Reinforcement Learning Tutorial 5, Week 10
—
Policy Gradients / Revision

Pavlos Andreadis

March 2020

Overview: The following tutorial questions relate to material taught in weeks 1 to 6 of the 2019-20 Reinforcement Learning course. They aim at encouraging engagement with the course material and facilitating a deeper understanding.

Problem 1 - Policy Gradients: REINFORCE

Consider the orchard problem from our last tutorial Andreadis and Rakshit [2020], and the trajectory representing its last harvest:

Concentration of A (ppm)	Concentration of B (ppm)	Concentration of C (ppm)	Action Taken	Profit/Reward (credits)
6	7	2	Wait	-1
0	5	2	Wait	-1
3	8	4	Harvest	19

Assume that these actions were taken using a policy parameterization with softmax in action preferences with linear action preferences, and a known parameter $\theta_0 = [0, 0]$. Using the REINFORCE algorithm with a step size of $\alpha = 10^{-4}$, update your policy given the above trajectory.

Problem 2 - Revision: MDP Modelling

[Adapted from RL exams in 2017-18]

Pamp the sailor was in a shipwreck and has been left stranded on an island. Though this island is now ’home’, it has no resources and Pamp occasionally sets out on a raft to scavenge for resources from the surrounding islands. Upon arriving on an island, Pamp accumulates a specific, known, amount of resources (except for the 'home' island which never has any resources). Each islasnd has a fixed amount of resources, which are replenished after each visit. Pamp can attempt to move from any island to any *other* island, but some times the sea currents will move the raft randomly to one of the islands (even the one Pamp is on at the moment).

1. Consider the control problem where the current state is specified by the current island Pamp is on, and the actions Pamp can take are to attempt a move towards another island. Assume that there are 2 other islands, except for 'home' which is always the starting island (so 3 islands in total). Moreover, assume that Pamp has full knowledge of the amount of resources on each island, and that there is always a 90% chance of Pamp transitioning towards the intended island, with the rest of the probability uniformly distributed across the remaining islands.

    (a) Formulate a Markov Decision Process (MDP) for the problem of controlling Pamp's actions in order to maximise the accumulation of resources during an episode/trip. (Give the transition and reward functions in tabular format, or give the transition graph with rewards).

    (b) If Pamp's trip ends upon returning to the 'home' island, how would you modify the above MDP? (Similarly, "How would an MDP for this modified problem differ from the MDP for the above question?").

    (c) Consider the discounted return from the state 'home' for a single episode. For which of the models above in i) and ii) could this number be an accurate representation of the sum of resources gathered during that episode? (Assuming your rewards have been defined to represent the quantity of resources gathered when visiting each island).

2. In the example at the beginning of this question, Pamp has access to a Transition and Reward function.

    (a) Assuming no access to the Reward and Transition functions, would Pamp be able to compute an optimal policy without leaving 'home', and why? Considering a Reinforcement Learning algorithm in general, what is the property of not needing these two functions as input called?

    (b) Consider any of the MDPs defined above, focusing on that your states are defined as the island Pamp is currently on. Which basic assumption of MDPs would be violated if the transition probabilities from one island to another also depended on the number of previously visited islands? If this assumption was violated, but you were asked to evaluate a plan for moving from island to island, which algorithm would you choose and why?

    (c) Is the algorithm you chose well defined for continuing (non-episodic) tasks?

Problem 3 - Revision: Function Approximation

[Adapted from RL exams in 2017-18]

Consider the problem with Pamp the sailor in Problem 2, but with an infinite number of islands. Moreover, assume that Pamp can only ever see and choose between 2 different islands (*left* and *right*) to move towards and that Pamp can observe an estimate of the amount of resources on each of those 2 islands. If you were to formulate the control problem as an MDP:

1. What would be a good representation of state if Pamp had no memory of previously visited islands?
2. What would be a good representation of state if Pamp could remember the previous island (in addition to the current one)?
3. Define the linear approximate state-value function for one of the above two cases.

References

P. Andreadis and S. Rakshit. Reinforcement Learning Tutorial 4, Week 8 — with solutions — Value Function Approximation / Eligibility Traces. https://www.learn.ed.ac.uk/webapps/blackboard/execute/content/file?cmd=view&mode=designer&content_id=_4609921_1&course_id=_70929_1, 2020.