ECE 4110/5110 Wednesday, 10/19/23

Lecture 14: Markov Decision Processes

Dr. Kevin Tang Handout 15

Basic Setup

We use A to denote the set of actions.

If a Markov process is in state i, and action $a \in A$ is chosen, then the transition probability is $p_{ij}(a)$.

A policy β is a set of numbers $\beta = \{\beta_i(a), a \in A, i = 1, \ldots, m\}$. It specifies that if the process is in state i, then action a is to be chosen with probability $\beta_i(a)$.

Under any given policy β, we have

$$p_{ij}(\beta) = \sum_a p_{ij}(a)\beta_i(a) \tag{1}$$

Let $R(i, a)$ denote the reward that is earned whenever action a is chosen in state i. We can then ask for the optimal policy to maximize the total reward of n steps, or to maximize the average reward per step over infinite time horizon.

Optimal Dynamic Policy (for finite number of steps)

Start from the boundary (last step), work backwards to find the optimal solution recursively.

Optimal Stationary Policy

Let π_{ia} be the steady-state probability of being in state i and choosing action a. Then we have

$$\beta_i(a) = \frac{\pi_{ia}}{\sum_a \pi_{ia}} \tag{2}$$

Furthermore, the optimal stationary policy will maximize the expected average reward $\sum_i \sum_a \pi_{ia} R(i, a)$. The optimization variable here is π_{ia}. Besides nonnegativity and normaliation, it also needs to satisfies the following constraint.

$$\sum_a \pi_{ja} = \sum_i \sum_a \pi_{ia} p_{ij}(a) \tag{3}$$