

# Artificial Neural Networks (Gerstner). Solutions 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 eligibility 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  $\theta$  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

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

In addition, consider the exponential value network

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

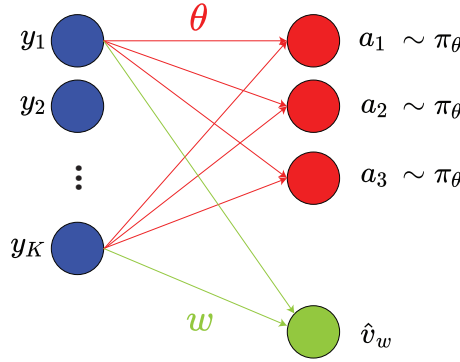


Figure 1: The network structure.

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

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

b. Interpret the update of the eligibility trace  $z_5^w$  in terms of a ‘presynaptic factor’ and a ‘postsynaptic factor’. Can the rule be implemented in biology?

c. Show that

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

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

- d. Interpret the update of the eligibility trace  $z_{35}^\theta$  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  $\theta_{35}$  in the framework of three factor learning rules. Can the rule be implemented in biology?

**Solution:**

a.

$$\frac{d}{dw_5} \hat{v}_w(x^t) = \frac{d}{dw_5} \exp\left(\sum_k w_k y_k\right) = y_5^t \exp\left(\sum_k w_k y_k\right) = y_5^t \hat{v}_w(x^t). \quad (5)$$

b. We have

$$z_5^w \leftarrow \lambda^w z_5^w + \frac{d}{dw_5} \hat{v}_w(x^t) = \lambda^w z_5^w + y_5^t \hat{v}_w(x^t). \quad (6)$$

The first term is a decay of the eligibility trace and is local (i.e. it is only function of  $z_5^w$ ). To interpret the 2nd term, we note that  $w_5$  connects the presynaptic neuron  $y_5$  in the input layer to the output of the value network  $\hat{v}_w(x^t)$ . Hence, the presynaptic factor is  $y_5^t$ , and the postsynaptic factor is  $\hat{v}_w(x^t)$ . Higher values of  $y_5^t$  and  $\hat{v}_w(x^t)$  lead to a greater increase of the eligibility trace  $z_5^w$ .

c. Assume that action  $i$  is taken at time  $t$ ; then we have  $a_i^t = \delta_{ji}$  for some  $i \in \{1, 2, 3\}$ , where  $\delta$  is the Kronecker delta. We first note that

$$\log(\pi_\theta(a^t|x^t)) = \log(\pi_\theta(a_i^t = 1|x^t)) = \sum_k \theta_{ik} y_k^t - \log\left(\sum_j \exp\left(\sum_k \theta_{jk} y_k^t\right)\right). \quad (7)$$

Therefore, we can compute the derivative as

$$\frac{d}{d\theta_{35}} \log(\pi_\theta(a_i^t = 1|x^t)) = \delta_{3i} y_5^t - \frac{\exp(\sum_k \theta_{3k} y_k^t)}{\sum_j \exp(\sum_k \theta_{jk} y_k^t)} y_5^t. \quad (8)$$

We then use [Equation 1](#) and the fact that  $a_3^t = \delta_{3i}$ :

$$\frac{d}{d\theta_{35}} \log(\pi_\theta(a^t|x^t)) = (a_3^t - \pi_\theta(a_3 = 1|x^t)) y_5^t. \quad (9)$$

d. We have

$$z_{35}^\theta \leftarrow \lambda^\theta z_{35}^\theta + \frac{d}{d\theta_{35}} \log(\pi_\theta(a^t|x^t)) = \lambda^\theta z_{35}^\theta + (a_3^t - \pi_\theta(a_3 = 1|x^t)) y_5^t. \quad (10)$$

The first term is a decay of the eligibility trace and is local (i.e. it is only function of  $z_{35}^\theta$ ). To interpret the 2nd term, we note that  $\theta_{35}$  connects the presynaptic neuron  $y_5$  in the input layer to the action neuron  $a_3$ . Hence, the presynaptic factor is  $y_5^t$ . The postsynaptic factor is  $(a_3^t - \pi_\theta(a_3 = 1|x^t))$ , where  $\pi_\theta(a_3 = 1|x^t)$  can be interpreted as the ‘drive’ or ‘membrane potential’ of the postsynaptic neuron  $a_3$  or, similarly, as its temporal average  $\langle a_3 \rangle$ .

Hence, if presynaptic and postsynaptic neuron are both active ( $a_3^t = 1$ ), the eligibility trace, after decay, is increased by an amount  $(a_3^t - \pi_\theta(a_3 = 1|x^t)) y_5^t$ . Second, if another action is taken, we have  $a_3^t = 0$ . Hence, the eligibility trace decreases by an amount which is proportional to  $y_5^t$  and  $\pi_\theta(a_3 = 1|x^t)$ .

Yes, the rule would be implementable in biology.

e. We have

$$\Delta w_5 = \alpha^w z_5^w \delta^t \quad (11)$$

$$\Delta \theta_{35} = \alpha^\theta z_{35}^\theta \delta^t \quad (12)$$

with  $\delta^t = r^{t+1} + \gamma \hat{v}_w(x^{t+1}) - \hat{v}_w(x^t)$  being the TD error. Hence, the weights get updated by an amount proportional to the global factor  $\delta^t$  and the value of their eligibility traces (i.e. their ‘flags’).

Yes, the rule would be implementable in biology.