CS 7642: Reinforcement Learning and Decision Making **Homework #2**

The λ-return

1 Problem

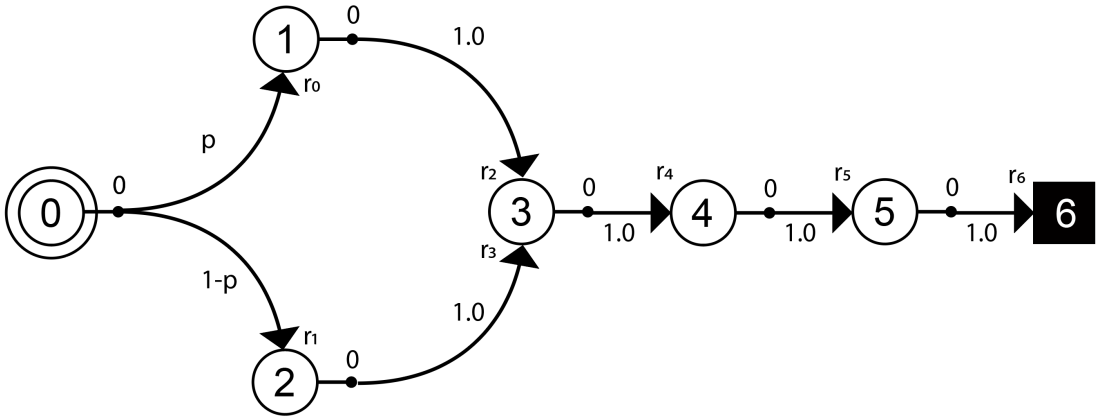
1.1 Description

Given an MDP and a particular time step t of a task (continuing or episodic), the λ-return, G_t^λ, $0 \leq \lambda \leq 1$, is a weighted combination of the n-step returns $G_{t:t+n}$, $n \geq 1$:

$$G_t^\lambda = \sum_{n=1}^{\infty} (1-\lambda)\lambda^{n-1} G_{t:t+n}.$$

Whereas the n-step return $G_{t:t+n}$ can be viewed as the target of an n-step TD update rule, the λ-return can be viewed as the target of the update rule for the TD(λ) prediction algorithm, which you will become familiar with in project 1. Note that the n-step return $G_{t:t+n}$ is functionally equivalent to the k-step estimator E_k from the lectures if we set $k = n$ and estimate the value of the state the agent is in at time t.

Consider the Markov reward process described by the following state diagram and assume the agent is in state 0 at time t (also assume the discount rate is $\gamma = 1$).



A Markov reward process can be thought of as an MDP with only one action possible from each state (denoted as action 0 in the figure above).

1.2 Procedure

- You will be given `p`, the probability of transitioning from state 0 to state 1, `V`, the estimate of the value function at time t, represented as a vector $[V(0), V(1), V(2), V(3), V(4), V(5), V(6)]$, and `rewards`, a vector of the rewards $[r_0, r_1, r_2, r_3, r_4, r_5, r_6]$ corresponding to the MDP.
- Your goal for this homework is to find a value of λ, strictly less than 1, such that the expected value of the λ-return equals the expected Monte-Carlo return at time t.
- Provide answers for the specific problems you are given on Canvas. Your answer must be correct to 3 decimal places, truncated (e.g. 3.14159265 becomes 3.141).

2 Examples

The following examples can be used to verify that your agent is implemented correctly.

- Input : `p` $= 0.81$; `V` $= [0.0, 4.0, 25.7, 0.0, 20.1, 12.2, 0.0]$; `rewards` $= [7.9, -5.1, 2.5, -7.2, 9.0, 0.0, 1.6]$
  Output : 0.622
- Input : `p` $= 0.22$; `V` $= [12.3, -5.2, 0.0, 25.4, 10.6, 9.2, 0.0]$; `rewards` $= [-2.4, 0.8, 4.0, 2.5, 8.6, -6.4, 6.1]$
  Output : 0.519
- Input : `p` $= 0.64$; `V` $= [-6.5, 4.9, 7.8, -2.3, 25.5, -10.2, 0.0]$; `rewards` $= [-2.4, 9.6, -7.8, 0.1, 3.4, -2.1, 7.9]$
  Output : 0.207

3 Resources

The concepts explored in this homework are covered by:

3.1 Lectures

- Lesson 3: TD and Friends

3.2 Readings

- Chapter 7 (7.1 n-step TD Prediction) and Chapter 12 (12.1 The λ-return) of R. S. Sutton and Barto 2020
- R. Sutton 1988

4 Submission Details

The due date is indicated on the Canvas page for this assignment.
Make sure you have set your timezone in Canvas to ensure the deadline is accurate.
Submit your answers on Canvas, as outlined in section 1.2. You will have a total of 10 submission attempts - only the highest score is kept.

References

[SB20] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. 2nd Ed. MIT press, 2020. URL: http://incompleteideas.net/book/the-book-2nd.html.

[Sut88] Richard Sutton. "Learning to Predict by the Method of Temporal Differences". In: *Machine Learning* 3 (Aug. 1988), pp. 9–44.