“The rotten tree-trunk, until the very moment when the storm-blast breaks it in two, has all the appearance of might it ever had.”

Isaac Asimov, *Foundation*

CMPUT 365 Introduction to RL

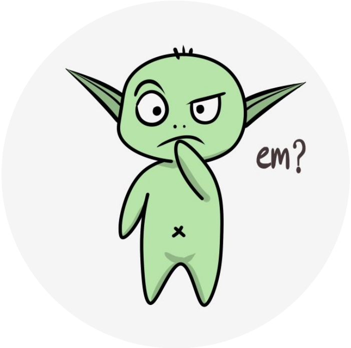
Marlos C. Machado

Class 7/ 35

Plan

- Finish Non-comprehensive overview of MDPs
    - Returns and Episodes
- Go over common errors in Coursera
- More exercises

Please, interrupt me at any time!



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

The ultimate goal: Maximize Returns

End of an episode

$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T$$

Continuing task

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

$$\begin{aligned} G_t &\doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \\ &= R_{t+1} + \gamma \left(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \cdots \right) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned}$$

Unifying Notation

$$G_t \doteq \sum_{k=0}^{T} R_{t+k+1} \qquad\qquad G_t \doteq \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

- We can't use the same notation for episodic and continuing tasks because:
  - We are not specifying the episodes in the indices of an episodic task, we should actually have $R_{t,i}$.
  - In continuing tasks we have a sum over infinite numbers and in episodic tasks we sum over finite numbers.

Unifying Notation

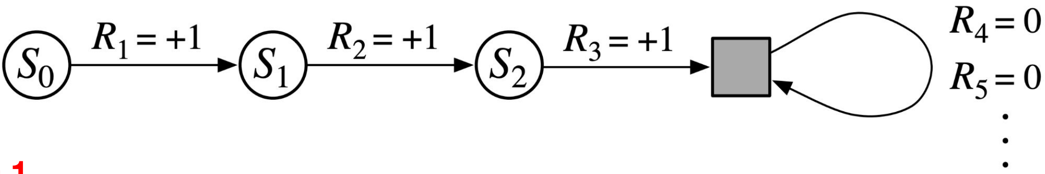
$$G_t \doteq \sum_{k=0}^{T} R_{t+k+1} \qquad\qquad G_t \doteq \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

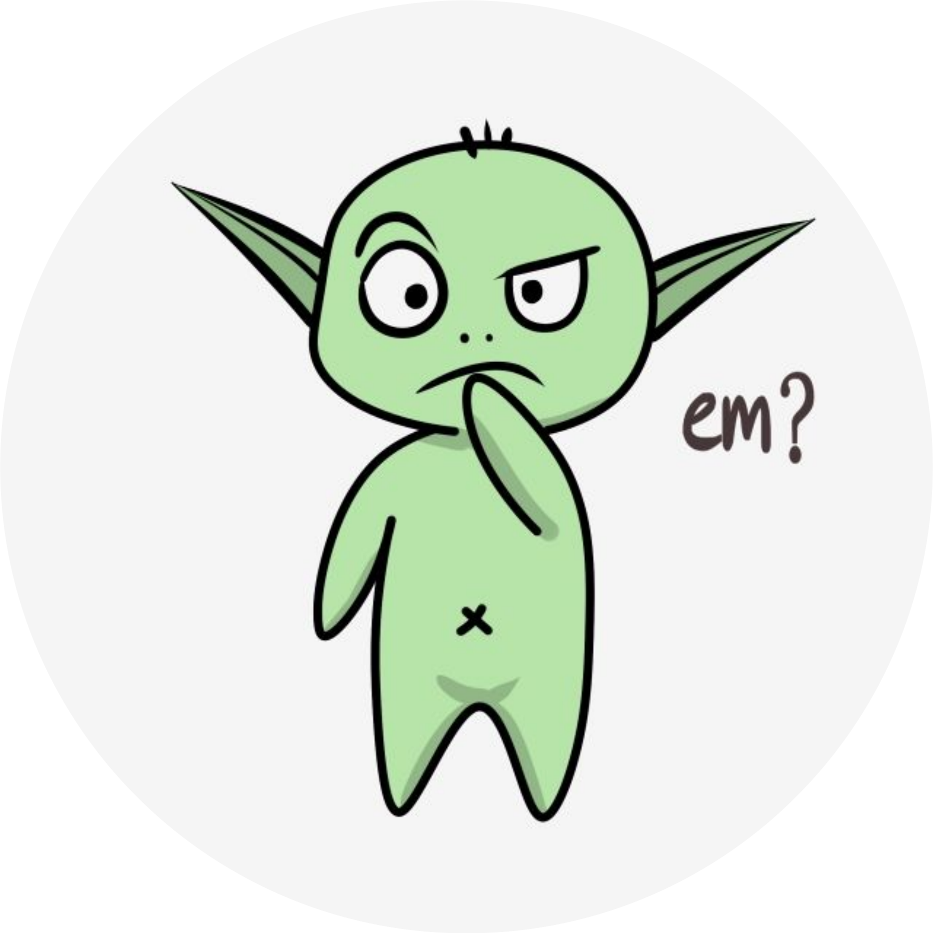
- We can't use the same notation for episodic and continuing tasks because:
  - We are not specifying the episodes in the indices of an episodic task, we should actually have $R_{t,i}$.
  - In continuing tasks we have a sum over infinite numbers and in episodic tasks we sum over finite numbers.

- Solution:
  - It is mostly fine to drop the episode number.
  - We create an absorbing state!

$$G_t \doteq \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$$

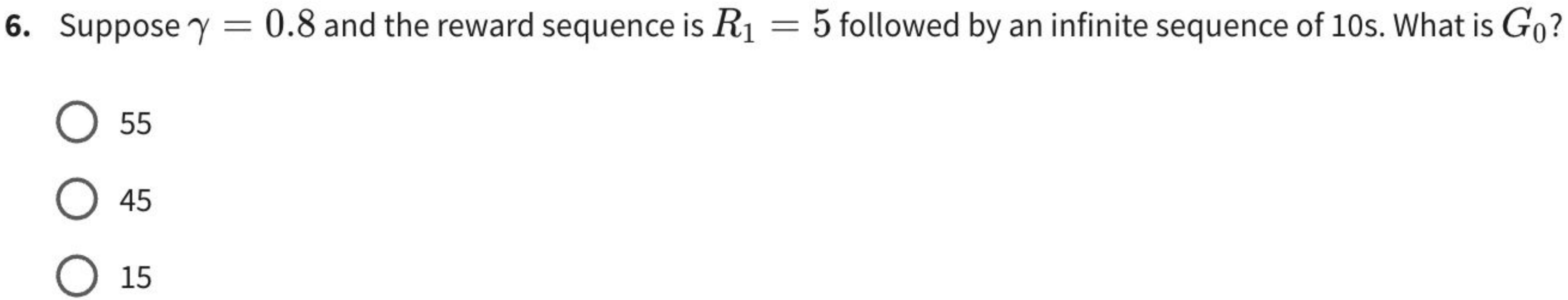
T = ∞ or γ = 1
(but not both)





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

Exercise 6 from Quiz on Coursera



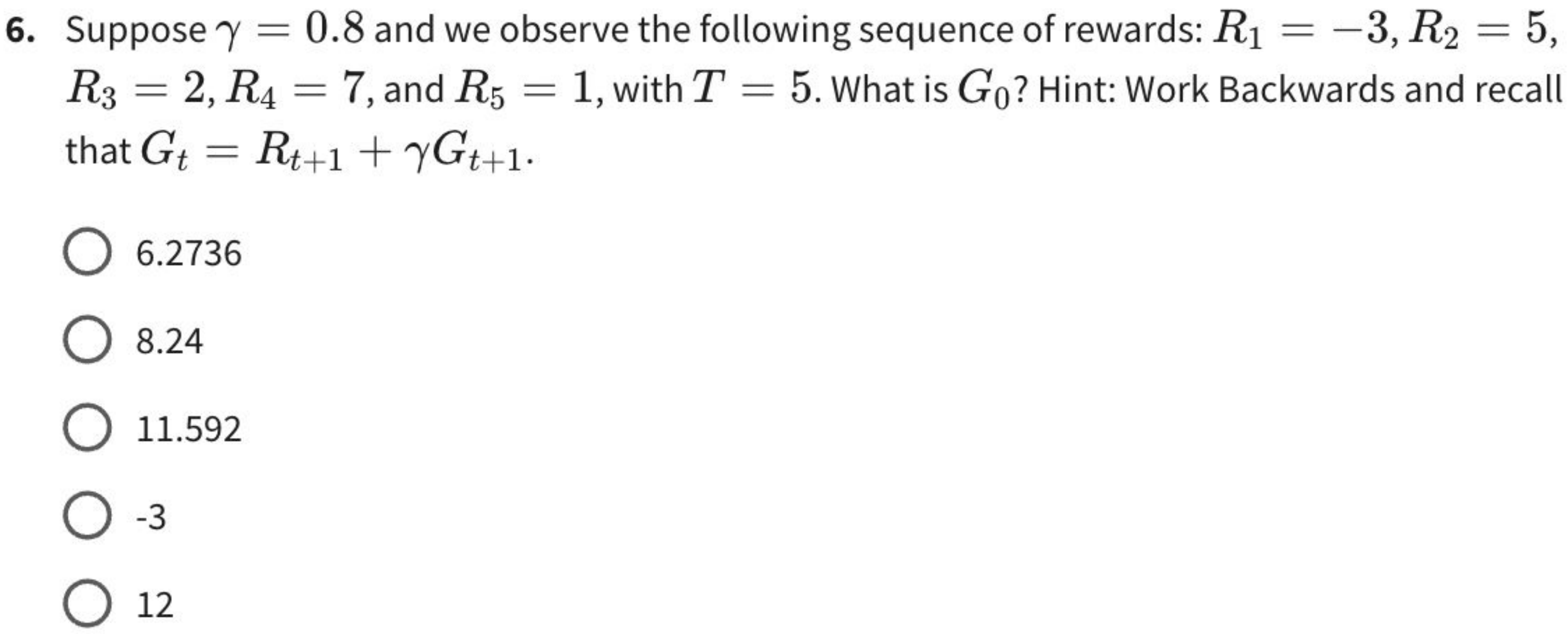
6. Suppose $\gamma = 0.8$ and the reward sequence is $R_1 = 5$ followed by an infinite sequence of 10s. What is G_0?

○ 55

○ 45

○ 15

Exercise 6 from Quiz on Coursera



6. 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 $G_t = R_{t+1} + \gamma G_{t+1}.$

○ 6.2736

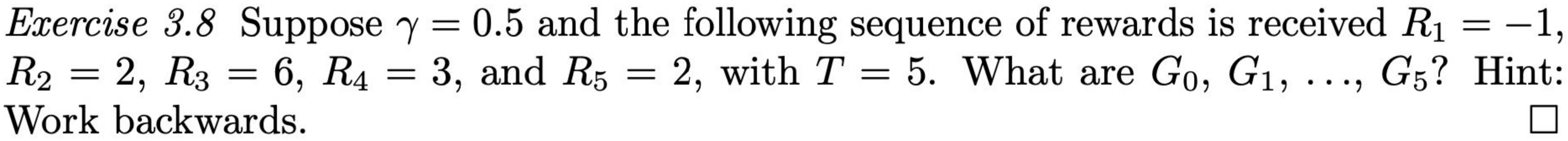
○ 8.24

○ 11.592

○ -3

○ 12

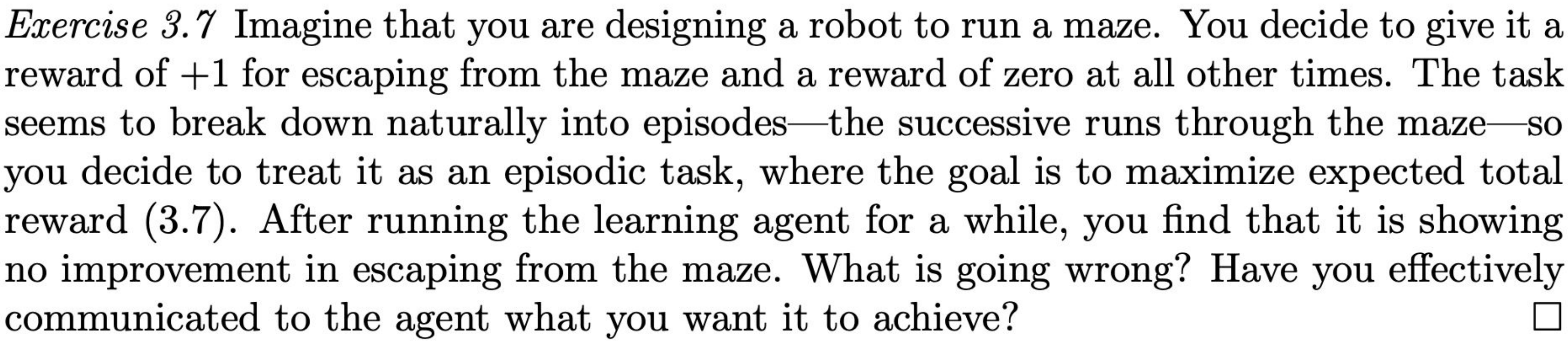
Exercise 3.8 of the Textbook



Exercise 3.8 Suppose $\gamma = 0.5$ and the following sequence of rewards is received $R_1 = -1$, $R_2 = 2$, $R_3 = 6$, $R_4 = 3$, and $R_5 = 2$, with $T = 5$. What are G_0, G_1, …, G_5? Hint: Work backwards. □

Solution Exercise 3.8 of the Textbook

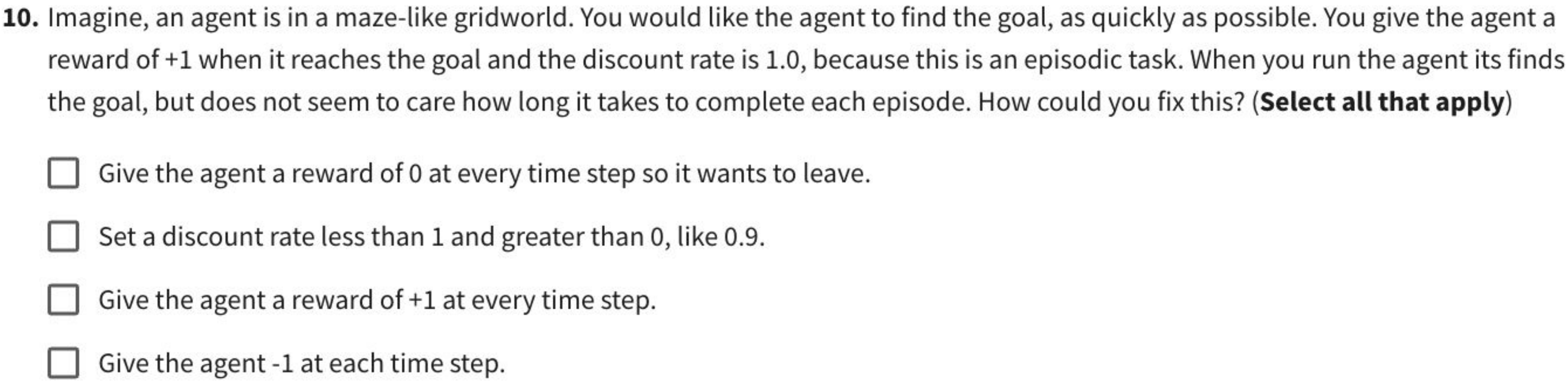
Exercise 3.7 of the Textbook



Exercise 3.7 Imagine that you are designing a robot to run a maze. You decide to give it a reward of +1 for escaping from the maze and a reward of zero at all other times. The task seems to break down naturally into episodes—the successive runs through the maze—so you decide to treat it as an episodic task, where the goal is to maximize expected total reward (3.7). After running the learning agent for a while, you find that it is showing no improvement in escaping from the maze. What is going wrong? Have you effectively communicated to the agent what you want it to achieve? □

Solution Exercise 3.7 of the Textbook

Exercise 10 from Quiz on Coursera



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**)

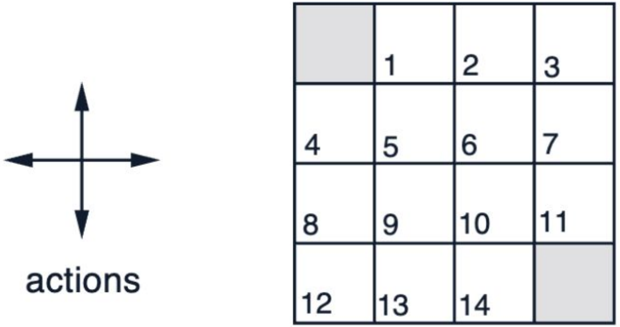
□ Give the agent a reward of 0 at every time step so it wants to leave.

□ Set a discount rate less than 1 and greater than 0, like 0.9.

□ Give the agent a reward of +1 at every time step.

□ Give the agent -1 at each time step.

Practice Exercise



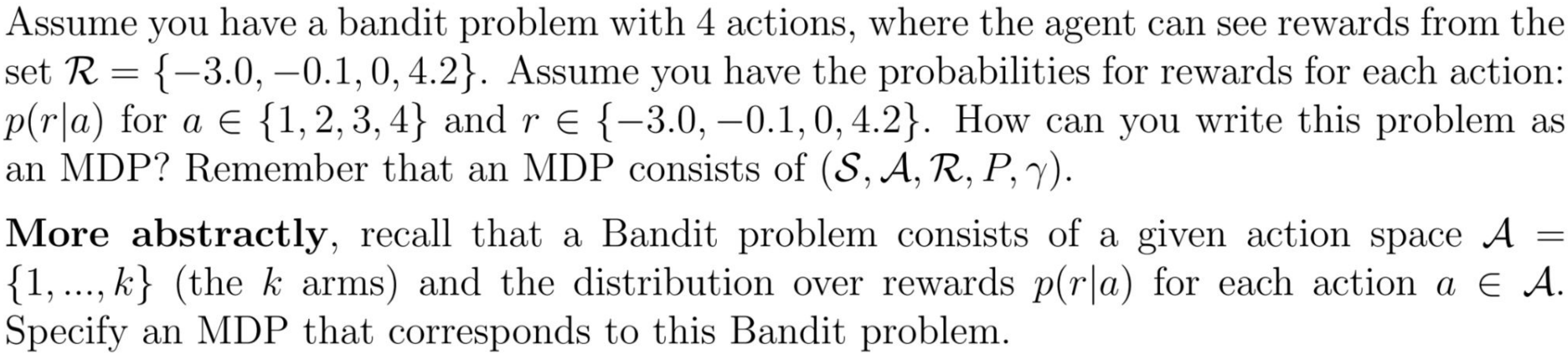
$R_t = -1$
on all transitions

$p(6, -1 \,|\, 5, \mathtt{right}) =$

$p(10, r \,|\, 5, \mathtt{right}) =$

$p(7, -1 \,|\, 7, \mathtt{right}) =$

Practice Exercise – Modeling

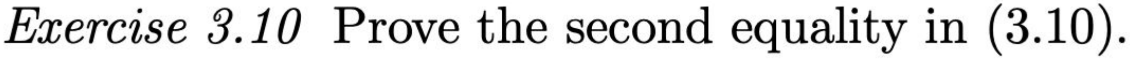


Assume you have a bandit problem with 4 actions, where the agent can see rewards from the set $\mathcal{R} = \{-3.0, -0.1, 0, 4.2\}$. Assume you have the probabilities for rewards for each action: $p(r|a)$ for $a \in \{1, 2, 3, 4\}$ and $r \in \{-3.0, -0.1, 0, 4.2\}$. How can you write this problem as an MDP? Remember that an MDP consists of $(\mathcal{S}, \mathcal{A}, \mathcal{R}, P, \gamma)$.

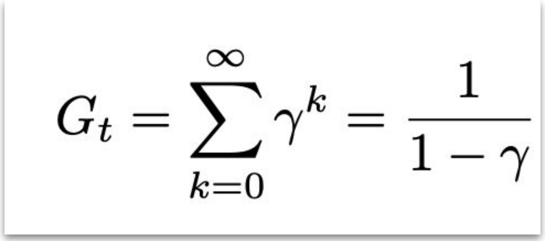
More abstractly, recall that a Bandit problem consists of a given action space $\mathcal{A} = \{1, ..., k\}$ (the k arms) and the distribution over rewards $p(r|a)$ for each action $a \in \mathcal{A}$. Specify an MDP that corresponds to this Bandit problem.

Practice Exercise – Modeling

Exercise 3.10 of the Textbook



Exercise 3.10 Prove the second equality in (3.10).



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

Solution Exercise 3.10 of the Textbook

Next class

- What **I** plan to do:
    - Discussion, answer questions, solve more exercises on MDPs.