Artificial Neural Networks (Gerstner). Exercises for week 11

Reinforcement Learning and the Brain

Exercise 1. A biological interpretation of the Advantage Actor-Critic with Eligibility traces

In this exercise you will show how applying Advantage Actor-Critic with eligibity traces to a softmax policy in combination with a linear read-out function leads to a biologically plausible learning rule.

Consider a policy and a value network as in Figure 1 with K input neurons $\{y_k = f(x - x_k)\}_{k=1}^{K}$. The policy network is parameterized by θ and has three output neurons corresponding to actions a_1, a_2 and a_3 with 1-hot coding. If $a_k = 1$ implies that action a_k is taken and we have $a_{k'} = 0$ for $k' \neq k$ The output neurons are sampled from a softmax policy: The probability of taking action a_i is given by

$$\pi_\theta(a_i = 1|x) = \frac{\exp\left(\sum_{k=1}^{K} \theta_{ik} y_k\right)}{\sum_j \exp\left(\sum_{k=1}^{K} \theta_{jk} y_k\right)}. \tag{1}$$

In addition, consider the exponential value network

$$\hat{v}_w(x) = \exp\left(\sum_{k=1}^{K} w_k y_k\right). \tag{2}$$

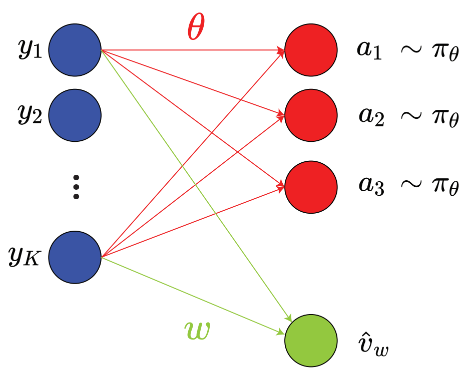

Figure 1: The network structure.

Assume the transition to state x^{t+1} with a reward of r^{t+1} after taking action a^t at state x^t. The learning rule for the Advantage Actor-Critic with Eligibility traces is

$$\begin{aligned}
\delta &\leftarrow r^{t+1} + \gamma \hat{v}_w(x^{t+1}) - \hat{v}_w(x^t) \\
z^w &\leftarrow \lambda^w z^w + \nabla_w \hat{v}_w(x^t) \\
z^\theta &\leftarrow \lambda^\theta z^\theta + \nabla_\theta \log \pi_\theta(a^t|x^t) \\
w &\leftarrow w + \alpha^w z^w \delta \\
\theta &\leftarrow \theta + \alpha^\theta z^\theta \delta
\end{aligned}$$

Your goal is to show that this learning rule applied to the network of Figure 1 has a biological interpretation.

a. Show that

$$\frac{d}{dw_5} \hat{v}_w(x^t) = y_5^t \hat{v}_w(x^t). \tag{3}$$

b. Interpret the update of the eligibity trace z_5^w in terms of a 'presynaptic factor' and a 'postsynaptic factor'. Can the rule be implemented in biology?

c. Show that

$$\frac{d}{d\theta_{35}} \log\left(\pi_\theta\left(a^t|x^t\right)\right) = \left(a_3^t - \pi_\theta\left(a_3 = 1|x^t\right)\right) y_5^t. \tag{4}$$

Hint: simply insert the softmax and then take the derivative.

d. Interpret the update of the eligibity trace z_{35}^{θ} in terms of a 'presynaptic factor' and a 'postsynaptic factor'. Can the rule be implemented in biology?

e. Interpret the update of the weights w_5 and θ_{35} in the framework of three factor learning rules. Can the rule be implemented in biology?