Artificial Neural Networks (Gerstner). Exercises 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-gaol 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 π, 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 realtionship 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 $p_\pi(s'|s) = \sum_{a \in \mathcal{A}} \pi(a|s)p(s'|s,a)$.

Exercise 2. The magic of seeking novelty.

Consider a special case of the environment in Exercise 1 with $N+2$ states: $\mathcal{S} = \{0, 1, \ldots, N, G\}$, where G is the goal (terminal) state. At each non-goal state $s \in \{0, \ldots, N\}$, 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

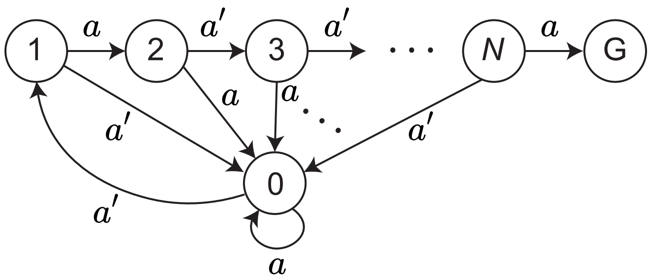

Figure 1: Environment of Exercise 2

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 $n < N$.

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., $N >> 1$.

- The agent starts in state 0 and explores the environment for $T_0 << \mu_\pi(0)$ 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 $n \in \{1, \ldots, N\}$, 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 $n \in \{1, \ldots, N\}$, 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.