CS 188
Fall 2022

Introduction to
Artificial Intelligence

Challenge Q13 HW4

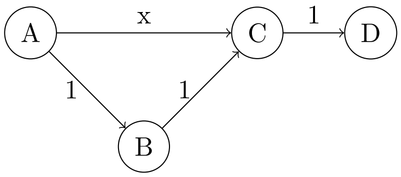
- **Due:** Friday 9/30 at 11:59pm.
- **Policy:** Can be solved in groups (acknowledge collaborators) but must be submitted individually.
- **Make sure to show all your work and justify your answers**.
- **Note:** This is a typical exam-level question. On the exam, you would be under time pressure, and have to complete this question on your own. We strongly encourage you to first try this on your own to help you understand where you currently stand. Then feel free to have some discussion about the question with other students and/or staff, before independently writing up your solution.
- Your submission on Gradescope should be a PDF that matches this template. Each page of the PDF should align with the corresponding page of the template (page 1 has name/collaborators, question begins on page 2.). **Do not reorder, split, combine, or add extra pages**. The intention is that you print out the template, write on the page in pen/pencil, and then scan or take pictures of the pages to make your submission. You may also fill out this template digitally (e.g. using a tablet.)

First name	
Last name	
SID	
Collaborators	

For staff use only:

Q13. [20 pts] Challenge Question (MDP)

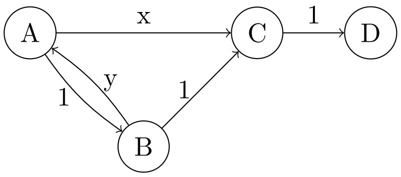
13.1) (5 pts) Consider the following deterministic MDP with four states A, B, C and D:



The edges designate actions between states, the weights on those edges are the rewards, and the discount factor is $\gamma = 1$. Let k be the **first** iteration of Value Iteration at which the value function converges for some x for a particular state (i.e. $V_k(s) = V^*(s)$). For each state A, B, C and D, list **all** possible values of k. In the case a value function for a particular state never converges, set $k = \infty$ for that state.

For A $k = 2$ or $k = 3$, for B $k = 2$, for C $k = 1$ and for D $k = 0$. C will find its optimal value in one iteration. B will find its optimal value one iteration after that (2 iterations total). A will find its optimal value one iteration after C (2 iterations total) or one iteration after B (3 iterations total), depending on the value of x.

Now for questions 13.2) and 13.3) consider the following deterministic MDP with four states A, B, C and D:



The edges designate actions between states, the weights on those edges are the rewards, and the discount factor is again $\gamma = 1$. Furthermore assume that $x, y \geq 0$.

13.2) (5 pts) Let k be the **first** iteration of Value Iteration for some nonnegative x and y at which the value function converges for a particular state ($V_k(s) = V^*(s)$). For each state A, B, C and D list **all** possible values of k. In case a value for a particular state never converges set $k = \infty$ for that state.

For A $k = \infty$, for B $k = \infty$, for C $k = 1$ and for D $k = 0$. A and B will never find their optimal value because they can get infinite value. C and D are the same as above.

13.3) (6 pts) Now consider that we perform Policy Iteration and that k is the **first** iteration for which the policy is optimal for a particular state (i.e. $\pi_k(s) = \pi^*(s)$). On top of $x, y \geq 0$ also assume that $x + y < 1$ and that tie-breaking during policy improvement is alphabetical. For each state A, B, C and D, find k; if the policy never converges set $k = \infty$ for that state. The initial policy is given in the table below.

State s	Policy $\pi_0(s)$
A	C
B	C
C	D
D	D

For A $k = 1$, for B $k = 2$ and for C, D $k = 0$. First, evaluate π_0 :

$$\begin{aligned}
&V^{\pi_0}(D) = 0 \\
&\Rightarrow V^{\pi_0}(C) = 1 + V^{\pi_0}(D) = 1 \\
&\Rightarrow V^{\pi_0}(B) = 1 + V^{\pi_0}(C) = 2 \\
&\Rightarrow V^{\pi_0}(A) = x + V^{\pi_0}(C) = x + 1
\end{aligned}$$

Now do policy improvement to obtain π_1:

$$\begin{aligned}
&\pi_1(D) = D\\
&\pi_1(C) = D\\
&\pi_1(B) = \text{argmax}_{A,C}\{A : x + y + 1, C : 2\} = C\\
&\pi_1(A) = \text{argmax}_{B,C}\{B : 3, C : x + 1\} = B
\end{aligned}$$

Now, evaluate π_1 :

$$\begin{aligned}
&V^{\pi_1}(D) = 0\\
&\Rightarrow V^{\pi_1}(C) = 1 + V^{\pi_1}(D) = 1\\
&\Rightarrow V^{\pi_1}(B) = 1 + V^{\pi_1}(C) = 2\\
&\Rightarrow V^{\pi_1}(A) = 1 + V^{\pi_1}(B) = 3
\end{aligned}$$

Now run policy improvement to obtain π_2 :

$$\begin{aligned}
&\pi_2(D) = D\\
&\pi_2(C) = D\\
&\pi_2(B) = \text{argmax}_{A,C}\{A : y + 3, C : 2\} = A\\
&\pi_2(A) = \text{argmax}_{B,C}\{B : 3, C : x + 1\} = B
\end{aligned}$$

Observe that this policy is optimal, because the value $V^{\pi_2}(A) = V^{\pi_2}(B) = \infty$. The other values are trivially optimal because the agent has only one choice of action.

The following two questions are conceptual.

13.4) (2 pts) Which of the following statements are guaranteed to be correct for any MDP? Select all that apply.

A. For all states s and for all policies π, $V^{\pi}(s) \leq V^*(s)$.

B. For no state s and for all policies π, $V^{\pi}(s) \leq V^*(s)$.

C. For some state s and some policy π, $V^{\pi}(s) \leq V^*(s)$.

D. None of the above.

A,C. A is a property of the optimal policy and C is just a special case of that. B cannot be true as the optimal policy cannot be worse than a random one.

13.5) (2 pts) Which of the following statements are guaranteed to be correct for Value Iteration? Select all that apply.

A. At each iteration, the value functions are at least as high as the values at the previous iteration for all states.

B. At each iteration, the value functions are higher than the values at the previous iteration for all states.

C. At each iteration, the value function can be lower than the earlier values for some state.

D. Once its converged, value iteration does not change the value function for any state.

E. None of the above.

C,D. For C before convergence values can fluctuate. For D upon convergence everything stays fixed and does not change anymore. For A, B there is no guarantee that this will happen, as values fluctuate until convergence.