Planning and Learning in Dynamical Systems (046194)
Homework 3

Question 1 – Worst Case Reward

Consider a finite horizon MDP with the following twist. For a stationary stochastic policy π, instead of looking at the expected cumulative reward, we look at the worst case reward until a given time T:

$$\underline{J}^{\pi,T}(s) = \min_{(s_0,a_0,s_1,a_1,s_2,a_2,\ldots,s_T,a_T):P_\pi(s_0,a_0,s_1,a_1,s_2,a_2,\ldots,s_T,a_T|s_0=s)>0} \sum_{t=0}^{T} r(s_t,a_t).$$

That is, the worst possible reward given that the initial state is s.
Similarly, we define the best case reward as

$$\overline{J}^{\pi,T}(s) = \max_{(s_0,a_0,s_1,a_1,s_2,a_2,\ldots,s_T,a_T):P_\pi(s_0,a_0,s_1,a_1,s_2,a_2,\ldots,s_T,a_T|s_0=s)>0} \sum_{t=0}^{T} r(s_t,a_t).$$

1. Write a backward recursion (dynamic programming on $T, T-1, \ldots$) formula for $\underline{J}^{\pi,T}$.
2. Suppose that we want to compute $\overline{J}^{*,T}(s) = \sup_\pi \overline{J}^{\pi,T}(s)$. Suggest an algorithm to do so.
3. Is there an optimal deterministic strategy that attains $\overline{J}^{*,T}(s)$? That is, does there exist a deterministic policy π such that $\overline{J}^{*,T}(s) = \overline{J}^{\pi,T}(s)$ for all s? Prove or give a counterexample.
4. We now consider the infinite horizon case. For some $\gamma < 1$ we define

$$\underline{J}^{\pi}(s) = \lim_{T\to\infty} \min_{(s_0,a_0,s_1,a_1,s_2,a_2,\ldots,s_T,a_T):P_\pi(s_0,a_0,s_1,a_1,s_2,a_2,\ldots,s_T,a_T|s_0=s)>0} \sum_{t=0}^{T} \gamma^t r(s_t,a_t).$$

   i. Write dynamic programming equations for $\underline{J}^{\pi,*}(s) = \sup_\pi \underline{J}^{\pi,T}(s)$.
   ii. Show that the resulting dynamic programming operator is a contraction.
   iii. (bonus: 5 points) Prove that there exists a stationary deterministic strategy that attains $\underline{J}^{\pi,*}(s)$.

5. (bonus: 10 points) Consider $J^{\pi}_{\frac{1}{2}}(s) = \frac{1}{2}\cdot\left(\underline{J}^{\pi}(s) + \overline{J}^{\pi}(s)\right)$. We now consider $J^{*}_{\frac{1}{2}}(s) = \sup_\pi J^{\pi}_{\frac{1}{2}}(s)$. Is there an optimal deterministic policy? Prove or give a counterexample.

Question 2 – The $c\mu$ rule

Assume N jobs are scheduled to run on a single server. At each time step (t=0,1,2,…), the sever may choose one of the remaining unfinished jobs to process. If job i is chosen, then with probability $\mu_i > 0$ it will be completed, and removed from the system; otherwise the job stays in the system, and remains in an unfinished state. Notice that

the job service is memoryless – the probability of a job completion is independent of the number of times it has been chosen.

Each job is associated with a waiting cost $c_i > 0$ that is paid for each time step that the job is still in the system. The server's goal is minimizing the total cost until all jobs leave the system.

a. Describe the problem as a Markov decision process. Write Bellman's equation for this problem.
b. Show that the optimal policy is choosing at each time step $i^* = \arg\max_i c_i \mu_i$ (from the jobs that are still in the system).

Hint: Compute the value function for the proposed policy and show that it satisfies the Bellman equation).

Remark: the $c\mu$ law is a fundamental result in queuing theory, and applies also to more general scenarios.

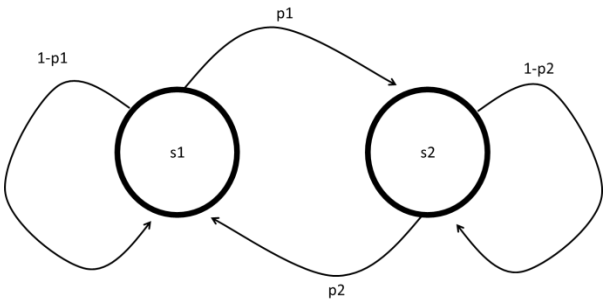
Question 3 - DP operator not contracting in Euclidean norm

Recall the fixed-policy DP operator T^π defined as (see Section 5.4 of the lecture notes)

$$\left(T^\pi\left(J\right)\right)\left(s\right) = r\left(s,\pi(s)\right) + \gamma \sum_{s'\in S} p\left(s' \mid s,\pi(s)\right) J\left(s'\right),$$

where $\gamma < 1$. We have seen that T^π is a contraction in the sup-norm. Show that T^π is not necessarily a contraction in the Euclidean norm. We will later see, in the learning part of the course, that the Euclidean norm is related to a an approximation error due to learning, which will require additional discussion.

Hint: one possible approach is to consider the following 2-state MDP, and choose appropriate values for p_1, p_2, γ to obtain a contradiction to the contraction property.



Question 4 – Stochastic Shortest Path

Consider an MDP with $\gamma = 1$, positive cost with a cost free termination state denoted by 0. Once the system reaches that state it remains there at no further cost. We are interested in the optimal policy in such an MDP; the policy with the minimal cost to the terminal state.

- Define the value function v^π as the expected sum of costs until reaching terminations, and similarly v^* as the optimal. Write Bellman equations for v^π, v^*. Note the difference between the terminal state and all other states.

- Write the Bellman operators T, T_π explicitly for the SSP problem.

  In the next part we will show that under certain assumptions, T_π is contractions in a suitably defined norm (the contracting property of T follows from the contraction of T_π).

Definition (1): A proper policy. A policy is proper if there exists a positive probability that the termination state, 0, will be reached.
Definition (2): Let $J, \xi \in R^S$. The maximum norm of J is defined as follows.

$$\|J\|_\xi = \max_{s \in S} \frac{|J(s)|}{\xi(s)}$$

- For the SSP problem, it holds $TJ^* = J^*$ for J^* the optimal cost vector and is finite if all stationary policies are proper. Explain why the proper polices assumption is needed?

- Prove: Assume all stationary policies are proper. Then, there exists a vector ξ with positive components such that the mapping T_π, for all stationary policies π, are contraction mappings with respect to the weighted maximum norm.

Guidelines to the proof.

- You first need to define ξ properly. Consider a new SSP with same transitions and costs all equal to -1, except at the termination state, 0. Define $\hat{J}(s)$ as the optimal value from state s in the new SSP problem, and define

  $$\xi(s) = -\hat{J}(s) .$$

  See that $\xi(s) \geq 1$.
- By writing the optimal Bellman equation of the new SSP for $\hat{J}$, and basic analysis, show the following:
  a) For any stationary policy π, $\sum_{s' \in S} p^\pi(s' \mid s)\xi(s') \leq \xi(s) - 1$.
  b) $\xi(s) - 1 \leq \beta\xi(s)$, where $\beta = \max_{s'} \frac{\xi(s')-1}{\xi(s')}$. Explain why $\beta < 1$.
- Prove that, for any π, T_π is a contraction in the max norm and its contraction coefficient is β. Hint: as usual, prove by showing that for any $J_1, J_2 \in R^S$,

  $$\frac{|T_\pi J_1(s) - T_\pi J_2(s)|}{\xi} \leq \beta \|J_1 - J_2\|_\xi .$$

Question 5

The Value Iteration produces a sequence of value functions V_i where V_0 is the initial value function supplied to the algorithm. Denote the greedy policy w.r.t. V_i by π_i. Does $\{\pi_i\}$ is a sequence of improving policies? i.e, does $V^{\pi_i} \geq V^{\pi_{i-1}}$ for all i? Prove or give a counter example.