EC 700, HOMEWORK 1

1. Given an MDP M with rewards that depend both on the state and action selected, describe how you would construct a new MDP M' with rewards depending only on state such that,from an optimal policy for M', you can construct an optimal policy for M.

2. Suppose you have a Markov decision process. We fix a randomized policy π which chooses the action as a function of the current state s. Prove that the sequence of states is a Markov process. This is a claim made in the lecture slides but not justified there.

3. Given a policy π, define

$$P_\pi(s'|s) = \sum_a \pi(a|s)P(s'|s,a).$$

Prove that the matrix P_π has all of its eigenvalues upper bounded by one in magnitude. Then explain why this justifies the expansion $(I - \gamma P_\pi)^{-1} = I + \gamma P_\pi + \gamma^2 P_\pi^2 + \cdots$.

4. Consider the so-called asynchronous value iteration update: this is the value iteration update, but we only update one coordinate of the vector at a time.

More formally, starting from some vector J_0, we have the following procedure for obtaining J_{k+1} from J_k: we compute $J_{\text{new}}(k) = TJ_k$, choose one coordinate s, and set

$$J_{k+1}(s) = J_{\text{new}}(s),$$
$$J_{k+1}(u) = J_k(u) \text{ for all } u \neq s.$$

Prove if every coordinate s is chosen infinitely often, this converges to J^*.

5. In some cases, you cannot compute the Bellman operator T exactly; instead given vector J, you can compute an approximate $\tilde{T}J$ which is a vector satisfying

$$||\tilde{T}J - TJ||_\infty \leq \epsilon$$

Given an upper bound on $||\tilde{T}^k J_0 - J^*||_\infty$ (where J^* is the vector of optimal values in the infinite-horizon problem) under the assumption that J_0 is the zero vector and all rewards are upper bounded by M.