Reinforcement Learning Tutorial 2, Week 3

—

MABs and MDPs

Pavlos Andreadis, Sanjay Rakshit, Adam Jelley

January 2025

Overview: The following tutorial questions relate to material taught in the first two weeks of the 2024-25 Reinforcement Learning course. They aim at encouraging engagement with the course material and facilitating a deeper understanding.

In this week of tutorials we will start looking into modelling "real-world" problems as Reinforcement Learning problems. I quote "real-world" as we will typically resort to what is often called a "toy-problem" in academia. Part of this is because they are easier to write, and much easier to limit in scope. Part of it is because it allows us to poke controllable holes into the problems and make the limitations and assumptions that much clearer. Rather than overwhelming you with potential considerations (and more realistic problems especially from your everyday experience would trigger many such personalised thoughts that we can't always predict), you are introduced to them one at a time. It is much easier for you to focus on what we are trying to communicate when you accept from the start that it is a "story", and not exactly reality.

The descriptions of the problems will still not always be a perfect specification; that is after all what is required of you to produce when modelling it. The description will however hint at the correct level of abstraction. For example, in these problems we are presenting a frog climbing onto a rock as if it were a trivial matter, and do not give any detailed information on that process.

[*Could be fun to consider what such information might be. By the way, you will see a lot of italicised brackets like these, especially in the solutions. These are just here for you to contemplate further, or perhaps discuss with your colleagues or tutor if you so want.*]

Problem 1 - Modelling: Frog on a Rock

A friendly frog, Hop Along was stranded on a rock surrounded by water. It needs to get to land without falling in. The only way to safety is for it to jump on to neighbouring rocks till it arrives on land.

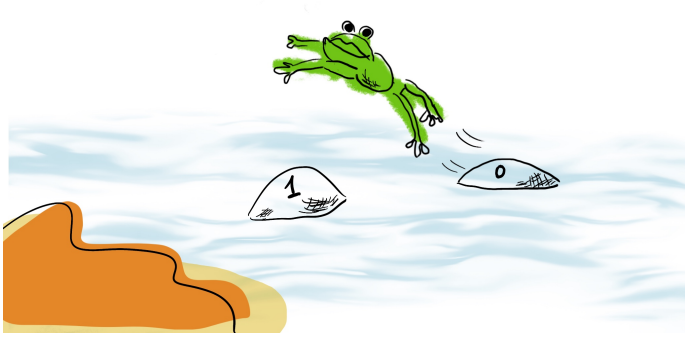

Figure 1: "Will He Make it?" (Image and title from Yana Knight [2021], with permission)

There are two rocks, and Hop Along can jump from a rock to another or from the final rock to land. However, sometimes it misses and ends up in the water, having to climb back onto the same rock it tried to jump from. The rocks are arrayed in a row leading from its starting rock, which we will name $rock_0$, to $rock_1$, and then finally land which can only be reached from $rock_1$.

Consider the control problem where the current state is specified by Hop Along's location and the actions Hop Along can take is to attempt to jump towards another reachable rock or land.

Assume that Hop Along's jumps always have a 90% chance of reaching the intended destination, while the rest of the time Hop Along falls in the water.

Formulate a Markov Decision Process (MDP, see Sutton and Barto [2018], Ch. 3) for the problem of deciding on Hop Along's actions in order to help it reach land. Why did you formulate it as you did? What additional assumptions did you have to make?

Can you define the uniform random policy on this MDP? What about the optimal policy?

Problem 2 - Modelling: Frog on a Rock 2

Though in Problem 1 we modelled our frog as always getting back on the rock it falls off, reality does not always comply with our assumptions (or conform to our expectations) and Hop Along the frog ended up being carried away by the river. It now resides in a calm little pond, with a nice rock in the middle, and surrounded by flies. It spends its time climbing onto this rock, and then jumping into one of the four cardinal directions (South, West, North, East) and swallowing the flies it comes across on its way down to the water.

Figure 2: "Frog with Problems." (Image and title from Yana Knight [2021], with permission)

Part a

Formulate Hop Along's attempt to catch as many flies as possible as a Multi-Armed Bandit (MAB) problem (see Sutton and Barto [2018], Ch. 2).

Part b

Can you define two simple exploration strategies that Hop Along could employ to try to maximise the total number of flies he catches? What is an advantage and disadvantage of each potential strategy?

Part c

An implied assumption in **Part a** is that the frog will never stop jumping. If there was a limited amount ϕ of jumps the frog could do (let's say $\phi = 100$ jumps), would it still make sense to model this problem as a MAB Problem? Why? If not, how would you go about modelling this scenario?

Part d

Let's go back to assuming an infinite number of jumps. Another implied assumption in **Part a** has been that we can control Hop Along's actions. If we couldn't though (perhaps because we are not Hop Along ourselves, but a researcher interested in how many flies it eats), would it still make sense to model this as a MAB problem? Why? If not, how would you go about modelling this scenario?

[*What if the much-troubled frog chasing flies was an oil-rig company sampling oil deposits, and the* 4 *cardinal directions were* 4 *different extraction sites?*]

References

Yana Knight. "Story of Yana". http://storyofyana.com/, 2021.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.