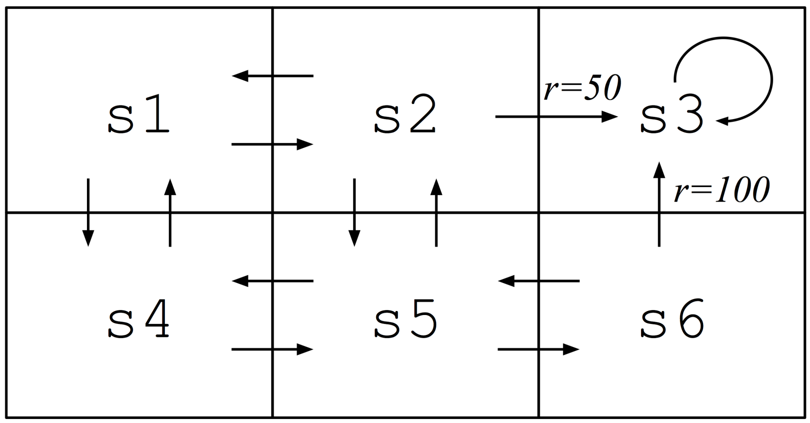
Robots

5. (16 points) Consider the following deterministic Markov Decision Process (MDP), describing a simple robot grid world. Notice that the values of the immediate rewards r for **two** transitions are written next to them; the other transitions, with no value written next to them, have an immediate reward of $r = 0$. **Assume the discount factor γ is 0.8.**



(a) For states $s \in \{s6, s5, s2\}$, write the value for $V_{\pi^*}(s)$, the discounted infinite horizon value of state s using an optimal policy π^*. It is fine to write a numerical expression—you don't have to evaluate it—but it shouldn't contain any variables.

Solution:

$$V_{\pi^*}(s6) = 100$$

Solution:

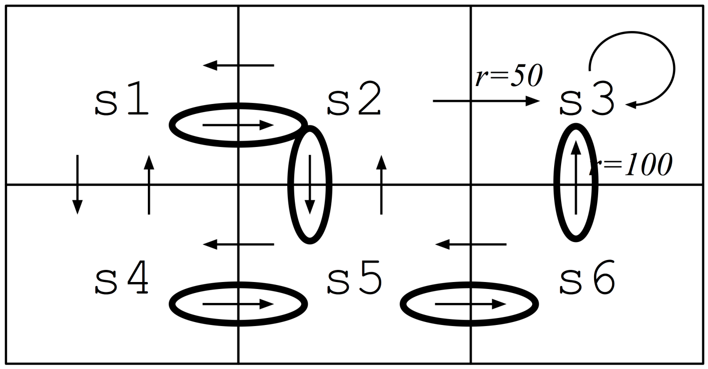
$$V_{\pi^*}(s5) = \gamma V_{\pi^*}(s6) = 80$$

Solution:

$$V_{\pi^*}(s2) = \gamma V_{\pi^*}(s5) = 64$$

(b) For each state in the state diagram below, circle exactly one outgoing arrow, indicating an optimal action $\pi^*(s)$ to take from that state. If there is a tie, it is fine to select any action with optimal value.

Solution:



(c) Give a value for γ (constrained by $0 < \gamma < 1$) that results in a different optimal policy, and describe the resulting policy by indicating which $\pi^*(s)$ values (i.e., which policy actions) change.

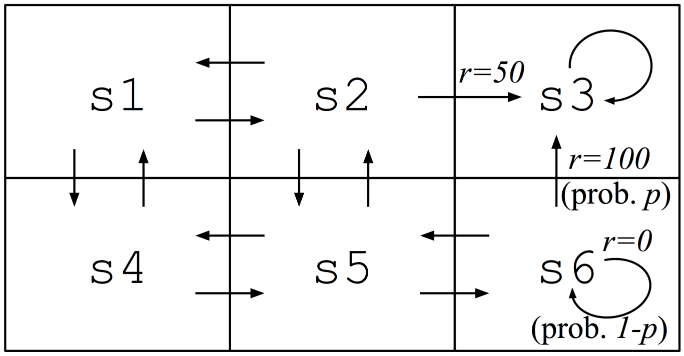
Solution: A small $\gamma = 0.001$ will make it not worthwhile to defer gains for very long. In this problem, if $\gamma^2 100 < 50$, then it will be better to directly take the 50 reward. So valid answers here are $0 < \gamma < \frac{\sqrt{2}}{2}$.

Solution: Now $\pi^*(s2)$ is to go right (east).

(d) Is it possible to change the immediate reward for each state in such a way that V_{π^*} changes but the optimal policy π^* remains unchanged? If yes, provide a new reward function, and explain how the resulting V_{π^*} changes but π^* does not. Otherwise, explain in at most two sentences why this is impossible.

Solution: Yes. We can establish small immediate rewards, say $r = 1$, for all of the transitions currently with $r = 0$. These are not enough to change the π^* decisions, but do change V_{π^*} for all of these states.

When winter comes, snow also appears on one path in the grid world, making exactly one of the actions non-deterministic. The resulting MDP is shown below. Specifically, the change is that now the result of the action "go north" from state `s6` results in one of two outcomes. With probability p, the robot succeeds in transitioning to state `s3` and receives immediate reward 100. However, with probability $(1 - p)$ it slips on the ice, and remains in state `s6` with 0 immediate reward. **Assume again that the discount factor** $\gamma = 0.8$.



(e) Assume $p = 0.75$. For each of the states $s \in \{s2, s5, s6\}$, write the value for $V_{\pi^*}(s)$. It is fine to write a numerical expression, but it shouldn't contain any variables.

Solution:

$$\begin{aligned} V_{\pi^*}(s6) &= 100p + (1-p)\gamma V_{\pi^*}(s6) \\ V_{\pi^*}(s6)(1-(1-p)\gamma) &= 100p \\ V_{\pi^*}(s6) &= \frac{100p}{1-(1-p)\gamma} = 93.75 \end{aligned}$$

Solution:

$$V_{\pi^*}(s5) = \gamma V_{\pi^*}(s6) = 75$$

Solution:

$$V_{\pi^*}(s2) = \gamma V_{\pi^*}(s5) = 60$$

(f) How bad does the ice have to get before the robot will prefer to completely avoid the ice? Let us answer the question by giving a value for p for which the optimal policy chooses actions that completely avoid the ice, i.e., choosing the action "go left" over "go up" when the robot is in the state `s6`. Approach this in four parts. The answer to each of the first three parts can be a numerical expression; the answer to the last part can be an expression involving numbers and p.

i. What is the value V of going right in state `s2`?

Solution: 50

ii. What is the value V of going up in state `s5`, if you're going to go right in state `s2`?

Solution: $\gamma \cdot 50 = 40$

iii. What is the value V of going left in state `s6`, if you're going to go up in state `s5` and right in state `s2`?

Solution: $\gamma^2 \cdot 50 = 32$

iv. Under what condition on p is it better to go left in state `s6` (then up in state `s5` and right in state `s2`) than it is to go up in state `s6`?

Solution:

$$\frac{p \cdot 100}{1 - (1-p) \cdot 0.8} < 32$$

$$p < \frac{8}{93} \approx 0.086$$