Q1 Numerical Answer Formatting

0 Points

Many of the questions in this homework have answers that are decimal numbers. Due to current limitations of Gradescope, your answers must be an exact string match to ours. In order to ensure an exact match, please carefully follow the following formatting for your numerical answers.

- Do not round decimals. None of the answers are infinite decimals, so include full precision (all answers should be less than 5 places after the decimal).
- Do not include any leading or trailing 0s unless they are necessary to show the location of the decimal
- If the number is an integer, do not include a decimal

Examples:

.1234

-.001

10.4

-10

0

Note: If you use the Python interpreter to do your math, floating-point error may lead to inexact decimal numbers. It is probably best to use another calculator, but if you do use Python you may need to adjust its output to get the actual exact answer.

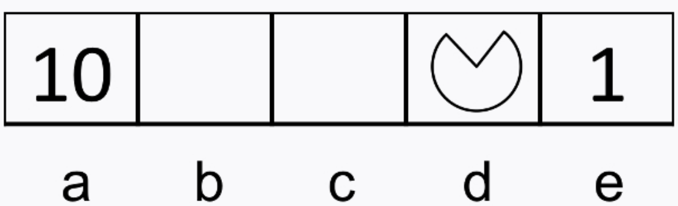
Q2 Solving MDPs

10 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.

$V_0(d) =$

0

$V_1(d)$

0

$V_2(d) =$

1

$V_3(d) =$

1

$V_4(d) =$

10

QUESTION 5
Value Iteration Convergence **0** / 10 pts
✔ + 0 pts Incorrect
+ 10 pts Correct

QUESTION 6
Policy Evaluation **10** / 10 pts

QUESTION 7
Policy Iteration **10** / 10 pts

QUESTION 8
Wrong Discount Factor **15** / 15 pts

QUESTION 9
MDP Properties **10** / 10 pts
9.1 (no title) **5** / 5 pts
9.2 (no title) **5** / 5 pts

QUESTION 10
Policies **10** / 10 pts

$V_5(d) =$

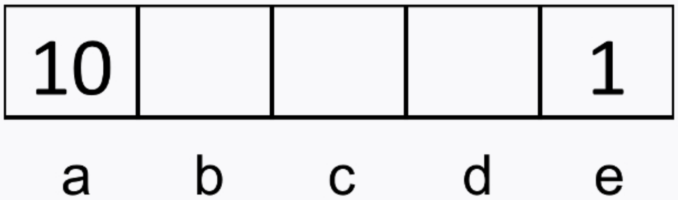
10

Q3 Value Iteration Convergence Values

15 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.

$V^*(a) = V_\infty(a)$

10

$V^*(b) = V_\infty(b) =$

2

$V^*(c) = V_\infty(c) =$

.4

$V^*(d) = V_\infty(d) =$

.2

$V^*(e) = V_\infty(e) =$

1

Q4 Value Iterations Properties

10 Points

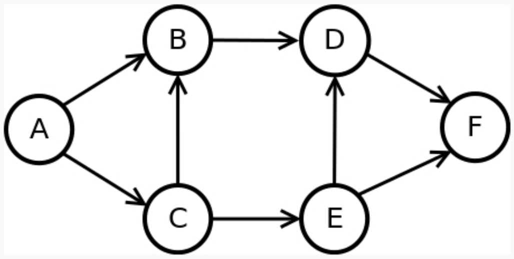
Which of the following are true about value iteration? We assume the MDP has a finite number of actions and states, and that the discount factor satisfies $0 < \gamma < 1$.

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

Q5 Value Iteration Convergence

10 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 one for all transitions, with one exception: staying in F gets a reward of zero. Assume a discount factor = 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.)



Part 1

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 to the true optimal.)

2

ANSWER

2

Part 2

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 to the true optimal.)

3

ANSWER

4

EXPLANATION

Because there are no moves from state F, we have the optimal value of F upon initializing. Since all the rewards are earned from transitions, finding the optimal value of a state amounts to finding the longest path from that state to F. For example, state D, whose longest path to F is only length 1, will find its optimal value after only one iteration.

$V^*(D) = V_1(D) = R(D, go, F) + \gamma V^*(F) = 1$

Similarly, the state A will find its optimal value after four iterations, because it will find out about its length 4 path to F after four iterations. Because A's length 4 path is the longest of the graph, it will take four iterations for all states to converge to their optimal values.

Q6 Policy Evaluation

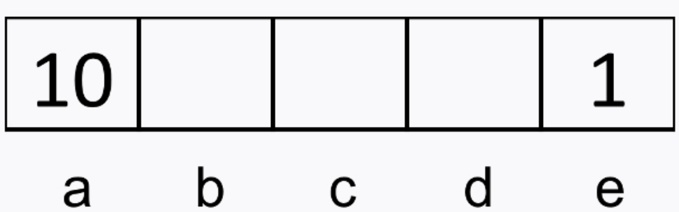
10 Points

Consider the gridworld where

Left and Right actions are successful 100% of the time.

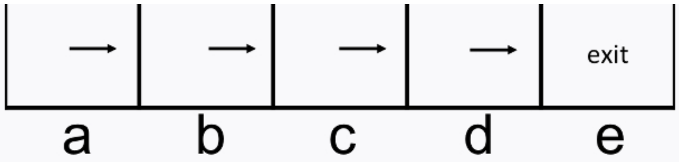
Specifically, the available actions in each state are to move to the neighboring grid squares. From state a, there is also an exit action available, which results in going to the terminal state and collecting a reward of 10. Similarly, in state e, the reward for the exit action is 1. Exit actions are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

The discount factor (γ) is 1.



Part 1

Consider the policy pi_1 shown below, and evaluate the following quantities for this policy.



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

1

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

1

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

1

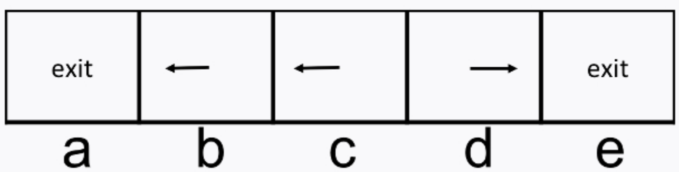
$V^{\pi_1}(d) =$

1

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

1

Part 2
Consider the policy π_2 shown below, and evaluate the following quantities for this policy.



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

10

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

10

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

10

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

1

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

1

Q7 Policy Iteration
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.
The discount factor (γ) is 0.9.

10				1
a	b	c	d	e

We will execute one round of policy iteration.

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

exit	←	→	←	exit
a	b	c	d	e

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

10

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

9

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

0

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

0

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

1

Q8 Wrong Discount Factor

15 Points

Bob notices value iteration converges more quickly with smaller $\gamma\gamma$ 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.
- [x] If the MDP's transition model is deterministic and the MDP has zero rewards everywhere, except for a single transition at the goal 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.
- [x] 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.

Q9 MDP Properties

10 Points

Q9.1

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.
- [x] For an infinite horizon MDP with a finite number of states and actions and with a discount factor \gammay 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

Q9.2
5 Points

Which of the following statements are true for an MDP?

- [x] 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.
- [x] 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

Q10 Policies
10 Points

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

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

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

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

Michael takes John'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 $\forall s \in S : V^{\pi_{\text{James}}}(s) \geq V^{\pi_{\text{Alvin}}}(s)$
- [x] It is guaranteed that $\forall s \in S : V^{\pi_{\text{Michael}}}(s) \geq V^{\pi_{\text{Alvin}}}(s)$
- [ ] It is guaranteed that $\forall s \in S : V^{\pi_{\text{Michael}}}(s) > V^{\pi_{\text{John}}}(s)$
- [ ] It is guaranteed that $\forall s \in S : V^{\pi_{\text{James}}}(s) > V^{\pi_{\text{John}}}(s)$
- [ ] None of the above.