Otro autobús mágico

We formalize a **Markov Decision Process** or **MDP** for short as:

- **S**tates: States S with starting state $s_{\text{start}} \in S$.
- **T**ermination State: $\texttt{isEnd}(s)$.
- **A**ctions: $a \in A(s)$, all possible actions at state s.
- **R**ewards: $R(s, a, s')$, the reward of going from state s to s' by taking action a.
- **T**ransitions: $T(s, a, s')$, the probability of going from state s to s' by taking action a.
- **D**iscount: $0 \leq \gamma \leq 1$, the discount factor (default 1) for computing utility.

$V_\pi(s)$ is the expected utility received by following policy π from state s.
$Q_\pi(s, a)$ is the expected utility of taking action a from state s, and then following policy π.

The following are **off-policy** algorithms that output the optimal Q-value, Q_{opt}:

- **Value Iteration**: $V_{\text{opt}}^{(t)}(s) \leftarrow \max_{a \in A(s)} Q_{\text{opt}}^{(t-1)}(s, a)$.
- **Model-Based Value Iteration**: Estimate T and R using Monte Carlo, then run value iteration using estimates $\hat{T}$ and $\hat{R}$.
- **Q-Learning**: Estimate $\hat{Q}_{\text{opt}}(s, a)$ based on (i) the reward to state s' and (ii) the estimated optimal max value of s'.

The following are **on-policy** algorithms that output the Q-value, Q_π, of a specific policy:

- **Policy Iteration**: $V_\pi^{(t)}(s) \leftarrow Q_\pi^{(t-1)}(s, \pi(s))$.
- **Model-Free Monte Carlo**: Estimate $\hat{Q}_\pi(s, a)$ from the utility, u_t, along the path.
- **SARSA**: Estimate $\hat{Q}_\pi(s, a)$ based on (i) the update (s, a, r, s', a') and (ii) the estimated $\hat{Q}_\pi(s', a')$.

Algorithm	**Estimating**	**Based On**
Model-Based Monte Carlo	$\hat{Q}_{\text{opt}}$	$s_0, a_1, r_1, s_1, a_2, r_2, s_2, \ldots \implies \hat{T}, \hat{R}$
Q-Learning	$\hat{Q}_{\text{opt}}$	$(s, a, r, s'), \hat{V}_{\text{opt}}(s')$
Model-Free Monte Carlo	$\hat{Q}_\pi$	$u_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots$
SARSA	$\hat{Q}_\pi$	$(s, a, r, s', a'), \hat{Q}_\pi(s', a')$

1) Problem 1: MDP for Riding the Bus

(a) Sabina wants to go from their house (located at location 1) to the gym (located at location n). At each location s, Sabina can either (i) deterministically walk forward to the next location $s+1$ (takes 1 unit of time) or (ii) wait for the bus.

The bus comes with probability ϵ, in which case, it will take Sabina to the gym in $1+\alpha(n-s)$ units of time, where α is some parameter. If the bus doesn't come, then Sabina stays put waiting for nothing, and that takes 1 unit of time.

Let our reward be negative time, which is equivalent to minimizing the time it takes to get to the gym.

1	2	3	4	$\dots$	n
House				$\cdots$	Gym

We have formalized the problem as an MDP for you:

- State: $s \in \{1, 2, \dots, n\}$ is Sabina's location
- $\text{Actions}(s) = \{\text{Walk}, \text{Bus}\}$
- $\text{Reward}(s, \text{Walk}, s') = \begin{cases} -1 & \text{if } s' = s+1 \\ -\infty & \text{otherwise} \end{cases}$
- $\text{Reward}(s, \text{Bus}, s') = \begin{cases} -1-\alpha(n-s) & \text{if } s' = n \\ -1 & \text{if } s' = s \\ -\infty & \text{otherwise} \end{cases}$
- $T(s'|s, \text{Walk}) = \begin{cases} 1 & \text{if } s' = s+1 \\ 0 & \text{otherwise} \end{cases}$
- $T(s'|s, \text{Bus}) = \begin{cases} \epsilon & \text{if } s' = n \\ 1-\epsilon & \text{if } s' = s \\ 0 & \text{otherwise} \end{cases}$
- $\text{IsEnd}(s) = \mathbf{1}[s = n]$

BEFORE YOU MOVE FORWARD: Make sure you understand *why* the MDP is formulated the way it is!

Compute closed form expressions for (i) the value of a policy where Sabina always walks at every location and (ii) the value of a policy where Sabina always waits for the bus at every location (using some or all of the variables ϵ, α, n). Assume a discount rate of $\gamma = 1$.

- $V_{\text{Walk}}(s) =$ ____

- $V_{\text{Bus}}(s) =$ ____

For what values of ϵ (as a function of α and n) is it advantageous to walk rather than take the bus?

(b) Unfortunately, Sabina's town is unable to provide transition probabilities or a reward function (i.e. a bus schedule), making the above MDP possibly (and likely) inaccurate. To get around this, Sabina decides to use reinforcement learning, specifically Q-learning to determine the best policy. Sabina starts going around town both by bus and by walking, recording the following data:

s_0	a_1	r_1	s_1	a_2	r_2	s_2	a_3	r_3	s_3	a_4	r_4	s_4	a_5	r_5	s_5
1	Bus	-1	1	Bus	-1	1	Bus	3	3	Walk	1	4	Walk	1	5

Run the Q-learning algorithm once over the given data to compute an estimate of the optimal Q-value $Q_{\text{opt}}(s, a)$. Process the episodes from left to right. Assume all Q-values are initialized to zero, and use a learning rate of $\eta = 0.5$ and a discount of $\gamma = 1$.

- $\hat{Q}(1, \text{Walk}) =$ ______________________
- $\hat{Q}(1, \text{Bus}) =$ ______________________
- $\hat{Q}(3, \text{Walk}) =$ ______________________
- $\hat{Q}(3, \text{Bus}) =$ ______________________
- $\hat{Q}(4, \text{Walk}) =$ ______________________
- $\hat{Q}(4, \text{Bus}) =$ ______________________