CS533
Intelligent Agents and Decision Making
Homework 4, Winter 2011

1. 17.4 c
2. 17.6
3. 17.9
4. 17.10 b,c
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.