Reinforcement Learning
Exercise 9

Jim Mainprice, Philipp Kratzer
Machine Learning & Robotics lab, U Stuttgart
Universitätsstraße 38, 70569 Stuttgart, Germany

July 8, 2020

Submission Instructions:

The submission deadline for this exercise sheet is 14.07., 23:55.

Put your answers into a single pdf. Your python code should be a single python script. Upload both files to ilias. Make sure that the code runs with *python3 yourscript.py* without any errors.

Group submissions of up to three students are allowed.

1 The Intra-Option Policy Gradient (5P)

a) The option value function is given by

$$Q_\Omega(s,\omega) = \sum_a \pi_{\omega,\theta}(a|s) Q_U(s,\omega,a) \tag{1}$$

$$Q_U(s,\omega,a) = r(s,a) + \gamma \sum_{s'} P(s'|s,a) U(\omega,s') \tag{2}$$

What is it's derivative $\frac{\partial Q_\Omega(s,\omega)}{\partial\theta}$? (You do not need to solve $\frac{\partial \pi_{\omega,\theta}(a|s)}{\partial\theta}$ and $\frac{\partial U(\omega,s')}{\partial\theta}$). (2P)

b) The value U of executing ω upon entering s' is given by:

$$U(\omega,s') = (1-\beta_{\omega,\vartheta(s')}) Q_\Omega(s',\omega) + \beta_{\omega,\vartheta}(s') V_\Omega(s') \tag{3}$$

What is the derivative of $\frac{\partial U(\omega,s')}{\partial\theta}$? (You do not need to solve $\frac{\partial Q_\Omega(s',\omega')}{\partial\theta}$). (2P)

c) What is the problem for calculating the full derivative, if you plug $\frac{\partial U(\omega,s')}{\partial\theta}$ (task b)) into $\frac{\partial Q_\Omega(s,\omega)}{\partial\theta}$ (task a))? How could it be solved? (1P)

2 Implement The Option-Critic Algorithm on the 4-Room example (5P)

The code template can be found on github (`https://github.com/humans-to-robots-motion/rl-course`) in *ex09-hier/ex09-hier.py*. For this exercise we will use the Fourroom environment as described in the lecture. You can have a look on the implementation in *ex09-hier/fourrooms.py*

The implementation of the softmax policy and the sigmoid terminations as well as their gradients are already given in the code.

a) Implement the options evaluation step. Update both, Q_U and Q_Ω using a td target. (2P)

b) Implement the Options improvement step. (2P)

c) Let the algorithm run for 1000 episodes and put the plots for the episode lengths and termination probabilities into your submission pdf. (1P)