Week-6

## L2 Stochastic Inventory

- When the demand of the prod. is not known with certainity - only probab. distribu$^n$ is known. At each time period, the decision maker obs. the current inventory level & decides how much to order. $C_o$ & $C_c$
- Demand is discrete ($D_t = 1, 2 \dots$), delivery is instantaneous. Demand is IID random var., with a stationary PMF of $p_j$ ($Pr(D_t = j) = P_j$). Warehouse capacity is M units. Inventory is non-perishable. Demand not met on the day is lost.
- $t = 1, 2 \dots N$, $s_t$, $a_t$

reward func$^n$

$$r_t(s_t, a_t) = E[Min(D_t, s_t + a_t) * price] - C_o(a_t) - C_c(s_t + a_t)$$

- terminal

$$r_N(s_N, a_N) = g(s_n) = \text{salvage value.}$$

- MDP

$$s_{t+1} = Max(s_t + a_t - D_t, 0)$$

- Obj. → Max. of reward over time horizon.
- State transition posibilities -

$$Pr(s' \mid s, a) = \begin{cases} 0 & \text{if } s' \in (s+a, M) \\ p_{s+a-s'} & \text{if } s' \in [0, s+a] \ \& \ s+a \leq M \\ \sum_{k \geq s+a} p_k & \text{if } s' = 0 \ \& \ s+a \leq M \end{cases}$$

Comp. of the obj. func$^n$ –

$$C_0(a_t) = \begin{cases} k + c(a_t), & \text{if } a_t > 0 \\ 0 & \text{if } a_t = 0 \end{cases}$$

Revenue,

$$F_t(s_t + a_t) - E[\text{Min}(D_t, s_t + a_t) * \text{price}]$$

- If $u = s_t + a_t > D_t = j$, let revenue $f(j) = \text{price} \times j$. This happens with probab. $p_j$.
If demand exceeds stock, then revenue is $f(u)$ with probability $q_u = \sum_{j=u}^{\infty} p_j$.

$$F_t(u) = F_t(s_t + a_t) = \sum_{j=0}^{u-1} f(j) * p(j) + f(u) \times q_u$$

- Policies → Markov decision rule: an order quantity to be ordered in each time period $(a_t)$ to each possible starting inventory for that time period. A policy is a seq. of such decision rules $(d_n = a_n(1), a_n(2)$
- Value Iteration

→ Finite horizon optimality (Bellman eq$^n$) –

$$v_n(s) = \max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} \lambda * p(j|s,a) * v_{n-1}(j) \right.$$

Passing to limit, the ∞ horizon eq$^n$ becomes –

$$v(i) = \max_{a \in A_s} \left\{ r(i,a) + \sum_{j \in S} \lambda * p(j|s,a) * v(j) \right\}$$

HO

- Q-learning

→ model-based MDP – when probab. transi" matrix is known ($Pa(s'|s,a)$), model-free – when these probab. are not known.

→ def. seq. of vectors $\{Q_t \in R^{S \times A} : t = 0, 1, 2 \cdots\}$

→ this is Q learning.