Reinforcement Learning Tutorial 6, Week 7

—

Policy Gradients: The REINFORCE algorithm* and Hyperparameters in RL

Pavlos Andreadis

March 2024

Overview: The following tutorial questions relate to material taught in weeks 5 and 6 of the 2023-24 Reinforcement Learning course. They aim at encouraging engagement with the course material and facilitating a deeper understanding.

We continue on the "AI controlled orchard" problem from tutorial 6 for a deep dive into a simple application of the REINFORCE algorithm. As you apply the algorithm, consider the changes to your policy and what they reflect. Also, consider the limitations in the suggested formulation of the problem (you are implicitly given that you have to use two state-action features, which you will have to define).

The second problem asks you to consider the use of the use of the learning rate and discount factor in Reinforcement Learning.

Problem 1 - Policy Gradients: REINFORCE

Consider the orchard problem from our last tutorial Andreadis [2021], 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

*with special thanks to **Ross McKenzie** for introducing a first version of problem 1

Assume that these actions were taken using a policy parameterization with soft-max in action preferences with linear action preferences, and a known parameter $\theta_0 = [0, 0]$ (i.e. assume the state-action space is parameterised by two features, here chosen to be defined by the action *only*). Using the REINFORCE algorithm with a step size of $\alpha = 10^{-4}$, update your policy given the above trajectory.

Problem 2 - Discussion

Part a

Considering a Reinforcement Learning algorithm in general, what is the overall effect of increasing the learning rate? What happens when you set it too high? What happens when you set it too low?

Part b

Is the discount factor γ:

1. Part of the definition of a Markov Decision Process? That is, a part of the definition of the problem to be solved; or
2. Is it external to the problem? That is, a hyperparameter for training the model.

When a discount factor is close to 1, we end up with a long horizon problem. That is, we plan for long-term gains. Assume we were training a Reinforcement Learning agent for a long-horizon problem. Could you think of a reason for which a method using short-horizon targets (cut-off at some horizon h) might outperform a method using long-term horizon targets on this problem? To aid in answering, consider searching online for "planning horizon reinforcement learning model accuracy" and look for related work.

Part c

What other parameters are relevant in Reinforcement Learning, and how do you set their values?

References

P. Andreadis. Reinforcement Learning Tutorial 5, Week 6 — Reward Shaping / Semi Gradient Monte Carlo. https://www.learn.ed.ac.uk/webapps/blackboard/execute/content/file?cmd=view&mode=designer&content_id=_5795113_1&course_id=_82627_1, 2021.