

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.

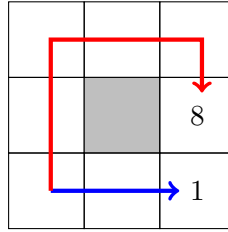


Figure 1: Gridworld

1. For the states $s = (1,1)$ and $s = (1,2)$, determine the transition probabilities $P(\cdot|s, a)$ for any $a \in \mathcal{A}$.
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?
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?
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?
5. Next, we consider two trajectories sampled from the initial softmax policy and truncated to a horizon $H = 7$. The first trajectory τ_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 τ_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 τ is $\Pr(\tau) = \rho(s_0) \pi_\theta(a_0|s_0) \prod_{t=1}^{H-1} P(s_t|s_{t-1}, a_{t-1}) \pi_\theta(a_t|s_t)$. What are the probabilities of choosing τ_1 and τ_2 ? What are the discounted rewards for these trajectories?

6. Calculate $\nabla_{\theta} \log \pi_{\theta}(a|s)$.
7. Consider a discount factor of $\gamma = 0.8$ and the parameter $\theta = \mathbf{0}$. Based on the two aforementioned trajectories, τ_1 and τ_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 $H = 7$.

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 $s = (1, 1)$.

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 γ to ensure that the blue trajectory is preferred over the red one?