- Assignment 4 -
Reinforcement Learning Spring 2025

Overview:

- This assignment can be solved after studying policy gradient methods.
- Unless otherwise stated, we parametrize the policy using Eqn. 13.2 and Eqn. 13.3.
- You may find Eqn 13.9 useful in multiple questions. For a proof of Eqn. 13.9 see the exercises on policy gradient methods.
- The quiz also includes various true-false questions on key concepts related to policy gradient methods, such as advantage function, actor-critic, gradient ascent.
- Whenever appropriate, studium quiz answers allow a certain percentage of error in the numerical values. Assuming that you're reasonably accurate in your calculations, this makes it highly unlikely to fail a question in studium due to the number of decimal points you keep track of.

Questions:

1. Consider the shortest possible example of random walk, i.e. a room with two exits. Hence, we have one starting state (S) and two terminal states: terminal left (TL) and terminal right (TR). The environment can be represented as: TL S TR. We start at state S and have only two actions available: LEFT (ActL) and RIGHT (ActR). If we choose ActL, we exit to TL. Similarly, if we choose ActR, we exit to TR. When we exit the room, we get the reward and the episode ends. Exiting to TR gives the reward $R_R = 3$ and exiting to the TL gives the reward $R_L = 6$. We define the features similar to Example 13.1: $x(S, ActR) = [1, 0]^T$ and $x(S, ActL) = [0, 1]^T$.

   We consider REINFORCE on page 328 with $\alpha = 1$ and $\gamma = 1$. We will keep track of $\boldsymbol{\theta}$ and $\pi(a)$. To refer to these quantities in the quiz, we will use the notation $\boldsymbol{\theta}(i)$ and $\pi(a, i)$ where i is the episode index. Note that $\boldsymbol{\theta}(i)$ and $\pi(a, i)$ refers to the value after episode i is terminated and the associated updates are performed.

   (a) Preliminaries: Find an expression for the gradients $\nabla \ln \pi(ActR|s, \theta)$ and $\nabla \ln \pi(ActL|s, \theta)$ in terms of $\pi(ActL)$ and $\pi(ActR)$.

   For this example these gradients have a simple expression in terms of $\pi(ActL)$ and $\pi(ActR)$. Although there will be no direct questions on this expression in the quiz, we suggest you first work on this item since it will make hand-tracing much easier.

   (b) Initialize $\boldsymbol{\theta}(0) = \mathbf{0}$. Find the initial policy i.e. find $\pi(ActL, 0)$ and $\pi(ActR, 0)$.

   (c) Suppose that our action in the first episode is ActR. Find $\boldsymbol{\theta}(1)$ and the resulting policy, i.e. $\pi(ActL, 1)$ and $\pi(ActR, 1)$.

   (d) Consider the second episode. Suppose that our action is ActL. Find $\boldsymbol{\theta}(2)$ and the resulting policy, i.e. $\pi(ActL, 2)$ and $\pi(ActR, 2)$.

2. We consider the corridor environment of 2 rooms. The environment is represented with 2 non-terminal states $\mathcal{S} = \{1, 2\}$ and one terminating state 3. The state labels correspond to the room numbers; for instance, state 2 corresponds to Room 2. The leftmost room is Room 1 and the rightmost room is Room 3. The actions are LEFT (ActL) and RIGHT (ActR) with their usual behaviour on the left edge.

   We approximate $v_\pi(s)$ for $s \in \mathcal{S}$ using value function approximation $\hat{v}(s, \boldsymbol{w}) = \sum_{i=1}^{2} w_i x_i(s)$, where $\boldsymbol{w} \in \mathbb{R}^2$. Note that $v_\pi(3) = 0$ since it is a terminating state. The features for value function approximation are given by $x_1(s) = 1$, $x_2(s) = s$.

   We approximate the policy using the preferences $h(s, a, \boldsymbol{\theta}) = \sum_{i=1}^{2} \theta_i z_i(s, a)$, where $\boldsymbol{\theta} \in \mathbb{R}^2$. The features are given by $z_1(s, ActL) = 2$, $z_2(s, ActL) = s$ and $z_i(s, ActR) = s^i$ for i=1, 2.

   The discount rate is $\gamma = 1$. The step sizes are $\alpha^{\boldsymbol{w}} = 0.9$ and $\alpha^{\boldsymbol{\theta}} = 0.8$. We initialize with $\boldsymbol{w} = \mathbf{0}$, $\boldsymbol{\theta} = \mathbf{0}$. We obtain the following episode using Reinforce with baseline from pg. 330:

t	S_t	A_t	R_{t+1}
0	Room 1	ActR	-1
1	Room 2	ActR	+10
2	Room 3 (terminate)	-	-

   In the quiz, you will be asked for $\boldsymbol{w}_{final}$ and $\boldsymbol{\theta}_{final}$ after the end of the episode and the associated quantities, such as $\pi(a|s, \boldsymbol{\theta}_{final})$, $\nabla \ln \pi(s, a, \boldsymbol{\theta}_{final})$, $\boldsymbol{z}(s, a)$ for a given s and a.

3. We now consider a simplified control task. We would like to determine the directional force $f \in \mathbb{R}$ that will be applied to a cart to bring it to a certain location. Hence, our action at time t is f_t. The state s is given by the position z_t. **We use the Gaussian policy parametrization presented in Section 13.7 of the textbook; see the textbook for the notation**. We will use one-step actor-critic from Chapter 13. Let $\alpha^{\boldsymbol{\theta}} = 1$, $\gamma = 0.5$, $\alpha^{\boldsymbol{w}} = 0.5$. We have

$$\boldsymbol{x}_\mu(s) = \begin{pmatrix} 1 \\ s \\ s^2 \end{pmatrix}, \qquad \boldsymbol{x}_\sigma(s) = \begin{pmatrix} s-1 \\ (s-1)^2 \end{pmatrix}, \quad \boldsymbol{\theta}_t = \begin{pmatrix} 3 \\ 3 \\ 3 \\ 1 \\ 1 \end{pmatrix}$$

   (a) Let $z_t = 2$. Find $\mu(z_t, \boldsymbol{\theta}_t)$ and $\sigma(z_t, \boldsymbol{\theta}_t)$.

   (b) We will now choose the action f_t using the policy induced by $\boldsymbol{\theta}_t$. State which of these is more likely: i) $17 \leq |f_t| \leq 18$, ii) $2 \leq |f_t| \leq 3$.
   Hint: To answer this question, you only need to have a general understanding of the shape of Gaussian distribution. No numerical calculations are needed.

   (c) We have $f_t = 10$. Find $\nabla \ln \pi(f_t|z_t, \boldsymbol{\theta}_t)$.
   Hint: See Exercise 13.4.

   (d) At time t, just before $\boldsymbol{\theta}$ update, we have $I = 0.25$, $\delta = 8$. Find $\boldsymbol{\theta}_{t+1}$.