# Reinforcement Learning Tutorial 2, Week 4

—

# Dynamic Programming / Monte Carlo

Pavlos Andreadis, Sanjay Rakshit

January 2020

**Overview**: The following tutorial questions relate to material taught in weeks 2 and 3 of the 2019-20 Reinforcement Learning course. They aim at encouraging engagement with the course material and facilitating a deeper understanding.

## Problem 1 - Modelling: Monkey-Banana

You are the manager for the local zoo, and it has come to your attention that the, one and only, zoo monkey has taken to begging for food from the visitors. Interestingly enough, the visitors will occasionally react to this by purchasing a banana from the zoo's kiosk, which they will then proceed to give to the monkey. Since the zoo is going through some hard times, you wonder whether this pattern of behaviour can be used to increase the zoo's income. Each banana nets a £1 income, after all.

Your resident monkey expert has informed you that the monkey's behaviour depends on whether or not it is sleeping, as well as on how hungry it is. The following 4 modes of behaviour can be distinguished:

- sleeping
- sated
- hungry
- furious

The expert has further provided you with information on how the monkey's behaviour will change depending on whether or not it was fed a banana during the previous 10 minutes:

- There is a 5% probability that a monkey will wake up during 10 minutes of sleeping. Monkeys are always hungry when they wake up.

- If it does not receive a banana, a sated monkey will fall asleep with a 50% probability, remain sated with 25% probability, and get hungry with a 25% probability. If a sated monkey does receive a banana, then it will either fall asleep, with 75% probability, or remain sated.
- A hungry monkey that does not receive a banana has a 30% chance of getting furious, and will otherwise remain hungry. If a hungry monkey receives a banana it has a 40% chance of becoming sated, and will otherwise remain hungry.
- A furious monkey that does not receive a banana will remain furious, while receiving a banana has a 40% chance of reducing it to a merely hungry state, and a 10% chance of sating it (50% of remaining furious).
- A furious monkey might scare a zoo visitor with a 20% chance. You are told that this has an expected negative impact on the zoo income of £10.

Another expert lets you know that the chance of a visitor purchasing and giving a banana to the monkey during a 10 minute period depends on the monkey's state:

- If the monkey is sleeping, visitors show no interest and won't bother it.
- If the monkey is sated, there is only a 10% probability that the visitors will give it a banana.
- If the monkey is hungry, then visitors have a 70% chance of buying it a banana.
- Furious monkeys occasionally manage to extort a banana; there is a 10% chance that they are fed one.

Lastly, you know that opening or closing the banana-kiosk takes 10 minutes, and that no bananas are sold when the kiosk is closed.

## Part 1

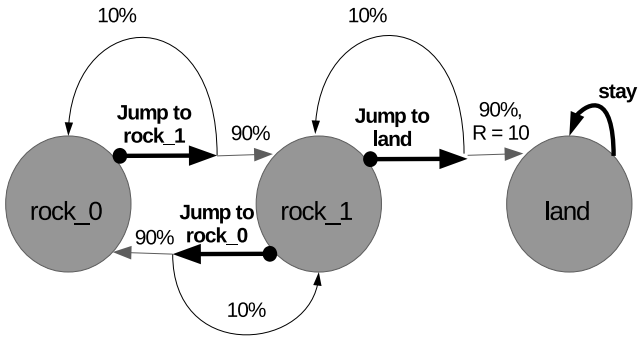
Assume we are concerned with maximising the zoo's net income (sales − costs). Use the information above to produce a finite state and action Markov Decision Process that could help you decide on a banana-kiosk policy.

## Part 2

Why did you pick the specific reward function for your model? Would another reward function have been just as good? Can we pick the reward function in a way that the value of a state could have a meaningful interpretation in terms of the zoo's income?

# Problem 2 - Dynamic Programming

Consider the Hop-a-long MDP from tutorial 1 Rakshit and Andreadis [2020] below, with a discount factor of $\gamma = 0.9$. Apply two iterations of *Policy Iteration*, starting from a uniform initial policy.




# Problem 3 - Monte Carlo

Compute the state-value function for a given policy $\pi$ and MDP without access to the MDP's model, using the following four episodes, in order:

$$\text{rock}_0, 0, \text{rock}_0, 0, \text{rock}_1, 10, \text{land} \tag{1}$$
$$\text{rock}_1, 0, \text{rock}_0, 0, \text{rock}_1, 10, \text{land} \tag{2}$$
$$\text{rock}_0, 0, \text{rock}_0, -100, \text{sea} \tag{3}$$
$$\text{rock}_1, 0, \text{rock}_0, -100, \text{sea} \tag{4}$$

## Part 1

Use first-time visit Monte Carlo to evaluate the state-value function at each state.

## Part 2

Use every-time visit Monte Carlo to evaluate the state-value function at each state.

## Part 3

Which are the absorbing states?

## References

S. Rakshit and P. Andreadis. Reinforcement Learning Tutorial 1, Week 2 — with solutions — Introduction / MDPs. https://www.learn.ed.ac.uk/webapps/blackboard/execute/content/file?cmd=view&mode=designer&content_id=_4543883_1&course_id=_70929_1, 2020.