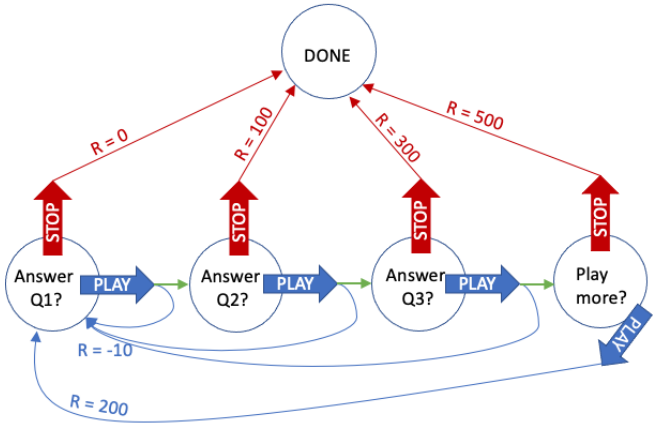
Learning Objectives

- To formulate a problem as an MDP
- To use value iteration to find the optimal policy for that MDP

Q1. Trivia Game Show

Suppose you sign up to play on a game show. The game show works like this. You start with $0. You'll be asked a series of questions of increasing difficulty and with increasing prize money for getting the question right. In other words, starting at the beginning with $0 in the pot, you can *Stop* and with probability 1.0 you win $0, or you can *Play* by answering question Q1 for the opportunity for $100. If you *Play*, you hear the question and answer it. If you get the answer correct, the show adds money in your pot (but you don't receive that money until you say *Stop*) and you get the opportunity to answer Q2. If you answer incorrectly, you must pay $10 and must start again and decide whether to answer Q1. The game ends when you decide to *Stop*, at which time you win a reward with probability 1.0 and end in a terminal Done state. The reward you receive for stopping depends on the number of questions answered.

The figure below shows the MDP we are trying to model and the transition matrix. Big arrows represent the action decisions. Small arrows represent the probabilistic transitions to different states. If there is only one small line from a big arrow, it means the transition probability is 1.0.



(s,a)	Q1	Q2	Q3	Done	More
(Q1,PLAY)	0.1	0.9			
(Q2,PLAY)	0.3		0.7		
(Q3,PLAY)	0.5				0.5
(MORE,PLAY)	1.0				
(Q1,STOP)				1.0	
(Q2,STOP)				1.0	
(Q3,STOP)				1.0	
(MORE,STOP)				1.0	

(a) Write the reward matrix $R(s, a, s')$.

(s,a)	Q1	Q2	Q3	Done	More
(Q1,PLAY)	-10				
(Q2,PLAY)	-10				
(Q3,PLAY)	-10				
(MORE,PLAY)	200				
(Q1,STOP)			0		
(Q2,STOP)			100		
(Q3,STOP)			300		
(MORE,STOP)			500		

Table 1: Rewards. All other values are 0.

(b) Suppose that $\gamma = 0.9$. Run value iteration for 3 iterations. What do you notice about the policy changing? Do you think the value will continue to change as we add more iterations? Do you think the policy will change after iteration 3?

Below shows the first 5 iterations of the value function. You can see that the values start changing less and less on each iteration and they will converge. The amount of extra money that gets added to the expected value over time is not enough to get above 300 and change the policy any more.

The policy for Q1, Q2, and Q3 is Play while the policy for Q4, Q5, and More is Stop. Why is this the case? The amount of extra points that the player receives versus the dropping likelihood that they will get the answer right means that it stops being advantageous to Play as the questions get harder.

	$V_0(*)$	$Q_1(*, Play)$	$Q_1(*, Stop)$	$V_1(*)$
Q1	0	(.1*(-10+.9*0))+(.9*(0+.9*0)) = -1	0	0
Q2	0	(.3*(-10+.9*0)+(.7*(0+.9*0)) = -3	100	100
Q3	0	(.5*(-10+.9*0)+(.5*(0+.9*0)) = -5	300	300
More	0	(1.0*(200+.9*0)) = 200	500	500
Done	0	0	0	0

Table 2: Value Iteration Iteration 1

	$V_1(*)$	$Q_2(*, Play)$	$Q_2(*, Stop)$	$V_2(*)$
Q1	0	(.1*(-10+.9*0))+(.9*(0+.9*100)) = 80	0	80
Q2	100	(.3*(-10+.9*0)+(.7*(0+.9*300)) = 186	100	186
Q3	300	(.5*(-10+.9*0)+(.5*(0+.9*500)) = 220	300	300
More	500	(1.0*(200+.9*0)) = 200	500	500
Done	0	0	0	0

Table 3: Value Iteration Iteration 2

	$V_2(*)$	$Q_3(*, Play)$	$Q_3(*, Stop)$	$V_3(*)$
Q1	80	(.1*(-10+.9*80))+(.9*(0+.9*186)) = 156.86	0	156.86
Q2	186	(.3*(-10+.9*80)+(.7*(0+.9*300)) = 207.6	100	207.6
Q3	300	(.5*(-10+.9*80)+(.5*(0+.9*500)) = 256	300	300
More	500	(1.0*(200+.9*80))=272	500	500
Done	0	0	0	0

Table 4: Value Iteration Iteration 3

	$V_3(*)$	$Q_4(*, Play)$	$Q_4(*, Stop)$	$V_4(*)$
Q1	156.86	(.1*(-10+.9*156.86))+(.9*(0+.9*207.6)) = 181.27	0	181.27
Q2	207.6	(.3*(-10+.9*156.86)+(.7*(0+.9*300)) = 228.35	100	228.35
Q3	300	(.5*(-10+.9*156.86)+(.5*(0+.9*500)) = 290.59	300	300
More	500	(1.0*(200+.9*156.86))=341.17	500	500
Done	0	0	0	0

Table 5: Value Iteration Iteration 4

	$V_3(*)$	$Q_4(*, Play)$	$Q_4(*, Stop)$	$V_4(*)$
Q1	181.27	(.1*(-10+.9*181.27))+(.9*(0+.9*228.35)) = 200.27	0	200.27
Q2	228.35	(.3*(-10+.9*181.27)+(.7*(0+.9*300)) = 234.94	100	234.94
Q3	300	(.5*(-10+.9*181.27)+(.5*(0+.9*500)) = 301.57	300	301.57
More	500	(1.0*(200+.9*181.27))=363.14	500	500
Done	0	0	0	0

Table 6: Value Iteration Iteration 5