Homework on MDPs
CSE 571 Fall 2010
Due: Oct 25th

17.10 Consider an undiscounted MDP having three states, (1, 2, 3), with rewards -1, -2, 0, respectively. State 3 is a terminal state. In states 1 and 2 there are two possible actions: a and b. The transition model is as follows:

- In state 1, action a moves the agent to state 2 with probability 0.8 and makes the agent stay put with probability 0.2.
- In state 2, action a moves the agent to state 1 with probability 0.8 and makes the agent stay put with probability 0.2.
- In either state 1 or state 2, action b moves the agent to state 3 with probability 0.1 and makes the agent stay put with probability 0.9.

Answer the following questions:

a. What can be determined *qualitatively* about the optimal policy in states 1 and 2?

b. Apply policy iteration, showing each step in full, to determine the optimal policy and the values of states 1 and 2. Assume that the initial policy has action b in both states.

c. What happens to policy iteration if the initial policy has action a in both states? Does discounting help? Does the optimal policy depend on the discount factor?

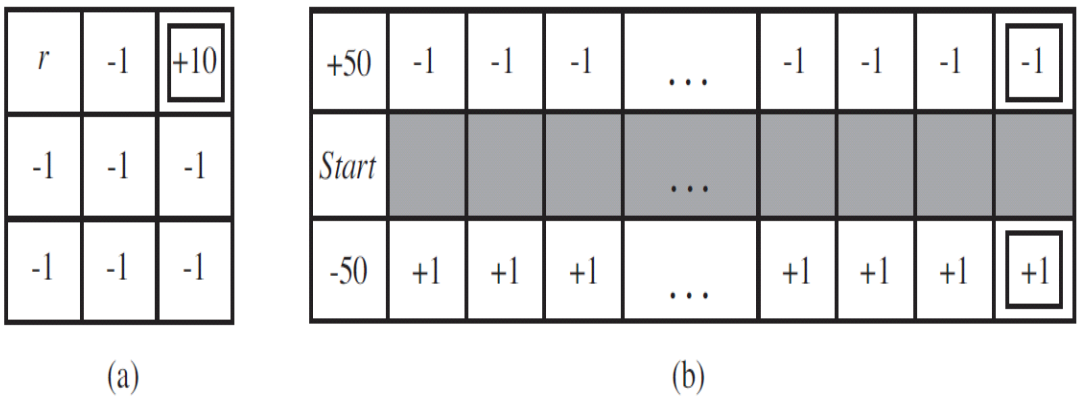

Figure 17.14 (a) 3×3 world for Exercise 17.8. The reward for each state is indicated. The upper right square is a terminal state. (b) 101×3 world for Exercise 17.9 (omitting 93 identical columns in the middle). The start state has reward 0.

17.9 Consider the 101×3 world shown in Figure 17.14(b). In the start state the agent has a choice of two deterministic actions, *Up* or *Down*, but in the other states the agent has one deterministic action, *Right*. Assuming a discounted reward function, for what values of the discount γ should the agent choose *Up* and for which *Down*? Compute the utility of each action as a function of γ. (Note that this simple example actually reflects many real-world situations in which one must weigh the value of an immediate action versus the potential continual long-term consequences, such as choosing to dump pollutants into a lake.)

5. **(Policy Evaluation.)** Given a policy π, let V_π be the infinite-horizon, discounted value function (as defined in class), which we know satisfies the following equation at all states s,

$$V_\pi(s) = R(s) + \gamma \sum_{s'} \Pr(s'|s, \pi(s)) \cdot V_\pi(s') \tag{1}$$

We can compute V_π by solving the above system of linear equations. However, there is also an iterative technique for computing V_π that is often more efficient. Consider the following value-function operator T_π,

$$T_\pi[V](s) = R(s) + \gamma \sum_{s'} \Pr(s'|s, \pi(s)) \cdot V(s')$$

note that $T_\pi[V]$ is simply a value function and $T_\pi[V](s)$ gives the value of state s. As described in your book in the discussion of modified policy iteration, this operator can be used to iteratively compute a sequence of value functions V^k that converge to V_π as follows:

$$\begin{aligned} V^0(s) &= 0, \text{for all } s \\ V^k &= T_\pi[V^{k-1}] \end{aligned}$$

Use the following steps to prove that the sequence does converge to the correct value function.

(a) Show that T_π is a contraction operator with respect to the max-norm. That is show that for any value functions V and V',

$$||T_\pi[V] - T_\pi[V']|| \leq \gamma ||V - V'||$$

(b) Use this fact to prove that $\lim_{k\to\infty} V^k = V_\pi$. You may use equation 1 if desired.

(c) Does the sequence still converge to V_π if we initialize V^0 to random values? Explain.

(d) What value of k is sufficient so that $||V^k - V_\pi|| \leq \epsilon$? Explain.

Consider the “over-subscription” planning problem as defined below:

--you have a deterministic planning domain, with each action a in the domain incurring a cost C(a)

--You have a set of n goals G1…Gn

--Each goal Gi has a reward R(Gi)

Given a plan P, we denote the cumulative cost of actions in P as C(P) the cumulative reward achieved by P is just the sum of the rewards of the goals that are made true by P in its final state. The “net-benefit” of a plan P is defined as its cumulative reward minus its cumulative cost.

Our task is to find a plan P with optimal net-benefit.

Show how this problem can be modeled as an MDP problem

You have to describe the States, actions, transition functions and rewards of the MDP (either formally or in english). Make sure that your formulation doesn’t double-dip (i.e., collect the reward for a goal more than once).