Lecture 07

- So far we have not 'introduced' Markov Decision Process!

Today we will formally introduce finite horizon/episodic Markov Decision Process (MDP).

Definition: A finite horizon MDP is given by a five tuple (S, U, N, p, g) where is the planning horizon, (S, U) denotes the set of states and actions in each step. State transitions are governed by a collection of transition kernels

$$p = \{p_k(\cdot|x,u)\}_{k\in[N], x\in S, u\in U}$$

where $p_k(\cdot|x,u)$ gives the distribution over states S if action u is taken in state x at step k. The instantaneous costs fully described by collection of cost functions $g = \{g_k : S\times U\times S \to \mathbb{R}\}_{k\in[N]}$ such that $g_k(x,u,y)$ is the cost incurred when the state at step k is x and upon taking action u the system state changes to y at step $k+1$.

* We can work with expected cost functions $\bar{g}_k : S\times U \to \mathbb{R}$ defined as,

$$\bar{g}_k(x,u) := \mathbb{E}\left[g_k(x,u,Y)\right] = \sum_{y\in S} p_k(y|x,u)\, g_k(x,u,y)$$

because we minimize the total 'expected' cost.

* Similar to cost function we can define reward function and then we want to maximize the total reward. The underlying problems are equivalent.

are equivalent.

* The dynamic programming equation of the finite horizon problem becomes from

$$J_N(x_N) := g_N(x_N)$$

$$J_k(x_k) := \min_{u_k \in U_k(x_k)} \mathbb{E}_{w_k} \left[g_k(x_k, u_k, w_k) + J_{k+1}\left(f_k(x_k, u_k, w_k)\right) \right]$$

to

$$J_N(x_N) := g_N(x_N)$$

$$J_k(x_k) := \min_{u_k \in U_k(x_k)} \sum_{y \in S} p_k(y \mid x_k, u_k) \left[g_k(x_k, u_k, y) + J_{k+1}(y) \right]$$

$$= \min_{u_k \in U_k(x_k)} \left[\bar{g}_k(x_k, u_k) + \sum_{y \in S} p_k(y \mid x_k, u_k)\, J_{k+1}(y) \right]$$

* The randomness incorporated by w_k is entirely captured by the transition probability kernels!

* Introduction to infinite horizon problems:

Infinite horizon problems differ from finite horizon problems in two aspects:

i) The number of stages is infinite.

ii) The system is stationary, i.e., the cost per stage, g is time independent and so is the random disturbances.

So, the instantaneous cost is fully known if the single cost function g is known and $\{w_k\}_{k=0}^{\infty}$ is iid stochastic process.

- This is equivalent of that the transitions are governed by a single stage independent kernel p.

* We will discuss 3 principal classes of infinite horizon problem. In first [illegible] we try to minimize the total cost over an infinite

* We will discuss 3 principal classes of infinite horizon problem. In first two, we try to minimize the total cost over an infinite number of stages, given by

$$J_\pi(x_0) = \lim_{N\to\infty} \mathbb{E}\left[\sum_{k=0}^{N-1} \alpha^k \bar{g}(x_k, \mu_k(x_k))\right]$$

Where $J_\pi(x_0)$ denotes the cost associated with an initial state x_0 and a policy $\pi = \{\mu_0, \mu_1, \ldots\}$, and α is a positive scalar with $\alpha \in (0,1]$ called the discount factor.

In first two following classes, finiteness of $J_\pi(x_0)$ is guaranteed through appropriate assumptions on the problem structure and discount factor.

In the third class, this sum need not to be finite for any policy so the cost is appropriately redefined.

a. Stochastic shortest path problems: there $\alpha = 1$ but there is a special cost free termination state, t. Once the system reaches t, it stays there for ever with no further cost.
- If we assume that termination is inevitable, the effective horizon will be finite.
  - We will see this in the next lecture.

b. Discounted problems with bounded cost per stage:
Here $\alpha < 1$ and $|g(x,u,y)| < M$ for some constant $M \in \mathbb{R}_+$. This makes $J_\pi(x_0)$ bounded because $|\bar{g}(x,u)| < M \;\; \forall\, x,u \in S \times U$ and so,

$$J_\pi(x_0) \leq \lim_{N\to\infty} \sum_{k=0}^{N-1} M \cdot \alpha^k = \frac{M}{1-\alpha} < \infty \quad \forall\, x_0 \in S$$

c. Average Cost per Stage problems:
In many real life problems discounted total cost optimization is not what is needed. If we withhold discounting, i.e., $\alpha = 1$

is not what is needed. If we withhold discounting, i.e., $\alpha = 1$ in general for every policy π and every initial state x_0, $J_\pi(x_0) = \infty$.

It turns out that in many such problems the average cost per stage, defined by

$$\lim_{N\to\infty} \frac{1}{N} \mathbb{E}\left[\sum_{k=0}^{N-1} g(x_k, \mu_k(x_k))\right]$$

is well defined and finite.