National University of Computer and Emerging Sciences

Peshawar Campus

Reinforcement Learning (CS3018)

Sessional-II Exam

Total Time (Hrs):	1
Total Marks:	30
Total Questions:	3

Date: 9th April, 2025

Course Instructor

Ms. Sara Rehmat

Roll No Section Student Signature

Do not write below this line

Attempt all the questions.

CLO 1: Explain the basic concepts and mathematical tools of reinforcement learning.

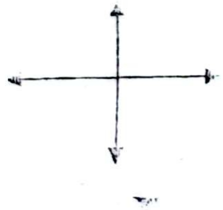
Q1: Differentiate between the following terms: **[Marks: 1*7=7]** **[Estimated mins: 15]**

a. On-policy and Off-policy algorithms
b. Control and Prediction
c. Model based and Model free algorithms
d. Value Iteration and Policy Iteration
e. State values and state-action values
f. Epsilon soft and epsilon greedy algorithms
g. Multi-armed bandits problem and Markov Decision Process

CLO 2: Implement basic RL algorithms to solve standard benchmark problems.

Q2: Consider an undiscounted Markov Decision Process in which states are represented by the following two-dimensional grid. The left diagram shows the labelling of the state and the right diagram shows the values of states under equiprobable random policy. Consider that the possible actions in each state are left, right, up or down. The next state is determined based on the corresponding action except that actions that would take the agent off the grid leave the state unchanged. The reward for each transition is -1. The MDP is episodic and the terminal states are grayed out.

[Marks: 3*2=6]
[Estimated mins: 10]



Actions

T	1	2
3	4	T
	5	

r = -1 on all transitions

State Labels

T	-4.5	-4
-5	?	T
	-10	

State Values

a. Utilizing the given information, find the value of state labelled 4.

b. Determine the optimal policy for given MDP and draw it i.e., label each state with an arrow representing best action in that state. No credit will be given if arrows are assigned without any working done.

Q3: Consider the following scenario to compute the state values for a robot operating in a small warehouse environment.

Imagine a small automated warehouse with three zones: **A**: Loading Area, **B**: Sorting Area and C: Packaging Area. It can choose to go Left or Right with equal probability, but due to physical limitations (e.g., slippery floor, tight spaces), its movement is not always successful. With 0.8 probability, it successfully moves in the intended direction. With 0.2 probability, it fails and remains in the same zone.

Consider the following rewards: +1 for short moves (A→B, B→C, B→A); +2 for long moves (A→C, C→A); 0 for all others. Since it's a continuous task, consider the discounting factor 0f 0.7.

[6+3+8=17 marks]

[Estimated mins: 25]

a. *Fill in the following table based on the given information for each combination of current state, action and next state.*

Current State (s)	Action (a)	Policy π(a|s)	Next state (s')	Transition probability p(s',r|s,a)	Reward (r)

b. Will p(s',r | s,a) be always same as p(s' | s,a) in the given case? Justify your answer.

c. Manually compute the state values for two iterations using the Dynamic Programming approach.