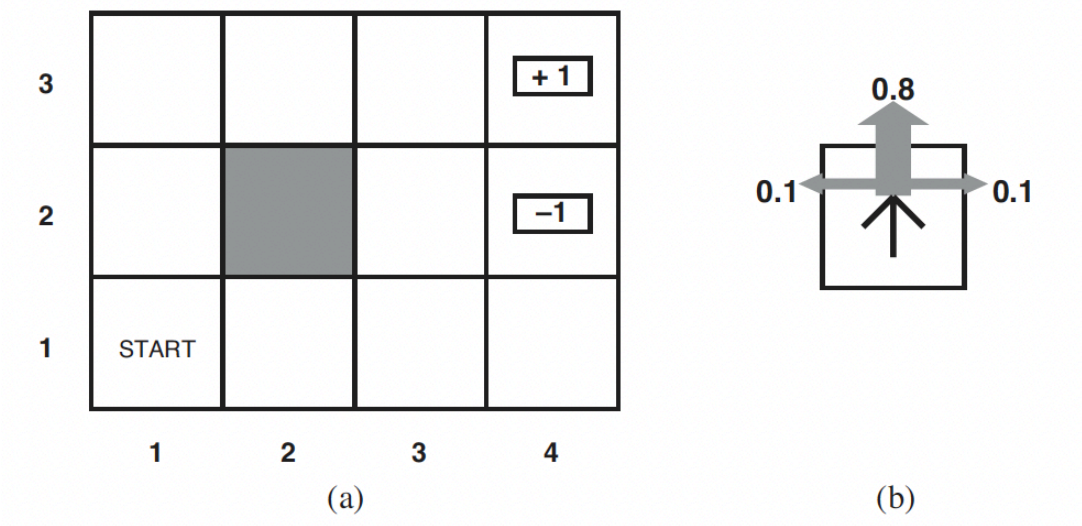
MDP	start time:

Before you start, share this document with your team member(s) and then complete the form below to assign the role of speaker.

Team Role	Team Member
Speaker: shares your team's ideas with the class.	Arogya, Tim, Nic

A. Grid World	start time:



The agent lives in the grid. Outside walls block the path. Also, there are walls around the shaded square (can't go there). There are two exits of the grid (goal state) with two rewards 1 and -1.

The position of the agent is a coordinate (column, row).

Noisy movement: actions do not always go as planned:

- 80% of time, agent goes in direction planned (e.g., North)
- 10% of time, agent goes in +90 deg direction (e.g., East)
- 10% of time, agent goes in –90 deg direction (e.g. West)
- If agent would run into a wall, it stays there.

Questions:

0a. Is this environment fully observable? yes

0b. Is this environment deterministic or stochastic?

stochastic

0c. Is the problem one-shot (episodic) or sequential?

sequential

1. In this simple simulated world, what is the state of the agent and what size is the state space?

- state of agent: position of the agent in the board
- size of state space: 11

2. For any given state of the agent, how many actions are at the state, including feasible and infeasible actions?

- 4, Each direction, N E S W

3. Let's temporarily forget about the noisy movement. Suppose the environment is deterministic. That means the agent always moves in the intended position 100% of the time. What is the sequence of actions for the agent to move from start to goal, in order to maximize the rewards with the fewest actions?

5, N N E E E or E E N N E

4. Now let's return to the stochastic environment. Assume the bottom left corner is position (1,1). For each of the following actions starting in position (1,1), what possible outcomes would you have:

a. North: Move North (80) , Move East (10), Wall West (10), 10 Wall
b. East: Move East (80), Move North (10), Wall South (10) 10 Wall
c. West: Wall West (80), Move North (10), Wall South (10), 90 Wall

5.1 Look at the possible outcomes for trying to move East. What percent of the time will you have each outcome? Give your answer in the form of T(s, a, s') = probability, where T means **transition model**, s is the current state, a is the action, s' is the new state after executing action a on state s.

T((1,1),EAST,(1,2)) = 0.8

T((1,1),EAST,(2,1)) = 0.1 (NORTH)

T((1,1),EAST,(1,1)) = 0.1 (SOUTH)

5.2 What if the action is moving to West?

T((1,1),WEST,(1,1)) = 0.8

T((1,1),WEST,(1,1)) = 0.1 (SOUTH WALL)

T((1,1),WEST,(2,1)) = 0.1 (NORTH)

6. Do you think the value of T(s, a, s') can be explained by the conditional probability P(s'| s, a)?

Yes

7. What is the probability for an agent to move along the path you answered in question 3 from start to goal? Before you run the calculation, given the 80% probability, what is your confidence that the agent will move to the +1 exit as planned?

T((1,1),NORTH,(2,1)) = 0.8

T((2,1),NORTH,(3,1)) = 0.8

T((3,1),EAST,(3,2)) = 0.8

T((3,2),EAST,(3,3)) = 0.8

T((3,3),EAST,(3,4)) = 0.8

Probability to get to +1: 0.33 Change of following intended path to goal state as intended (0.8 * 0.8 * 0.8 * 0.8 * 0.8)

8. The agent receives a reward for each move, with a big reward (or penalty) at the end.

a. What position gives the big reward, and what is it?

   3,4

b. What position gives the big penalty, and what is it?
   2,4

9. The agent also receives a small "living" reward (or penalty if negative). In this case, assume a penalty of R(s,a,s') = -0.04, which means every move from s to s' via action a costs -0.04. We here call R(s, a, s') as **reward function**. How does it affect the agent's behavior that it has a small penalty for every step?

When the agent takes the perfect path it has a fixed living penalty (-0.04 * 5). Any time the agent takes more steps (suboptimal path or is getting stuck on walls) it has a higher living cost.

10. The **utility** is defined as the total rewards from a sequence of transitions. What is the utility value of the path you answered in question 3 now with R(s,a,s') = -0.04, if the agent always moves as planned?

-0.2 (LIVING) + 1 (GOAL) = 0.8

11. Consider position (3, 3), the state immediately to the left of the +1 terminal state
    a. What is the best direction to try to move in from that position?

       EAST

    b. What are the possible outcomes? Give your answer in the form of T(s, a, s') = probability.

       T((3,3),EAST,(3,4)) = 0.8

       T((3,3),EAST,(3,3)) = 0.1 (NORTH WALL)

       T((3,3),EAST,(3,2)) = 0.1 (SOUTH)

    c. Give a mathematical equation for the expected utility of that state moving to next state, if the agent tried to move the best possible direction.

       -0.04 (MOVE) + 1 (GOAL) * 0.8 = 0.76

    d. Find the expected utility of that state moving to the next state.
       Undecided, it depends on the action.

12. Consider position (2, 3), the state left of the –1 terminal state?

    a. What is the best direction to try to move in from that position?
       North.

    b. What are the possible outcomes? Give your answer in the form of T(s, a, s') = probability.
       T((3,2), NORTH, (3,1)) = 0.1 (EAST) Penalty: -1
       T((3,2),NORTH, (3,2)) = 0.1 (WEST)
       T((3,2),NORTH, (3,3)) = 0.8 (NORTH)

    c. Give a mathematical equation for the expected utility of that state moving to next state, if the agent tried to move the best possible direction.

       -0.04 + -1 * 0.1 = -0.14

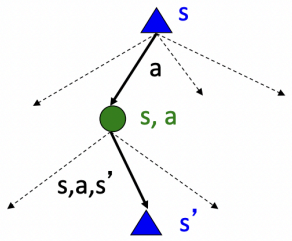
13. Repeated for some other nearby states.

B. Markov Decision Processes (MDP)	start time:

"Markov" generally means that given the present state, the future and the past are independent

Markov Decision Processes are **non-deterministic** and **sequential** search problems where actions' outcomes are **dependent only on the current state**, rather than the entire history prior to the current state.

- Markov decision processes:
  - Set of states S
  - Start state s_0
  - Set of actions A
  - Transitions P(s'|s,a) (or T(s,a,s'))
  - Rewards R(s,a,s') (and discount γ)



- MDP quantities so far:
  - Policy = Choice of action for each state
  - Utility = sum of (discounted) rewards

14. Is the Grid World in part A an example of MDP?

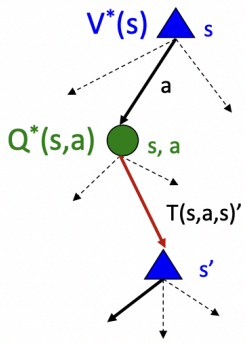
Yes

15. What consists of an MDP model? (Hint: use terminologies learned in part A)

B. Optimal Policy	start time:

Ultimately, the problem is to answer the question, **what move should the agent make in each square?** This set of state/action rules would be known as the agent's **policy**.

- The value (utility) of a state s:
  $V^*(s)$ = expected utility starting in s and acting optimally
- The value (utility) of a q-state (s,a):
  $Q^*(s,a)$ = expected utility starting out having taken action a from state s and (thereafter) acting optimally
- The optimal policy:
  $\pi^*(s)$ = optimal action from state s

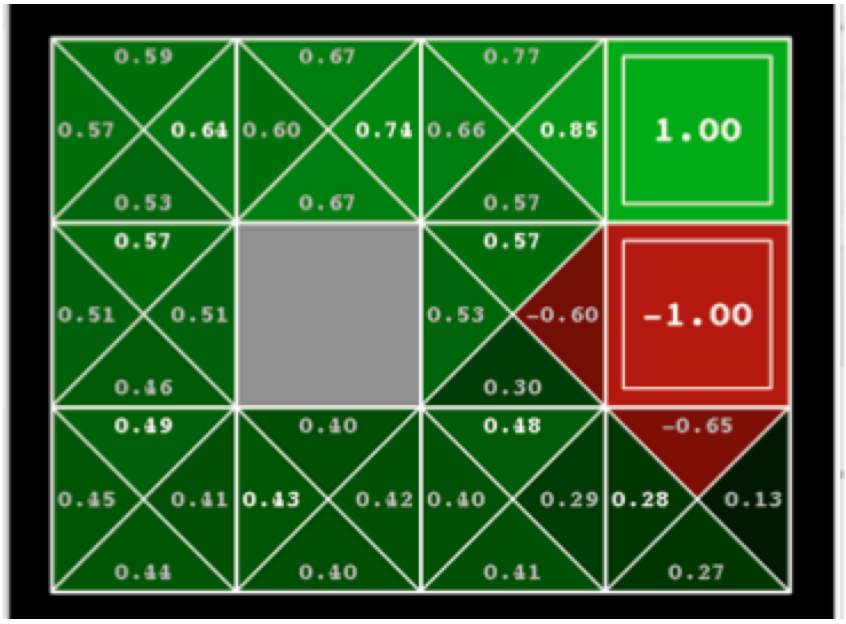


16. The goal of an MDP is to identify an **optimal policy** for every state in the problem. To calculate the policy for a Markov Decision Process, we need to calculate values from q-values:

values = expected future utility from a **state**

q-values = expected future utility from a **q-state** (when you've committed to an action, but you don't know what action will actually occur)

Here is a table of given q-values: (We will learn calculating the values in the next activity.)



a. Why are there four q-values for each state?
indicating 4 directions that the agent could take

b. Examine the state immediately left of the large reward, the state (3,3). Which of the four q-values is the largest, and what is therefore the optimal action to be taken from the state

(3,3)?

0.85 is the largest so EAST should be taken by the agent

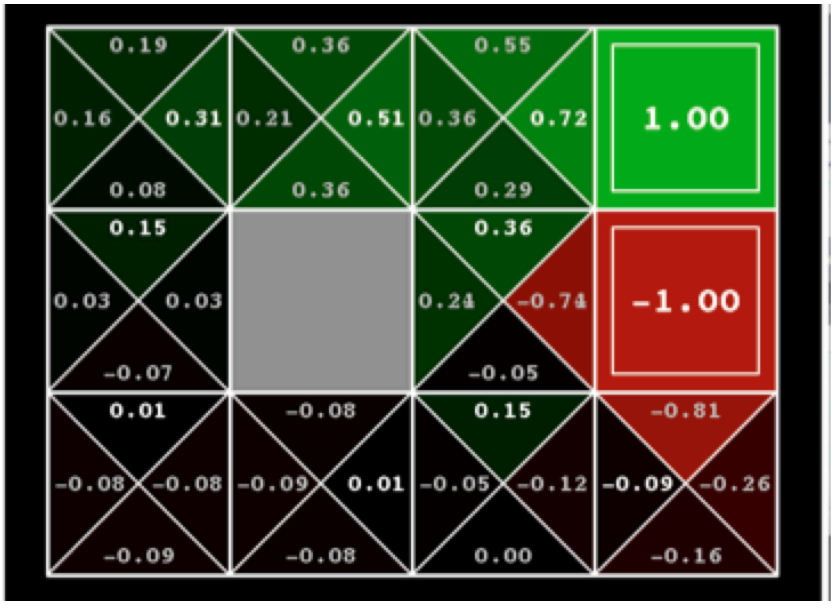
c. For each state, the largest q-value becomes the value for that state and identifies the ideal action. Fill in the value for each state in this gridworld:

0.64	0.74	0.85	**1.0**
0.57	**BLANK**	0.57	**-1.0**
0.49	0.4	0.48	0.28

d. And fill in the ideal action for each state in this gridworld (north, south, east, west). This is the optimal policy:

EAST	EAST	EAST	**1.0**
NORTH	**BLANK**	NORTH	**-1.0**
NORTH	WEST	NORTH	WEST

17. Here is another set of q-values:



a. Fill in the values for this version of gridworld:

0.31	0.51	0.72	**1.0**
0.15	**BLANK**	0.36	**-1.0**
0.01	0.01	0.15	-0.09

b. Fill in the optimal policy:

EAST	EAST	EAST	**1.0**
NORTH	**BLANK**	NORTH	**-1.0**
NORTH	EAST	NORTH	WEST

c. Identify the differences between the optimal policies between the two policies (question 16.d versus 17.b).

Below blank spot, there is EAST instead of WEST in 17. b

The probabilities or penalties seem to be different between the two games. Game two is harsher

d. What do you think is causing the difference in the q-values? Would it be the:
   - **Transition** function = T(s, a, s') = probability of state s' given that you take action a from state s
   - **Reward** function = R(s, a, s') = living penalty (or reward)

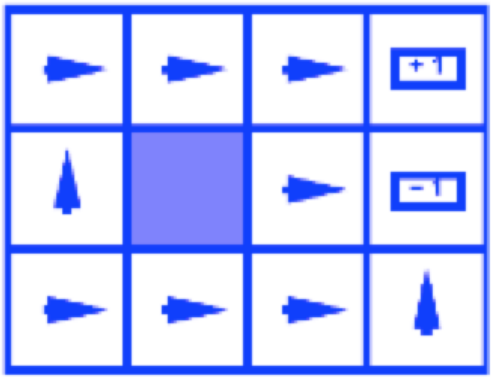
Living reward.

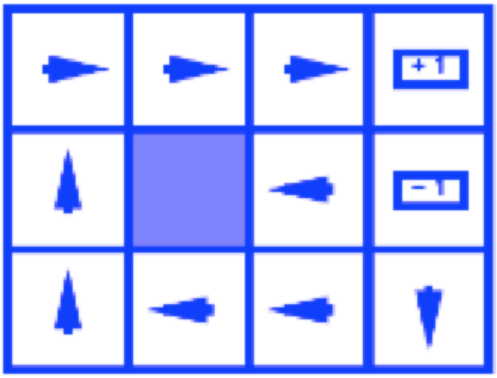
18. Below are four policies. When the reward function is negative, it serves as a small penalty for surviving another time step.

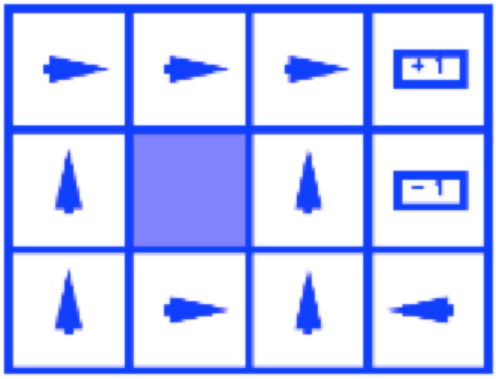
R(s, a, s')

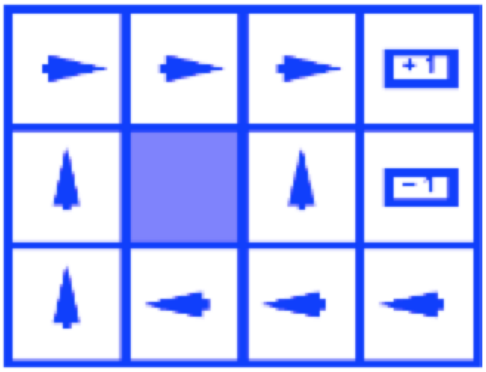
The following four policies are optimal policies for Gridworld when the living reward functions of -0.01, -0.03, -0.4, and -2.0.

Match the policy to the reward function and explain your reasoning.









Reward Function	Policy (paste the image here)	Rationale
-0.01	Top Right	Prioritizes getting the +1 over number of moves
-0.03	Bottom Right	Prioritizes +1 however is willing to take some risk
-0.4	Bottom Left	Prioritizes +1 however is willing to take higher risks
-2.0	Top Left	Prioritizes finishing the game as soon as possible. -1 is more reward than -2

19. What will happen for the agent if the R(s, a, s') is positive?

The agent will prioritize never finishing the game because it can maximize its reward.

20. Please draw the optimal policy map when the R(s, a, s') is positive?

Fill in the optimal policy:

N E W S	N E W S	W	**1.0**
N E W S	**BLANK**	W	**-1.0**
N E W S	N E W S	N E W S	S

D. Discounting in MDPs	start time:

MDPs typically use additive or multiplicative discounting.

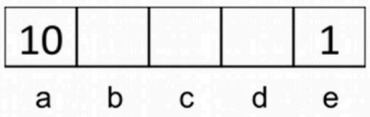
- With additive discounting (given below as an **additive** utility), there is typically a small negative reward for moving to non-terminal states.
- With multiplicative discounting (given below as an **discounted** utility), the value of the final reward is discounted by a factor γ (gamma), which is less than one, that is applied to each step.

The following formulas show the utility of receiving a series of rewards starting with r_0 and ending in some terminal reward:

- Additive utility: $U([r_0, r_1, r_2, \ldots]) = r_0 + r_1 + r_2 + \cdots$
- Discounted utility: $U([r_0, r_1, r_2, \ldots]) = r_0 + \gamma r_1 + \gamma^2 r_2 \cdots$

An example. Consider the following MDP with five states {a, b, c, d, e}, two terminal reward states {a, e}, deterministic actions {east, west, exit}, and multiplicative discounting.

- Given:



  - Actions: East, West, and Exit (only available in exit states a, e)
  - Transitions: deterministic

21. What does it mean that the transitions in this MDP are deterministic?

The next action depends on previous action and is predictable.

22. What are the possible actions from:

a. state a?
   east, exit
b. state b?
   east, west

23. Assume γ (gamma) is 1. What is the optimal policy and value for each state (a-e)?

c. State a: 10
d. State b: 10
e. State c: 10
f. State d: 10
g. State e: 1

24. Assume γ is 0.1. What is the optimal policy and value for each state (a-e)?

h. State a: 10
i. State b: 1
j. State c: 0.1
k. State d: 0.1
l. State e: 1

25. For which γ are West and East equally good when in state d? Hint: you'll need to do some algebra to solve this problem.

10 , 3.16227766, 1 ? 1

10, 3.16, 0.999, 0.316, 1

10*someval*someval*someval = 1*someval

10*someval*someval = 1

someval = 0.316

when gamma = d

26. Please list at least two reasons explaining that we typically include discounted rewards in MDPs.

- preventing the infinity maximization of the utility / preventing infinite life
- to reduce the uncertainity in the future

27. How does a large or a small discounting rate influence the agent decision in general?

- When the discount rate is large, the agent thinks about long term future rewards. If the discount rate is small, the agent wants to collect the rewards as soon as possible.

28. With discounted rewards, do you think the utility of an infinite sequence is finite? (Hint: assume the maximum reward in the environment is R_max)

- the total reward is converging

r + r^2 + r^3 + r^4 + … + r^n

r (1-r^n)/(1-r)

when n approaches infinity r^n is 0

r/(1-r) is a constant.