EE6106 LECTURE 21 Dated: 12th April 2024

AVERAGE COST ~~OF~~ MDPs

Goal : Minimize

$$\phi_\pi(i) = \limsup_{n\to\infty} \frac{E_\pi\left[\sum_0^n C(X_t, A_t) \,\middle|\, X_0 = i\right]}{n+1}$$

(S, A, P, C) $\qquad$ $\phi^*(i) = \inf_\pi \phi_\pi(i)$

$\downarrow$

S : countable
A : finite

π^* is <u>optimal</u> if $\phi_{\pi^*}(i) = \phi^*(i)\ \forall\ i$

<u>Eg</u>

$S = \{1, 1', 2, 2', 3, 3', \dots\}$

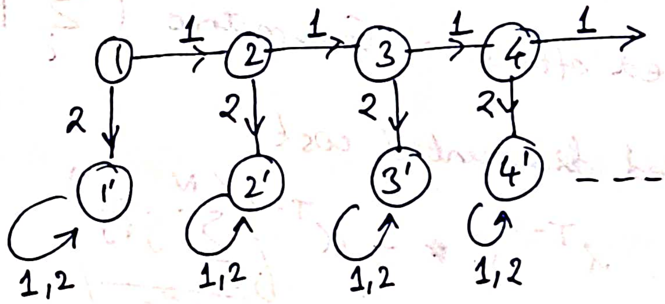
$A = \{1, 2\}$

$P_{i,i+1}(1) = P_{i,i'}(2) = 1$

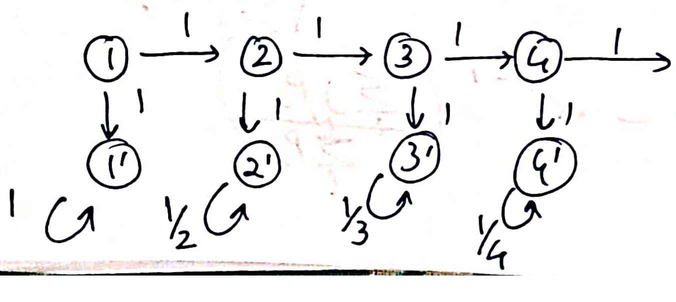
$P_{i'i'}(1) = P_{i'i'}(2) = 1$

$C(i, \cdot) = 1$

$C(i', \cdot) = 1/i$



transitions are represented over arrows
The number denotes action chosen



costs associated with the transitions are on the arrows

Claim: This MDP has no optimal policy

For any policy π, $\phi_\pi(1) > 0$

But can make $\phi_\pi(1)$ arbitrarily close to zero

Note that if we throw in a discount factor, there will exist a stationary optimal policy

Eg $S = \mathbb{N}$ $\quad P_{i,i+1}(1) = P_{i,i}(2) = 1$

$A = \{1, 2\}$ $\quad C(i,1) = 1$; $C(i,2) = 1/i$

Under any stationary policy π, $\phi_\pi(1) > 0$

In state i, play action 2 i times, then action 1

Cost seq. $1, 1, \frac{1}{2}, \frac{1}{2}, 1, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 1, \ldots$

This policy will have a time-average of 0 as $t \to \infty$

$$\phi_\pi(1) = 0$$

Ex
The can also find a randomized stationary policy which is optimal.

Thm: If there exists $h \in B(S)$ and constant g s.t.

$$g + h(i) = \min_a \left[C(i,a) + \sum P_{ij}(a) h(j) \right] \forall i \qquad (1)$$

then there exists an optimal stationary policy π^*, where

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

$$\text{and } \phi_{\pi^*}(i) = g \ \forall i$$

Note: The correct interpretation of g is the optimal time-average cost starting from the state i. associated with the optimal policy

(1) becomes the Bellman Equation for a Time-average cost MDP.

The g has to be unique.
h is not unique, because we can add an additive constant to h and it will still satisfy Bellman equation. This was not the case for discounted MDP, because the α acted as a counterweight.

Proof

Under any policy π,

$$\mathbb{E}_\pi[h(X_t) \mid H_{t-1}] = \sum h(j) P_{X_{t-1},j}(A_{t-1})$$

$$= C(X_{t-1}, A_{t-1}) + \sum h(j) P_{X_{t-1},j}(A_{t-1}) - C(X_{t-1}, A_{t-1})$$

$$\geq \min_a \left[c(X_{t-1}, a) + \sum h(j) P_{X_{t-1}, j}(a) \right] - c(X_{t-1}, A_{t-1})$$

$$= g + h(X_{t-1}) - c(X_{t-1}, A_{t-1})$$

$$\Rightarrow \quad 0 \leq E_\pi\left[\sum_1^n \left(h(X_t) - h(X_{t-1}) - g + c(X_{t-1}, A_{t-1}) \right) \right]$$

$$\Rightarrow \quad g \leq \frac{E_\pi[h(X_n)]}{n} - \frac{E_\pi[h(X_0)]}{n} + \frac{E_\pi\left[\sum_1^n c(X_{t-1}, A_{t-1}) \right]}{n}$$

Letting n tend to infinity and noting that h is a bounded function

$$g \leq \phi_\pi(X_0)$$

recall $\displaystyle \phi_\pi(i) = \limsup_{n \to \infty} \frac{E_\pi\left[\sum_0^n c(X_t, A_t) \mid X_0 = i \right]}{n+1}$

There exists only one place in the whole proof, where there was a inequality, hence if we choose the policy π^*, all inequalities turn into equalities and hence we have

$$g = \phi_{\pi^*}(X_0)$$

α-optimal policy (optimal with discount factor α) "solves" (when)

$$V_\alpha^*(i) = \min_a \left[c(i,a) + \alpha \sum P_{ij}(a) V_\alpha^*(j) \right]$$

<u>Idea</u>: Consider

$$\lim_{\alpha \uparrow 1} \left[c(i,a) + \alpha P_{ij}(a) V_\alpha^*(j) \right]$$

An issue is that the limit may not exist

$h_\alpha(i) = V_\alpha^*(i) - V_\alpha^*(0)$ where 0 is some reference state

$$(1-\alpha) V_\alpha^*(0) + h_\alpha(i) = \min_a \left[c(i,a) + \alpha \sum P_{ij} h_\alpha(j) \right]$$

It is more likely that things will work out better using this format

The "h" are now called not value functions, but DIFFERENTIAL VALUE FUNCTIONS

The hope is that $(1-\alpha) V_\alpha^*(0) \longrightarrow g$ as $\alpha \to 1$

Thm X: If $\exists\, N < \infty$ s.t. $|V_\alpha^*(i) - V_\alpha^*(0)| < N \;\; \forall\, \alpha, i$, then

i) There exist g, h satisfying $\boxed{1}$ → Bellman equation for time-avg MDP]

ii) For some seq. $\alpha_n \to 1$,

$$h(i) = \lim_{n\to\infty} \left[V_{\alpha_n}^*(i) - V_{\alpha_n}^*(0) \right]$$

iii) $\lim_{\alpha \to 1} (1-\alpha) V_\alpha^*(0) = g$

~~Proof~~

Def An MDP is unichain if it is irreducible under any stationary policy.

↓ it is possible to reach every other state from every state

Thm On a finite unichain MDP, hypothesis of Thm X holds.