

# Reinforcement Learning

## Exercise 9

Mathias Niepert, Vinh Tong

April 28, 2023

### 1 REINFORCE on the Cart-Pole (10P)

The code template can be found on Ilias in `ex09-pg/ex09-pg.py`. It has been tested with gym version 0.18.0 (but should also be stable with version 0.18.3). For this exercise we will use the Cart-Pole environment from gym ([https://www.gymlibrary.ml/environments/classic\\_control/cart\\_pole/](https://www.gymlibrary.ml/environments/classic_control/cart_pole/)). The task is to apply forces to a cart moving along 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 continuous 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  $\theta$  per action): Give the equation for  $\pi(a|s, \theta)$  for the cart-pole (2 actions) and its derivative with respect to  $\theta$ . (2P)

- b) What is the equation of the gradient  $\nabla_\theta \log \pi(A_t | S_t, \theta)$  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  $\geq 495$ . How many episodes do you need? Plot the mean over the episode count. (4P)
- d) Implement the *REINFORCE with baseline* algorithm. Use basic linear function approximation  $\hat{v}(s_t, w) = w^\top s_t$ . Compare the results to c). Which algorithm learns faster? (2P)
- e) Mention possibilities/extensions that you think could improve the performance of the algorithm. (1P)