ECE586RL: MDPs and Reinforcement Learning **Spring 2020**

Homework 2

Instructor: Bin Hu Due date: May 4, 2020

1.Concepts (60 points in total, 15 points for each subproblem)

(a) What is the distribution shift issue in imitation learning? How does DAGGER address that issue?

(b) What is the iterative LQR (iLQR) algorithm? Write out the algorithm and explains the rationale behind it.

(c) Prove the following relative policy performance identity:

$$J(\pi') - J(\pi) = \mathbb{E}_{\tau\sim\pi'}\left[\sum_{t=0}^{\infty}\gamma^t A^{\pi}(s_t, a_t)\right]$$

where (π, π') are two policies with finite cost, A^{π} is the advantage function for policy π, and γ is the discounting factor.

(d) What is transfer learning? What does "0-shot transfer learning" mean?

2. Averaged-Cost LQR (40 points in total, 10 points for each subproblem)
Consider the linear time-invariant system

$$x_{k+1} = Ax_k + Bu_k + w_k$$

where x_k is the state, u_k is the control action, and the process noise w_k is sampled from a Gaussian distribution in an IID manner, i.e. $w_k \sim \mathcal{N}(0, W)$. The objective is to choose u_k to minimize the following cost

$$\mathcal{C} = \lim_{N\to\infty}\frac{1}{N+1}\sum_{k=0}^{N}\mathbb{E}(x_k^\mathsf{T} Q x_k + u_k^\mathsf{T} R u_k)$$

The matrices Q and R are positive definite.

(a) Policy evaluation: Suppose we are using a linear policy $u_k = -Kx_k$. How to calculate the relative state value function? How to calculate the relative Q-function? Derive the Bellman equations for both cases.

(b) Optimal Bellman equation: Derive the optimal Bellman equation for the above average-cost LQR.

(c) Approximate Policy Iteration: Write out the policy iteration algorithm for the above problem. In the policy evaluation step, is there a way to modify the LSTD algorithm to estimate relative Q-function from sample trajectories of $\{x_k, u_k\}$?

(d) Policy Gradient: For the above problem, how to evaluate the policy gradient for a linear policy?