Lecture 04

Saturday, 14 January 2023 11:09 AM

* So far

1. Review of Probability theory
2. Stochastic processes: finite state DTMC.

* Today:

Dynamic Programming Algorithm for finite horizon problem:

- We are given a discrete-time dynamic system

$$x_{k+1} = f_k(x_k, u_k, w_k) \quad , \quad k = 0, 1, \ldots, N-1$$

where x_k is state at time k, from state space S_k.
u_k is the control i/p at time k, from control space C_k.
w_k is random disturbance takes value from D_k.

$U(x_k) \subseteq C_k$ is the set of allowable actions when the state is x_k.

- Since w_k, $k = 0, 1, \ldots, N-1$ are random variable, so are x_k, $k = 0, \ldots, N$ and if u_k is chosen based on past observation of states then u_k, $k = 0, \ldots, N-1$ are random variables too.
- w_k is characterized by prob. dist $P_k(\cdot \mid x_k, u_k)$.
- We assume that w_k is independent of $w_{k-1}, \ldots, w_0$.
- We also consider the class of policies that consist of a seq. of functions

$$\pi := \{\mu_0, \ldots, \mu_{N-1}\}$$

- If $\forall\ k \in \{0, \ldots, N-1\}$, $\mu_k(x_k) \in U(x_k)\ \forall\ x_k \in S_k$ we call π is an admissible policy.

Given an initial condition x_0 and an admissible policy $\pi = \{\mu_0, \ldots, \mu_{N-1}\}$, the system equation

$$x_{k+1} = f_k(x_k, \mu_k(x_k), w_k) \quad , \quad k = 0, 1, \ldots, N-1$$

- Let $g_k : S_k \times U_k \times D_k \to \mathbb{R}$, $k = 0, 1, \ldots, N-1$ and $g_N : S_N \to \mathbb{R}$ are instantaneous cost function. So the total expected cost of the system given initial state x_0 and policy π is

$$J_\pi(x_0) = \mathop{\mathbb{E}}_{w_k,\, k=0,\ldots,N-1} \left[g_N(x_N) + \sum_{k=0}^{N-1} g(x_k, \mu_k(x_k), w_k) \right]$$

- This is a well-defined quantity.
- For a given initial state x_0, an optimal policy π^* is one that minimizes this cost $J_\pi(x_0)$ over all admissible policies.

$$J_{\pi^*}(x_0) = \min_{\pi \in \Pi} J_\pi(x_0)$$

and $\pi^* \in \arg\min_{\pi \in \Pi} J_\pi(x_0)$

- Π : Set of all admissible policies.
- Though π^* is associated with fixed x_0, it is typically possible to find a policy π^* that is simultaneously optimal for all initial states.
- However the optimal cost depend on x_0 and is denoted by

$$J^*(x_0) := \min_{\pi \in \Pi} J_\pi(x_0)$$

and is called as optimal value function.

* The Dynamic Programming (DP) Algo.

Principle of Optimality : Let $\pi^* = (\mu_0^*, \mu_1^*, \ldots, \mu_{N-1}^*)$ be an optimal policy for the basic problem, and assume that when using π^* a given state x_i occurs at time i with positive probability. Consider the subproblem whereby we are at x_i at time i and wish to minimize the "cost-to-go" from time i to time N

$$\mathbb{E}\left[g_N(x_N) + \sum_{k=i}^{N-1} g_k(x_k, \mu_k(x_k), w_k) \right]$$

Then the truncated policy $\{\mu_i^*, \mu_{i+1}^*, \ldots, \mu_{N-1}^*\}$ is optimal for this sub-problem.

Then the truncated policy $\{\mu_i^*, \mu_{i+1}^*, \ldots, \mu_{N-1}^*\}$ is optimal for this sub-problem.

Motivation for Dynamic Programming Algo:

Consider the inventory control example we discussed in lecture 1. We will follow the following procedure to determine the optimal ordering policy starting from the last period and proceeding backward in time.

Period N-1: Assume that at the beginning of period N-1 the stock is x_{N-1}.

- Inventory manager should order $u_{N-1} \in \{0, 1, \ldots, C - x_{N-1}\}$ that minimizes $cu_{N-1} + \mathbb{E}[R(x_N)] = cu_{N-1} + \mathbb{E}[R(x_{N-1} + u_{N-1} - w_{N-1})]$
- Adding to the holding/shortage cost for the last period (plus the terminal cost) is

$$J_{N-1}(x_{N-1}) = r(x_{N-1}) + \min_{u_{N-1} \geq 0} \left[cu_{N-1} + \mathbb{E}_{w_{N-1}}\left[R(x_{N-1} + u_{N-1} - w_{N-1})\right]\right]$$

- So, $J_{N-1} : S_{N-1} \to \mathbb{R}$. In the process of computing $J_{N-1}(x_{N-1})$ we also find u_{N-1}^* that minimizes rhs. and we can define μ_{N-1}^* s.t $\mu_{N-1}^*(x_{N-1}) = u_{N-1}^*$

Period N-2: Assume that at the beginning of period N-2 the stock is x_{N-2}. Inventory manager should order the amount of inventory that minimizes

(expected cost of period N-2) + (expected cost of period N-1 given that an optimal policy will be used at period N-1)

which is equal to

$$r(x_{N-2}) + cu_{N-2} + \mathbb{E}_{w_{N-2}}\left[J_{N-1}(x_{N-1})\right]$$

So define.

$$J_{N-2}(x_{N-2}) := r(x_{N-2}) + \min_{u_{N-2} \geq 0} \left[cu_{N-2} + \mathbb{E}_{w_{N-2}}\left[J_{N-1}(x_{N-2} + u_{N-2} - w_{N-2})\right]\right]$$

- Again $J_{N-2}(x_{N-2})$ is calculated for every x_{N-2}. At the same

$u_{N-2} \geq 0$ $\quad \underset{w_{N-2}}{\mathbb{E}}\left[J_{N-1}(x_{N-2} + u_{N-2} - w_{N-2}) \right]$

- Again $J_{N-2}(x_{N-2})$ is calculated for every x_{N-2}. At the same time the optimal policy $\mu_{N-2}^*(x_{N-2})$ is also computed.

Period k: Similarly, we have that at period k, when the stock is x_k, the inventory manager should order u_k to minimize

(expected cost of period k) + (expected cost of periods $k+1, \ldots, N-1$ given that an optimal policy will be used for these periods)

Define

$$J_k(x_k) = r(x_k) + \min_{u_k \geq 0} \left[cu_k + \underset{w_k}{\mathbb{E}}\left[J_{k+1}(x_k + u_k - w_k) \right] \right]$$

which is actually the dynamic programming equation for this problem.

Proposition 1: For every initial state x_0, the optimal cost $J^*(x_0)$ of the basic problem is equal to $J_0(x_0)$ where the function J_0 is given by the last step of the algorithm, which proceeds backward in time from period $N-1$ to period 0:

$$J_N(x_N) = g_N(x_N) \qquad -(1)$$

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

$$k = 0, 1, \ldots, N-1 \qquad -(2)$$

where the expectation is taken w.r.t. the prob. dist. of w_k, which depends on x_k and u_k. Furthermore, if $u_k^* = \mu_k^*(x_k)$ minimizes the r.h.s of (2) for each x_k and k, the policy $\pi^* = \{\mu_0^*, \ldots, \mu_{N-1}^*\}$ is optimal.