Felix Bühler - 2973140
Jan Leusmann - 2893121
Jamie Ullerich - 3141241

Exercise 8

Task 1 - REINFORCE on the Cart-Pole

a)

Action a_1: $\pi(a_1|s,\theta) = \frac{e^{h(s,a_1,\theta)}}{e^{h(s,a_1,\theta)} + e^{h(s,a_2,\theta)}}$

Action a_2: $\pi(a_2|s,\theta) = \frac{e^{h(s,a_2,\theta)}}{e^{h(s,a_1,\theta)} + e^{h(s,a_2,\theta)}}$

Derivative:

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

$\rightarrow$ gradient:

$$\begin{aligned}\nabla_\theta \log(\pi(a|s,\theta)) &= \nabla_\theta \log(e^{h(s,a,\theta)}) - \nabla_\theta \log(\sum_b e^{h(s,b,\theta)}) \\ &= \nabla_\theta(h(s,a,\theta)) - \frac{\nabla_\theta \sum_b e^{h(s,b,\theta)}}{\sum_b e^{h(s,b,\theta)}} \\ &= \nabla_\theta(\theta_a^T s) - \nabla_\theta \sum_b \theta_b^T s \ \pi(b|s,\theta)\end{aligned}$$

b)

$$\begin{aligned}\nabla_\theta \log \pi(A_t|S_t,\theta) &= ... \\ &= x(s,a) - \sum_b (q_\pi(s,a) - b(s))\nabla(a|s,\theta)\end{aligned}$$

c)

After 2700 episodes it reached over the score 495.

d)

- REINFORCE with Baseline
- reduce learning rate over time
- different reward function where the amount of movement is lowering the reward

Felix Bühler - 2973140
Jan Leusmann - 2893121
Jamie Ullerich - 3141241

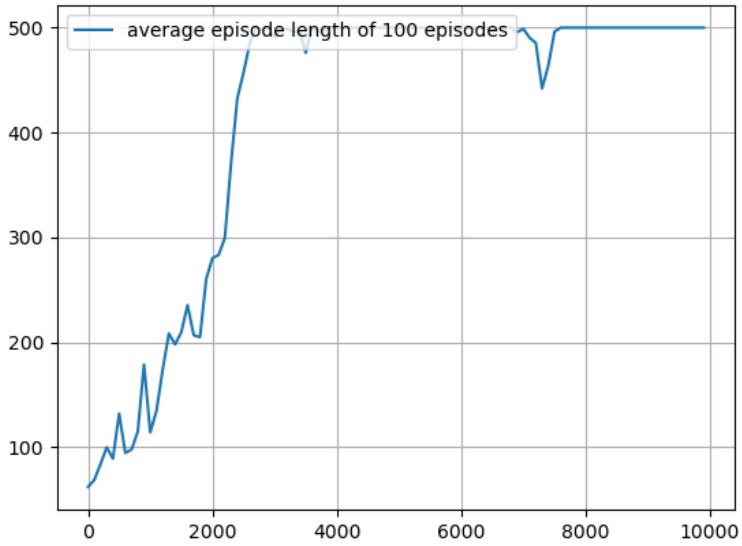

Figure 1: average episode lengths