EE6106 LECTURE 20 | Dated: 10^{th} April 2024

Discounted Average Cost MDPs

$$(S, A, P, C, \alpha)$$

- V^* satisfies

$$V^*(i) = \min_a \left[C(i,a) + \alpha \sum P_{ij}(a) V^*(j) \right] \quad ①$$

Step

- Given stationary policy f, $T_f : B(S) \longrightarrow B(S)$

$$T_f^n u \xrightarrow{n \uparrow \infty} V_f \quad ②$$

- Stationary policy that minimizes RHS of ① is optimal

The infimization is over all policies.

① can be thought of as a certain fixed point operation

In ②, u can be thought of somewhat like a termination cost

We are not in the learning setting right now, but in the "PLANNING" setting

Def A mapping $T: B(S) \longrightarrow B(S)$ is a contraction mapping if, for $u, v \in B(S)$

$$\|Tu - Tv\| \leq \beta \|u - v\| \quad \text{for } \beta \in (0,1)$$

$$\|u\| = \max_i |u(i)|$$

Contraction mapping Theorem:

Say $T: B(S) \longrightarrow B(S)$ is a contraction mapping. Then it has a unique fixed point. g i.e. $Tg = g$

Moreover, for any $u \in B(S)$, $T^n u \xrightarrow{n \uparrow \infty} g$

Lemma $T_\alpha : B(S) \longrightarrow B(S)$ is defined as

$$(T_\alpha u)(i) = \min_a \left[C(i,a) + \alpha \sum P_{ij}(a) u(j) \right]$$

is a contraction mapping

The interpretation of T_α is the minimum cost starting from state i.

Further, V^* is a fixed point of T_α

Our Bellman equations state that $T_\alpha V^* = V^*$

But since T_α is a contraction mapping, V^* is THE unique fixed point.

Algorithm

Initialize $V(1)$ arbitrarily

For $t \geq 1$

$$V(t+1) = T_\alpha V(t)$$

$$V(t) \xrightarrow{t \uparrow \infty} V^*$$

$\boxed{\text{value iteration}}$

Proof to be shared on Moodle

V^* solves the LP :

$$\boxed{\begin{aligned} &\max \sum_i V(i) \\ &\text{s.t.} \quad V(i) \le C(i,a) + \alpha \sum P_{ij}(a) V(j) \quad \forall\, i \in S, a \in A \end{aligned}}$$

linear objective and linear constraints

The dual of this LP has some nice interpretations.

For stationary policy π,

$$Q_\pi(i,a) = C(i,a) + \alpha \sum P_{ij}(a) V_\pi(j)$$

Action value function

Interpretation : Start at a state, take given action & then play the move as dictated by the policy.

$$V_\pi(i) = Q_\pi(i, \pi(i))$$

Similarly,

$$Q^*(i,a) \equiv$$

Given stationary policy f,

$\hookrightarrow V_f, A_f$

$$f^*(i) = \arg\min_a \left[C(i,a) + \alpha \sum P_{ij}(a) V_f(j) \right]$$

$$= \arg\min_a Q_\pi(i,a)$$

If we observe, there is some sort of a mapping between T_f and this quantity.

Lemma

$$V_{f^*}(i) \leq V_f(i) \quad \forall i$$

An iteration would give something which is strictly better

Proof

$$T_{f^*} V_f(i) = C(i, f^*(i)) + \alpha \sum P_{ij}(f^*(i)) V_f(j)$$

... by definition

Proof $T_{f^*} V_f \ (i) = C(i, f^*(i)) + \alpha \sum P_{ij} (f^*(i)) V_f(j)$

... by defenition

By construction, since the RHS is the minimum

$$T_{f^*} V_f(i) \leq C(i, f(i)) + \alpha \sum P_{ij}(f(i)) V_f(j)$$

$$= V_f(i)$$

$$\Rightarrow T_{f^*} V_f(i) \leq V_f(i)$$

$$\Rightarrow T_{f^*}^{\,n} V_f(i) \leq V_f(i)$$

POLICY ITERATION

At $t \geq 1$

- Compute V_{π_t} $\longrightarrow$ Policy Evaluation
- $\pi_{t+1} = \pi_t^*$ $\longrightarrow$ Policy Improvement
- If $\pi_{t+1} = \pi_t$, STOP

REFERENCE :

Application : Selling an asset

- iid offers taking values $\{0, 1, \ldots, N\}$

P_i = probability of offer i

- Daily maintenance cost C
- Discount future cost by α
- Once sold, no cost going forward

$S = \{0, 1, 2, \ldots, N\} \cup \{\infty\}$

$A = \{0, 1\}$ OR $\{S_{(SELL)}, W_{(WAIT)}\}$

$$V^*(i) = \min \left\{ -i, \; C + \alpha \sum_{j=0}^{N} P_j V^*(j) \right\}$$

Note that the second term is not dependent on the state we are in.

Let $i^* = \min \{ i : -i < C + \alpha \sum P_j V^*(j) \}$

The optimal policy is then : sell if the bid is greater than i^*, else hold.

Note that we first need to obtain V^*. The way to do this is use the fact that we know the structure of this policy, use it to sweep over the CLASS of all such policies, find their value functions, and then optimize. This is now a 1-D optimization problem.

According to JKN, this is also a method to solve ARRANGED MARRIAGES

$f_i \sim$ policy of accepting offers $\geq i$

One of f_i is optimal.

The bids are i.i.d. Hence, the number of days before we accept the offer is a Geometrically distributed variable.

T = # of considered offers $\sim$ Geometric $\left(\sum_i^N P_i\right)$

Given T, conditional expected discounted cost

$$= C + \alpha C + \cdots \alpha^{T-2} C - \alpha^{T-1} \left(\frac{\sum_{j=i}^{N} j P_j}{\sum_{j=1}^{N} P_j} \right).$$

$$= \frac{C(1 - \alpha^{T-1})}{1 - \alpha} - \alpha^{T-1} \left(\frac{\sum_i^N j P_j}{\sum_i^N P_j} \right)$$

T is geometric, so we need to compute the following

<u>Exercise</u> : $\mathbb{E}[\alpha^{T-1}] = \dfrac{p}{1-\alpha(p-1)}$ where $p = \sum_{i}^{N} p_j$

The expected discounted cost under f_i is the following

$$\frac{c\sum_{0}^{i-1} p_j - \sum_{i}^{N} j p_j}{1 - \alpha \sum_{0}^{i-1} p_j}$$