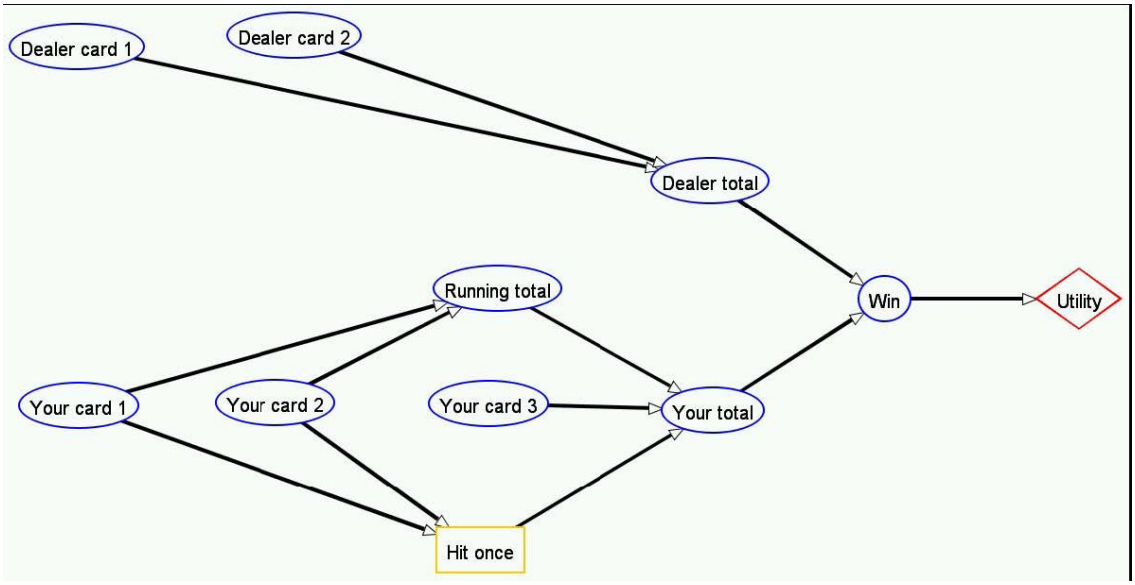
UBC-CS CPSC 422, 2017

Homework 1 (Due Mon, Oct 2, 12pm – either Connect or in class**) (tot: 110 points)**

Question 1 – Value of Information and Value of Control [20 points]
Download the "Blackjack.xml" file from the course web page or from connect. Open it with the Belief and Decision Networks applet in AIspace. It should appear as this decision network (you may have already seen this in 322).



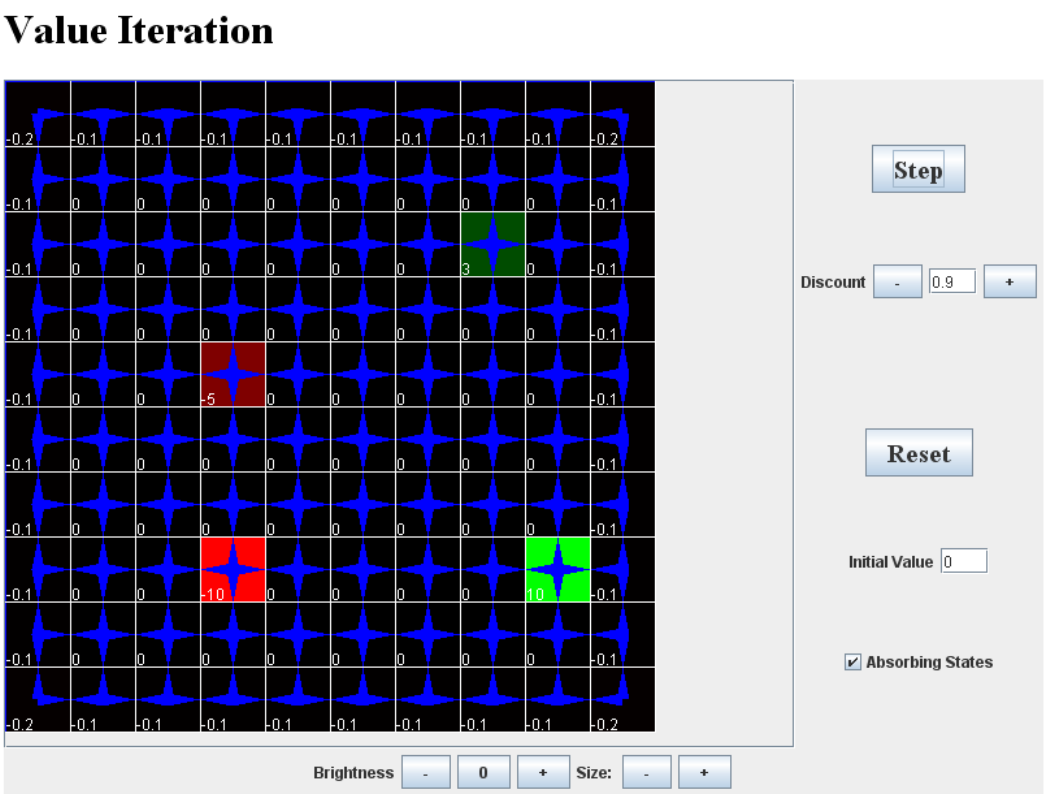
If needed, read the tutorial on that applet. Then, in AIspace, examine the details of the model. Make sure that you understand the rationale for all the nodes and their connections. Finally, use the model to answer the following questions.
(a) [6 points] What is the value of information for Dealer card 1? How do you calculate this?
(b) [6 points] What is the value of control for Running Total? How do you 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 http://artint.info/demos/mdp/vi.html (see piazza post for browsers' compatibility)

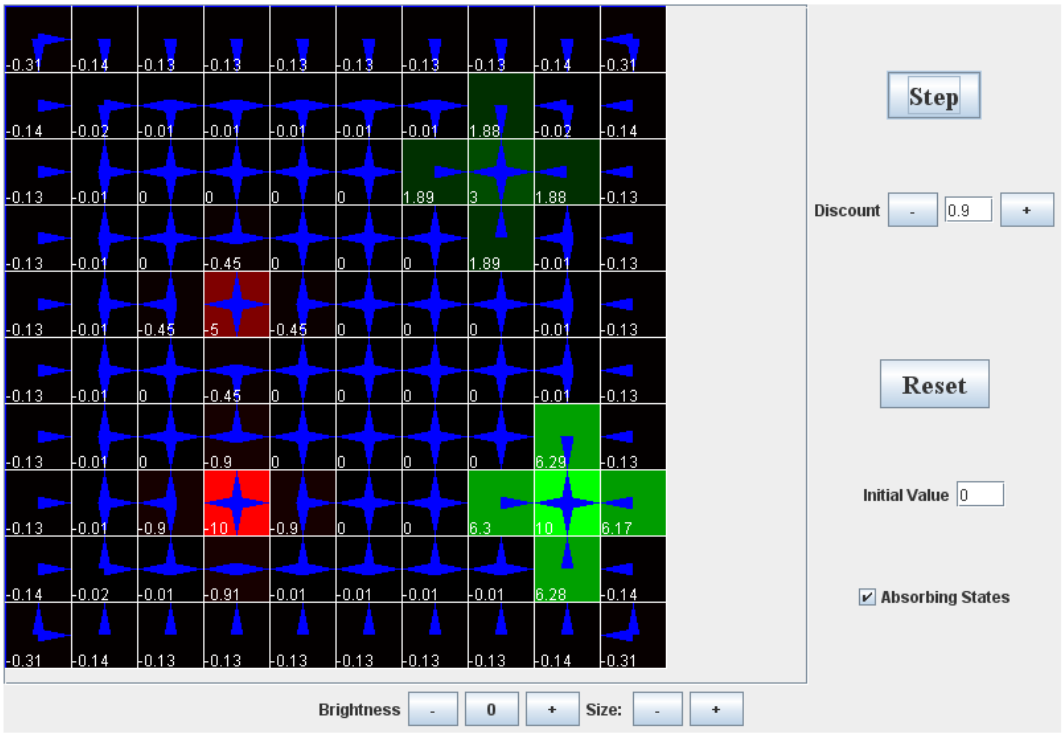
There are some questions listed on that website; for this assignment, please disregard those questions and only answer the following ones. In this assignment, we are using a discount factor of 0.9, initial values of $U^{(0)}(s) = 0$ for all s, and the "absorbing states" option (explained in detail on the website with the applet).
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 below 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.
Show in detail how $U^{(2)}((10,8))$ is computed using the values $U^{(1)}(s)$.
Intuitively, why is $U^{(2)}((9,9))$ larger than $U^{(2)}((10,8))$?
Intuitively, why is $U^{(2)}((9,7))$ larger than $U^{(2)}((9,9))$?

Value Iteration



(c) 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.
(c.1) (5 points) What are the optimal policies for these three discount factors (just hand in screen shots of the applet after value iteration converged.)
(c.2) (5 points) Why do the optimal policies change for the states around the absorbing state with reward 3 at (8,3), depending on the discount factor?
(c.3) (5 points) 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 slide 28 (lecture 5) and the following observation model, with three possible observations 1-wall, 2-walls, end.

	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 the language you prefer) that given as 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).
- Test your program on the following four sequences. When specified the agent knows that it is starting in the given $S_0 = (x,y)$ (where x is the column starting from the left and y is the row

starting from the bottom), otherwise the agent has no idea where it is (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 $S_0 = (2,3)$
- (up, right, right, right) (2,2,1,1) with $S_0 = (1,1)$

Hand in your program and its output on the four sequences. Also for each sequence justify qualitatively why the output of your program makes sense.

Question 4 (10 points)

[Note that this question is worth marks, so don't forget to do it.]
Fill out the following google form. Thanxs!
https://docs.google.com/forms/d/e/1FAIpQLSdeh8t2NXZfm11OqZETxGkc5Q909oOjjph0fV2I5GvUJdw03g/viewform?usp=sf_link