Q3. [9 pts] A Multiplayer MDP

Alice and Bob are playing a game on a 2-by-2 grid, shown below. To start the game, a puck is placed on square A.

At each time step, the available actions are:

- Up, Left, Down, Right: These actions move the puck deterministically. Actions that move the puck off the grid are disallowed. These actions give both players a reward of 0.
- Exit: This action ends the game. Note that the Exit action can be taken no matter where the puck is. This action gives Alice and Bob rewards depending on the puck's final location, as shown below.

Each player tries to maximize their own reward, and does not care about what rewards the other player gets.

X	Y
W	Z

Figure 1: The grid that the puck moves on.

Puck location when "exit" is taken	Alice's reward	Bob's reward
W	−1	+3
X	0	0
Y	+1	+2
Z	−1	+2

Figure 2: Exit rewards for Alice and Bob.

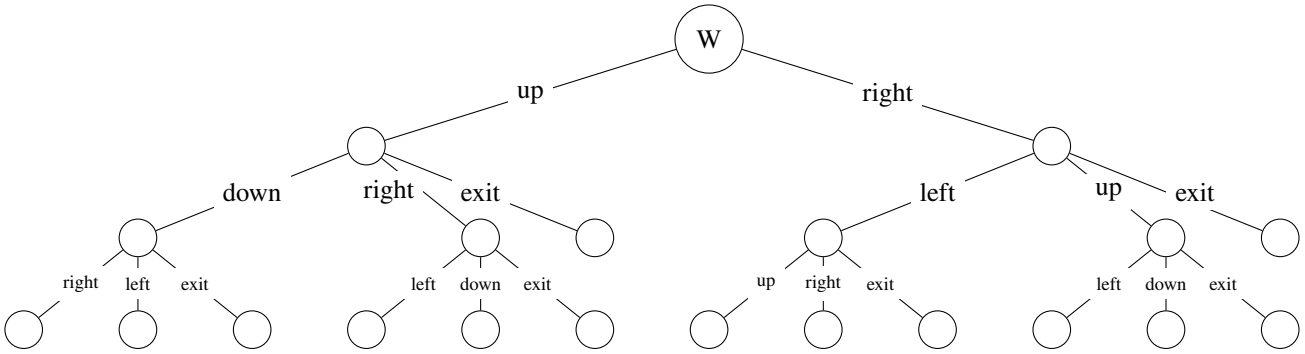
(a) [2 pts] Suppose Bob is the only player taking actions in this game. Bob models this game as an MDP with a discount factor $0 < \gamma \leq 1$.

For which of the following states does Bob's optimal policy depend on the value of γ? Select all that apply.

☐ W ☐ X ☐ Y ☐ Z ○ None of the above.

For the rest of the question, Alice and Bob alternate taking actions. Alice goes first.

Alice models this game with the following game tree, and wants to run depth-limited search with limit 3 turns. She uses an evaluation of 0 for any leaf node that is not a terminal state. (Note: Circles do not necessarily represent chance nodes.)



(b) [1 pt] This is a zero-sum game.

○ True ○ False

(c) [1 pt] What is Alice's value of the root node of the tree?

(d) [1 pt] What is Bob's value of the root node of the tree?

(e) [2 pts] Assuming that Alice and Bob play optimally, what are possible sequences of actions that they might play? Select all that apply.

- ☐ up, right, exit
- ☐ up, right, down
- ☐ right, left, right
- ☐ right, up, exit
- ☐ right, exit
- ◯ None of the above.

Alice instead decides to model the game as an MDP. Assumptions:

- $\gamma = 0.5$
- Alice knows Bob's policy is π.
- $D(s, a)$ represents the new state you move into when you start at state s and take action a.

(f) [2 pts] Fill in the blanks to derive a modified Bellman equation that Alice can use to compute the values of states. Let $s' = D(s, a)$ and $s'' = D(s', \pi(s'))$.

$$V(s) = \max_a R(s, a, s') + \textbf{(i)}\ \textbf{(ii)} + \textbf{(iii)}\ \textbf{(iv)}$$

(i)	◯ 1	◯ 0.5	◯ 0.25
(ii)	◯ 0	◯ $R(s, \pi(s), s')$	◯ $R(s', \pi(s'), s'')$
(iii)	◯ 1	◯ 0.5	◯ 0.25
(iv)	◯ $V(s)$	◯ $V(s')$	◯ $V(s'')$

Q5. [13 pts] MDPs: Jim & Pam Part 2

Jim and Pam are on a $1x5$ grid, where Jim starts at square 1, and Pam is fixed at square 5.

Jim				Pam

In each time step, Jim chooses to either move right or to rest. Choosing to move succeeds with probability p and fails with probability $1-p$, in which case Jim stays in his original square (Jim received 0 utility regardless of success or failure). Choosing to rest always succeeds, and gives $R(d) = 4^{5-d}$ utility, where d is the distance between Jim and Pam. For example, at the start, where $d = 4$, if Jim decides to rest, he gets 4 utility. We represent this as an infinite horizon MDP with no terminal state.

(a) [3 pts] Jim is considering two policies:

Policy 1: Rest at the start forever.

Policy 2: Attempt to move right once, and then, regardless of success or failure, rest forever.

Assuming that Jim starts in square 1, for what values of p is Policy 1 superior to Policy 2 when the discount factor $\gamma = 0.5$? Hint: the sum S of an infinite geometric series with starting value a and ratio r is $S = \frac{a}{1-r}$

Show your work. $0 \leq \square \leq p \leq \square \leq 1$

(b) [3 pts]

Now assume that $p = 1$. Still assuming that Jim starts in square 1, for what values of γ is Policy 1 superior to Policy 2?

Show your work. $0 \leq \square < \gamma < \square \leq 1$

(c) [7 pts] For the following subparts, assume that $\gamma = 0.5$ and $p = 1$.

(i) [3 pts] Perform two iterations of value iteration, for the following locations of Jim. Show your work.

$States$	$s_J = 1$	$s_J = 2$	$s_J = 3$	$s_J = 4$	$s_J = 5$
V_0	0	0	0	0	0
V_1					
V_2					

(ii) [3 pts] Perform two iterations of policy iteration for the following locations of Jim.

$States$	$s_J = 1$	$s_J = 2$	$s_J = 3$	$s_J = 4$	$s_J = 5$
π_i	$a_J = right$	$a_J = rest$	$a_J = rest$	$a_J = rest$	$a_J = rest$
V^{π_i}					
π_{i+1}					
$V^{\pi_{i+1}}$					
π_{i+2}					

(iii) [1 pt] Assuming that policy iteration has converged, Jim argues it isn't guaranteed values have converged yet, so they need to run value iteration to get the correct values. Pam agrees that policy convergence doesn't guarantee value convergence, but thinks that we don't need to switch to value iteration, as if we continue running policy iteration, eventually the values will converge as well. Who is correct and why?

◯ Jim ◯ Pam

Q4. [13 pts] Slightly Different MDPs

Each subpart of this question is independent.

For all MDPs in this question, you may assume that:

- The maximum reward for any single action is some fixed number $R > 0$, and the minimum reward is 0.
- The discount factor satisfies $0 < \gamma \leq 1$.
- There are a finite number of possible states and actions.

(a) [2 pts] Which statements are always true? Select all that apply.

- ☐ $\sum_{s \in S} T(s, a, s') = 1$.
- ☐ $\sum_{s' \in S} T(s, a, s') = 1$.
- ☐ $\sum_{a \in A} T(s, a, s') \leq 1$.
- ☐ For all state-action pairs (s, a), there exists some s', such that $T(s, a, s') = 1$.
- ◯ None of the above

(b) [2 pts] Which statements are always true? Select all that apply.

- ☐ Every MDP has a unique set of optimal values for each state.
- ☐ Every MDP has a unique optimal policy.
- ☐ If we change the discount factor γ of an MDP, the original optimal policy will remain optimal.
- ☐ If we scale the reward function $r(s, a)$ of an MDP by a constant multiplier $\alpha > 0$, the original optimal policy will remain optimal.
- ◯ None of the above

(c) [2 pts] Which statements are true? Select all that apply.

- ☐ Policy iteration is guaranteed to converge to a policy whose values are the same as the values computed by value iteration in the limit.
- ☐ Policy iteration usually converges to exact values faster than value iteration.
- ☐ Temporal difference (TD) learning is off-policy.
- ☐ An agent is performing Q-learning, using a table representation of Q (with all Q-values initialized to 0). If this agent takes only optimal actions during learning, this agent will eventually learn the optimal policy.
- ◯ None of the above

(d) [2 pts] We modify the reward function of an MDP by adding or subtracting at most ϵ from each single-step reward. (Assume $\epsilon > 0$.)

We fix a policy π, and compute $V_\pi(s)$, the values of all states s in the original MDP under policy π.

Then, we compute $V'_\pi(s)$, the values of all states in the modified MDP under the same policy π.

What is the maximum possible difference $\left|V'_\pi(s) - V_\pi(s)\right|$ for any state s?

- ◯ ϵ
- ◯ $\gamma\epsilon$
- ◯ $\sum_{n=0}^{\infty} \epsilon(\gamma)^n = \epsilon/(1-\gamma)$
- ◯ ϵR
- ◯ None of the above

(e) [3 pts] We modify the reward function of an MDP by adding exactly C to each single-step reward. (Assume $C > 0$.) Let V and V' be the optimal value functions for the original and modified MDPs. Select all true statements.

- ☐ For all states s, $V'(s) - V(s) = \sum_{n=0}^{\infty} C(\gamma)^n = C/(1-\gamma)$.
- ☐ For all states s, $V'(s) - V(s) = C$.
- ☐ For some MDPs, the difference $V'(s) - V(s)$ may vary depending on s.
- ◯ None of the above

(f) [2 pts] We notice that an MDP's state transition probability $T(s, a, s')$ does **not** depend on the action a.

Can we derive an optimal policy for this MDP without computing exact or approximate values (or Q-values) for each state?

- ◯ Yes, because the optimal policy for this MDP can be derived directly from the reward function.
- ◯ Yes, because policy iteration gives an optimal policy and does not require computing values (or Q-values).
- ◯ No, because we need a set of optimal values (or Q-values) to do policy extraction.
- ◯ No, because there is no optimal policy for such an MDP.
- ◯ None of the above