

**Exercise 1. (Gridworld)**

Consider the gridworld setup, in which an agent (for example, a robot) is moving in a 2D plane. The plane is modeled by a discrete grid, as shown in Figure 1. We refer to the bottom left corner as  $(1,1)$  and the top right corner as  $(3,3)$ . The initial distribution is  $\rho((1,1)) = 1$ , which means that the agent starts in cell  $(1,1)$  with probability 1. The agent can choose from four actions:  $\mathcal{A} = \{\text{'up'}, \text{'down'}, \text{'left'}, \text{'right'}\}$ . When the agent arrives at cell  $(3,1)$  for any action  $a \in \mathcal{A}$ , the agent receives a reward of 1. When the agent arrives at cell  $(3,2)$  for any action  $a \in \mathcal{A}$ , the agent receives a reward of 8. The gray walls and the gridworld boundary block the agent's path, specifically the wall on cell  $(2,2)$ . The agent's actions do not always go as planned:

- 1) 85% of the time, the agent takes the intended action.
  - 2) 15% of the time, the agent uniformly takes any of the other three actions.
  - 3) If there is a wall in the direction the agent would have taken, the agent remains in place.
- Once the agent reaches a rewarded cell, it stays there forever, i.e., those are terminal states.

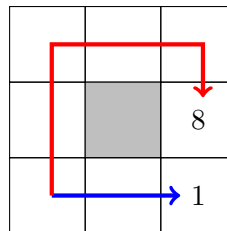


Figure 1: Gridworld

1. For the states  $s = (1,1)$  and  $s = (1,2)$ , determine the transition probabilities  $P(\cdot|s, a)$  for any  $a \in \mathcal{A}$ .

*Solution:* For the state  $s = (1, 1)$ , the transition probabilities are

| $P(s' s, a)$  | $a = \text{'up'}$ | $a = \text{'down'}$ | $a = \text{'left'}$ | $a = \text{'right'}$ |
|---------------|-------------------|---------------------|---------------------|----------------------|
| $s' = (1, 1)$ | 0.1               | 0.9                 | 0.9                 | 0.1                  |
| $s' = (1, 2)$ | 0.85              | 0.05                | 0.05                | 0.05                 |
| $s' = (2, 1)$ | 0.05              | 0.05                | 0.05                | 0.85                 |

For the state  $s = (1, 2)$ , the transition probabilities are

| $P(s' s, a)$  | $a = \text{'up'}$ | $a = \text{'down'}$ | $a = \text{'left'}$ | $a = \text{'right'}$ |
|---------------|-------------------|---------------------|---------------------|----------------------|
| $s' = (1, 2)$ | 0.1               | 0.1                 | 0.9                 | 0.9                  |
| $s' = (1, 3)$ | 0.85              | 0.05                | 0.05                | 0.05                 |
| $s' = (1, 1)$ | 0.05              | 0.85                | 0.05                | 0.05                 |

2. Using direct parametrization of the policy,  $\pi_\theta(a|s) = \theta_{s,a}$ , how many parameters are there? What is the possible range for each parameter?

*Solution:* We need a parameter for each state action pair  $(s, a) \in \mathcal{S} \times \mathcal{A}$ . There are 8 states ( $\mathcal{S} = \{(1, 1), (1, 2), (1, 3), (2, 1), (2, 3), (3, 1), (3, 2), (3, 3)\}$ ) and four actions. Hence, there are  $|\mathcal{S}||\mathcal{A}| = 32$  parameters.

For each parameter  $\theta_{s,a}$ , we need to ensure it satisfies the following condition:

$$0 \leq \theta_{s,a} \leq 1, \sum_{a' \in \mathcal{A}} \theta_{s,a'} = 1.$$

*Comment:* It would also be correct to use 36 parameters where we consider 9 states including the gray cell.

3. Using softmax parametrization of the policy,  $\pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})}$ , how many parameters are there?

*Solution:* Similarly to Problem 2, we need 32 parameters (or 36 when counting the gray grid cell as a state).

4. From now on, we will always assume the policy is parameterized with softmax. We initialize the parameters as  $\theta = \mathbf{0} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ . What is the initial policy?

*Solution:* It is a uniform distribution over the action set  $\{\text{'up'}, \text{'down'}, \text{'left'}, \text{'right'}\}$  as

$$\pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})} = \frac{\exp(0)}{\sum_{a' \in \mathcal{A}} \exp(0)} = \frac{1}{4}.$$

5. Next, we consider two trajectories sampled from the initial softmax policy and truncated to a horizon  $H = 7$ . The first trajectory  $\tau_1$  is given by

$$(1, 1) \xrightarrow{\text{up}} (1, 2) \xrightarrow{\text{up}} (1, 3) \xrightarrow{\text{right}} (2, 3) \xrightarrow{\text{right}} (3, 3) \xrightarrow{\text{right}} (3, 3) \xrightarrow{\text{down}} (3, 2) \xrightarrow{\text{down}} (3, 2) \xrightarrow{\text{left}} .$$

The second trajectory  $\tau_2$  is given by

$$(1, 1) \xrightarrow{\text{right}} (2, 1) \xrightarrow{\text{right}} (1, 1) \xrightarrow{\text{right}} (2, 1) \xrightarrow{\text{right}} (3, 1) \xrightarrow{\text{right}} (3, 1) \xrightarrow{\text{right}} (3, 1) \xrightarrow{\text{right}} (3, 1) \xrightarrow{\text{up}} .$$

For each trajectory  $\tau := \{s_0, a_0, s_1, a_1, \dots, s_H, a_H\}$ , the discounted reward is computed as  $R(\tau) := \sum_{t=0}^H \gamma^t r(s_t, a_t)$ , where  $r(s_t, a_t)$  denotes the reward at each step along the trajectory. The probability of choosing  $\tau$  is  $\Pr(\tau) = \rho(s_0) \pi_\theta(a_0|s_0) \prod_{t=1}^{H-1} P(s_t|s_{t-1}, a_{t-1}) \pi(a_t|s_t)$ . What are the probabilities of choosing  $\tau_1$  and  $\tau_2$ ? What are the discounted rewards for these trajectories?

Continue here

*Solution:* The probability of choosing the first trajectory  $\tau_1$  is

$$\begin{aligned}
 \Pr(\tau_1) = & \underbrace{\rho((1,1))}_1 \underbrace{\pi(\text{up}|(1,1))}_{0.25} \underbrace{P((1,2)|(1,1), \text{up})}_{0.85} \underbrace{\pi(\text{up}|(1,2))}_{0.25} \underbrace{P((1,3)|(1,2), \text{up})}_{0.85} \underbrace{\pi(\text{right}|(1,3))}_{0.25} \times \\
 & \times \underbrace{P((2,3)|(1,3), \text{right})}_{0.85} \underbrace{\pi(\text{right}|(2,3))}_{0.25} \underbrace{P((3,3)|(2,3), \text{right})}_{0.85} \underbrace{\pi(\text{right}|(3,3))}_{0.25} \times \\
 & \times \underbrace{P((3,3)|(3,3), \text{right})}_{0.9} \underbrace{\pi(\text{down}|(3,3))}_{0.25} \underbrace{P((3,2)|(3,3), \text{down})}_{0.85} \underbrace{\pi(\text{down}|(3,2))}_{0.25} \times \\
 & \times \underbrace{P((3,2)|(3,2), \text{down})}_1 \underbrace{\pi(\text{left}|(3,2))}_{0.25} \approx 6.1 \cdot 10^{-6}.
 \end{aligned}$$

Similarly, for trajectory  $\tau_2$  we have

$$\begin{aligned}
 \Pr(\tau_2) = & 1 \cdot 0.25 \cdot 0.85 \cdot 0.25 \cdot 0.05 \cdot 0.25 \cdot 0.85 \cdot 0.25 \cdot 0.85 \cdot 0.25 \cdot 1 \cdot 0.25 \cdot 1 \cdot 0.25 \cdot 1 \cdot 0.25 \\
 & \approx 4.7 \cdot 10^{-7}.
 \end{aligned}$$

The discounted reward for the first trajectory  $\tau_1$  is

$$R(\tau_1) = 8(\gamma^6 + \gamma^7).$$

The discounted reward for the second trajectory  $\tau_2$  is

$$R(\tau_2) = \gamma^4 + \gamma^5 + \gamma^6 + \gamma^7 = \frac{\gamma^4(1 - \gamma^4)}{1 - \gamma}.$$

6. Calculate  $\nabla_{\theta} \log \pi_{\theta}(a|s)$ .

*Solution:* The components of  $\nabla_{\theta} \log \pi_{\theta}(a|s) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$  are  $\frac{\partial \log \pi_{\theta}(a|s)}{\partial \theta_{s',a'}}$  for all  $s' \in \mathcal{S}$  and all  $a' \in \mathcal{A}$ . The partial derivatives with respect to the parameter  $\theta_{s',a'}$  can be computed as

$$\begin{aligned}
& \frac{\partial \log \pi_{\theta}(a|s)}{\partial \theta_{s',a'}} \\
&= \frac{\partial}{\partial \theta_{s',a'}} \left( \log \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})} \right) \quad (\text{by definition of softmax policy parameterization}) \\
&= \frac{\partial}{\partial \theta_{s',a'}} \left( \theta_{s,a} - \log \sum_{a'} \exp(\theta_{s,a'}) \right) \quad (\text{by additive property of logarithmic function}) \\
&= \mathbf{1}_{(s,a)=(s',a')} - \mathbf{1}_{s=s'} \frac{\exp(\theta_{s,a'})}{\sum_{a'} \exp(\theta_{s,a'})} \quad (\text{basic derivative rules}) \\
&= \mathbf{1}_{(s,a)=(s',a')} - \mathbf{1}_{s=s'} \pi_{\theta}(a'|s) \quad (\text{by definition of softmax policy parameterization}) \\
&= \mathbf{1}_{s=s'} (\mathbf{1}_{a=a'} - \pi_{\theta}(a'|s)) \quad (\text{take } \mathbf{1}_{s=s'} \text{ outside of the equation}).
\end{aligned}$$

Thus,

$$\nabla_{\theta} \log \pi_{\theta}(a|s) = \{\mathbf{1}_{s=s'} (\mathbf{1}_{a=a'} - \pi_{\theta}(a'|s))\}_{(s',a') \in \mathcal{S} \times \mathcal{A}}. \quad (0.1)$$

*Reminder:* Here, we denote  $\mathbf{1}_{\mathcal{A}}$  as an indicator function, which equals 1 when event  $\mathcal{A}$  happens and 0 otherwise.

*Example:*

$$\begin{aligned}
\nabla_{\theta_{(1,1),\text{up}}} \log \pi_{\theta}(\text{up}|(1,2)) &= \mathbf{1}_{(1,1)=(1,2)} (\mathbf{1}_{\text{up}=\text{up}} - \pi_{\theta}(\text{up}|(1,1))) = 0, \\
\nabla_{\theta_{(1,1),\text{up}}} \log \pi_{\theta}(\text{up}|(1,1)) &= \mathbf{1}_{(1,1)=(1,1)} (\mathbf{1}_{\text{up}=\text{up}} - \pi_{\theta}(\text{up}|(1,1))) = 1 - \pi_{\theta}(\text{up}|(1,1)), \\
\nabla_{\theta_{(1,1),\text{down}}} \log \pi_{\theta}(\text{up}|(1,1)) &= \mathbf{1}_{(1,1)=(1,1)} (\mathbf{1}_{\text{down}=\text{up}} - \pi_{\theta}(\text{down}|(1,1))) = -\pi_{\theta}(\text{down}|(1,1)).
\end{aligned}$$

7. Consider a discount factor of  $\gamma = 0.8$  and the parameter  $\theta = \mathbf{0}$ . Based on the two aforementioned trajectories,  $\tau_1$  and  $\tau_2$ , provided in Problem 5, compute the stochastic policy gradient

$$\hat{\nabla}_{\theta_{(1,1),a}} J(\pi_{\theta}) = \frac{1}{2} \sum_{i=1}^2 \underbrace{\left( \sum_{t=0}^H \gamma^t r(s_t^i, a_t^i) \right)}_{R(\tau_i)} \left( \sum_{t=0}^H \nabla_{\theta} \log \pi_{\theta}(a_t^i | s_t^i) \right),$$

for the state  $s = (1,1)$ , actions  $a \in \mathcal{A}$ , and the horizon  $H = 7$ .

Continue here:

Continue here:

*Solution:* Using the trajectories defined in Problem 5, we first calculate for  $i = 1, 2$ :

$$\left( \hat{\nabla}_{\theta_{(1,1),a}} J(\pi_\theta) \right)_i = \underbrace{\left( \sum_{t=0}^H \gamma^t r(s_t^i, a_t^i) \right)}_{R(\tau_i)} \left( \sum_{t=0}^H \nabla_\theta \log \pi_\theta(a_t^i | s_t^i) \right),$$

For the first trajectory, we get

$$\begin{aligned} \begin{pmatrix} \hat{\nabla}_{\theta_{(1,1),\text{up}}} J(\pi_\theta) \\ \hat{\nabla}_{\theta_{(1,1),\text{down}}} J(\pi_\theta) \\ \hat{\nabla}_{\theta_{(1,1),\text{left}}} J(\pi_\theta) \\ \hat{\nabla}_{\theta_{(1,1),\text{right}}} J(\pi_\theta) \end{pmatrix}_1 &= R(\tau_1) \times \left( \sum_{t=0}^7 \begin{pmatrix} \nabla_{\theta_{(1,1),\text{up}}} \log \pi(a_t | s_t) \\ \nabla_{\theta_{(1,1),\text{down}}} \log \pi(a_t | s_t) \\ \nabla_{\theta_{(1,1),\text{left}}} \log \pi(a_t | s_t) \\ \nabla_{\theta_{(1,1),\text{right}}} \log \pi(a_t | s_t) \end{pmatrix} \right) \\ &= 8(\gamma^6 + \gamma^7) \times \begin{pmatrix} \nabla_{\theta_{(1,1),\text{up}}} \log \pi(\text{up} | (1,1)) \\ \nabla_{\theta_{(1,1),\text{down}}} \log \pi(\text{up} | (1,1)) \\ \nabla_{\theta_{(1,1),\text{left}}} \log \pi(\text{up} | (1,1)) \\ \nabla_{\theta_{(1,1),\text{right}}} \log \pi(\text{up} | (1,1)) \end{pmatrix} \\ &= 8(\gamma^6 + \gamma^7) \begin{pmatrix} 0.75 \\ -0.25 \\ -0.25 \\ -0.25 \end{pmatrix} \approx \begin{pmatrix} 2.83 \\ -0.94 \\ -0.94 \\ -0.94 \end{pmatrix}, \end{aligned}$$

The second equation utilizes the formula from (0.1), which equals 0 if the state  $s_t \neq (1,1)$ . In the first trajectory, only the initial state visits the state (1,1). Hence, the summation from  $t = 1$  to  $t = 7$  amounts to computing  $t = 1$  since the remaining terms are equal to 0. The third equation involves substituting the policy values into the aforementioned equation (0.1).

And for the second trajectory

$$\begin{aligned} \begin{pmatrix} \hat{\nabla}_{\theta_{(1,1),\text{up}}} J(\pi_\theta) \\ \hat{\nabla}_{\theta_{(1,1),\text{down}}} J(\pi_\theta) \\ \hat{\nabla}_{\theta_{(1,1),\text{left}}} J(\pi_\theta) \\ \hat{\nabla}_{\theta_{(1,1),\text{right}}} J(\pi_\theta) \end{pmatrix}_2 &= R(\tau_2) \times \left( \sum_{t=0}^7 \begin{pmatrix} \nabla_{\theta_{(1,1),\text{up}}} \log \pi(a_t | s_t) \\ \nabla_{\theta_{(1,1),\text{down}}} \log \pi(a_t | s_t) \\ \nabla_{\theta_{(1,1),\text{left}}} \log \pi(a_t | s_t) \\ \nabla_{\theta_{(1,1),\text{right}}} \log \pi(a_t | s_t) \end{pmatrix} \right) \\ &= \frac{2\gamma^4(1 - \gamma^4)}{1 - \gamma} \times \begin{pmatrix} \nabla_{\theta_{(1,1),\text{up}}} \log \pi(\text{right} | (1,1)) \\ \nabla_{\theta_{(1,1),\text{down}}} \log \pi(\text{right} | (1,1)) \\ \nabla_{\theta_{(1,1),\text{left}}} \log \pi(\text{right} | (1,1)) \\ \nabla_{\theta_{(1,1),\text{right}}} \log \pi(\text{right} | (1,1)) \end{pmatrix} \\ &= \frac{2\gamma^4(1 - \gamma^4)}{1 - \gamma} \begin{pmatrix} -0.25 \\ -0.25 \\ -0.25 \\ 0.75 \end{pmatrix} \approx \begin{pmatrix} -0.60 \\ -0.60 \\ -0.60 \\ 1.81 \end{pmatrix}. \end{aligned}$$

The factor of 2 in the second equation arises because the agent visits state (1,1) twice and takes the same action, 'right,' on both occasions.

Hence, the stochastic policy gradient is given by

$$\begin{pmatrix} \hat{\nabla}_{\theta_{(1,1),\text{up}}} J(\pi_\theta) \\ \hat{\nabla}_{\theta_{(1,1),\text{down}}} J(\pi_\theta) \\ \hat{\nabla}_{\theta_{(1,1),\text{left}}} J(\pi_\theta) \\ \hat{\nabla}_{\theta_{(1,1),\text{right}}} J(\pi_\theta) \end{pmatrix} = \frac{1}{2} \sum_{i=1}^2 \begin{pmatrix} \hat{\nabla}_{\theta_{(1,1),\text{up}}} J(\pi_\theta) \\ \hat{\nabla}_{\theta_{(1,1),\text{down}}} J(\pi_\theta) \\ \hat{\nabla}_{\theta_{(1,1),\text{left}}} J(\pi_\theta) \\ \hat{\nabla}_{\theta_{(1,1),\text{right}}} J(\pi_\theta) \end{pmatrix}_i \approx \begin{pmatrix} 1.12 \\ -0.77 \\ -0.77 \\ 0.44 \end{pmatrix}.$$

8. Assuming  $\theta = \mathbf{0} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ , compute the stochastic policy gradient update

$$\theta_{(1,1),a} \leftarrow \theta_{(1,1),a} + \alpha \hat{\nabla}_{\theta_{(1,1),a}} J(\pi_\theta), \theta_{(1,1),a} \in \mathbb{R}^1,$$

for the state  $s = (1, 1)$  and every action  $a \in \mathcal{A}$  with  $\alpha = 0.1$ . Moreover, compute the updated policy for  $s = (1, 1)$ .

*Solution:* According to the above update rule, we can compute the updated  $\theta_{(1,1),a}$ ,  $a \in \mathcal{A}$  as

$$\begin{pmatrix} \theta_{(1,1),\text{up}} \\ \theta_{(1,1),\text{down}} \\ \theta_{(1,1),\text{left}} \\ \theta_{(1,1),\text{right}} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + 0.1 \begin{pmatrix} \hat{\nabla}_{\theta_{(1,1),\text{up}}} J(\pi_{\theta}) \\ \hat{\nabla}_{\theta_{(1,1),\text{down}}} J(\pi_{\theta}) \\ \hat{\nabla}_{\theta_{(1,1),\text{left}}} J(\pi_{\theta}) \\ \hat{\nabla}_{\theta_{(1,1),\text{right}}} J(\pi_{\theta}) \end{pmatrix} \approx \begin{pmatrix} 0.112 \\ -0.077 \\ -0.077 \\ 0.044 \end{pmatrix}.$$

This yields to the following updated policy for the state  $s = (1, 1)$

$$\begin{aligned} \pi_{\theta}(\text{up}|(1, 1)) &= \frac{\exp(\theta_{(1,1),\text{up}})}{\sum_{a'} \exp(\theta_{(1,1),a'})} \approx 0.28, \\ \pi_{\theta}(\text{down}|(1, 1)) &\approx 0.23, \pi_{\theta}(\text{left}|(1, 1)) \approx 0.23, \pi_{\theta}(\text{right}|(1, 1)) \approx 0.26. \end{aligned}$$

9. Now, assume that there is no noise, i.e., the agent always moves in the intended direction if the action leads to a free cell and otherwise stays in its previous cell. What are the discounted sums of rewards for the red and blue trajectory (for an infinite horizon)? How to choose  $\gamma$  to ensure that the blue trajectory is preferred over the red one?

*Solution:* The discounted sum of rewards for the blue trajectory is

$$R(\tau_{\text{blue}}) = \sum_{t=2}^{\infty} \gamma^t = \frac{\gamma^2}{1-\gamma}.$$

The discounted sum of rewards for the red trajectory is

$$R(\tau_{\text{red}}) = 8 \sum_{t=5}^{\infty} \gamma^t = \frac{8\gamma^5}{1-\gamma}.$$

If  $R(\tau_{\text{blue}}) > R(\tau_{\text{red}})$ , then the blue trajectory is preferred compared to the red trajectory, otherwise the red trajectory is preferred compared to the blue trajectory. Therefore, the red trajectory is preferred when  $1 > \gamma > \frac{1}{2}$  and the blue trajectory is preferred when  $\frac{1}{2} > \gamma > 0$ .