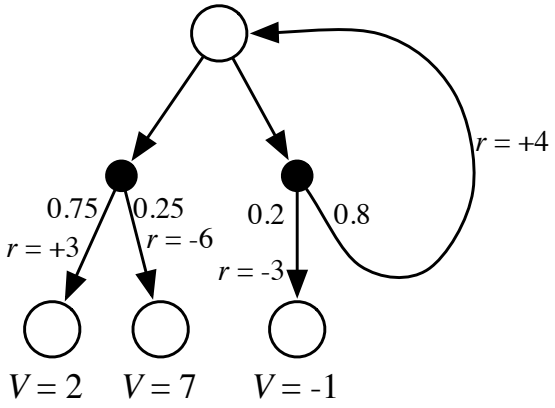
Homework Assignment # 2

Due: Tuesday, February 25, 2016, 11:59 p.m.
Total marks: 100

Question 1. [10 MARKS]

Consider the following fragment of an MDP graph. The fractional numbers indicate the world's transition probabilities and the whole numbers indicate the expected rewards. The three numbers at the bottom indicate what you can take to be the value of the corresponding states. The discount rate is 0.9. What is the value of the top node for the equiprobable random policy (all actions equally likely) and for the optimal policy? Show your work.

$v_\pi =$ $v_* =$



Question 2. [4 MARKS]

In Example 4.1 (from the SB textbook), if π is the equiprobable random policy, what is $q_\pi(4, \text{up})$? What is $q_\pi(5, \text{left})$? Please explain how you arrived at your answer.

Question 3. [8 MARKS]

Exercise 4.2 in Sutton and Barto 2nd Ed., 2016. Answer both parts and explain your answer.

Question 4. [4 MARKS]

Exercise 4.10 in Sutton and Barto 2nd Ed., 2016. (value iteration equation for action values). Explain your answer.

Question 5. [24 MARKS]

Programming Question

Part 1, worth 20 marks: Implement value iteration (figure 4.5) for the Gambler's problem. Recreate the plot of $v_\star$ and $\pi_\star$ from the book (figure 4.6). We will make one change to the problem as described in the book: let $\mathbf{pr}(heads) = 0.25$. Use whatever programming language you wish.

For this simple setting we do not have to use RL-glue. This should only require a simple function in a high-level language like matlab or python.

Please submitted your plot and ALL your code (including any scripts and data processing). I am aware that there are implementations of this problem available on the internet. **It is not acceptable to find the code online, modify it and submit it as your own. I expect you to implement this yourself, from scratch.**

Part 2, worth 4 marks: Describe why your plot of $v_\star$ differs from figure 4.6 in the textbook.

Question 6. [5 MARKS]

Exercise 5.1 in Sutton and Barto 2nd Ed., 2016. [there are three subquestions] (discuss blackjack value function)

Question 7. [5 MARKS]

Exercise 5.7 in Sutton and Barto 2nd Ed., 2016.

Question 8. [40 MARKS]

Programming: **Part 1, worth 35 marks**: Implement *Every-visit Monte Carlo Control with Exploring Starts* for state-action-values (Q(s,a)) on the *Gambler's problem* from Chapter 4. We will make two changes from the problem specification in the book. As in Question 4 above, let $\mathbf{pr}(heads) = 0.25$. In addition we will **not allow the zero bet action**; the agent must always make a bet of one or greater. Plot the $V_\pi(s)$ ($V_\pi(s) = \max_a Q_\star(s,a)$) for each state (on the x-axis). Plot $V_t(s)$ after 100 episodes,10000 episodes, and 10000000 episodes, producing a graph similar to the top plot in Figure 4.6. $V_t(s)$ should be produced by averaging over 30 runs. That is, your plot should contain 3 lines: $V_{t=100}$ averaged over 30 runs, $V_{t=10000}$ averaged over 30 runs, and $V_{t=10000000}$ averaged over 30 runs. My implementation takes approximately 15 minutes to run. It is running on a new macbook pro, but could probably be further speedup.

If you use RL-Glue you will notice you cannot send a query to the agent program from the experiment program; that is, the experiment program cannot ask the agent to send back V_t. However, you need to average V_t over runs. You could do this by writing V_t to a file from inside your agent on episodes 100,10000, and 10000000, on each run. Afterwards you could load the files into a data processing program like excel and compute the average and plot. Alternatively, you could also use RL_agent_message and return a string to the experiment program. The string could contain the pointer to V_t created by the agent program. I will provide code for this as a note on canvas.

This will require you to implement three things:

1. a simulation of the gamblers problem, that allows exploring starts (an Environment program)
2. the Every-visit Monte Carlo Control Algorithm with Exploring Starts (an Agent program)
3. code to run the experiment (Environment program)

Please submit your plot and ALL your code (including any scripts and data processing).

Part 2, worth 5 marks: Discuss and compare how the plot produced by Monte Carlo Control with Exploring Starts is similar and different from the plot produced by value iteration (in

question 4). Discuss why Dynamic Programming is suitable for the Gambler's problem and why the Monte Carlo method with exploring starts is less suitable for the Gambler's problem.

Homework policies:

Your assignment will be submitted as a single pdf document and a zip file with code, on canvas. The questions must be typed; for example, in Latex, Microsoft Word, Lyx, etc. or must be written legibly and scanned. Images may be scanned and inserted into the document if it is too complicated to draw them properly.

Policy for late submission assignments: Unless there are legitimate circumstances, late assignments will be accepted up to 5 days after the due date and graded using the following rule:

on time: your score 1

1 day late: your score 0.9

2 days late: your score 0.7

3 days late: your score 0.5

4 days late: your score 0.3

5 days late: your score 0.1

For example, this means that if you submit 3 days late and get 80 points for your answers, your total number of points will be $80 \times 0.5 = 40$ points.

All assignments can be done in collaboration, however, you must write your own answers, write your own programs, and generate your own results (data and graphs). All the sources used for problem solution must be acknowledged, e.g. web sites, books, research papers, personal communication with people, etc. Academic honesty is taken seriously; for detailed information see Indiana University Code of Student Rights, Responsibilities, and Conduct.

Good luck!