CS 470 Homework #8
Markov Decision Processes
80 Points

---

1. **Markov Decision Process [40 points]**

   Consider the following MDP:

   - **States:** 6 states $\{1, 2, \ldots, 6\}$
   - **Initial state:** state 1
   - **Actions:** The agent has 4 actions available, actions A, B, C, and D. They induce probabilistic transitions according to the transition model (which also indicates which actions are possible in which states)
   - **Transition Model:** Actions A and B can only be performed when the agent is in state 1. If A is executed, the agent moves to state 2 or 3, with probabilities 0.8 and 0.2, respectively. If B is executed, the agent moves to state 2 or 3 with probabilities 0.4 and 0.6, respectively.

     We represent these transition rules by:

A:	$1 \mapsto \{2(0.8), 3(0.2)\}$
B:	$1 \mapsto \{2(0.4), 3(0.6)\}$

     Similarly, the transition rules for C and D are defined as follows:

C:	$2 \mapsto \{4(0.8), 5(0.2)\}$
D:	$2 \mapsto \{4(0.6), 6(0.4)\}$
C:	$3 \mapsto \{6(1)\}$
D:	$3 \mapsto \{5(0.6), 6(0.4)\}$

     The states 4, 5 and 6 are terminal states and have no actions available in them.
   - **Rewards:** The reward for each state is 0, with the exception of the three terminal states which provide different rewards as follows: R(4) = 100, R(5) = 200, R(6) = 50.

   (a) **[5 points]** Assume that the agent has first executed action A and is now in state 2. What is the (expected) utility of action C? What is the (expected) utility of action D?

   (b) **[25 points]** Assign states 1, 2 and 3 an initial utility of 0. Use value iteration to update the utilities for each state. Use a discount factor of $\gamma = 1$. Report the utility of every state after each step of value iteration until the values converge.

   (c) **[10 points]** A policy is a complete mapping of every non-terminal state into an action executable in that state. For example $\pi = \{1(A), 2(C), 3(C)\}$ is a policy that tells the agent to perform A in state 1, and C in states 2 and 3. What is the optimal policy π^* for the above problem?

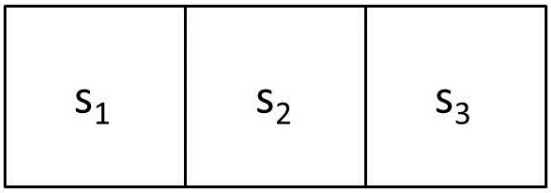

Figure 1: State space for Question 2

2. **Markov decision processes [40 points]**

   Consider an MDP with three states s_1, s_2 and s_3, as shown in Figure 1, and two actions a_1 and a_2. We will model a problem in which an agent would like to get to a "goal state" s_3 as quickly as possible, and suffers a negative reward on each step until it manages to do so. Further, action a_1 makes the agent stay in the same state, whereas action a_2 has a 50% change of advancing it towards the goal state.

   In detail, let the reward function be $R(s_1) = -1$, $R(s_2) = -1$, $R(s_3) = 0$. Let the discount factor be $\gamma = 1$. Further, the state transition probabilities $P(s'|s, a)$ are given as follows:

   - In state s_1, if you take action a_1, you transition back to (i.e. stay in) state s_1 with probability 1. If you take action a_2, you stay in s_1 with probability 0.5, and advance (transition) to state s_2 with probability 0.5
   - In state s_2, if you take action a_1, you stay in state s_2 with probability 1. If you take action a_2, you stay in s_2 with probability 0.5, and advance to s_3 with probability 0.5
   - In state s_3, regardless of what action you take, you stay in s_3 with probability 1.

   (a) **[15 points]** What is the optimal policy? For each of the three states, give an optimal action. No justification needed.
   $\pi^*(s_1) =$
   $\pi^*(s_2) =$
   $\pi^*(s_3) =$

   (b) **[15 points]** What is the optimal value of each state? Give your answer as three real numbers, one for each state. Please show your work/justification.

   (c) **[10 points]** If we were to change γ from 1 to 0.99, would $V^*(s_1)$ increase or decrease? **Don't guess.** Briefly justify your answer. (Credit given only for a correct answer with correct justification.)

3. **EXTRA CREDIT: MDPs with Random Stopping Times [25 points]**

   Suppose we have a Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \gamma, R)$, where $\mathcal{S}$ is a discrete state space with n states, and the rewards are discounted by a factor γ. (Recall that $P(s' \mid s, a)$ is the "transition model" and $R(s)$ is the reward function.) We can view this process as a game where we begin in some state $s_0 \in \mathcal{S}$ and take turns selecting actions and transitioning to new states, accumulating rewards along the way. At the nth turn, we first receive some (discounted) reward, $\gamma^n R(s)$, for the current state s. Then, we select an action, $a \in \mathcal{A}$ and transition, randomly, to a new state s' according to the probabilities, $P(s' \mid s, a)$. Since the discount factor is $\gamma < 1$, our rewards become smaller and smaller as the game goes on. (Hence, the optimal strategy will try to accumulate big rewards early.)

   Now consider a slight modification of this game. At the start of each turn we receive an *undiscounted* reward, $R(s)$, and then flip a biased coin that lands `heads` with probability ϵ, $0 < \epsilon \leq 1$. If the coin lands `heads` then the game is stopped and we are left with whatever reward we have accumulated thus far. Otherwise, we choose our action and we transition to the next state according to $P(s' \mid s, a)$, as usual. We will now show that this new game can be expressed as an MDP. In addition, we'll also show that the value of this game (i.e., the largest reward we expect to gain from playing it) is equivalent to the *discounted* reward in the original MDP, $\mathcal{M}$.

   To do this you will define a new MDP, $\tilde{\mathcal{M}} = (\tilde{\mathcal{S}}, \mathcal{A}, \tilde{P}, 1, \tilde{R})$. This MDP will have the same action space as $\mathcal{M}$, but the discount factor will be 1, and we have a different state space, transition model, and reward function. You will construct the MDP $\tilde{\mathcal{M}}$ so that it is just like the MDP $\mathcal{M}$, but with the necessary modifications to include the coin-flipping rules that were described above.

   In particular, we are going to add a new state called the "sink" state, which we'll denote e. If the coin toss comes up `heads`, then we'll transition, always, to this state and remain there forever (accumulating 0 reward each turn). If the coin toss is `tails`, then we'll just transition according to P as before, with no chance of entering the sink state, e.

   (a) **[10 points]** Complete the construction by explicity specifying each of the remaining components of the MDP $\tilde{\mathcal{M}}$: [*Hint: use the components of the original MDP* $\mathcal{M}$]

      i. **[3 points]** What is $\tilde{\mathcal{S}}$?

      ii. **[4 points]** What is $\tilde{P}(s' \mid s, a)$?

      iii. **[3 points]** What is $\tilde{R}$?

   (b) **[5 points]** Show that $\tilde{V}^*(e) = 0$, where $\tilde{V}^*$ is the optimal value function for the MDP $\tilde{\mathcal{M}}$.

   (c) **[10 points]** Now show that when $s \in \mathcal{S}$, the Bellman equation for $\tilde{\mathcal{M}}$ is the same as the Bellman equation for $\mathcal{M}$, but with a different discount factor.