0/39 Questions Answered

HW 8 (Electronic Component)

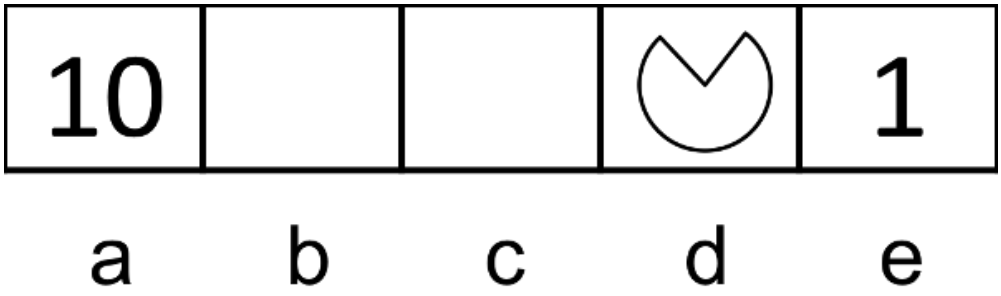
STUDENT NAME

Search students by name or email...

Q1 Solving MDPs

12 Points

Consider the gridworld MDP for which Left and Right actions are 100% successful. Specifically, the available actions in each state are to move to the neighboring grid squares. From state a, there is also an exit action available, which results in going to the terminal state and collecting a reward of 10. Similarly, in state e, the reward for the exit action is 1. Exit actions are successful 100% of the time.



Let the discount factor $\gamma = 1$. Fill in the following quantities, where $U_i(x)$ means the utility value of state x at iteration i of the value iteration algorithm (AIMA4e Figure 17.6).

Q1.1

2 Points

$U_0(d) =$

Enter your answer here

Save Answer

Q1.2
2 Points

$U_1(d) =$

Enter your answer here

Save Answer

Q1.3
2 Points

$U_2(d) =$

Enter your answer here

Save Answer

Q1.4
2 Points

$U_3(d) =$

Enter your answer here

Save Answer

Q1.5
2 Points

$U_4(d) =$

Enter your answer here

Save Answer

Q1.6
2 Points

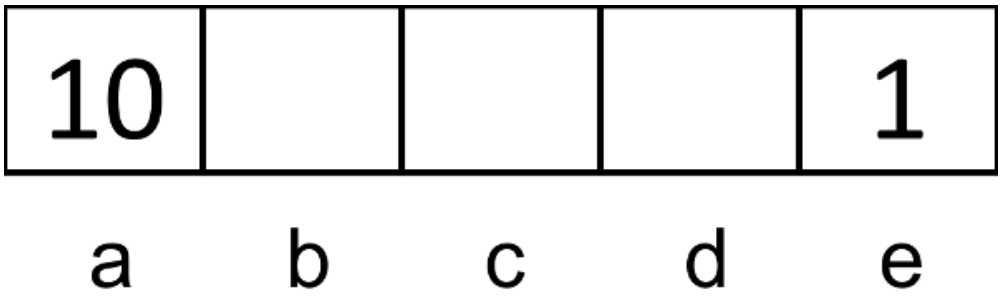
$U_5(d) =$

Enter your answer here

Save Answer

Q2 Value Iteration Convergence Values
10 Points

Consider the gridworld where Left and Right actions are successful 100% of the time. Specifically, the available actions in each state are to move to the neighboring grid squares. From state a, there is also an exit action available, which results in going to the terminal state and collecting a reward of 10. Similarly, in state e, the reward for the exit action is 1. Exit actions are successful 100% of the time.



Let the discount factor $\gamma = 0.2$. Fill in the following quantities.

Q2.1

2 Points

$U^*(a) =$

Enter your answer here

Save Answer

Q2.2

2 Points

$U^*(b) =$

Enter your answer here

Save Answer

Q2.3

2 Points

$U^*(c) =$

Enter your answer here

Save Answer

Q2.4

2 Points

$U^*(d) =$

Enter your answer here

Save Answer

Q2.5
2 Points

$U^*(e) =$

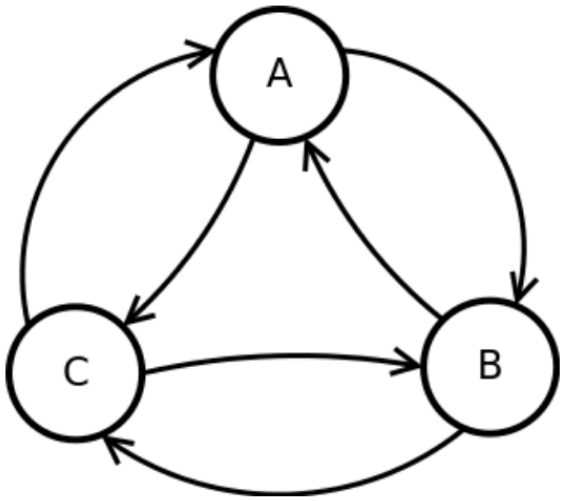
Enter your answer here

Save Answer

Q3
12 Points

We recommend you work out the solutions to the following questions on a sheet of scratch paper, and then enter your results into the answer boxes.

Consider the following transition diagram, transition function and reward function for an MDP. V has the same meaning as U in the lecture.



Discount Factor, $\gamma = 0.5$

s	a	s'	T(s,a,s')	R(s,a,s')
A	Clockwise	B	1.0	0.0
A	Counterclockwise	C	1.0	-2.0
B	Clockwise	A	0.4	-1.0
B	Clockwise	C	0.6	2.0
B	Counterclockwise	A	0.6	2.0
B	Counterclockwise	C	0.4	-1.0
C	Clockwise	A	0.6	2.0
C	Clockwise	B	0.4	2.0
C	Counterclockwise	A	0.4	2.0
C	Counterclockwise	B	0.6	0.0

Q3.1
3 Points

Suppose that after iteration k of value iteration we end up with the following values for V_k :

$V_k(A)$	$V_k(B)$	$V_k(C)$
0.400	1.400	2.160

What is $V_{k+1}(A)$?

Enter your answer here

Save Answer

Q3.2

3 Points

Now, suppose that we ran value iteration to completion and found the following value function, V^*.

$V^*(A)$	$V^*(B)$	$V^*(C)$
0.881	1.761	2.616

What is Q^*(A, clockwise)?

Enter your answer here

Save Answer

Q3.3

3 Points

Following Part 2, what is Q^*(A, counterclockwise)?

Enter your answer here

Save Answer

Q3.4
3 Points

Following Part 2, what is the optimal action from state A?

- ○ Clockwise
- ○ Counterclockwise

Save Answer

Q4 Value Iteration Properties
7 Points

Which of the following are true about value iteration? We assume the MDP has a finite number of actions and states, and that the discount factor satisfies $0 < \gamma < 1$.

- ☐ Value iteration is guaranteed to converge.
- ☐ Value iteration will converge to the same vector of values (U^*) no matter what values we use to initialize U.
- ☐ None of the above

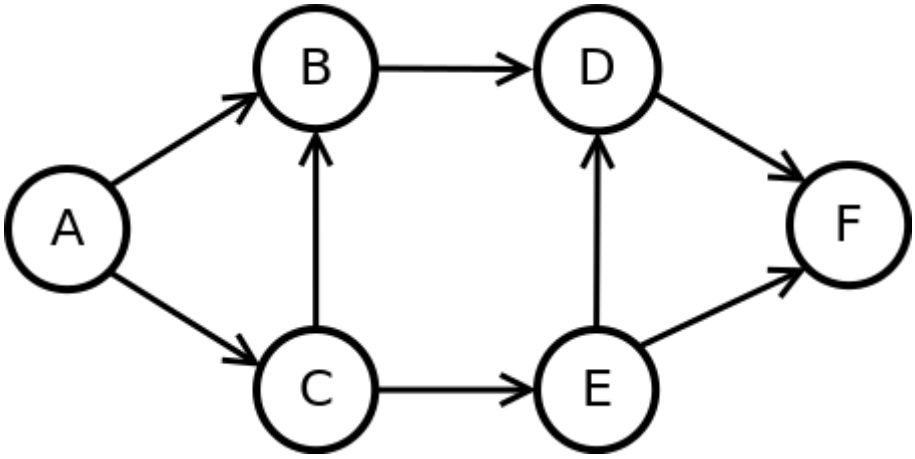
Save Answer

Q5 Value Iteration Convergence
6 Points

We will consider a simple MDP that has six states, A, B, C, D, E, and F. Each state has a single action, go. An arrow from a state x to a state y

indicates that it is possible to transition from state x to next state y when go is taken. If there are multiple arrows leaving a state x, transitioning to each of the next states is equally likely.

The state F has no outgoing arrows: once you arrive in F, you stay in F for all future times. The reward is 1 for all transitions, with one exception: staying in F gets a reward of 0. Assume a discount factor $\gamma = 0.5$. We assume that we initialize the value of each state to 0. (Note: you should not need to explicitly run value iteration to solve this problem.)



Q5.1

3 Points

After how many iterations of value iteration will the value for state E have become exactly equal to the true optimum? (Enter inf if the values will never become equal to the true optimal but only converge asymptotically to the true optimal.)

Enter your answer here

Save Answer

Q5.2

3 Points

How many iterations of value iteration will it take for the values of all states to converge to the true optimal values? (Enter inf if the values

will never become equal to the true optimal but only converge asymptotically to the true optimal.)

Enter your answer here

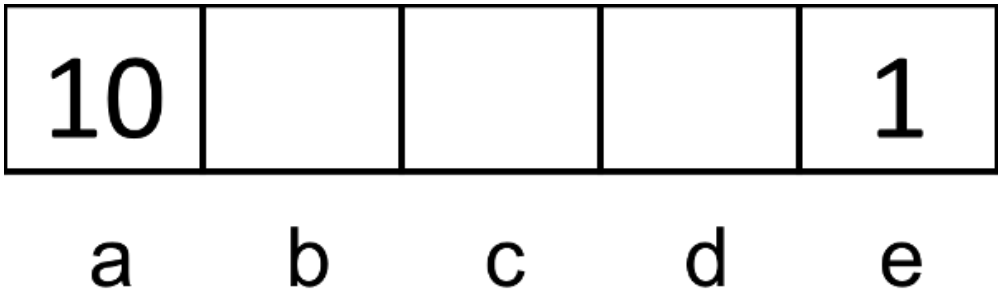
Save Answer

Q6 Policy Evaluation

10 Points

Consider the following gridworld where Left and Right actions are successful 100% of the time. Specifically, the available actions in each state are to move to the neighboring grid squares. From state a, there is also an exit action available, which results in going to the terminal state and collecting a reward of 10. Similarly, in state e, the reward for the exit action is 1. Exit actions are successful 100% of the time.

The discount factor (γ) is 1.

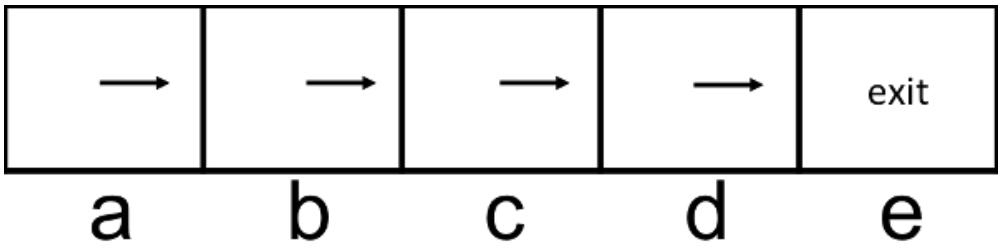


Q6.1

5 Points

Consider the policy π_1 shown below, and evaluate the following quantities

for this policy.



$U^{\pi_1}(a) =$

Enter your answer here

$U^{\pi_1}(b) =$

Enter your answer here

$U^{\pi_1}(c) =$

Enter your answer here

$U^{\pi_1}(d) =$

Enter your answer here

$U^{\pi_1}(e) =$

Enter your answer here

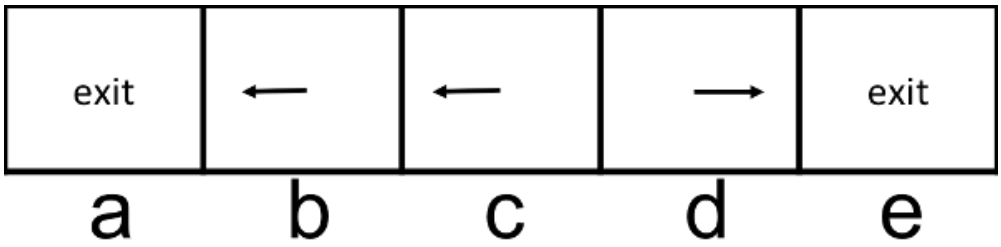
Save Answer

Q6.2

5 Points

Consider the policy π_2 shown below, and evaluate the following quantities

for this policy.



$U^{\pi_2}(a) =$

Enter your answer here

$U^{\pi_2}(b) =$

Enter your answer here

$U^{\pi_2}(c) =$

Enter your answer here

$U^{\pi_2}(d) =$

Enter your answer here

$U^{\pi_2}(e) =$

Enter your answer here

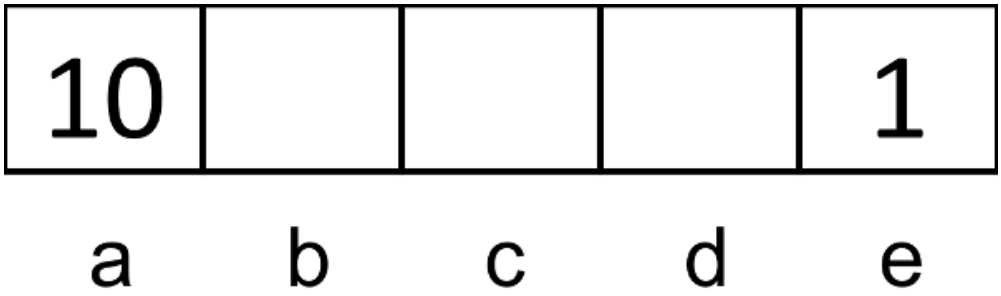
Save Answer

Q7 Policy Iteration

11 Points

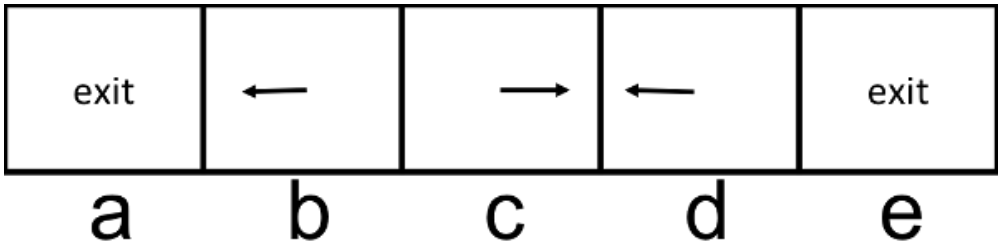
Consider the following grid world where Left and Right actions are successful 100% of the time. Specifically, the available actions in each state are to move to the neighboring grid squares. From state a, there is also an exit action available, which results in going to the terminal state and collecting a reward of 10. Similarly, in state e, the reward for the exit action is 1. Exit actions are successful 100% of the time.

The discount factor (γ) is 0.9.



We will execute one round of policy iteration.

Consider the policy π_i shown below, and evaluate the following quantities for this policy.



Q7.1
1 Point

$U^{\pi_i}(a) =$

Enter your answer here

Save Answer

Q7.2
1 Point

$U^{\pi_i}(b) =$

Enter your answer here

Save Answer

Q7.3
1 Point

$U^{\pi_i}(c) =$

Enter your answer here

Save Answer

Q7.4
1 Point

$U^{\pi_i}(d) =$

Enter your answer here

Save Answer

Q7.5
1 Point

$U^{\pi_i}(e) =$

Enter your answer here

Save Answer

Q7.6
1 Point

After the policy improvement, what is the new policy π_{i+1}? The 3 possible actions are {left, right, exit}.

$\pi_{i+1}(a) =$

Enter your answer here

Save Answer

Q7.7
1 Point

$\pi_{i+1}(b) =$

Enter your answer here

Save Answer

Q7.8
1 Point

$\pi_{i+1}(c) =$

Enter your answer here

Save Answer

Q7.9
1 Point

$\pi_{i+1}(d) =$

Enter your answer here

Save Answer

Q7.10
1 Point

$\pi_{i+1}(e) =$

Enter your answer here

Save Answer

Q7.11
1 Point

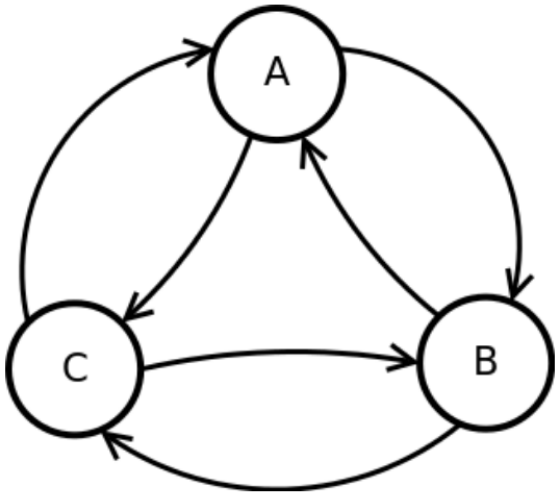
Is $\pi_{i+1} = \pi^*$ for all states?

- True
- False

Save Answer

Q8 Policy Iteration: Cycle
12 Points

Consider the following transition diagram, transition function and reward function for an MDP. V has the same meaning as U in the lecture.



Discount Factor, $\gamma = 0.5$

s	a	s'	T(s,a,s')	R(s,a,s')
A	Clockwise	B	0.8	0.0
A	Clockwise	C	0.2	2.0
A	Counterclockwise	B	0.4	1.0
A	Counterclockwise	C	0.6	0.0
B	Clockwise	C	1.0	-1.0
B	Counterclockwise	A	0.6	-2.0
B	Counterclockwise	C	0.4	1.0
C	Clockwise	A	1.0	-2.0
C	Counterclockwise	A	0.2	0.0
C	Counterclockwise	B	0.8	-1.0

Q8.1

3 Points

Suppose we are doing policy evaluation, by following the policy given by the left-hand side table below. Our current estimates (at the end of some iteration of policy evaluation) of the value of states when following the current policy is given in the right-hand side table.

A	B	C
Counterclockwise	Counterclockwise	Counterclockwise

$V_k^\pi(A)$	$V_k^\pi(B)$	$V_k^\pi(C)$
0.000	-0.840	-1.080

We recommend you work out the solutions to the following questions on a sheet of scratch paper, and then enter your results into the answer boxes.

What is $V_{k+1}^\pi(A)$?

Enter your answer here

Save Answer

Q8.2

3 Points

Suppose that policy evaluation converges to the following value function, V_∞^π.

$V_\infty^\pi(A)$	$V_\infty^\pi(B)$	$V_\infty^\pi(C)$
-0.203	-1.114	-1.266

Now let's execute policy improvement.

What is Q_∞^π(A, clockwise)?

Enter your answer here

Save Answer

Q8.3
3 Points

Following Part 2, what is Q^{π}_{∞}(A, counterclockwise)?

Enter your answer here

Save Answer

Q8.4
3 Points

Following Part 2, what is the updated action for state A?

- Clockwise
- Counterclockwise

Save Answer

Q9 Wrong Discount Factor
7 Points

Bob notices value iteration converges more quickly with smaller γ and rather than using the true discount factor γ, he decides to use a discount factor of $\alpha\gamma$ with $0 < \alpha < 1$ when running value iteration. Mark each of the following that are guaranteed to be true:

- [ ] While Bob will not find the optimal value function, he could simply rescale the values he finds by $\frac{1-\gamma}{1-\alpha}$ to find the optimal value function.
- [ ] If the MDP's transition model is deterministic and the MDP has zero rewards everywhere, except for a single exit transition at the unique goal state with a positive reward, then Bob will still find the optimal policy.
- [ ] If the MDP's transition model is deterministic, then Bob will still find the optimal policy.
- [ ] Bob's policy will tend to more heavily favor short-term rewards over long-term rewards compared to the optimal policy.
- [ ] None of the above.

Save Answer

Q10 MDP Properties

10 Points

Q10.1

5 Points

Which of the following statements are true for an MDP?

- [ ] If the only difference between two MDPs is the value of the discount factor then they must have the same optimal policy.
- [ ] For an infinite horizon MDP with a finite number of states and actions and with a discount factor γ that satisfies $0 < \gamma < 1$, value iteration is guaranteed to converge.
- [ ] When running value iteration, if the policy (the greedy policy with respect to the values) has converged, the values must have converged as well.
- [ ] None of the above

Save Answer

Q10.2
5 Points

Which of the following statements are true for an MDP?

- [ ] If one is using value iteration and the values have converged, the policy must have converged as well.
- [ ] Expectimax will generally run in the same amount of time as value iteration on a given MDP.
- [ ] For an infinite horizon MDP with a finite number of states and actions and with a discount factor γ that satisfies $0 < \gamma < 1$, policy iteration is guaranteed to converge.
- [ ] None of the above

Save Answer

Q11 Policies

7 Points

Jane, Gemma, Alvin and Michael all get to act in an MDP $(S, A, T, \gamma, R, s_0)$.

Jane runs value iteration until she finds U^* which satisfies $\forall s \in S$: $U^*(s) = \max_{a \in A} \sum_{s'} T(s, a, s')(R(s, a, s') + \gamma U^*(s'))$ and acts according to $\pi_{\text{Jane}} =$
$\arg\max_{a \in A} \sum_{s'} T(s, a, s')(R(s, a, s') + \gamma U^*(s'))$.

Gemma acts according to an arbitrary policy π_{Gemma}.

Alvin takes Gemma's policy π_{Gemma} and runs one round of policy iteration to find his policy π_{Alvin}.

Michael takes Jane's policy and runs one round of policy iteration to find his policy π_{Michael}.

Note: One round of policy iteration =
performing policy evaluation followed by performing policy improvement.

Mark all of the following that are guaranteed to be true:

- [ ] It is guaranteed that $\forall s \in S : U^{\pi_{\text{Gemma}}}(s) \geq U^{\pi_{\text{Alvin}}}(s)$
- [ ] It is guaranteed that $\forall s \in S : U^{\pi_{\text{Michael}}}(s) \geq U^{\pi_{\text{Alvin}}}(s)$
- [ ] It is guaranteed that $\forall s \in S : U^{\pi_{\text{Michael}}}(s) > U^{\pi_{\text{Jane}}}(s)$
- [ ] It is guaranteed that $\forall s \in S : U^{\pi_{\text{Gemma}}}(s) > U^{\pi_{\text{Jane}}}(s)$
- [ ] None of the above.

Save Answer

Save All Answers

Submit & View Submission ›