Homework 3 Electronic

June 10rd, 2020 at 11:59pm

1 Nash Equilibria

Compute all (mixed and pure) Nash equilibria for each of the following normal-form games:

(a)

	L	R
T	1, 4	2, 3
B	0, 2	3, 5

(b)

	L	R
T	1, 4	2, 5
B	7, 1	2, 2

(c)

	L	R
T	2, 1	5, 1
B	5, 1	3, 1

(a) (B,R),(T,L),((3/4,1/4),(1/2,1/2))
(b) (B,R),(T,R)
(c) (B,L),(T,R),((any),(2/5,3/5))

Sample Answer:
(T,L),(T,R),((1/3,2/3),(3/4,1/4))

Especially, if no mixed equilibria, write as:
(T,L),(T,R)

If one player is indifferent for any mixed strategy, write as:
((any),(P,1-P))

2 Strategies

Consider the following game in matrix form with two players. Payoffs for the row player Izzy are indicated first in each cell, and payoffs for the column player Jack are second.

	X	Y	Z
S	5, 2	10, 6	25, 10
T	10, 12	5, 6	0, 0

(a) This game has two pure strategy Nash equilibria. What are they (justify your answer)? Of the two pure equilibria, which would Izzy prefer? Which would Jack prefer?

Sample Answer:
(S,X),(S,Y)
Izzy(S,X)
Jack(S,Y)

(S,Z),(T,X)
Izzy(S,Z)
Jack(T,X)

(b) Suppose Izzy plays a strictly mixed strategy, where both S and T are chosen with positive probability. With what probability should Izzy choose S and T so that each of Jack's three pure strategies is a best response to Izzy's mixed strategy.

Sample Answer:
(1/2,1/2) (3/5,2/5)

Note that the former is the probability of S and the latter is the probability of T.

(c) Suppose Jack wants to play a mixed strategy in which he selects X with probability 0.7. With what probability should Jack plays actions Y and Z so both of Izzy's pure strategies is a best response to Jack's mixed strategy?

Sample Answer:
(1/4,3/4) (1/5,1/10)

Note that the former is the probability of Y and the latter is the probability of Z.

(d) Based on your responses above, describe a mixed strategy equilibrium for this game in which both Jack and Izzy play each of their actions (pure strategies) with positive probability (you can rely on the quantities computed in the prior parts of this question).

Sample Answer:
((1/4,3/4),(7/10,3/20,3/20)) ((3/5,2/5),(7/10,1/5,1/10))

Note that the first tuple is the strategy of Izzy and the second one is the strategy of Jack.

(e) If we swap two of Izzy's payoffs in this matrixin other words, if we replace one of his payoffs r in the matrix with another of his payoffs t from the matrix, and replace t with r, we can make one of his strategies dominant. What swap should we make, which strategy becomes dominant?

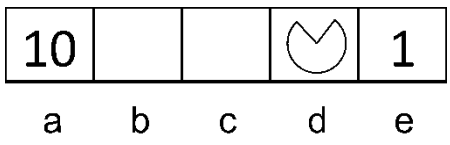
Sample Answer:
(S,X),(S,Y),S (S,X),(T,X),S

Note: The two tuples are strategies of which you want to exchange Izzy's payoffs. And please write in dictionary order, e.g. (S, Y) before (T, X). The S is the dominant strategy after the payoff exchange.

3 Solving MDPs

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.



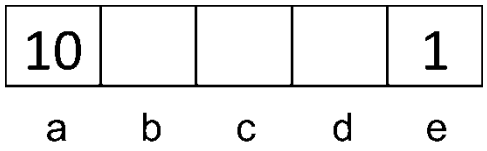
Let the discount factor $\gamma = 1$. Write the following quantities in one line.
$V_0(d), V_1(d), V_2(d), V_3(d), V_4(d), V_5(d)$

Sample Answer:
0,0,0,0,1,10 0,0,1,1,10,10

4 Value Iteration Convergence Values

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 =0.2. Fill in the following quantities.

$$V^*(a) = V_\infty(a) = \_\_\_\_\_$$
$$V^*(b) = V_\infty(b) = \_\_\_\_\_$$
$$V^*(c) = V_\infty(c) = \_\_\_\_\_$$
$$V^*(d) = V_\infty(d) = \_\_\_\_\_$$
$$V^*(e) = V_\infty(e) = \_\_\_\_\_$$

Write the following quantities in one line.
Sample Answer:
0,0,0,0,1

10,2,0.4,0.2,1

5 Value Iteration Properties

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 $0 < \gamma < 1$.

A. Value iteration is guaranteed to converge.

B. Value iteration will converge to the same vector of values (V^*) no matter what values we use to initialize V.

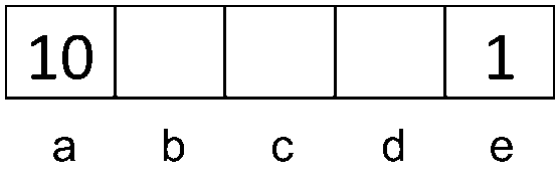
C. None of the above.

6 Policy Iteration

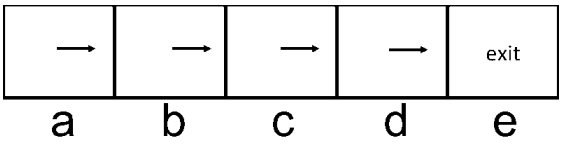
Consider the gridworld where Left and Right actions are successful 100% of the time.

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.



(a) Consider the policy π_1 shown below, and evaluate the following quantities for this policy. Write your answers in one line.

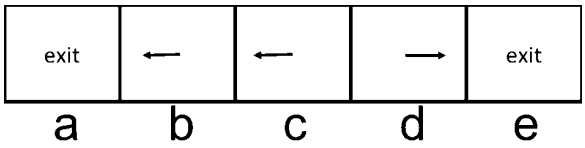


$V^{(\pi_1)}(a) =$ _____
$V^{(\pi_1)}(b) =$ _____
$V^{(\pi_1)}(c) =$ _____
$V^{(\pi_1)}(d) =$ _____
$V^{(\pi_1)}(e) =$ _____

Sample Answer:
0,0,0,0,1

1,1,1,1,1

(b) Consider the policy π_2 shown below, and evaluate the following quantities for this policy. Write your answers of the same format in Part 1 in one line.



$V^{(\pi_2)}(a) =$ _____
$V^{(\pi_2)}(b) =$ _____
$V^{(\pi_2)}(c) =$ _____
$V^{(\pi_2)}(d) =$ _____
$V^{(\pi_2)}(e) =$ _____

Sample Answer:
0,0,0,0,1

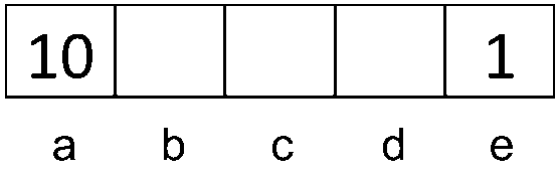
10,10,10,1,1

7 Policy Iteration

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

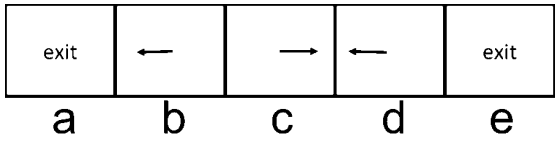
Specifically, the available actions in each state are to move to the neighboring grid squares. From state a, there is also an exit action available, which results in going to the terminal state and collecting a reward of 10. Similarly, in state e, the reward for the exit action is 1. Exit actions are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

The discount factor (γ) is 0.9.



(a) Policy Iteration

Consider the policy π_i shown below, and evaluate the following quantities for this policy. Write your answers in one line.



$V^{(\pi_i)}(a) =$ _____

$V^{(\pi_i)}(b) =$ _____

$V^{(\pi_i)}(c) =$ _____

$V^{(\pi_i)}(d) =$ _____

$V^{(\pi_i)}(e) =$ _____

Sample Answer:
0,0,0,0,1 10,9,0,0,1

(b) Policy Improvement

Perform a policy improvement step. The current policy's values are the ones from Part 1 (so make sure you first correctly answer Part 1 before moving on to Part 2). Write your answers in one line.

(i). $\pi_{(i+1)}(a) =$
- A. Exit
- B. Right

(ii). $\pi_{(i+1)}(b) =$
- A. Left
- B. Right

(iii). $\pi_{(i+1)}(c) =$
- A. Left
- B. Right

(iv). $\pi_{(i+1)}(d) =$
- A. Left
- B. Right

(v). $\pi_{(i+1)}(e) =$
- A. Left
- B. Exit

Sample Answer:
A,A,A,A,B

8 Wrong Discount Factor

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. Write the options that are guaranteed to be true:

A. 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.

B. 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.

C. If the MDP's transition model is deterministic, then Bob will still find the optimal policy.

D. Bob's policy will tend to more heavily favor short-term rewards over long-term rewards compared to the optimal policy.

E. None of the above.

9 MDP Properties

(a) Which of the following statements are true for an MDP?

A. If the only difference between two MDPs is the value of the discount factor then they must have the same optimal policy.

B. 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.

C. When running value iteration, if the policy (the greedy policy with respect to the values) has converged, the values must have converged as well.

D. None of the above.

(b) Which of the following statements are true for an MDP?

A. If one is using value iteration and the values have converged, the policy must have converged as well.

B. Expectimax will generally run in the same amount of time as value iteration on a given MDP.

C. 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.

D. None of the above.

10 Policies

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.

Write all of the options that are guaranteed to be true:

A. It is guaranteed that $\forall s \in S: \ V^{\pi_{\text{James}}}(s) \geq V^{\pi_{\text{Alvin}}}(s)$.

B. It is guaranteed that $\forall s \in S: \ V^{\pi_{\text{Michael}}}(s) \geq V^{\pi_{\text{Alvin}}}(s)$.

C. It is guaranteed that $\forall s \in S: \ V^{\pi_{\text{Michael}}}(s) > V^{\pi_{\text{John}}}(s)$.

D. It is guaranteed that $\forall s \in S: \ V^{\pi_{\text{James}}}(s) \geq V^{\pi_{\text{John}}}(s)$.

E. None of the above.