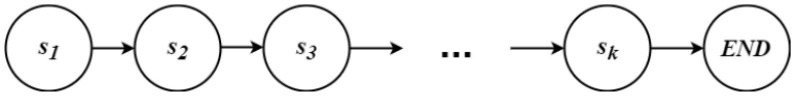
Name: ______________________________

## Go Positive, or Go Negative

5. (16 points) Consider the following simple MDP: **Positive Reward**




First consider the case where the MDP has positive reward. In this scenario, there is only one action ($next$); we name this decision policy $\pi_A$ with $\pi_A(s) = next$ for all $s$. The reward is $R(s, next) = 0$ for all states $s$, except for state $s_k$ where reward is $R(s_k, next) = 10$. We always start at state $s_1$ and each arrow indicates a deterministic transition probability $p = 1$. There is no transition out of the end state $END$, and 0 reward for any action from the end state.

(a) Calculate $V_\pi(s)$ for each state in the finite-horizon case with horizon $h = 1$, $k = 4$, and discount factor $\gamma = 1$.

**Solution:** The values for horizon 1 are just the rewards (base case). From textbook notes (Reinforcement Learning chapter), V^1_π(s) = R(s, π(s)) + 0.

$$V_\pi^1(s_4) = 10$$
$$V_\pi^1(s_3) = 0$$
$$V_\pi^1(s_2) = 0$$
$$V_\pi^1(s_1) = 0$$

(b) Calculate $V_\pi(s)$ for each state in the infinite horizon case with $k = 4$ and discount factor $\gamma = 0.9$.

**Solution:**

$$V_\pi(s_4) = 10$$
$$V_\pi(s_3) = 0 + \gamma * 10 = 0.9 * 10 = 9$$
$$V_\pi(s_2) = 0.9 * 9 = 8.1$$
$$V_\pi(s_1) = 0.9 * 8.1 = 7.29$$

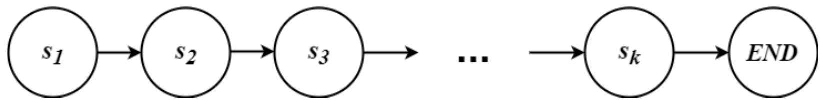
(c) Derive a formula for $V_\pi(s_1)$ that works for any value of (is expressed as a function of) $k$ and $\gamma$ for the above positive reward MDP, in the infinite horizon case.

**Solution:** At each step, we receive a reward of 0, except after the $k^{th}$ step, when we get a reward of 10. Therefore, the summation is

$$\sum_{i=0}^{k-1} 0 * \gamma^i + 10 * \gamma^{k-1} = 0 * \gamma^0 + 0 * \gamma^1 + 0 * \gamma^2 + 0 * \gamma^3 + ... + 10 * \gamma^{k-1} = 10\gamma^{k-1}.$$

**Negative Reward**

Now consider the case where this MDP has negative reward. In this scenario, the reward is $R(s, next) = -1$ for all states, except for state $s_k$ where the reward is $R(s_k, next) = 0$. Again, there is only one action, $next$, and the decision policy remains $\pi_A(s) = next$ for all $s$. We always start at state $s_1$ and each arrow has a deterministic transition probability $p = 1$. There is no transition out of the end state $END$, and zero reward for any action from the end state, i.e., $R(END, next) = 0$.




(d) Calculate $V_\pi(s)$ for each state in the finite-horizon case with horizon $h = 1$, $k = 4$, and discount factor $\gamma = 1$.

**Solution:**

$$
\begin{aligned}
V_\pi^1(s_4) &= 0 \\
V_\pi^1(s_3) &= -1 \\
V_\pi^1(s_2) &= -1 \\
V_\pi^1(s_1) &= -1
\end{aligned}
$$

(e) Calculate $V_\pi(s)$ for each state in the infinite horizon case with $k = 4$ and discount factor $\gamma = 0.9$.

**Solution:**

$$
\begin{aligned}
V_\pi(s_4) &= 0 \\
V_\pi(s_3) &= -1 + \gamma * 0 = -1 \\
V_\pi(s_2) &= -1 + 0.9(-1) = -1.9 \\
V_\pi(s_1) &= -1 + 0.9(-1.9) = -2.71
\end{aligned}
$$

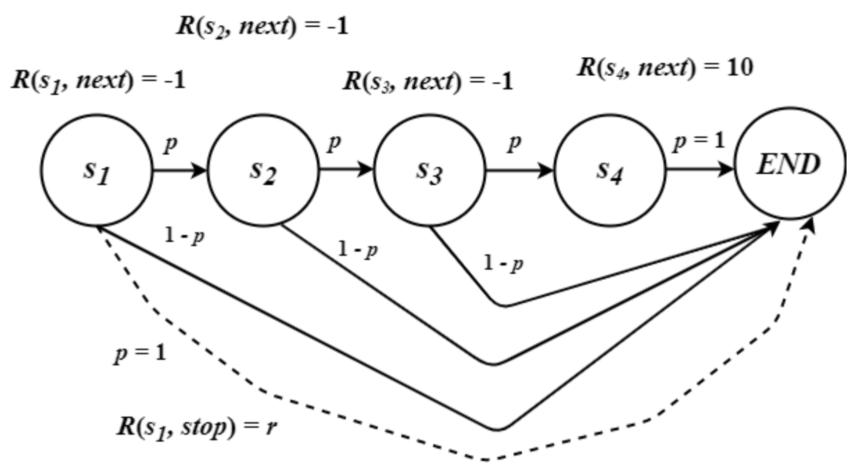
(f) Derive a formula for $V_\pi(s_1)$ that works for any value of (is expressed as a function of) $k$ and $\gamma$ for this negative reward MDP with infinite horizon. Recall that $\sum_{i=0}^{n} \gamma^i = \frac{(1-\gamma^{n+1})}{(1-\gamma)}$.

**Solution:** At every step, we receive a reward of -1, except for the $k^{th}$ step, where we receive a reward of 0. Therefore, the summation is

$$\sum_{i=0}^{k-1} -1 * \gamma^i + 0 * \gamma^{k-1} = -1 * \gamma^0 - 1 * \gamma^1 - 1 * \gamma^2 + ... - 1 * \gamma^{k-2} + 0 * \gamma^{k-1} = -\frac{1 - \gamma^{k-1}}{1 - \gamma}.$$

**Positive and Negative Reward**

Consider the MDP below with negative rewards for some $R(s, a)$ and positive rewards for others. Now there are two actions, *next* and *stop*. The solid arrows show the probabilities of state transitions under action *next*; the dashed arrows show the probability of state transitions under action *stop*. (If there is no dashed arrow from a state, that indicates a probability $p = 0$ of transitioning out of that state under action *stop*.) The corresponding rewards $R(s_i, a)$ are also indicated on the figure below. Note that the rewards are $R(s_i, next) = -1$ for all $s_i$, except for state $s_4$, where the reward is $R(s_4, next) = 10$. Finally, under action *stop*, we have reward $R(s_1, stop) = r$ (some unknown value $r$), and $R(s, stop) = 0$ for all other states. As before, we always start in state $s_1$. There is no transition out of the end state $END$, and zero reward for any action from the end state, i.e., $R(END, next) = R(END, go) = 0$. Assume discount factor $\gamma$ and infinite horizon.




(g) We consider two possible policies: $\pi_A(s) = next$ for all $s$, and $\pi_B(s) = stop$ for all $s$. Your goal is to maximize your reward. When you start at $s_1$, you have reward 0 before taking any actions. Determine what $r$ should be, so that it is best to run this MDP under policy $\pi_B$ rather than policy $\pi_A$. Give your answer as an expression for $r$ involving $p$ and $\gamma$.

**Solution:** Under policy $\pi_A$:

$$
\begin{aligned}
V_\pi(s_4) &= 10 \\
V_\pi(s_3) &= -1 + p\gamma V_\pi(s_4) + (1-p)\gamma V_\pi(end) = -1 + p\gamma \cdot 10 \\
V_\pi(s_2) &= -1 + p\gamma V_\pi(s_3) = -1 - p\gamma + (p\gamma)^2 \cdot 10 \\
V_\pi(s_1) &= -1 + p\gamma V_\pi(s_2) = -1 - p\gamma - (p\gamma)^2 + (p\gamma)^3 \cdot 10
\end{aligned}
$$

Under policy $\pi_B$, we simply have $V_\pi(s_1) = r$. So we should choose policy $\pi_B$ when

$$r > -1 - p\gamma - (p\gamma)^2 + (p\gamma)^3 \cdot 10$$

As an example, for $\gamma = 1$ and $p = 0.9$, $r$ is 4.58.