Reinforcement Learning
Exercise 8

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

June 23, 2020

Submission Instructions:

The submission deadline for this exercise sheet is 07.07., 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 REINFORCE on the Cart-Pole (9P)

The code template can be found on github (`https://github.com/humans-to-robots-motion/rl-course`) in *ex08-pg/ex08-pg.py*. For this exercise we will use the Cart-Pole environment from gym: `https://gym.openai.com/envs/CartPole-v1/` The task is to apply forces to a cart moving a long a track in order to keep the pole balanced. If the pole falls apart a given angle or an episode length of 500 is reached, the episode terminates. The state consists of 4 continous variables (position and velocity of cart and pole). There are 2 actions corresponding to left and right.

a) For discrete actions often a softmax action selection strategy is chosen:

$$\pi(a|s,\theta) = \frac{e^{h(s,a,\theta)}}{\sum_b e^{h(s,b,\theta)}}$$

Using simple linear features of the form $h(s,a,\theta) = \theta_a^\top s$ (with one set of parameters θ per action): Give the equation for $\pi(a|s,\theta)$ for the cart-pole (2 actions) and its derivative with respect to θ. (2P)

b) What is the equation of the gradient $\nabla_\theta \log \pi(A_t \mid S_t, \theta)$ for this example? (1P)

c) Implement the REINFORCE algorithm on the Cart-Pole example using the softmax action selection strategy. Track the mean of the 100 latest episode lengths. Tune the parameters and try to achieve a mean ≥ 495. How many episodes do you need? Plot the mean over the episode count and include the plot into your submission pdf. (4P)

d) Mention possibilities/extensions that you think could improve the performance of the algorithm. (2P)