University at Buffalo
Department of Computer Science and Engineering
CSE 574: Introduction to Machine Learning
Quiz 10: Reinforcement Learning (Online)
Due date: 11 December 2024 (11:59 pm) Points: 10

The following questions are about Markov Decision Process. Read the case carefully before answering the questions.

Think about a simple game:

a. *Each round, you can either continue or quit.*
b. *If you quit, you receive $5 and the game ends.*
c. *If you continue, you receive $3 and roll a 6-sided die. If the die comes up as 1 or 2, the game will end. Otherwise, the game continues onto the next round.*

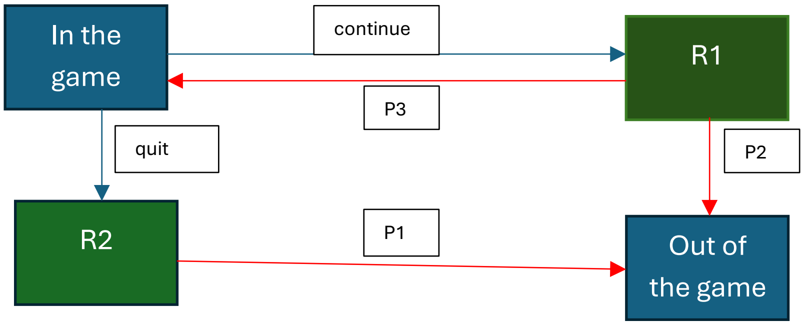
There is a clear trade-off here. For one, we can trade a deterministic gain of $2 for the chance to roll dice and continue to the next round.

To create an MDP to model this game, first we need to define a few things:

We can formally describe a Markov Decision Process as m = (S, A, P, R, gamma), where:

- *S represents the set of all states.*
- *A represents the set of possible actions.*
- *P represents the transition probabilities.*
- *R represents the rewards.*
- *Gamma is known as the discount factor. In this case the discount factor is 2/3.*

Question 1: Complete the probability values of P1, P2, and P3 for the following diagram which shows the MDP of the above scenario. Also, state the values of rewards R1 and R2. The red arrows show the probability for each possible scenario and green boxes show the rewards. (2.5 points)



Ans:

If we continue the game, we receive a reward of $3 so R1 = $3
If we quit the game, we receive a reward of $5 so R2 = $5

If we quit the game, we are out of the game there is no possibility of continuing the game so P1 is the probability of ending the game when choosing to quit is 1, P1 = 1

If we continue the game, we choose to roll the die. The probability of ending game when rolling 1 or 2 on a 6-sided die is 2/6 = 1/3. Hence P2 = 1/3

If we continue the game, we choose to roll the die. The probability of continuing game when rolling {3,4,5,6} on a 6-sided die is 4/6 = 2/3. Hence P3 = 2/3

Question 2: Now take the discount factor into account. There are two possible states, continue and quit in the above MDP. At each step, we can either quit and receive an extra $5 in expected value, or stay and receive an extra $3 in expected value. Each new round, the expected value is multiplied by 2/3, which is the discount factor. Draw the flow chart to show what will be the total reward for the two states after 4 rounds. Also identify the series of actions (for example: continue->continue->quit, etc) showing the maximization of the reward after 4 rounds. (5 points)

Ans: The player only goes to the next round if he continues to play. Discount Factor : 2/3

Round 1: Quit - $5 or Continue - $3

Round 2: Quit (Continue -> Quit) – 3 + (2/3) *5 = 6.3 or Continue (Continue -> Continue) – 3+(2/3) *3 = 5

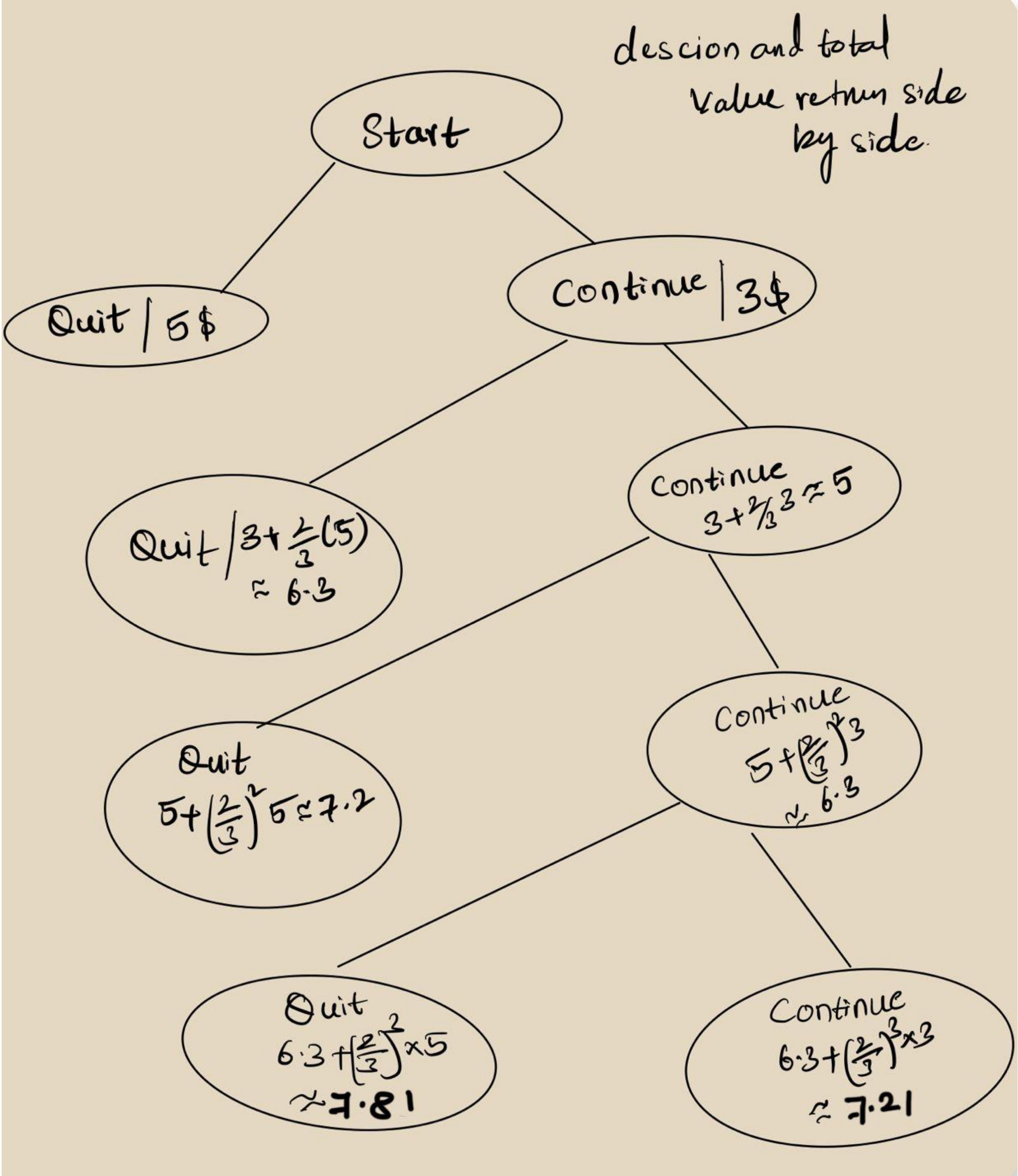
Round 3: Quit (Continue -> Continue -> Quit) – 5 + (2/3) ^2 * 5 = 7.2 or Continue (Continue -> Continue -> Continue) – 5+(2/3) ^2 * 3 = 6.3

Round 4: Quit (Continue -> Continue -> Continue -> Quit) – 6.3 + (2/3) ^3 * 5 = 7.81 or Continue (Continue -> Continue -> Continue-> Continue) – 6.3 + (2/3) ^3 * 3 = 7.21

Optimal Strategy: Continue -> Continue -> Continue -> Quit.

The maximum reward after 4 rounds, achieved using the optimal strategy, is approximately $7.81.

Flow Chart


descion and total
Value retnun side
by side.
Start
Quit | 5$
Continue | 3$
Quit | $3+\frac{2}{3}(5)$
≈ 6.3
Continue
$3+\frac{2}{3}3 \approx 5$
Quit
$5+\left(\frac{2}{3}\right)^2 5 \approx 7.2$
Continue
$5+\left(\frac{2}{3}\right)^2 3$
≈ 6.3
Quit
$6.3+\left(\frac{2}{3}\right)^2 \times 5$
≈ 7.81
Continue
$6.3+\left(\frac{2}{3}\right)^3 \times 3$
≈ 7.21

Question 3: Can you write down two rules that will mathematically show the computation for the two states "continue" and "quit" = using the vision of MDP which **takes the result of the previous step into consideration**? (2.5 points)

Ans:

R1 – reward received if we continue to play which is $3

R2 – reward received if we quit which is $5

From above the probabilities are

P1 = 1, P2 = 1/3, P3 = 2/3

Discount factor (γ) = 2/3

Mathematical rules for both states:

1. **Continue Action Rule**:

V_continue(k) = V_continue(k-1) + (γ)^(k-1) * R1

as we know γ and R1 by substituting values in the above equation we get

V_continue(k) = V_continue(k-1) + (2/3)^(k-1) * 3

Example:

Round 1: V_continue(0) = 0, without playing he won't get any thing
so k = 1 by substituting in the above equation we get V_continue(1) = 3

Round 2: k = 2

V_continue(2) = V_continue(1) + 2/3 * 3 = 3 + 2 = 5

2. Quit Action Rule:

V_quit(k) = V_continue(k-1) + (γ)^(k-1) *R2

as we know γ and R2 by substituting values in the above equation we get

V_quit(k) = V_continue(k-1) + (2/3) ^(k-1) * (5)