

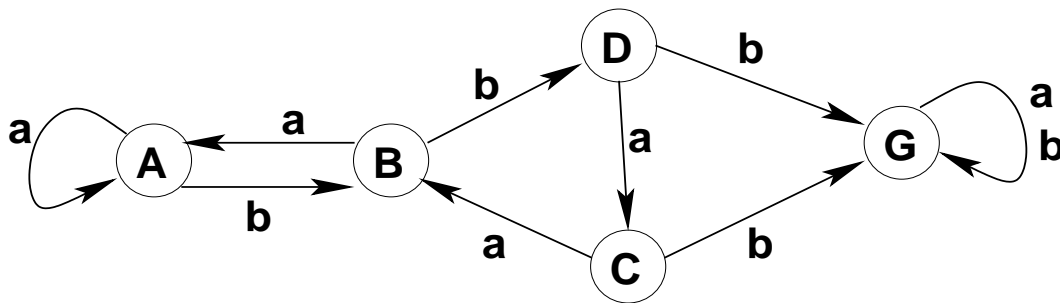
COS 402: Artificial Intelligence

Written exercises W5
MDP's

Fall 2011
Due: Thursday, December 8

Approximate point values are given in brackets. Be sure to show your work and justify all of your answers. See the course home page for information on when and where to submit written exercises, and grading criteria.

1. [10] Exercise 15.4 in R&N (either edition).
2. [15] Consider the following MDP:



There are five states: A , B , C , D and G . The reward at every state is -1 , except at G where the reward is 0 . There are two actions, a and b , and the effect of each action is deterministic as indicated in the figure. For instance, executing a in state B leads to state A . Assume $\gamma = 1$ in this problem.

[Note: If you understand the algorithms, this problem can (and should) be solved without a lot of tedious calculations, and without the use of a computer or even a calculator. You do not need to show easy calculations in detail, but should nevertheless justify your reasoning.]

- a. Show the sequence of utility estimates U_i that would result from executing value iteration on this MDP. Also show the optimal policy that is computed using the final utility estimate.
- b. Show the sequence of policies π_i and corresponding utility functions U^{π_i} that would result from executing policy iteration on this MDP. Assume that you start with a policy that assigns action a to every state. The utility functions U^{π_i} should be computed exactly; note that these utilities may be infinite for some states. Also, assume that all ties between the actions a and b in the policy improvement step are always broken in favor of a .
- c. Generalizing this example, suppose we are given a graph with a distinguished node (i.e., state) G , and k edges emanating from every node corresponding to k (deterministic) actions. As in this example, all of the edges emanating from G are self-loops, the node G is assigned reward 0 , and all other nodes are assigned reward -1 . In terms of properties of the graph, what is the optimal utility function U^* , and what is the optimal policy π^* ? If value iteration is applied to this graph (viewed as an MDP), exactly how many iterations will be needed until the algorithm converges? How about for policy iteration?

3. [10] Sometimes MDP's are formulated with a reward function $R(s, a)$ that depends on the action taken (so that reward $R(s, a)$ is received when action a is executed from state s), or a reward function $R(s, a, s')$ that also depends on the outcome state s' (so that this reward is received when state s' is reached after executing action a from state s). For each of these formulations, show how to appropriately modify each of the following:

- the Bellman equation (Eq. (17.5) in R&N);
- the formula for converting the optimal utility U^* (denoted simply U in R&N) into an optimal policy π^* (Eq. (17.4) in R&N);
- the value iteration algorithm;
- the policy iteration algorithm.

4. [15] Let $B(U)$ and $\|\cdot\|_\infty$ be as defined in class. (This is the same as BU and $\|\cdot\|$ defined in Section 17.2 of R&N.) The purpose of this exercise is to prove that B is a *contraction*, i.e., that $\|B(U) - B(U')\|_\infty \leq \gamma\|U - U'\|_\infty$. As discussed in the book and lecture, this is the key step in showing that value iteration converges to the right answer.

We will begin by proving some basic facts. Be sure to give genuine mathematical proofs for each part of this problem. Also, your proofs should use elementary facts — in other words, do not give proofs that rely on mathematical sledge-hammers like the Cauchy-Schwartz inequality.

- a. Let u_1, \dots, u_n and v_1, \dots, v_n be any sequences of real numbers. Prove that if $u_i \leq v_i$ for all i then

$$\max_i u_i \leq \max_i v_i.$$

- b. Let x_1, \dots, x_n and y_1, \dots, y_n be any sequences of real numbers. Prove that

$$\left(\max_i x_i\right) - \left(\max_i y_i\right) \leq \max_i (x_i - y_i),$$

and also that

$$\max_i (x_i - y_i) \leq \max_i |x_i - y_i|.$$

(Hint: both of these inequalities can be proved using part (a) for an appropriate choice of u_i and v_i .)

Finally, use these facts to prove that

$$\left| \left(\max_i x_i\right) - \left(\max_i y_i\right) \right| \leq \max_i |x_i - y_i|.$$

- c. Let x_1, \dots, x_n be any real numbers, and suppose that p_1, \dots, p_n are nonnegative real numbers such that $\sum_i p_i = 1$. Use the fact that $|a + b| \leq |a| + |b|$ for any real numbers a and b to prove that

$$\left| \sum_i p_i x_i \right| \leq \max_i |x_i|.$$

- d. Now let s be any state, and let $(B(U))(s)$ denote the value of $B(U)$ at state s . By plugging in the definition of B , and using the properties proved above, prove that

$$|(B(U))(s) - (B(U'))(s)| \leq \gamma\|U - U'\|_\infty.$$

Conclude that

$$\|B(U) - B(U')\|_\infty \leq \gamma\|U - U'\|_\infty.$$