UBC CPSC 422 2021W1

Homework 1

Due Oct 1 @ 11:59pm

Submit your assignment on Gradescope in PDF format. Recall that you may work in groups of up to 2 people, but you must submit jointly (i.e. one submission for the partnership).

Make sure you include the following on the first page of your submission:

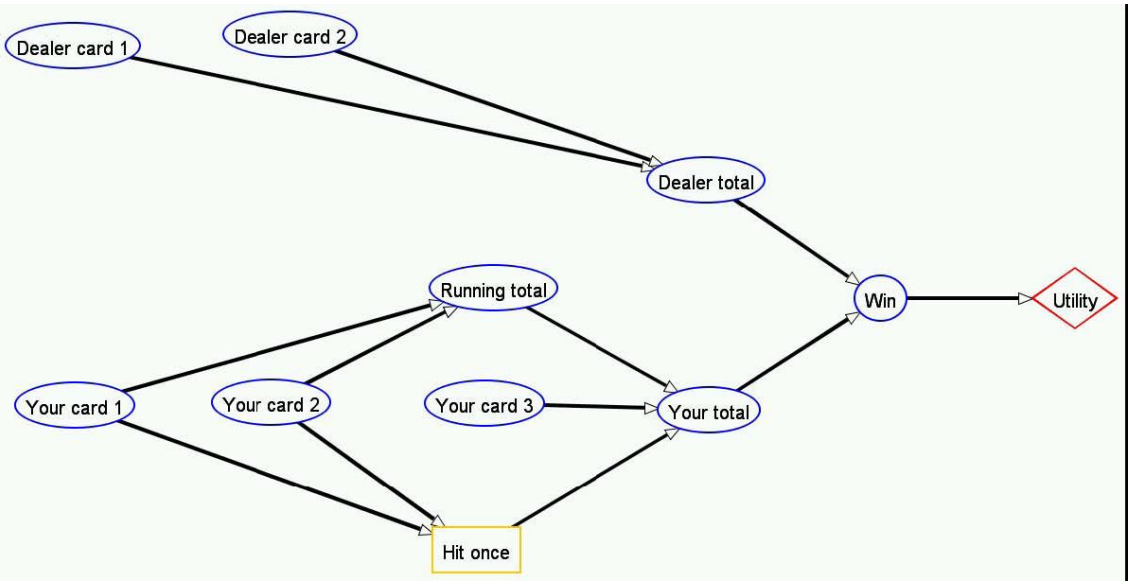
- The number of late days claimed (if any)
- The names of any students (outside your partnership, if applicable) with whom you have discussed the assignment

By uploading a submission, you are declaring that said submission is **your work** and has been created in accordance with the academic conduct policies of the university and those laid out in the CPSC 422 course syllabus.

- If you make use of any sources outside of the provided course material, ensure that those sources are cited in the body of your submission in an accessible format of your choice.

Question 1 – Value of Information and Value of Control [20 points]

Download the **Blackjack.xml** file from Canvas. Open it with the Belief and Decision Networks applet in AIspace. It should appear as this decision network (which you may have already seen in 322).



If needed, read the tutorial on that applet. Then, in AIspace, examine the details of the model. Make sure that you understand the rationale for all the nodes and their connections. Finally, use the model to answer the following questions.

(a) [6 points] What is the value of information for **Dealer card 1**? How do you calculate this? What changes (if any) do you make to the network in order to calculate this?

(b) [6 points] What is the value of control for **Running Total**? How do you calculate this? What changes (if any) do you make to the network in order to calculate this?

(c) [8 points] What is the value of control for **Dealer Card 2**? What is an optimal policy when Dealer Card 2 is controlled? Assume that you control the dealer's card after observing your own. Note: There are many parts of policy space that aren't reachable or don't affect the outcome (e.g. what to do when your first 2 cards result in a bust). You don't need to describe these parts of your policy.

Question 2 – Value Iteration [35 points]

In this question, you will be using an applet to improve your understanding of value iteration. You can find the applet at https://artint.info/demos/mdp/vi.html

Note: modern browsers don't seem to like Java. There are workarounds, but a vastly less painful way to access the applet is to make sure you have the Java appletviewer installed (which should be the case if you have the JDK installed). You might need an older version of Java, unfortunately: https://www.oracle.com/java/technologies/javase/javase8-archive-downloads.html

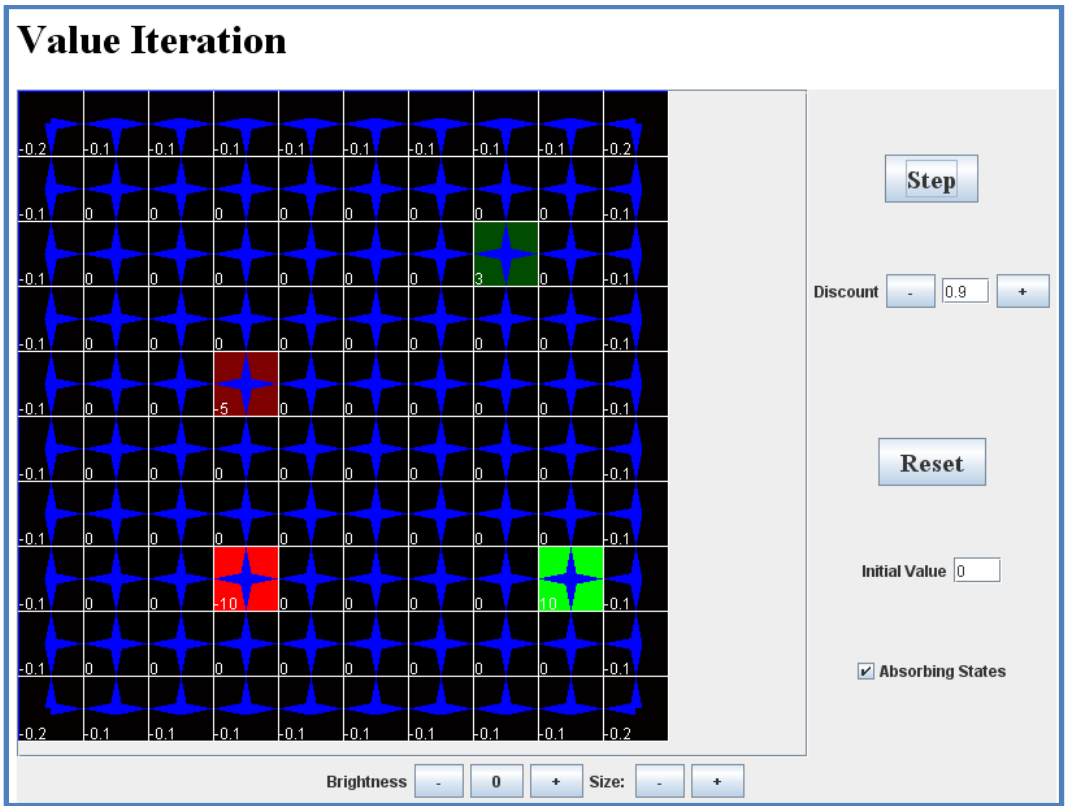
With the appletviewer installed, from your command line run

```
appletviewer https://artint.info/demos/mdp/vi.html
```

You may need to first navigate to the directory where the appletviewer program is located.

If running the appletviewer on the URL fails, follow the instructions provided at the artint.info site linked above to download the code and run the appletviewer on it locally.

In this assignment, we are using a discount factor of 0.9, initial values of $U^{(0)}(s) = 0$ for all s, and the "absorbing states" option (explained in detail on the website with the applet).
In contrast to the convention we have been using in lecture, we will refer to states as (x,y), meaning the state in the x-th column and the y-th row: e.g. (1,1) for the state at the top left, and (10,1) for the state at the top right.

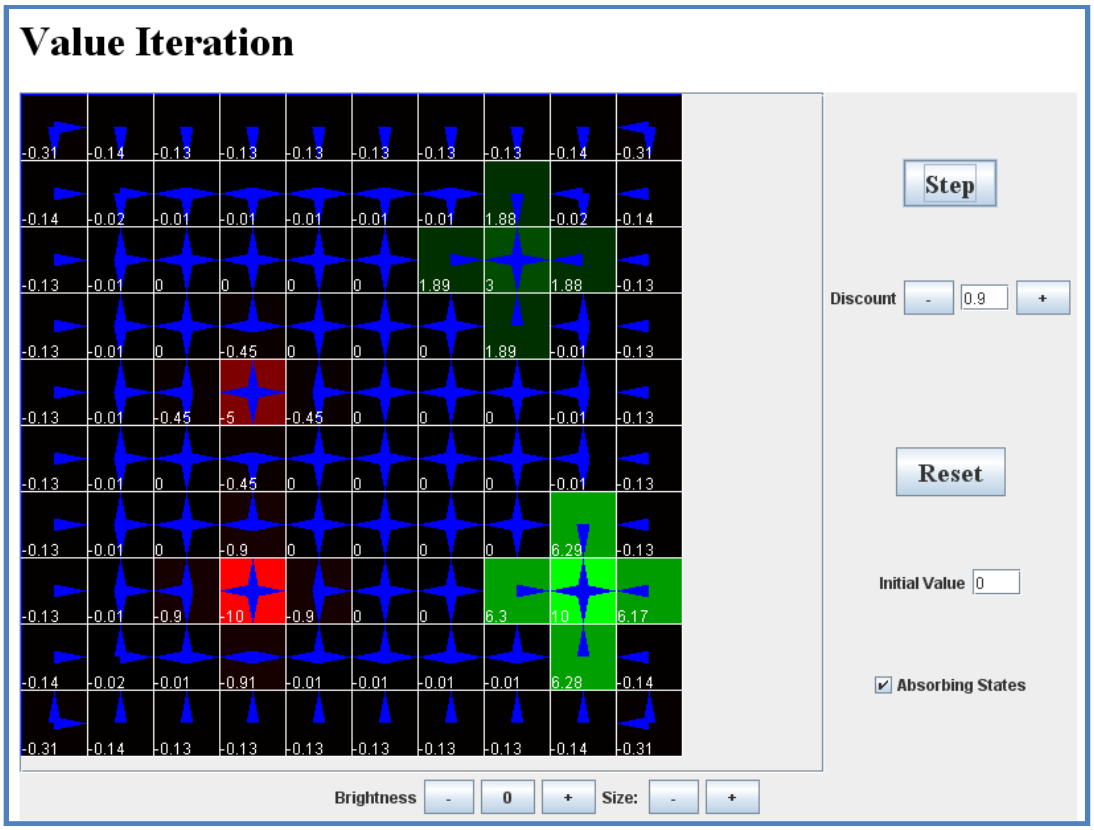


(a) (10 points) The figure above shows the values $U^{(1)}(s)$ in each state – that is, the values after one step of value iteration. We will focus on the entry in a single state, namely state (10,8), the state to the right of the absorbing state with reward 10 (which is located at (9,8)).

Show in detail how $U^{(1)}((10,8))$ is computed using the values $U^{(0)}(s)$.

(b) (10 points) The figure below shows the values $U^{(2)}(s)$ in each state, that is, the values after two steps of value iteration.

1. Show in detail how $U^{(2)}((10,8))$ is computed using the values $U^{(1)}(s)$.
2. Intuitively, why is $U^{(2)}((9,9))$ larger than $U^{(2)}((10,8))$?
3. Intuitively, why is $U^{(2)}((9,7))$ larger than $U^{(2)}((9,9))$?



(c) (15 points) We now study the importance of the discount factor.
With the option "absorbing states" enabled and discount factors of 0.8, 0.9, and 0.999, repeatedly perform steps until value iteration converges.

1. Provide the optimal policies for these three discount factors (just hand in screen shots of the applet after value iteration converged).
2. Why do the optimal policies change for the states around the absorbing state with reward 3 at (8,3), depending on the discount factor?
3. Why do the optimal policies change for the states (2,6), (3,6), (2,7), (3,7), depending on the discount factor?

Question 3: Belief State Update in POMDPs (programming) (45 points)

Consider the grid world example we have used in class to discuss MDPs and POMDPs. Let's focus on its interpretation as a POMDP with a transition model specified in the lectures, and the following observation model, with three possible observations: 1-wall, 2-walls, end.

State	1-wall	2-walls	end
Non-terminal in third column	.9	.1	0
All other non-terminal	.1	.9	0
Terminal	0	0	1

Write a program (in whatever language you prefer) that, given input

- an initial belief state b(s)
- a sequence of actions $a_{1:n}$
- a sequence of observations $e_{1:n}$

computes and prints out the belief state of the agent after performing $a_{1:n}$ and observing $e_{1:n}$ (i.e., observing each e_i after performing the corresponding a_i).

Please ensure that your code is readable (appropriate variable names, commenting, etc.).

Run your program on the following four sequences. When specified, the agent knows that it is starting in the given $S_0 =$ (row,column) *(with row and column starting from the bottom left – i.e. the same convention we use in the lecture slides)*; otherwise the agent has no idea where it is at the start *(i.e., uniform belief state on non-terminal states)*.

- (up, up , up) (2,2,2)
- (up, up, up) (1,1,1)
- (right, right, up) (1,1,end) with $S_0 =$ (3,2)
- (up, right, right, right) (2,2,1,1) with $S_0 =$ (1,1)

For each of the four sequences, provide:

- The output of your program.
    - While you may format the belief state however you see fit, it must be clear which probabilities correspond to which states. One easy way of doing this is by arranging the values in a grid as we have often done in lecture.
- A brief qualitative justification of why the output of your program makes sense.

Also include your program code at the end of your PDF submission (I recommend using a fixed-width font such as Courier New for your code to improve readability).