policy.
- ☐ They compute optimal value functions.

9. Consider the undiscounted, episodic MDP below. There are four actions possible in each state, A = {up, down, right, left}, which deterministically cause the corresponding state transitions, except that actions that would take the agent off the grid in fact leave the state unchanged. The right half of the figure shows the value of each state under the equiprobable random policy. If π is the equiprobable random policy, what is $q(7, \text{down})$?

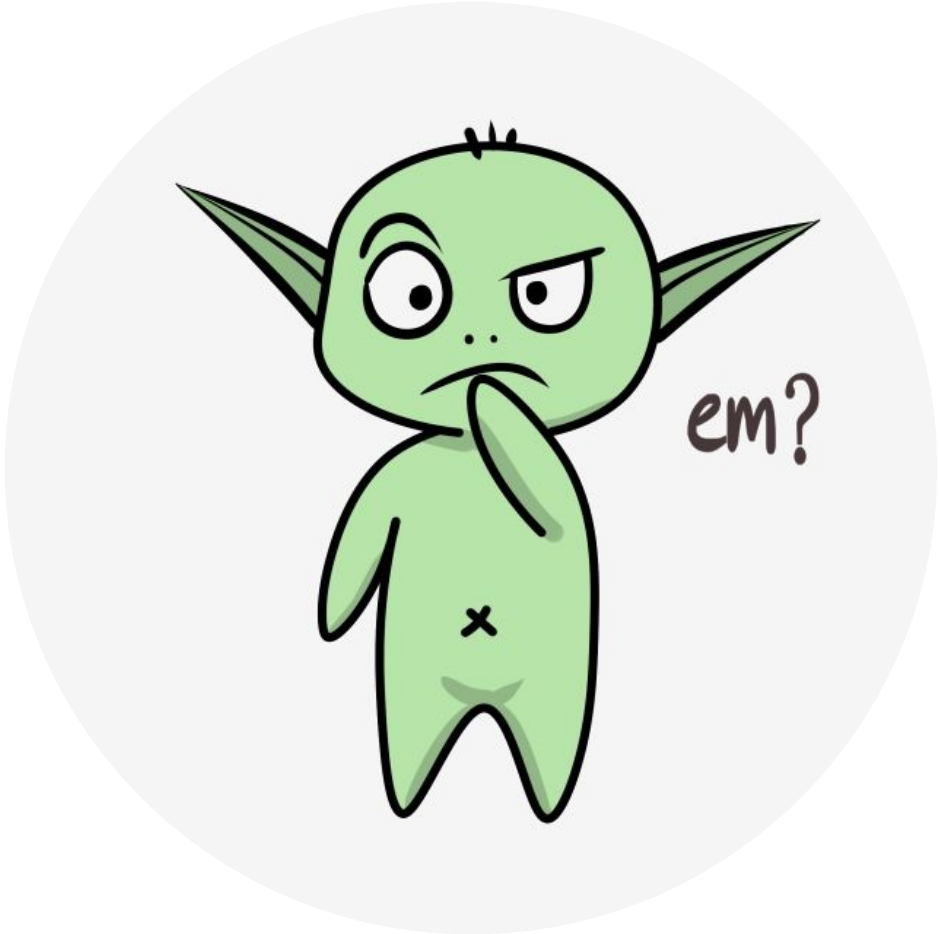


- $q(7, \text{down}) = -14$
- $q(7, \text{down}) = -20$
- $q(7, \text{down}) = -21$
- $q(7, \text{down}) = -15$

Exercise 3.29 Rewrite the four Bellman equations for the four value functions (v_π, v_*, q_π, and q_*) in terms of the three argument function p (3.4) and the two-argument function r (3.5). □

$$p(s'|s,a) \doteq \Pr\{S_t=s' \mid S_{t-1}=s, A_{t-1}=a\} = \sum_{r\in\mathcal{R}} p(s',r|s,a). \tag{3.4}$$

$$r(s,a) \doteq \mathbb{E}[R_t \mid S_{t-1}=s, A_{t-1}=a] = \sum_{r\in\mathcal{R}} r \sum_{s'\in\mathcal{S}} p(s',r|s,a), \tag{3.5}$$



https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht