Artificial Neural Networks (Gerstner). Solutions for week 6

From Policy Gradient to Actor-Critic

Exercise 1. Computer exercises: Environment 2 (part 2)[1]

Complete the computer exercise for environment 2.

Exercise 2. From Policy Gradient to eligibility traces

In this exercise you will show that eligibility traces appear naturally in any policy gradient algorithm. Eligibility traces are nice because they lead to a transparent and easy–to–interpret algorithm. Moreover, eligibility traces enable a direct online implementation of the algorithm in distributed hardware (or biology).

Consider a discrete multistep reinforcement learning problem with the usual graph, the usual notations and transitions: an action a_t leads you (stochastically) from state s_t to s_{t+1} and on this transition you collect the reward r_t. Suppose that you always start in state $s_{t=0} = s_{\text{start}}$. We assume that there is a simple terminal state s_{target}. You get a particularly strong positive reward when you reach s_{target}.

Your policy $\pi(a_t|s_t;\theta)$ depends on parameters θ. For the moment your aim is to optimize the parameters of the policy such that you maximize the expected discounted reward

$$\mathbb{E}_\theta[\text{Return}(s_{\text{start}} \to s_{\text{target}})] = \mathbb{E}_\theta[r_0 + \gamma r_1 + \gamma^2 r_2 + ...].$$

We proceed in five steps.

a. Derive a batch version of the policy gradient algorithm over multiple time steps by optimizing $\mathbb{E}_\theta[\text{Return}(s_{\text{start}} \to s_{\text{target}})]$ through gradient descent.

   Hint: Use the log-likelihood trick and take the derivative with respect to parameter θ_j.

b. A batch algorithm means averaging over many episodes. Transform the batch algorithm into an online algorithm where you consider one episode at a time. Assume that in one episode you traverse the state-action sequence: $s_0, a_0, r_0; s_1, a_1, r_1; s_2, a_2, r_2; s_3, a_3, r_3; s_4, a_4, r_4; s_5 = s_{\text{target}}$.

   Show that the parameter updates can be written as

$$\begin{aligned}
\Delta\theta_j = \quad & [r_0 + \gamma r_1 + \gamma^2 r_2 + \gamma^3 r_3 + \gamma^4 r_4] \frac{\partial}{\partial \theta_j} \log[\pi(a_0|s_0;\theta)] \\
+ \quad & [\gamma r_1 + \gamma^2 r_2 + \gamma^3 r_3 + \gamma^4 r_4] \frac{\partial}{\partial \theta_j} \log[\pi(a_1|s_1;\theta)] \\
+ \quad & [\gamma^2 r_2 + \gamma^3 r_3 + \gamma^4 r_4] \frac{\partial}{\partial \theta_j} \log[\pi(a_2|s_2;\theta)] \\
+ \quad & [\gamma^3 r_3 + \gamma^4 r_4] \frac{\partial}{\partial \theta_j} \log[\pi(a_3|s_3;\theta)] \\
+ \quad & \gamma^4 r_4 \frac{\partial}{\partial \theta_j} \log[\pi(a_4|s_4;\theta)] \qquad (1)
\end{aligned}$$

c. So far we were only interested in maximizing the discounted future reward from the INITIAL state, with the discount factor computed relative to that state ($t = 0$). However, while you move along the trajectory you pass by other states s_1, s_2, s_3, s_4. For each of these states s_t, you should now also optimize the future expected discounted reward starting from s_t; that is you want to maximize

$$\mathbb{E}_\theta[\text{Return}(s_t \to s_{\text{target}})] = \mathbb{E}_\theta[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + ...].$$

[1] Start this exercise in the first exercise session of week 6.

More generally, you should optimize the future discounted returns from every step t, assuming that the discounting started at the current step or at any possible step m in the past (i.e. $m \leq t$). Assume that m runs from $-\infty$ to t.

Redo the calculation in (b), but calculate the parameter update resulting from returns starting in arbitrary states.

Hint: Copy, but time-shift the results from (b).

d. Sum all the updates from (b) and (c) and reorder all terms such that updates that are multiplied with the same reward are grouped together.

Show that this results in updates of the form

$$\Delta\theta_j = \tag{2}$$

$$c \sum_t r_t \left[\frac{\partial}{\partial \theta_j} \log[\pi(a_t|s_t;\theta)] + \gamma \frac{\partial}{\partial \theta_j} \log[\pi(a_{t-1}|s_{t-1};\theta)] + \gamma^2 \frac{\partial}{\partial \theta_j} \log[\pi(a_{t-2}|s_{t-2};\theta)] + ... \right] \tag{3}$$

with some constant c. What is this constant?

e. Now we introduce eligibility traces by defining for each parameter θ_j a 'shadow variable' z_j which, in each time step t, decreases by a factor $\lambda < 1$

$$z_j \longleftarrow \lambda z_j \tag{4}$$

and then (in the same time step) increase by an amount

$$z_j \longleftarrow \frac{\partial}{\partial \theta_j} \log[\pi(a_t|s_t;\theta)] \tag{5}$$

where a_t is the action taken in time step t.

What is the relation of λ and γ? What is the final weight update?

f. Suppose that all rewards are zero, except the reward in the final time step $r_4 > 0$. Furthermore suppose that parameter θ is only sensitive to a_2, s_2. To be specific, say $\frac{\partial}{\partial \theta_j} \log[\pi(a_2|s_2;\theta)] > 0$ and $\frac{\partial}{\partial \theta_j} \log[\pi(a_t|s_t;\theta)] = 0$ for $t \neq 2$.

How can you interpret the resulting algorithm? How much will the parameter θ_j change?

Solution:

a. We show the total discounted reward from time $t = 0$ by

$$G_0 = r_0 + \gamma r_1 + \gamma^2 r_2 + \ldots + \gamma^T r_T,$$

where we assume that the episode had T steps. Our goal is to maximize the expected return under the current policy

$$V_\theta(s_0) = \mathbb{E}_\theta[G_0|s_0].$$

We can use the law of total expectation and write

$$\begin{aligned} \mathbb{E}_\theta[G_0|s_0] &= \mathbb{E}_\theta\left[\mathbb{E}[G_0|s_{0:T}, a_{0:T-1}]|s_0\right] \\ &= \int \mathbb{E}[G_0|s_{0:T}, a_{0:T-1}] \prod_{\tau=0}^{T-1} p(s_{\tau+1}|a_\tau, s_\tau)\pi(a_\tau|s_\tau;\theta)\, da_{0:T-1}\, ds_{1:T}, \end{aligned}$$

where we define $a_{0:T-1} = \{a_0, \ldots, a_{T-1}\}$ and $s_{0:T} = \{s_0, \ldots, s_T\}$ and note that the term $\mathbb{E}[G_0|s_{0:T}, a_{0:T-1}]$ does not depend on the parameters θ.

We can now use log-likelihood trick and write

$$\frac{\partial \mathbb{E}_\theta[G_0|s_0]}{\partial \theta_j} = \int \mathbb{E}[G_0|s_{0:T}, a_{0:T-1}] \frac{\partial}{\partial \theta_j} \prod_{\tau=0}^{T-1} p(s_{\tau+1}|a_\tau, s_\tau)\pi(a_\tau|s_\tau;\theta)\, da_{0:T-1}\, ds_{1:T} \tag{6}$$

$$= \int \mathbb{E}[G_0|s_{0:T}, a_{0:T-1}] \prod_{\tau=0}^{T-1} p(s_{\tau+1}|a_\tau, s_\tau)\pi(a_\tau|s_\tau;\theta) \left[\sum_{\tau=0}^{T-1} \frac{\partial}{\partial \theta_j} \log \pi(a_\tau|s_\tau;\theta)\right] da_{0:T-1}\, ds_{1:T} \tag{7}$$

$$= \mathbb{E}_\theta \left[G_0 \sum_{\tau=0}^{T-1} \frac{\partial}{\partial \theta_j} \log \pi(a_\tau|s_\tau;\theta) \Big| s_0 \right]. \tag{8}$$

For a given t, the action a_t given state s_t is independent of the reward values $r_0, \ldots, r_{t-1}$. This implies that, for $\tau < t$,

$$\mathbb{E}_\theta \left[r_\tau \frac{\partial}{\partial \theta_j} \log \pi(a_t|s_t;\theta) \Big| s_0 \right] = \int \left(r_\tau \frac{\partial}{\partial \theta_j} \log \pi(a_t|s_t;\theta) \right) p_\theta(r_\tau, s_t, a_t|s_0)\, dr_\tau\, ds_t\, da_t,$$

$$= \int \left(r_\tau \frac{\partial}{\partial \theta_j} \log \pi(a_t|s_t;\theta) \right) p_\theta(r_\tau|s_0) p_\theta(s_t|r_\tau, s_0) \pi(a_t|s_t;\theta)\, dr_\tau\, ds_t\, da_t,$$

$$= \int r_\tau p_\theta(r_\tau|s_0) p_\theta(s_t|r_\tau, s_0) \underbrace{\left(\int \pi(a_t|s_t;\theta) \frac{\partial}{\partial \theta_j} \log \pi(a_t|s_t;\theta)\, da_t \right)}_{= \frac{\partial}{\partial \theta_j} \int \pi(a_t|s_t;\theta) da_t = \frac{\partial}{\partial \theta_j} \cdot 1 = 0} dr_\tau\, ds_t = 0.$$

Hence, interestingly, we can write

$$\mathbb{E}_\theta \left[G_0 \frac{\partial}{\partial \theta_j} \log \pi(a_t|s_t;\theta) \Big| s_0 \right] = \mathbb{E}_\theta \left[\left(r_0 + \gamma r_1 + \gamma^2 r_2 + \ldots. + \gamma^T r_T\right) \frac{\partial}{\partial \theta_j} \log \pi(a_t|s_t;\theta) \Big| s_0 \right]$$

$$= \mathbb{E}_\theta \left[\left(\gamma^t r_t + \gamma^{t+1} r_{t+1} + \ldots. + \gamma^T r_T\right) \frac{\partial}{\partial \theta_j} \log \pi(a_t|s_t;\theta) \Big| s_0 \right]$$

$$= \mathbb{E}_\theta \left[\gamma^t G_t \frac{\partial}{\partial \theta_j} \log \pi(a_t|s_t;\theta) \Big| s_0 \right].$$

Therefore, Equation 8 can be simplified further as

$$\frac{\partial \mathbb{E}_\theta[G_0|s_0]}{\partial \theta_j} = \mathbb{E}_\theta \left[\sum_{\tau=0}^{T-1} \gamma^\tau G_\tau \frac{\partial}{\partial \theta_j} \log \pi(a_\tau|s_\tau;\theta) \Big| s_0 \right].$$

To do the batch update, we run M episodes. We use s_t^i and a_t^i to denote the state and the selected action at time t in episode i and use G_t^i to denote the discounted return collected from time t onwards in episode i. Therefore, we have

$$\Delta\theta_j = \frac{1}{M} \sum_{i=1}^{M} \sum_{t=0}^{T_i-1} \gamma^t G_t^i \frac{\partial}{\partial \theta_j} \log \pi(a_t^i|s_t^i;\theta)$$

b. Transforming the batch algorithm into an online algorithm can be done by simply removing the averaging over M, i.e.

$$\Delta\theta_j = \sum_{t=0}^{T-1} \gamma^t G_t \frac{\partial}{\partial \theta_j} \log \pi(a_t|s_t;\theta).$$

For the given episode, we have

$$\Delta\theta_j = \gamma^0 G_0 \frac{\partial}{\partial \theta_j} \log \pi(a_0|s_0;\theta) + \gamma^1 G_1 \frac{\partial}{\partial \theta_j} \log \pi(a_1|s_1;\theta)$$
$$+ \gamma^2 G_2 \frac{\partial}{\partial \theta_j} \log \pi(a_2|s_2;\theta) + \gamma^3 G_3 \frac{\partial}{\partial \theta_j} \log \pi(a_3|s_3;\theta) + \gamma^4 G_4 \frac{\partial}{\partial \theta_j} \log \pi(a_4|s_4;\theta).$$

Evaluating $G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} \ldots$ gives the result above.

c. Optimizing for the returns starting from an arbitrary step m on the trajectory gives us

$$\Delta\theta_j^m = \sum_{t=m}^{T-1} \gamma^{t-m} G_t \frac{\partial}{\partial\theta_j} \log \pi(a_t|s_t;\theta).$$

d. Summing over all possible values of m gives us

$$\begin{aligned}\Delta\theta_j = \sum_{m=-\infty}^{T-1} \Delta\theta_j^m &= \sum_{m=-\infty}^{T-1} \sum_{t=m}^{T-1} \gamma^{t-m} G_t \frac{\partial}{\partial\theta_j} \log \pi(a_t|s_t;\theta) \\ &= \sum_{t=-\infty}^{T-1} \sum_{m=-\infty}^{t} \gamma^{t-m} G_t \frac{\partial}{\partial\theta_j} \log \pi(a_t|s_t;\theta)\end{aligned}$$

which can then be simplified further

$$\begin{aligned}\Delta\theta_j &= \sum_{t=-\infty}^{T-1} \sum_{m=0}^{\infty} \gamma^{m} G_t \frac{\partial}{\partial\theta_j} \log \pi(a_t|s_t;\theta) \\ &= \frac{1}{1-\gamma} \sum_{t=-\infty}^{T-1} G_t \frac{\partial}{\partial\theta_j} \log \pi(a_t|s_t;\theta).\end{aligned}$$

We can now replace G_t by $\sum_{\tau=t}^{T} \gamma^{\tau-t} r_\tau$ and write

$$\begin{aligned}\Delta\theta_j &= \frac{1}{1-\gamma} \sum_{t=-\infty}^{T-1} \sum_{\tau=t}^{T} \gamma^{\tau-t} r_\tau \frac{\partial}{\partial\theta_j} \log \pi(a_t|s_t;\theta) \\ &= \frac{1}{1-\gamma} \sum_{\tau=-\infty}^{T} r_\tau \sum_{t=-\infty}^{\tau} \gamma^{\tau-t} \frac{\partial}{\partial\theta_j} \log \pi(a_t|s_t;\theta),\end{aligned}$$

where we assumed a dummy action a_T with $\frac{\partial}{\partial\theta_j} \log \pi(a_T|s_T;\theta) = 0$. The expression above can be re-written as, with a change of variable $n = \tau - t$,

$$\Delta\theta_j = \frac{1}{1-\gamma} \sum_{\tau=-\infty}^{T} r_\tau \left(\sum_{n=0}^{\infty} \gamma^n \frac{\partial}{\partial\theta_j} \log \pi(a_{\tau-n}|s_{\tau-n};\theta) \right), \tag{9}$$

which is identical to the expression in the exercise with $c = \frac{1}{1-\gamma}$.

e. We can expand the shadow variables as

$$\begin{aligned}z_j^t &= \lambda z_j^{t-1} + \frac{\partial}{\partial\theta_j} \log \pi(a_t|s_t;\theta) \\ &= \lambda^2 z_j^{t-2} + \lambda \frac{\partial}{\partial\theta_j} \log \pi(a_{t-1}|s_{t-1};\theta) + \frac{\partial}{\partial\theta_j} \log \pi(a_t|s_t;\theta) \\ &= \sum_{n=0}^{\infty} \lambda^n \frac{\partial}{\partial\theta_j} \log \pi(a_{t-n}|s_{t-n};\theta).\end{aligned}$$

With $\gamma = \lambda$, we note that this is equivalent to the last sum in Equation 9. In this case, we can express the policy gradient update using our shadow variables as

$$\Delta\theta_j = \frac{1}{1-\gamma} \sum_{t=-\infty}^{T} r_t z_j^t \tag{10}$$

f. In this case, Equation 10 simplifies to

$$\Delta\theta_j = \frac{1}{1-\gamma} r_4 z_j^4 = \frac{1}{1-\gamma} r_4 \sum_{n=0}^{4} \gamma^n \frac{\partial}{\partial\theta_j} \log \pi(a_{4-n}|s_{4-n};\theta) = \frac{\gamma^2}{1-\gamma} r_4 \frac{\partial}{\partial\theta_j} \log \pi(a_2|s_2;\theta).$$

Since it is assumed that $\frac{\partial}{\partial\theta_j} \log \pi(a_2|s_2;\theta) > 0$, an increase in the value of the parameter θ_j will increase the probability of taking a_2 in s_2 again. In addition, since $r_4 > 0$, all terms are positive and the value of θ_j will increase.

The magnitude of increase depends on the magnitude of r_4. In other words, θ_j will increase more if it contributed to a larger reward, due to its effect on the policy 2 steps before receiving the reward.

The magnitude of increase also depends on $\frac{\gamma^2}{1-\gamma}$. If the discount factor γ is small, it suggests that earlier actions contribute little to later rewards; as a result, the gradient will also be small since it relates to the policy several steps before actually receiving the reward.

Exercise 3. Recap and prepration for the next week: Why target networks help

States $s^{(j)}$ are represented by three-dimensional vectors $(s_1^{(j)}, s_2^{(j)}, s_3^{(j)})$. Actions are labeled by a 1-dimensional index $a = \{1, 2\}$. We look at semi-gradient Q-learning with linear function approximation, i.e. $Q(s^{(j)}, a) = \sum_{i=1}^{3} w_{ai} s_i^{(j)}$. We start with $w_{ai} = 0$ for all a and i.

Assume we observe state $s^{(1)} = (1, 1, 0)$, take action $a = 1$, receive reward $r = 1$ and observe the next state $s^{(2)} = (0, 1, 1)$.

a. Compute $Q(s^{(1)}, 1)$ with the semi-gradient learning rule $\Delta w_{ai} = \eta(r + \gamma \max_{a'} Q(s', a') - Q(s^{(1)}, a)) s_i^{(1)}$ with $\gamma = 1$ and $\eta = 0.1$.

b. Show that $Q(s^{(2)}, 1)$ has also changed.

c. Assume $\hat{Q}(s, a) = \sum_i w_{ai} s_i + \epsilon$, where ϵ is a Gaussian noise term with mean 0 and variance σ^2. Show that $\langle \max_a \hat{Q}(s, a) \rangle > \max_a \langle \hat{Q}(s, a) \rangle$.

Hint: Evaluations are for fixed state s. Expectations run over the Gaussian variable ϵ. The noise term ϵ is drawn independently for each action. Exploit that the mean of the Gaussian vanishes and that expectations can be easily evaluated for linear operators.

Solution:

a. $\Delta w_{11} = 0.1 \cdot (1 + 1 \max_{a'} 0 - 0) \cdot 1 = 0.1$, similarly $\Delta w_{12} = 0.1$, $\Delta w_{13} = 1 \cdot (1 + 1 \max_{a'} 0 - 0) \cdot 0 = 0$. With these updates we get $Q(s^{(1)}, 1) = \sum_i w_{1i} s_i^{(1)} = 0.2$

b. $Q(s^{(2)}, 1)$ was 0 before the update and is now $Q(s^{(2)}, 1) = \sum_i w_{1i} s_i^{(2)} = 0.1$.

c. Let's call the maximal expected Q-value $Q(s, a^*) = \max_a \langle \hat{Q}(s, a) \rangle$. If the noise terms where always such that $\arg\max_a \hat{Q}(s, a) = a^*$, $\langle \max_a \hat{Q}(s, a) \rangle$ would be equal to $Q(s, a^*) = \max_a \langle \hat{Q}(s, a) \rangle$. However, for all cases where $\arg\max \hat{Q}(s, a) = \hat{a} \neq a^*$ we have $\hat{Q}(s, \hat{a}) > \hat{Q}(s, a^*)$ and averaging both sides, we conclude: $\langle \max_a \hat{Q}(s, a) \rangle > Q(s, a^*)$.