ECE 598: Interplay between Control and Machine Learning **Fall 2020**

Homework 2

Instructor: Bin Hu Due date: November 6, 2020

Each problem is worth 5 points, and the total points in HW2 is 20.

1. MDPs on Finite State Space

Consider a finite state MDP $\langle \mathcal{S}, \mathcal{A}, P, c, \gamma \rangle$. Suppose the state takes value on the set $\{1, 2, \ldots, n\}$ and the action takes value on the set $\{a_1, a_2\}$.

(a) Consider a stochastic policy that takes the action a_1 with probability p_1 for any state. In other words, the policy selects the action a_2 with probability $(1-p_1)$ for any state. Suppose the transition matrix is know. How to evaluate the state value function for this policy? How to evaluate the state-action value function for this policy? How about the cases where the transition matrix is not known? Provide one method here.

(b) Suppose we use the policy in (a) as our behavior policy and apply $\mathcal{Q}$-learning to solve the above MDP. What is the update rule here? What is the size for the $\mathcal{Q}$ table?

(c) Now suppose we want to apply SARSA. What is the difference between SARSA and $\mathcal{Q}$-learning?

2. Linear Quadratic Regulator

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

$$\mathcal{C} = \sum_{k=0}^{\infty} \gamma^k \mathbb{E}(x_k^\mathsf{T} Q x_k + u_k^\mathsf{T} R u_k)$$

where $0 < \gamma < 1$ is the discount factor. 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 state value function $\mathcal{C}_K(x) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k (x_k^\mathsf{T} Q x_k + u_k^\mathsf{T} R u_k) \middle| x_0 = x\right]$? How to calculate the state-action value function $\mathcal{Q}_K(x, u) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k (x_k^\mathsf{T} Q x_k + u_k^\mathsf{T} R u_k) \middle| x_0 = x,\, u_0 = u\right]$?

Derive the Bellman equations for both cases. How to estimate these value functions when the model is unknown?

(b) Approximate Policy Iteration: Write out the policy iteration algorithm for the above problem. In the policy evaluation step, how to estimate $\mathcal{Q}_K$ from sample trajectories of $\{x_k, u_k\}$? Write out the specific update rules for the algorithms you provide.

(c) What is the optimal state-action value function $\mathcal{Q}^*$? How can we calculate $\mathcal{Q}^*$? What is the difference between fitted $\mathcal{Q}$-iteration and approximate policy iteration?

3. Policy Gradient

Consider a nonlinear system $x_{k+1} = f(x_k, u_k, w_k)$ where x_k is the state, u_k is the action, and w_k is the process noised sampled from an IID Gaussian distribution. The objective is to choose u_k to minimize the cost

$$\mathcal{C} = \mathbb{E} \sum_{k=0}^{\infty} \gamma^k c(x_k, u_k) \tag{1}$$

where $0 < \gamma < 1$ is the discount factor. Suppose f is unknown, and we want to learn policy from sampled trajectories of $\{(x_k, u_k)\}$.

(a) What is the policy gradient theorem? Write out the statement.

(b) Suppose we are using a three-layer neural network to parameterize the Gaussian policy, i.e. $u_t \sim \mathcal{N}(W^2\sigma(W^1\sigma(W^0 x_t)), \sigma I)$ where σ is the elementwise activation, how to estimate the policy gradient? (Write out an explicit formula for $\log \pi_\theta(u_t|x_t)$ and substitute it into the gradient formula.)

(c) Suppose f is linear, i.e. $f(x_k, u_k, w_k) = Ax_k + Bu_k + w_k$, and $c(x_k, u_k) = x_k^\mathsf{T} Q x_k + u_k^\mathsf{T} R u_k$. If (A, B, Q, R) is known and a linear policy is used, i.e. $u_k = -Kx_k$, can we calculate the gradient $\nabla \mathcal{C}(K)$ using the model information directly? Derive a closed-form formula here.

4. Implementation Assignment

In this problem, you are asked to implement codes for a LQR example with the following parameters

$$A = \begin{bmatrix} 0.99 & 0.01 & 0 \\ 0.01 & 0.95 & 0.01 \\ 0.5 & 0.14 & 0.97 \end{bmatrix}, B = \begin{bmatrix} 1 & 0.1 \\ 0 & 0.2 \\ 1.1 & 0.7 \end{bmatrix}, Q = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0.5 \end{bmatrix}, R = \begin{bmatrix} 1.3 & 0 \\ 0 & 0.7 \end{bmatrix}, W = \begin{bmatrix} 0.3 & 0 & 0 \\ 0 & 0.15 & 0 \\ 0 & 0 & 0.1 \end{bmatrix}.$$

Here W is the covariance matrix of the process noise which is sampled from zero mean Gaussian process. We fix the discounting factor as $\gamma = 0.98$. Implement both the LSPI algorithm and the fitted Q-iteration to learn the optimal policy from the sampled trajectories of the above system. Do they work? Summarize your findings.