

Exercise 8

Task 1 - REINFORCE on the Cart-Pole

a)

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

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

Derivative:

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

→ gradient:

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

b)

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

c)

d)

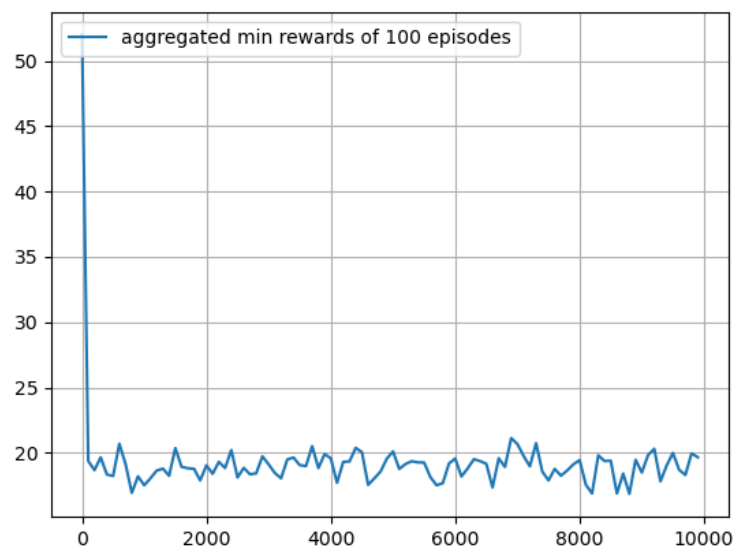


Figure 1: average episode lengths