Reinforcement Learning Tutorial 4, Week 5
— with solutions —
MDP Modelling

Pavlos Andreadis

February 2025

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

We take a breather this week from the algorithms to look at a slightly harder modelling problem. There are a lot of assumptions here and, topic aside, a lot of considerations also encountered in real world problems. As many such problems, there can be different ways to go about it. Give it a go before your tutorial session, and use this as an opportunity to consider the limitations of modelling this as a finite MDP.

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

Figure 1: "Zoe and Phill"
Image and title used with permission from Yana Knight [2021]

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.

Answer:

Define 2 state variables:

- $m \in \{0, 1, 2, 3\}$ or monkey state $\in \{'sleeping', 'sated', 'hungry', 'furious'\}$, equivalently; and
- $k \in \{0, 1\}$ or kiosk state $\in \{'closed', 'open'\}$, equivalently.

We can index states as s_{mk} to refer to the state where the monkey is in state m and the kiosk in state k.

We could define 3 actions for $'wait', 'close\ kiosk', 'open\ kiosk'$ but since we can only ever either open or close the kiosk, we would have an easier time just using 2 actions:

- a_0 or $'wait'$; and
- a_1 or $'open/close\ kiosk'$.

I give the transition and reward functions on Figure 3, and the transition function in tabular form in Figure 2. Note that I have omitted a number of transitions with probability 0 from the latter. The reward function is 0 everywhere except for:

- $R^{a}_{s_{30}s'} = -2, \forall a, s'$
- $R^{a_1}_{s_{31}s'} = -2, \forall s'$
- $R^{a_0}_{s_{31}s'} = -2 + 0.1, \forall s'$
- $R^{a_0}_{s_{21}s'} = 0.7, \forall s'$
- $R^{a_0}_{s_{11}s'} = 0.1, \forall s'$

a_0	s_{00}	s_{10}	s_{20}	s_{30}
s_{00}	0.95	0	0.05	0
s_{10}	0.50	0.25	0.25	0
s_{20}	0	0	0.70	0.30
s_{30}	0	0	0	1

a_0	s_{01}	s_{11}	s_{21}	s_{31}
s_{01}	0.95	0	0.05	0
s_{11}	0.1*0.75 + 0.9*0.50	0.1*0.25 + 0.9*0.25	0.9*0.25	0
s_{21}	0	0.7*0.40	0.7*0.60 + 0.3*0.70	0.3*0.30
s_{31}	0	0.1*0.1	0.1*0.4	0.1*0.50 + 0.9*1

a_1	s_{01} / s_{00}	s_{11} / s_{10}	s_{21} / s_{20}	s_{31} / s_{30}
s_{00} / s_{01}	0.95	0	0.05	0
s_{10} / s_{11}	0.50	0.25	0.25	0
s_{20} / s_{21}	0	0	0.70	0.30
s_{30} / s_{31}	0	0	0	1

Figure 2: Transition function in tabular form for the Banana-Kiosk problem. I have taken a shorthand in writing the tables for a_1, since when taking this action we can only ever transition from a state s_{m0} to a state $s_{m'1}$, and vice versa. Also note that the probabilities there are identical to those for taking action a_0 from a state s_{m0}.

We have made an additional assumption to those specified in the Question here; specifically that we can't sell a banana while the kiosk is closing. There are therefore no positive rewards when choosing action a_1 in states s_{m1}.

Since we don't always know whether a banana has been sold when making a transition, we can generally only reward for putting the world in a state where we could receive some gain or penalty. I have chosen to match this reward with the expectation of bananas sold (1 banana = £1) (e.g. 10% chance of selling 1 banana gives a reward of 0.1) or money lost (e.g. 20% chance of scaring a visitor for a loss of £10 gives a reward 0f -2).

One could try to match transitions closer to the probability of having sold a banana given the transition, but this is an obvious waste of time; the expectation of a reward given a state-action pair would still remain the same (as long as you calculated your probabilities correctly). That being said, if we didn't have a model, and had to sample trajectories, our signals would obviously not be of this nature (we would e.g. receive a signal of £1 10% of the time), but we would still converge to the same optimal 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?

Answer:

If you followed my lead, then you will have defined the rewards so that they represent an expected influx or outflux of £. However, should it really matter if these were expressed in the model, say, in $? You can multiply all your rewards together by the same positive number and they won't change our relevant preference among policies. The value function will change, but not how good one state is in relation to other states, given a policy. Computationally, of course, you will need to make sure you are not making these big/small enough to lose important information, and give special consideration to any convergence checks for your algorithms. You can even add any constant to all rewards, and this would still be the case. Note that I do mean ALL rewards, so this positive affine transformation would have to be applied over $R^a_{ss'}$. Oh, and this result does not hold for MDPs with absorbing states. For those who are adventurous, you can have a look first at Weng [2012] and then at Weng and Zanuttini [2013].

Now, what I was mostly asking for is much simpler. Since this would have to be a discounted MDP (there are no absorbing states), the value at each state or state-action would be the *discounted* return, and since the reward has been shaped such that the return, with no discounts, is the expected influx/outflux of £, these two numbers are obviously not the same.

References

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

Paul Weng. Ordinal decision models for markov decision processes. In *ECAI*, pages 828–833, 2012.

Paul Weng and Bruno Zanuttini. Interactive value iteration for markov decision processes with unknown rewards. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

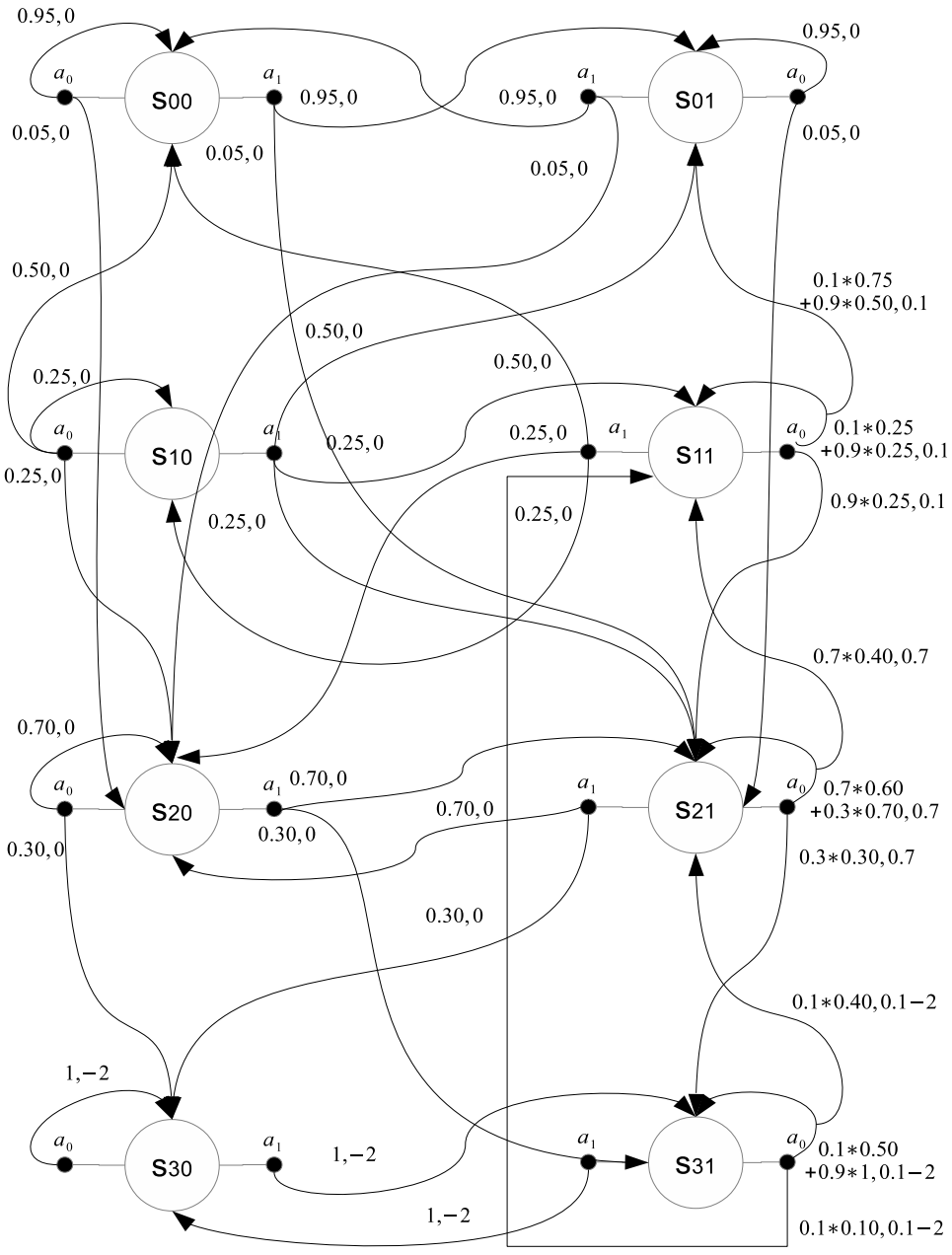

Figure 3: Transition graph for the Banana-Kiosk problem. We only get a chance to sell a banana if we wait (a_0) while the kiosk is open (s_{m1}) and the monkey isn't sleeping ($m \neq 0$). Since we are not tracking whether a banana is sold, or a visitor scared, we reward according to the probability of these things happening when we have spent 10 minutes in a state. Our reward function ends up only depending on the current state and action taken.