

# Artificial Neural Networks (Gerstner). Solutions for week 12

## Intrinsically motivated exploration

### Exercise 1. How fast can we find the goal state with a stationary policy?

Consider an environment with the state space  $\mathcal{S}$ , a goal (terminal) state  $G \in \mathcal{S}$ , and an action space  $\mathcal{A}$  in non-goal states (i.e.,  $\mathcal{S} - \{G\}$ ). After taking action  $a \in \mathcal{A}$  in state  $s \in \mathcal{S}$ , the agent moves to state  $s' \in \mathcal{S}$  with the transition probability  $p(s'|s, a)$ . These transition probabilities are unknown to the agent. We use  $T$  to denote the first time an agent find the goal state  $G$ , i.e.,  $s_T = G$ . If we assume that the agent uses a stationary policy  $\pi$ , then we can define the average of  $T$  given each initial state  $s \in \mathcal{S}$  as

$$\mu_\pi(s) := \mathbb{E}_\pi[T | s_0 = s],$$

where  $s_0$  is the state at time  $t = 0$ . In this exercise, we study  $\mu_\pi(s)$  in its most general case.

- a. What is the value of  $\mu_\pi(G)$ ?

*Hint:* Note that  $T$  is equal to the smallest  $t \geq 0$  when we have  $s_t = G$ .

- b. What is the relationship between  $\mathbb{E}_\pi[T | s_1 = s]$  and  $\mu_\pi(s)$ ?

*Hint:* Note that  $\mu_\pi(s)$  is the average of  $T$  if the agent starts in state  $s$  at time  $t = 0$ , whereas  $\mathbb{E}_\pi[T | s_1 = s]$  is the average of  $T$  if the agent starts in state  $s$  at time  $t = 1$ .

- c. Find a system of linear equations for finding  $\mu_\pi(s)$  for  $s \in \mathcal{S} - \{G\}$ .

*Hint:* Use the fact that  $p_\pi(s'|s) = \sum_{a \in \mathcal{A}} \pi(a|s)p(s'|s, a)$ .

### Solution:

- a. By definition, we have  $\mu_\pi(G) = 0$ .

- b. Using the Markovian property of the environment, we have

$$\mathbb{E}_\pi[T | s_1 = s] = 1 + \mathbb{E}_\pi[T | s_0 = s] = 1 + \mu_\pi(s).$$

- c. We use the law of total expectation as well as the Markovian property of the environment and write

$$\mu_\pi(s) = \mathbb{E}_\pi[T | s_0 = s] = \mathbb{E}_\pi \left[ \mathbb{E}_\pi [T | s_1] \mid s_0 = s \right] = \sum_{s' \in \mathcal{S}} p_\pi(s'|s) \mathbb{E}_\pi [T | s_1 = s'] .$$

We can then use part b and write

$$\mu_\pi(s) = 1 + \sum_{s' \in \mathcal{S}} p_\pi(s'|s) \mu_\pi(s') = 1 + \sum_{s' \in \mathcal{S} - \{G\}} p_\pi(s'|s) \mu_\pi(s'), \quad (1)$$

where we used the fact that  $\mu_\pi(G) = 0$ .

### Exercise 2. The magic of seeking novelty.

Consider a special case of the environment in [Exercise 1](#) with  $N + 2$  states:  $\mathcal{S} = \{0, 1, \dots, N, G\}$ , where  $G$  is the goal (terminal) state. At each non-goal state  $s \in \{0, \dots, N\}$ , two actions  $a$  and  $a'$  are available that connect different states through deterministic transitions shown in [Figure 1](#). In this exercise, we study how fast an agent that does not know the environment's structure can find the goal state  $G$ .

**Part I. Random exploration.** First, we consider purely random exploration:  $\pi(a|s) = \pi(a'|s) = 0.5$ . Because of the particular structure of the environment in [Figure 1](#), solving the system of linear equations that you found in [Exercise 1](#) for  $\mu_\pi(s)$  becomes exceptionally easy:

- a. Find  $\mu_\pi(N)$  as a function of  $\mu_\pi(0)$ .

*Hint:* Use the system of linear equations you that found in [Exercise 1c](#).

- b. Find  $\mu_\pi(n)$ , for  $n < N$  as a function of  $\mu_\pi(0)$  and  $n$ .

*Hint:* Repeatedly apply the trick of part a for state  $N - 1$ ,  $N - 2$ , down to  $n < N$ .

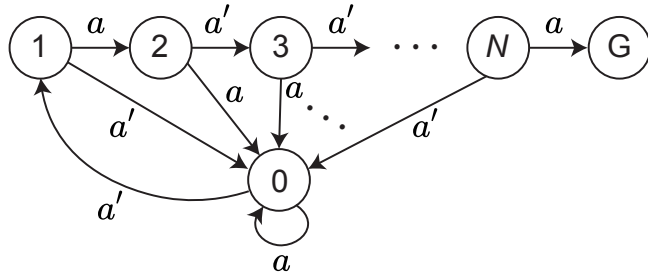


Figure 1: Environment of [Exercise 2](#)

c. Find  $\mu_\pi(0)$  as a function of  $N$ . How does it scale with  $N$  for large  $N$ ?

*Hint:* Use part b and write  $\mu_\pi(0)$  as a function of itself. Then solve the equation.

**Part II. Novelty-seeking.** To gain intuition about novelty-seeking, we consider a simple cartoon example: We assume

- The state space is very big, i.e.,  $N \gg 1$ .
- The agent starts in state 0 and explores the environment for  $T_0 \ll \mu_\pi(0)$  steps with *random exploration*.
- The agent does not find the goal state in these  $T_0$  steps of random exploration.
- $s_{T_0} = 0$ .

By the end of the initial  $T_0$  steps of random exploration, the agent has encountered state 0 many times, so the novelty of state 0 is on average much smaller than novelty of other states. This implies that, at the end of the initial  $T_0$  steps of random exploration, state 0 is considered as a ‘bad’ state by an agent that seeks novelty.

Starting from  $t = T_0$ , we consider the following simple *novelty-seeking* policy:

- $t \leftarrow T_0$
- While  $s_t \neq G$ :
  - If it is the first time in state  $s_t$  **after** the first  $T_0$  steps:
    - \* Pick action  $a_t \in \{a, a'\}$  at random.
    - \* Observe state  $s_{t+1}$ .
    - \* If  $s_{t+1} = 0$ 
      - $a_{\text{bad}}(s_t) \leftarrow a_t$  and  $a_{\text{good}}(s_t) \leftarrow !a_t$ ,  
where  $!a_t$  is the non-chosen action (e.g., if  $a_t = a$ , then  $!a_t = a'$ ).
    - else
      - $a_{\text{good}}(s_t) \leftarrow a_t$  and  $a_{\text{bad}}(s_t) \leftarrow !a_t$ ,  
where  $!a_t$  is the non-chosen action (e.g., if  $a_t = a$ , then  $!a_t = a'$ ).
  - If it is **not** the first time in state  $s_t$  **after** the first  $T_0$  steps:
    - \* Pick action  $a_t = a_{\text{good}}(s_t)$ .
    - \* Observe state  $s_{t+1}$ .
  - $t \leftarrow t + 1$

Let  $T(s) \geq T_0$  be the 1st time after  $T_0$  that the agent visit state  $s$ , e.g.,  $T(0) = T_0$ .

a. For  $n \in \{1, \dots, N\}$ , what is the minimum value of  $T(n)$  for the novelty-seeking policy described above? We denote this value  $T_{\min}(n)$ .

*Hint:*  $T_{\min}(n)$  corresponds to the case where the random action-selection step of novelty-seeking always picks the ‘good’ action.

b. For  $n \in \{1, \dots, N\}$ , what is the maximum value of  $T(n)$  for the novelty-seeking policy described above? We denote this value  $T_{\max}(n)$ .

*Hint:*  $T_{\max}(n)$  corresponds to the case where the random action-selection step of novelty-seeking always picks the ‘bad’ action.

- c. Find the corresponding values for  $T_{\min}(G)$  and  $T_{\max}(G)$ . How do these values scale with  $N$  for large  $N$ ? Compare your results with the scaling of  $\mu_\pi(0)$  for random exploration.

**Solution:**

**Part I. Random exploration.**

- a. Using Equation 1, we have

$$\mu_\pi(N) = 1 + \frac{\mu_\pi(0)}{2}.$$

- b. By repeating the same procedure as in a for  $N - 1$ ,  $N - 2$ , down to  $n < N$ , we have

$$\mu_\pi(n) = \left(1 + \frac{\mu_\pi(0)}{2}\right) \left(1 + \frac{1}{2} + \dots + \frac{1}{2^{N-n}}\right) = (2 + \mu_\pi(0)) \left(1 - \frac{1}{2^{N+1-n}}\right).$$

- c. Using the result of b, we have

$$\mu_\pi(0) = (2 + \mu_\pi(0)) \left(1 - \frac{1}{2^{N+1}}\right) \Rightarrow \mu_\pi(0) = 2^{N+2} - 2 = \mathcal{O}(e^{N \log 2}).$$

It scales exponentially.

**Part II. Novelty seeking.**

- a. The minimum value of  $T(n)$  corresponds to the case where the random action-selection step of novelty-seeking always picks the ‘good’ action, resulting in the sequence of states

$$0 \rightarrow 1 \rightarrow 2 \rightarrow 3 \rightarrow \dots \rightarrow n.$$

Hence, we have

$$T_{\min}(n) = T_0 + n.$$

- b. The maximum value of  $T(n)$  corresponds to the case where the random action-selection step of novelty-seeking always picks the ‘bad’ action, resulting in the sequence of states

$$0 \xrightarrow{\text{2 steps}} 0 \rightarrow 1 \xrightarrow{\text{3 steps}} 0 \rightarrow 1 \rightarrow 2 \xrightarrow{\text{4 steps}} 0 \rightarrow 1 \rightarrow 2 \rightarrow 3 \rightarrow 0 \rightarrow \dots \rightarrow n.$$

Hence, we have

$$T_{\max}(n) = T_0 + 2 + 3 + \dots + (n + 1) = \frac{n(n + 3)}{2} = T_0 + \frac{n^2 + 3n}{2}.$$

- c. Using the structure in Figure 1, we have

$$T_{\min}(G) = T_{\min}(N + 1) = T_0 + N + 1 = \mathcal{O}(N),$$

and

$$T_{\max}(G) = T_{\max}(N + 1) = T_0 + \frac{(N + 1) \cdot (N + 4)}{2} = T_0 + \frac{N^2 + 5N + 4}{2} = \mathcal{O}(N^2).$$

Switching from random exploration to novelty-seeking decreases the average search time of  $\mathcal{O}(e^{N \log 2})$  to a maximum search time of  $\mathcal{O}(N^2)$