30th Sept:

recap LP: $\min_{\vartheta} \sum \alpha(s)\vartheta(s)$ where $\vartheta(s) \geq r(s,a) + \lambda \sum P(j|s,a)\vartheta(j)$ is the primal, $\alpha(s) > 0 \ \forall s$, $\sum \alpha(s) = 1$ $\forall s,a$

from above LP we get optimal value

dual: $\max_{x} \sum_{s,a} r(s,a)x(s,a)$ s.t $\sum_{a \in S_j} x(j,a) - \lambda \sum_{s,a} P(j|s,a)x(s,a) = \alpha(j) \ \forall j$

$x(j,a) \geq 0 \ \forall j,a$

this LP gives optimal policy $(d^*)^\infty$, $Z \rightarrow$ feasible region

Note: If $\sum_{s,a} c(s,a)x(s,a) \leq \beta$ is for constraint MDP

now, $\lim_{T\to\infty} \sum_{k} \alpha(k) \frac{\sum_{t=1}^{T} P^{d^\infty}(X_t = s, Y_t = a \mid X_1 = k)}{T}$

$= \frac{\mathbb{E}[N(X_t = s, Y_t = a \text{ in } t = 1, \ldots T \text{ epochs})]}{T}$

(If this was markov chain it will converge to stationary state)

$\sum_{k} \alpha(k) \sum_{t=1}^{\infty} \lambda^{t-1} P^{d^\infty}(X_t = s, Y_t = a \mid X_1 = k) = x_d(s,a)$ ← so, $x(s,a)$ is just discounted occupancy measure

now, $\sum_{s,a} x_d(s,a) r(s,a) = \sum_{s} \vartheta^{d^\infty}(s)\alpha(s)$ as:

$$\sum_{s,a} x_d(s,a) r(s,a) = \sum_{k} \alpha(k) \sum_{t=1}^{\infty} \lambda^{t-1} \sum_{s,a} P^{d^\infty}(X_t = s, Y_t = a \mid X_1 = k) r(s,a)$$

$$= \sum_{k} \alpha(k) \sum_{t=1}^{\infty} \lambda^{t-1} \mathbb{E}[r(X_t, Y_t \mid X_1 = k)]$$

(different notations)

$$= \sum_{s} \vartheta^{d^\infty}(s)\alpha(s) = \vartheta^{d^\infty}(\alpha) = J^{d^\infty}_{\alpha,\lambda} = J(d^\infty, \alpha)$$

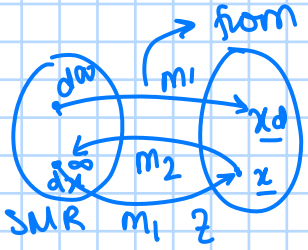
(LP)

from below theorem, (a) $\Rightarrow \vartheta^{d^\infty}(\alpha) = \sum_{s,a} r(s,a) x_d(s,a)$ as $x_d \in Z$ ← from below theorem

(b) $\Rightarrow \sum r(s,a)x(s,a) = \vartheta^{d^\infty_x}(\alpha)$ as $x = x_{d_x}$ ← from below theorem

we have $\sum x(s,a)r(s,a) = \sum x_{d_x}(s,a) r(s,a)$

$= \vartheta^{d^\infty_x}(\alpha)$



$m_1(d^\infty_x) = x$ ← from (b)

given any $d \rightarrow x_d$ given any $x \rightarrow d^*$ so one-one, onto

m_1, m_2 are mappings

so, max of feasible region gives us optimal policy

Note: $r(s,a)$ is not special here, we can use $c(s,a)$ to do same for constraint MDP

Hernandez result: $x^*(s,a) = 0 \ \forall a \neq a^s_1, s \neq \tilde{s}$

$x^*(\tilde{s}, a^{\tilde{s}}_1) > 0$, $x^*(\tilde{s}, a^{\tilde{s}}_2) > 0$ then constraint

Theorem: (a) If $d^\infty \in SMR$, then define

$$x_d(s,a) \triangleq \sum_{k \in S} \alpha(k) \sum_{t=1}^{\infty} \lambda^{t-1} P^{d^\infty}(X_t = s, Y_t = a \mid X_1 = k)$$

then $x_d \in Z$

(b) If $x \in Z$ define $d_x(s,a) = \frac{x(s,a)}{\sum_{a'} x(s,a')} \ \forall s,a$

now d^∞_x is SMR

let x_{d_x} be s.t $x_{d_x}(s,a) \triangleq \sum \alpha(k) \sum \lambda^{t-1} P^{d^\infty_x}(X_t = s, Y_t = a \mid X_1 = k)$

then $x_{d_x} = x$

proof: (a) given d^∞, defined x_d, we have to show $x_d \in \mathcal{Z}$

now $\lambda \sum_{s,a} p(j|s,a)\, x_d(s,a)$ (part LHS from dual of LP)

$$= \lambda \sum_{s,a} p(j|s,a) \sum_k \alpha(k) \sum_{t=1}^{\infty} \lambda^{t-1} p^{d^\infty}(X_t = s, Y_t = a \mid X_1 = k)$$

$$= \lambda \sum_k \alpha(k) \sum_{t=1}^{\infty} \lambda^{t-1} \sum_{s,a} p(j|s,a)\, p^{d^\infty}(X_t = s, Y_t = a \mid X_t = k)$$

$$= \sum_k \alpha(k) \sum_{t=1}^{\infty} \lambda^t \sum_{\tilde{a}} p^{d^\infty}(X_{t+1} = j, Y_{t+1} = \tilde{a} \mid X_1 = k)$$

$t+1 = n$, $n = 2 \to \infty$

$$= \sum_k \alpha(k) \sum_{n=2}^{\infty} \lambda^{n-1} \sum_{\tilde{a}} p^{d^\infty}(X_n = j, Y_n = \tilde{a} \mid X_1 = k)$$

$$= \sum_k \alpha(k) \sum_{n=1}^{\infty} \lambda^{n-1} \sum_{\tilde{a}} p^{d^\infty}(X_n = j, Y_n = \tilde{a} \mid X_1 = k)$$

$$- \sum_k \alpha(k) \lambda^0 \sum_{\tilde{a}} p^{d^\infty}(X_1 = j, Y_1 = \tilde{a} \mid X_1 = k)$$

$$= \sum_{\tilde{a}} x_d(j, \tilde{a}) - \alpha(j) \sum_k \delta(j = k)$$ → will be 1

So, we get: $\sum_a x(j,a) - \alpha(j) = \lambda \sum_{s,a} p(j|s,a)\, x(s,a) \quad \forall j$

this is same as $x \in \mathcal{Z}$ from dual LP

Exe: show (b) of theorem → done down

THEOREM: (i) The two LPs have optimal solution

(ii) say x^*, v^* are some optimal solution then

(a) $v^*(s)$ value vector of MDP

(b) $d^*(s,a) = \dfrac{x^*(s,a)}{\sum_{\tilde{a}} x^*(s,\tilde{a})}$ then $(d^*)^\infty$ is an optimal policy for MDP

assuming $|r(s,a)| \le M$

proof:

Feasibility will be true as for any $d \in SMR$ from (a) of above theorem we get $x_d \in \mathcal{Z}$

now as $\sum_{s,a} x_d(s,a)\, r(s,a) = \sum_{k \in S} \alpha(k) v^{d^\infty}(s) \le \dfrac{M}{1-\lambda}$

as $-\dfrac{M}{1-\lambda} \le v^{d^\infty}(s) \le \dfrac{M}{1-\lambda}$

$$\text{so, } -\frac{M}{1-\lambda} \le \sum_{s,a} x_d(s,a)\, r(s,a) \le \frac{M}{1-\lambda}$$

so its feasible, bounded and so max will exist and so will min

(i) of theorem follows from this as feasible and bounded

(ii) we get x^*, v^* are optimal solution

$\sum \alpha(s) v^*(s) = \sum x^*(s,a)\, r(s,a)$ (∵ duality)

$\Rightarrow \sum \alpha(s) v^*(s) = \sum x^*(s,a) r(s,a)$ ← from solution, we can use $v d_{x^*}^\infty$

$= \sum \alpha(s) v d_{x^*}^\infty(s)$ from previous theorem (b) part

now we have to show $v^*(s) = v d_{x^*}^\infty(s)$ $\forall s$, then we are done using below lemma we have

$v^*(s) \geq v d_{x^*}^\infty(s)$ ($\because d_{x^*}^\infty$ is also a policy)

now as $\sum \alpha(s) v^*(s) = \sum \alpha(s) v d_{x^*}^\infty(s)$

and $\alpha(s) > 0$

$\Rightarrow \sum \alpha(s) v^*(s) \geq \sum \alpha(s) v d_{x^*}^\infty(s)$

now, as $\sum \alpha(s) v^*(s) = \sum \alpha(s) v d_{x^*}^\infty(s)$

if $\exists s$ s.t $v(s) > v d_{x^*}^\infty(s)$

then $\sum \alpha(s) v^*(s) > \sum \alpha(s) v d_{x^*}^\infty(s)$

this is a contradiction

so $\Rightarrow v^*(s) = v d_{x^*}^\infty(s)$ $\forall s$

so $d_{x^*}^\infty$ is optimal policy for MDP

lemma: $v^*(s) \geq v d^\infty(s)$ $\forall s$, $\forall d^\infty$ where $v^*(\cdot)$ is solution of primal

proof:

as v^* is solution to primal it is in $\mathcal{F}$ i.e

$$v^*(s) \geq r(s,a) + \lambda \sum_{j \in S} P(j|s,a) v^*(j) \quad \forall s,a$$

now for any $d^\infty \in SMR$ we have

$\sum_a q_d(s,a) = 1$ ($\because$ probability distribution)

$\Rightarrow v^*(s) = \sum v^*(s) q_d(s,a)$

$\geq \sum_a (r(s,a) + \lambda \sum_{j \in S} P(j|s,a) v^*(j)) q_d(s,a)$

$= \sum_a [q_d(s,a) r(s,a)] + \sum_a [\lambda \sum_{j \in S} P(j|s,a) q_d(s,a) v^*(j)]$

$= r_d(s) + \lambda \sum_j P_d(j|s) v^*(j)$ $\forall s$

$\Rightarrow v^* \geq r_d + \lambda P_d v^*$

as $v d^\infty = r_d + \lambda P_d v d^\infty$

$\Rightarrow v^* \geq v d^\infty - \lambda P_d v d^\infty + \lambda P_d v^*$

$\Rightarrow v^*(I - \lambda P_d) \geq v d^\infty (I - \lambda P_d)$

$\Rightarrow v^* \geq v d^\infty$ ($\because$ seen before, P_d is positive maps)

Ene: show (b) of theorem

Ans: $x(s,a) \in \mathcal{Z}$ let

$u(s) = \sum_a x(s,a)$ then $u(s) > 0$ (from dual constraints)

and to d_x is valid $\Rightarrow d_x^\infty$ is valid SMR

now to show $\alpha = x_a$

$$\alpha(j) = u(j) - \sum_{s,a} \lambda P(j|s,a) x(s,a)$$

$$= u(j) - \sum_{s,a} \lambda P(j|s,a) x(s,a) \frac{u(s)}{\sum x(s,a')}$$

$$= u(s) - \sum_{s,a} \lambda P(s'|s,a) q_d(s,a) u(s)$$

$$= u(s) - \sum_{s} \lambda P_{d_\chi}(s'|s) u(s)$$

i.e $\underbrace{(\alpha(s_1), \dots \alpha(s_n))}_{\alpha^T} = \underbrace{(u(s_1), \dots u(s_n))}_{u^T} - \lambda (u(s_1), \dots u(s_n)) \times \begin{bmatrix} P_{d_\chi}(s_1|s_1), \dots & \vdots \\ \vdots & \vdots \\ P_{d_\chi}(s_1|s_n), \dots & \end{bmatrix}$

$$\Rightarrow \alpha^T = u^T - \lambda u^T P_{d_\chi}$$
$$= u^T (I - \lambda P_{d_\chi})$$
$$\Rightarrow \alpha^T (I - \lambda P_{d_\chi})^{-1} = u^T$$
$$\Rightarrow \alpha^T \left(\sum_{n=1}^{\infty} (\lambda P_{d_\chi})^{n-1} \right) = u^T$$

so, $u(s) = \sum_{s'} \alpha(s') \sum_{n=1}^{\infty} \lambda^{n-1} \sum_{a \in A_s} p_{d_\chi^\infty}(X_n = s, Y_n = a \mid X_1 = s')$

$$= \sum_{a \in A_s} x\, d_\chi(s,a)$$

$$\Rightarrow \sum_{a \in A_s} x(s,a) = \sum_{a \in A_s} x\, d_\chi(s,a) \quad \text{——} (1)$$

now, $x_{d_\chi}(s,a) = \sum_{s'} \alpha(s') \sum_{n=1}^{\infty} \lambda^{n-1} p_{d_\chi^\infty}(X_n = s \mid X_1 = s') q_{d_\chi(s)}(a)$

$$= \sum_{s'} \alpha(s') \sum_{n=1}^{\infty} \lambda^{n-1} p_{d_\chi^\infty}(X_n = s \mid X_1 = s') \frac{x(s,a)}{\sum_{a' \in A_s} x(s,a')} \quad \text{——} (2)$$

and $\sum_a x\, d_\chi(s,a) = \sum_{s'} \alpha(s') \sum_{n=1}^{\infty} \lambda^{n-1} p_{d_\chi^\infty}(X_n = s \mid X_1 = s')$ —— (3)

so, (2) becomes: $x_{d_\chi}(s,a) = \left(\sum_{a_1} x\, d_\chi(s,a_1) \right) \left(\frac{x(s,a)}{\left(\sum_{a_2} x(s,a_2) \right)} \right)$ (∵ (3))

$$= \left(\sum_{a_1} x\, d_\chi(s,a_1) \right) \left(\frac{x(s,a)}{\left(\sum_{a_2} x\, d_\chi(s,a_2) \right)} \right)$$ (∵ (1))

$$= x(s,a)$$

so, $x_{d_\chi}(s,a) = x(s,a) \;\forall\, s,a$

$$\Rightarrow \underline{x} d_\chi = \underline{x}$$

3rd out:

LPs: ① $\min_{\underline{v}} \sum_s \alpha(s) v(s)$, $v(s) \geq r(s,a) + \lambda \sum_j P(j|s,a) v(j) \quad \forall s,a$

② $\max_{\underline{x}} \sum_{s,a} x(s,a) r(s,a)$ $\quad \sum_a x(s',a) = \lambda \sum_{s,a} P(s'|s,a) x(s,a) + \alpha(s') \quad \forall s'$

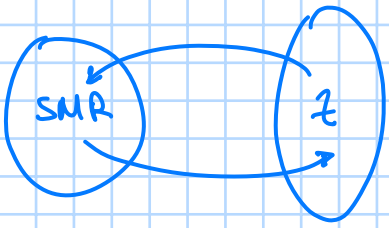
we also saw $\forall \underline{x} \in \mathcal{Z}$, $\exists d_{\underline{x}}(s,a) = \frac{x(s,a)}{\sum_{a'} x(s,a')}$

s.t $d_{\underline{x}}^{\infty}$ is SMR

and similarly $\forall d \in SMR$, $\exists \underline{x}_d \in \mathcal{Z}$

$$x_d(s,a) = \sum_k \alpha(k) \sum_{t=1}^{\infty} \lambda^{t-1} p^{d^{\infty}}(X_t = s, Y_t = a \mid X_1 = k)$$

so, given any vector $\in \mathcal{Z}$ $\exists$ SMR



$\exists f: SMR \to \mathcal{Z}$
f is one-one and onto
and so, $\exists f^{-1}$

we have seen, $\forall \underline{x}$, $\exists d_{\underline{x}}^{\infty}$ and $\forall d^{\infty}$, $\exists \underline{x}_d \in \mathcal{Z}$

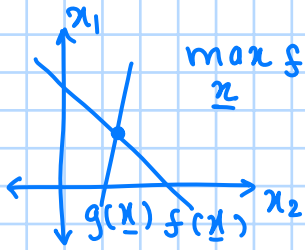
also using $d_{\underline{x}}^{\infty}$, if we define $\underline{x}_{d_{\underline{x}}}$ we get $\underline{x}_{d_{\underline{x}}} = \underline{x}$

Note: $O_r(x) = \sum_{s,a} r(s,a) x(s,a)$

then, $\underline{v}^{\pi}$ is s.t $\underline{v}^{d_{\underline{x}}^{\infty}}(\alpha) = O_r(x)$

$\hookrightarrow \sum_s \alpha(s) \underline{v}^{d_{\underline{x}}^{\infty}}(s)$

Note: if we have constraint MDP, $\sum x(s,a) c(s,a) \leq B$ will represent constrained fuction



$\max_{\underline{x}} f(\underline{x})$ s.t $x_1 + x_2 = 1$
let $g(\underline{x}) \leq B$
↑ constraint

$\sup_{\pi} \sum \lambda^{t-1} \mathbb{E}_{\alpha}^{\pi}[R_t]$ s.t $\sum \lambda^{t-1} \mathbb{E}_{\alpha}^{\pi}[C_t] \leq B \ \forall \pi$
↘ want to find this

Note: we also saw if $\underline{v}^*$, $\underline{x}^*$ are solution to ① and ②, then we saw given $\underline{x}^* \to d_{\underline{x}^*}^{\infty} \in SMR$

and $\underline{v}^* = \underline{v}^{d_{\underline{x}^*}^{\infty}}$ for any value of α i.e it does not depend on α

we only need
$\alpha(s) > 0 \ \forall s$
$\sum_s \alpha(s) = 1$

we get this as there is potential to optimise all of them in every direction (as $\alpha(s) > 0$) and so does not matter on α

any d^{∞}, we can show $\underline{v}^* \geq \underline{v}^{d^{\infty}}$

$\Rightarrow \underline{v}^* \geq \sup_{\pi} v^{d^{\infty}} = v_x^*$

$\Rightarrow \underline{v}^* \geq v_x^*$ point wise

and $\underline{v}^* \geq \underline{v}_{d^\infty}$ as for every $d^\infty \in SMR$

so, $\underline{v}^*(s) \geq v_{d^\infty}(s) \quad \forall s$

$\underline{v}_{d^\infty}$ satisfies ① with equality

and $\underline{v}_{d^\infty}$ so, $\in$ feasible primal LP

and as $\underline{v}^*$ is optimal

so, $\sum_s \alpha(s) v^*(s) \leq \sum_s \alpha(s) v_{d^\infty}(s)$

↑ as minimal value as $\underline{v}^*$ is optimal

if $\exists s$ s.t $v^*(s) > v_{d^\infty}(s)$

then

$\Rightarrow \sum_s \alpha(s) v_{d^\infty}(s) < \sum_s \alpha(s) v_{d^\infty}(s)$

$\Rightarrow 1 < 1$ this is a contradiction

so, $v^*(s) = v_{d^\infty}(s) \quad \forall s$

so, finally we get $\underline{v}^* = \underline{v}_{d^\infty} = \underline{v}^*_\lambda$

as $\underline{v}^* \geq \underline{v}^*_\lambda \Rightarrow \underline{v}_{d^\infty} \geq \underline{v}^*_\lambda$

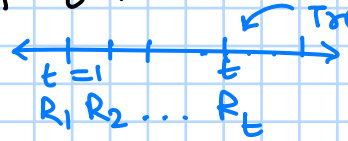
$\Rightarrow \underline{v}_{d^\infty} = \underline{v}^*_\lambda$ as $d^\infty \in SMR$

so, d^∞ is optimal policy

(also $\underline{v}^* = \underline{v}^*_\lambda$ and so $\underline{v}^*$ does not depend on α
↳ so by $\underline{v}^*$ we can get optimal policy)

Now, we saw discrete cost mdps where we want to maximise

$\sup_\pi \sum_{t=1}^\infty \lambda^{t-1} \mathbb{E}^\pi[R_t]$



Eg: for a queing system we are seeing something for 12 hours
we keep looking at toyetion cost every 5 min
here $\lambda = 1$ to model an MDP as reputation on future
interval is important ($\lambda = 1$) ← lot of importance to study state

Average cost MDPs:

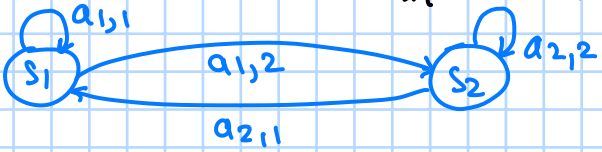
we now put $\lambda = 1$, and try to maximise $\lim_{T\to\infty} \frac{\sum_{t=1}^{T} \mathbb{E}^\pi[R_t]}{T}$

also, $\exists \bar{\lambda} \in (0,1)$ s.t $\bar{\lambda}$ is some big value

$d^*_\lambda = d^*_{av}$ ↱ average cost $\forall \lambda \geq \bar{\lambda}$

i.e $(1-\lambda) v_{*,\lambda} \to v_{*,avg}$

Eg:



$r(s_1, a) = 2 \quad \forall a \in A_{s_1}$
$r(s_2, a) = -2 \quad \forall a \in A_{s_2}$

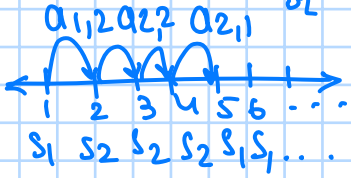
$a_{1,1}$
$a_{1,2}$
$P(s_1|s_1, a_{11}) = 1$
$P(s_2|s_1, a_{12}) = 1$

$a_{2,1}$
$a_{2,2}$
$P(s_2|s_2, a_{22}) = 1$
$P(s_1|s_2, a_{2,1}) = 0$

now if we start with S_1 always

$\pi \in HD$, $S_1 \to 3^0 \leftarrow$ 3^0 times on S_1

$S_2 \to 3^1 \leftarrow 3^1$ times on S_2



$S_1 \to 3^2$ time

$S_1 \to 3^{2k}$

$S_2 \to 3^{2k+2}$, twice at end of $2k$ switches, $2k+1$ switches

$$N_k = \sum_{n=0}^{2k} (3)^n \qquad N'_k = \sum_{n=0}^{2k+1} (3)^n$$

$$= \frac{3^{2k+1}-1}{3-1} \qquad \Rightarrow N_k + 3^{2k+1}$$

now, let $\frac{v_{N_k}}{N_k}$ = avg over N_k epos

$$\Rightarrow \frac{v_{N_k}}{N_k} = \frac{\sum_{n=0}^{k} 3^{2n}(2) - \sum_{n=0}^{k-1}(2)\,3^{2n+1}}{N_k}$$

$$= \frac{2}{N_k}\left(\sum_{n=0}^{k} 3^{2n} - \sum_{n=0}^{k-1} 3^{2n+1}\right)$$

$$= \frac{4}{3^{2k+1}-1}\left(\frac{(3^2)^{k+1}-1}{3^2-1} - \frac{(3^2)^k-1}{3^2-1}\right)$$

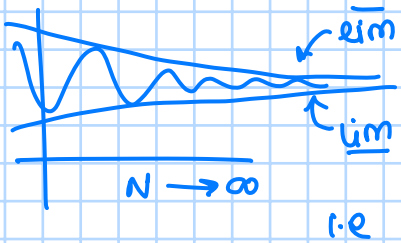
for $k \to \infty$ we get $\lim_{k\to\infty} \frac{v_{N_k}}{N_k} = \frac{1}{2}$

similarly for $N_k + 3^{2k+1}$ part

$$\lim_{k\to\infty} \frac{v_{N_k+3^{2k+1}}}{N_k+3^{2k+1}} = -\frac{1}{2}$$

So, $\overline{\lim} \frac{v_N}{N} = \frac{1}{2}$ $\underline{\lim} \frac{v_N}{N} = -\frac{1}{2}$ $\Rightarrow$ not convergent

Note: From above example we see that not all policy will converge to avg



for some π, but $\lim$ may not exist, we can still be interested in upper bound

i.e $\inf_{\pi} \overline{\lim}_{N\to\infty} \frac{v_N^{\pi}}{N}$ or $\sup_{\pi} \overline{\lim}_{N\to\infty} \frac{v_N^{\pi}}{N}$ $\to$ depending on use case of MDP

now, $\forall s \in \mathcal{S}$ $P(s'|s, d(s))$ is some markov chain by fixing d^∞

Defn: (irreducible MC) $\exists n$ s.t $P^n(s'|s) > 0$ if $s \to s'$, $\forall s, s'$
↖ Return in finite time $x \to x \to x \cdots$

Defn: If $P_x(\tau_x < \infty) = 1$, $\tau_x \to$ return time

$$\tau_x = \inf\{n > 1 ; X_n = x\}$$

then x is recurrent

Note: If irreducible, and $\exists x \in \mathcal{S}$ s.t x is recurrent $\Rightarrow \forall s \in \mathcal{S}$ s is recurrent

Defn: If $\mathbb{E}_x[\tau_x] < \infty$ then x is called positive recurrent

Note: If irreducible, and $\exists x$ s.t x is positive recurrent $\Rightarrow$ all positive recurrent

now, if we have a markov chain which is positive recurrent
$\Leftrightarrow \exists$ S.D s.t $\pi = \pi P_d$
$$\pi(x) = \frac{1}{\mathbb{E}_x[\tau_x]} \quad \forall x$$

π is dist, $\pi(x) \to$ prob of being in x, $x \in \mathcal{S}$

$$\pi P_d(s') = \sum_s \pi(s) P_d(s'|s)$$
$$\underbrace{P(X_2 = s | X_1 \sim \pi)}$$

now, $\dfrac{\sum_{n=1}^{N} \mathbb{I}_{\{X_n = s\}}}{N} \to \pi(s)$ as $N \to \infty$ almost surely

and also $\dfrac{\mathbb{E}\left[\sum_{n=1}^{N} \mathbb{I}_{\{X_n = s\}}\right]}{N} \to \pi(s)$ as $N \to \infty$

↖ expected value

i.e $\dfrac{\sum_{n=1}^{N} P_\mu^n(X_n = s)}{N} \to \pi(s)$ $\forall \mu$ as $N \to \infty$

we want to look at $\dfrac{\sum_{n=1}^{N} \mathbb{E}[r(X_n)]}{N}$

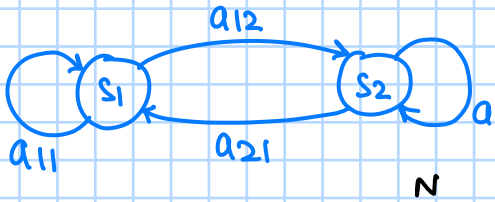
$$= \sum_s \frac{\sum_{n=1}^{N} P_\mu^n(X_n = s)\, r(s)}{N}$$

$$\to \sum_s \pi(s) r(s)$$

where μ: $\sum_{s_1} \mu(s_1) P(X_n = s | X_1 = s_1) = P_\mu(X_n = s)$
↑ initial distribution

7th Oct :

Eg:



$r(s_1, a) = 2$
$r(s_2, a) = -2$ $\forall a$

$P(s_p | s_i, a_{ij}) = 1$ $i \neq j$
$P(s_i | s_i, a_{ij}) = 0$

we are interested in $\lim_{N\to\infty} \frac{\sum_{n=1}^{N} \mathbb{E}[r(X_n, Y_n)]}{N}$ if it exist, in above

example, we can make a policy which is history dependent

$X_0 = s_1$, a → 1 3 3^2 ....
s_1 s_2 s_1

this is like pseudo stability

.... were we saw

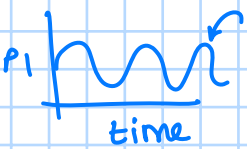
$$\overline{\lim} \frac{\sum_{n=1}^{N} \mathbb{E}[r(X_n, Y_n)]}{N} = 2$$

$$\underline{\lim} \frac{\sum_{n=1}^{N} \mathbb{E}[r(X_n, Y_n)]}{N} = -2$$

Eg: let P_1 = proportion of type 1 batteries

if P_1 is small then → 0

but if P_1, P_2 large both $\frac{P_1}{P_2} = 1$, then as time → ∞



any policy $\pi \in \Pi^{HR}$, we want to find

$$\overline{v}^{\pi}(s) = \lim_{N\to\infty} \frac{\sum_{n=1}^{N} \mathbb{E}_s^{\pi}[r(X_n, Y_n)]}{N}$$ ← this may not exist

and then $\sup_{\pi} \overline{v}^{\pi}(s)$

we know, $\pi \in \Pi^{HR}$ and $s = x_0$, $\exists \pi' \in \Pi^{MR}$ such that all marginal distribution are the same

$\pi \xleftrightarrow{}_{N} (X_n, Y_n) \overset{d}{=} \pi' \xleftrightarrow{} (X'_n, Y'_n)$ $\forall n$ ← only at time epoch n

so, $v_N^{\pi}(s) = \sum_{n=1}^{N} \mathbb{E}_s^{\pi}[r(X_n, Y_n)] = v_N^{\pi'}(s)$ $\forall N$

$\Rightarrow \overline{\lim_{N\to\infty}} \frac{v_N^{\pi}(s)}{N} = \overline{\lim_{N\to\infty}} \frac{v_N^{\pi'}(s)}{N}$ and similarly $\underline{\lim_{N\to\infty}} \frac{v_N^{\pi}(s)}{N} = \underline{\lim_{N\to\infty}} \frac{v_N^{\pi'}(s)}{N}$

If both $\overline{\lim}$ and $\underline{\lim}$ same then above limit will exist

Def: $g_+^{\pi}(s) = \overline{\lim_{N\to\infty}} \frac{v_N^{\pi}(s)}{N}$

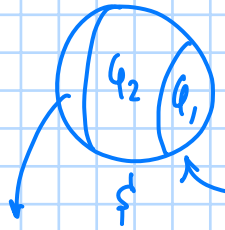
$g_-^{\pi}(s) = \underline{\lim_{N\to\infty}} \frac{v_N^{\pi}(s)}{N}$

Note: $\sup_{\pi\in\Pi^{HR}} \overline{v}^{\pi}(s) = \sup_{\pi\in\Pi^{MR}} \overline{v}^{\pi}(s)$, instead of Π^{HR} we use use Π^{MR}

Finite state MC:

$d^{\infty} \to$ SMR ← for every d^{∞} we will have one markov chain

$P_d(s'|s) = \sum_a P(s'|s,a) q_d(s,a)$ $|S|<\infty$ ← finite states

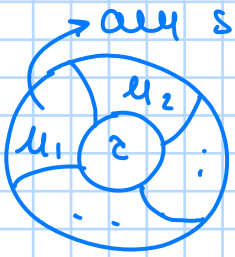


as finite states, there is atleast one recurrent state and will be positively recurrent (as finite many)

2 classes

finitely many classes

transient class (states we stop visiting after some time)



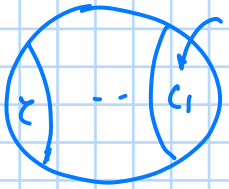
any state inside u_1 can be reached from u_1

so, $s\in\tau$, $\exists\, \sigma_{s,c_1}$ → prob of getting absorbed in c_1

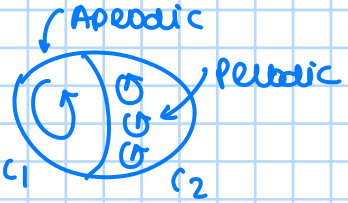
= prob of hitting class c_1 before class

Note: Now we will assume stationary distribution exist for every class as S.D exist let $(\mu^{(c)}(s))$ for $s\in c$, $\text{support}(\mu^{(c)}) = \{s: \mu^{(c)}(s)>0\} = C_c$, $\forall c$

↳ support on C_c



c_1 has unique S.D as irred + positive rec



now if S.D exist, $\mu_d P_d = \mu_d$, so $\mu_d^c P_d = \mu_d^c$ $\forall c\in$ class ← distribution

also, $P^n_{s,s'} = P(X_n=s' | X_0=s)$ ← prob of going from $s\to s'$ in n steps

if MC is aperiodic ← periodicity = 1 then

$P^n_{s,s'} \to \mu^{(c)}(s')$

if MC is periodic then above not garuetted but

$$\frac{\sum_{k\in n} P^k_{s,s'}}{n} \longrightarrow \mu^{(c)}(s') \quad \text{(cesaro avg)}$$

Note: $d^\infty \to$ SMR if $s' \in \tau$ then

$$\frac{\sum_{k\in n} P^k_{s,s'}}{n} \longrightarrow 0 \text{ as } n\to\infty \text{ infact } P^k_{s,s'} \to 0$$

Note: $d^\infty \to$ SMR$_1$ if $s' \in C_c$, $\mu^{(c)}$ #class, $\forall c$

then $$\frac{\sum_{k\in n} P^k_{s,s'}}{n} \longrightarrow \begin{cases} 0 & ; s \in C_{c'} \text{ with } c' \neq c \\ \sigma_{s,c}\mu^{(c)}(s') & ; s \in \tau \\ \mu^{(c)}(s') & ; s \in C_c \end{cases}$$

now, we are interested in $g^\pi_+(s)$, $g^\pi_-(s)$

$$\frac{\mathbb{E}_s\left[\sum_{n=1}^N r(X_n)\right]}{N} = \frac{\sum_{n=1}^N \mathbb{E}_s[r(X_n)]}{N}$$

$$= \frac{\sum_{n=1}^N \sum_{s'} r(s') P(X_n = s' \mid s)}{N}$$

$$= \sum_{s'} r(s') \frac{\sum_{n=1}^N P^n_{s,s'}}{N} \longrightarrow \begin{cases} \sum_{s' \in C_c} r(s') \mu^c(s') & ; s \in C_c \\ \sum \sigma_{s,c} \mu_c r & ; s \in T \end{cases}$$

($\mu_c r$: μ_c is a distribution)

Note: $\mu_c f \triangleq \sum_s \mu_c(s) f(s) = \mathbb{E}[f(X^\infty)]$ where $X^\infty \sim \mu_c$

(stationary, limit always exist)

now, $$g^\pi(s) = \lim_{N\to\infty} \frac{\vartheta^\pi_N(s)}{N}$$

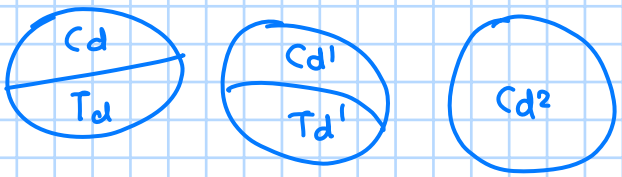
$$= \begin{cases} \mu_c r & ; s \in C_c \\ \sum_c \sigma_{s,c} \mu_c r & ; s \in T \end{cases}$$

Note: so if only c_1 and no T, then $g^\pi(s)$ does not depend on s

Theorm: $\mathcal{S}$ is finite states, $\forall$ SMR policy $d^\infty = \pi$, $g^\pi_+ = g^\pi_-$ and limit exist

proof: proof follows from above results directly

① unichain assumption, $\forall d^\infty$, $\exists$ single positive recurrent class



② multichain assumption, where we have multiple classes

Note: under ①, $g^\pi(s)$ will be constant $\forall s$, as $\sigma_{s,c} = 1 \ \forall s \in \mathcal{S}$ as only one c

now, for ①, $\forall s$, $g^\pi(s)$ is same, so we want to try writing DP equation

10th Oct:

No class on tuesday, tuesday 5:30 class
Next to next tuesday 5:30 Quiz-2

Quiz-2 : 1 or 2 questions, LP solving questions

we know $g^\pi(s) = \lim_{N\to\infty} \frac{v_N^\pi(s)}{N}$ (if limit exist) where $v_N^\pi(s) = \sum_{n\le N} E_s^\pi[r(X_n, Y_n)]$

also, $g_-^\pi(s) = \underline{\lim}_{N\to\infty} \frac{v_N^\pi}{N}$

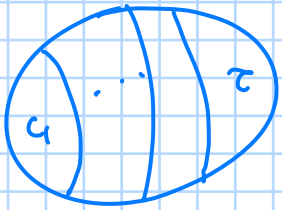
$g_+^\pi(s) = \overline{\lim}_{N\to\infty} \frac{v_N^\pi}{N}$

if $d^\infty \to$ SMR then $g^\pi(s)$ always exist (finite state space)

$\pi^* \leftarrow$ optimal if $g_-^{\pi^*}(s) \ge g_+^\pi(s)\ \forall \pi$ → criteria when g^π does not exist

Note: when limits don't exist, we can use above criteria

Finite state models: $\pi = d^\infty$



$\forall \pi = d^\infty$, we have $\{T, C_1, C_2 \dots\}$ classes

if $s \in C_c^{(\pi)}$, $g^\pi(s) = \mu_c r$ ← depends on π ← depends on π

if $s \in C_c^{(\pi)}$, $g^\pi(s) = \mu_c r = \sum_{s,a} \mu_c(s,a) r(s,a)$

we also saw: $d^\infty : P_d \longrightarrow \mu_{d,c}^{(x)}$ ← extra X is state, $Z = (X, Y)$

d^∞ ↕ SMR

$\mu_{d,c}^{(x)}$ ← stationary distribution

$Z_n = (X_n, Y_n)$ ← $Z = (X, Z)$ both state and action

prob dist of Z_n is $\mu_{d,c}^{(z)}(s,a) = \underbrace{\mu_{d,c}^{(x)}(s)}_{\text{Prob of state} = s}\ \underbrace{d(s,a)}_{\text{prob of selecting } a \text{ given } s}$

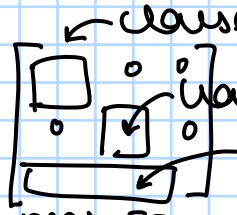
if $s \in T$: $g^\pi(s) = \sum_c \sigma_{s,c} \mu_c r$

in extreme state some π s.t whole space is irreducible then only on μ so, $\pi = d^\infty$

$g^\pi(s) = g\ \forall s$ where $g = \mu r$

Note: as $g^\pi(s)$ does not depend on s, we don't have normal DP equations

for $d^\infty \to X_n$ under d^∞: $P_d \sim$



P_d in block matrix form becomes $\begin{bmatrix} [\ ]_{n\times m} & [\ ]_{n\times n} \\ r\times m & r\times n \end{bmatrix}_{(n+r)\times(n+m)}$

if Pd has 2 cases, then $n_1 = |c_1|$, $n_2 = |c_2|$, $q \rightarrow$ transient states
so total $n_1 + n_2 + q$

$$P_d = \begin{bmatrix} P_{d_1} & 0 & 0 \\ n_1 \times n_1 & n_1 \times n_2 & n_1 \times q \\ 0 & P_{d_2} & 0 \\ n_2 \times n_1 & n_2 \times n_2 & n_2 \times q \\ * & * & * \\ q \times (n_1 + n_2 + q) & & \end{bmatrix} \quad (n_1 + n_2 + q) \times (n_1 + n_2 + q)$$

← this need not be zero

now, we can do jordan-canonical form

$$P x_i^\circ = \lambda_i^\circ x_i^\circ$$

$$W = \begin{bmatrix} \cdots x_1 \cdots \\ \cdots x_2 \cdots \\ \vdots \\ \cdots x_n \cdots \end{bmatrix} \qquad W^{-1} P W = \Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}$$

if some eigevalues are repeating, then

$$\begin{bmatrix} \begin{pmatrix} \lambda_1 & 1 \\ & \lambda_1 \end{pmatrix} & & & \\ & (\lambda_2) & & \\ & & (\lambda_2) & \\ & & & \begin{pmatrix} \lambda_3 & 1 & \\ & \lambda_3 & 1 \\ & & \lambda_3 \end{pmatrix} \end{bmatrix}_{n \times n}$$

The jordan canonical form of P_d:

m eigevetors of $\lambda = 1$, if $\{c_1, c_2 \ldots c_m, T\}$
cases, ∃ unique stationary distribution

Stationary dist of m cases:

$$\mu_1 = [\underbrace{\quad}_{n_1 \neq 0} \; 0 \cdots 0]$$

$$\mu_2 = [0 \cdots 0 \underbrace{\quad}_{n_1} \; \underbrace{\quad}_{n_2 \neq 0} \; 0 \cdots 0]$$

$$\vdots$$

$$\mu_m$$

corresponding to each μ_i we get: $\mu_i P_d = \mu_i \Rightarrow \mu_i$ is eigevetor of 1

now, $W = \begin{bmatrix} * \\ \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{bmatrix}$ ← any basis (ortogonal by gram-smsuitz) from other eigevetors; only one vetor per case

as stocastic matrix, other eigevalues < 1

$$P_d = W^{-1} \begin{bmatrix} Q & 0 \\ 0 & I \end{bmatrix} W$$

← jordan canonical form; $m \times m$ identity matrix

Note: Now we have $P_d = W^{-1}\begin{bmatrix} Q & 0 \\ 0 & I_{m\times m} \end{bmatrix} W$

(d^∞)

now, $v_\lambda^\pi = (I - \lambda P_d)^{-1} r$ (our discounted case)

but $(I - P_d)^{-1}$ does not exist as eigenvalues $= 0$ (m many)
we use pseudo inverse of $(I - P_d)$

in long run we get gain g if $m = 1$, then

$$g^\pi(s) = \mu_r \quad \forall s$$

Now we want to know how fast error i.e $\delta(X_t, Y_t)$ and $g(X_t)$ decreases

i.e $\sum_{t\in N} \mathbb{E}[r(X_t, Y_t) - g(X_t)]$ ← some form of bias

as, $\frac{\sum \mathbb{E}[r(X_n, Y_n)]}{N} \longrightarrow g(s) \longrightarrow \mathbb{E}[r(x^\infty, y^\infty)]$ ← LLN for markov chains

Pseudo inverse:

$$P = W^{-1}\begin{bmatrix} Q & 0 \\ 0 & I_{m\times m} \end{bmatrix} W \quad \text{now } P^2 = W^{-1}\begin{bmatrix} Q & 0 \\ 0 & I \end{bmatrix}^2 W$$

$$\Rightarrow P^2 = W^{-1}\begin{bmatrix} Q^2 & 0 \\ 0 & I \end{bmatrix} W$$

$$\vdots$$

$$P^k = W^{-1}\begin{bmatrix} Q^k & 0 \\ 0 & I \end{bmatrix} W$$

now, $I - P = W^{-1}W - W^{-1}\begin{bmatrix} \theta & 0 \\ 0 & I \end{bmatrix} W$

$$= W^{-1}\begin{bmatrix} I-\theta & 0 \\ 0 & 0 \end{bmatrix} W$$

Pseudo inverse → $(I-P)^{\#} = W^{-1}\begin{bmatrix} (I-\theta)^{-1} & 0 \\ 0 & 0 \end{bmatrix} W$ ← this is invtible as eigenvalues of $\theta < 1$

now, $P^* = \lim_{N\to\infty} P^N$ if aperiodic

and $P^* = \lim_{N\to\infty} P^N$ if periodic ← cesaro limit

$$= \lim_{N\to\infty} \frac{1}{N} \sum_{n\in N} P^N$$

now, $\frac{\sum_{K\in N} \theta^K}{N} = \frac{(I-\theta)^{-1}\theta^N}{N} \longrightarrow 0$ ← bounded

so, $P^* = W^{-1}\begin{bmatrix} 0 & 0 \\ 0 & I_{m\times m} \end{bmatrix} W$ in both cases

Eg: If 2 users n_1+n_2+q users: $n_1 \to \mu_1$ dist

$$P^* = \begin{bmatrix} \cdots \mu_1 \cdots \\ \cdots \mu_1 \cdots \\ \cdots \mu_1 \cdots \\ \cdots \mu_2 \cdots \\ \cdots \mu_2 \cdots \\ \cdots \end{bmatrix} \begin{matrix} \} n_1 \text{ times} \\ \\ \\ \} n_2 \text{ times} \\ \\ \} q \text{ times} \end{matrix} = \begin{bmatrix} (*)_{n_1 \times n_1} & 0 & 0 \\ 0 & (*)_{n_2 \times n_2} & 0 \\ \sigma_{s,1}\mu_1 + \sigma_{s,2}\mu_2 & & \end{bmatrix}_{(n_1+n_2+q) \times (n_1+n_2+q)}$$

From above example we get $P^* = \begin{bmatrix} \cdots \mu_1 \cdots \\ \cdots \mu_2 \cdots \\ \sigma_1\mu_1 + \sigma_2\mu_2 \end{bmatrix}$

$$\text{now, } (I-P+P^*) = W^{-1}\begin{bmatrix} I-\Theta & 0 \\ 0 & 0 \end{bmatrix} W \begin{bmatrix} \cdots \mu_1 \cdots \\ \cdots \mu_2 \cdots \\ \sigma_1\mu_1 + \sigma_2\mu_2 \end{bmatrix}$$

$$+ W^{-1}\begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} W$$

$$(I-P+P^*) = W^{-1}\begin{bmatrix} I-\Theta & 0 \\ 0 & I \end{bmatrix} W$$

$$(I-P+P^*)^{-1} = W^{-1}\begin{bmatrix} (I-\Theta)^{-1} & 0 \\ 0 & I \end{bmatrix} W$$

now we see $PP^* = P^*P = P^*P^* = P$ (from P^* def also we get same)

$$PP^* = \lim_{N\to\infty} \frac{\sum_{n\le N} P P^N}{N}$$

$$\Rightarrow PP^* = \lim_{N\to\infty} \frac{\sum_{n\le N+1} P^n - P}{N}$$

$$= P^*$$

now, $(I-P+P^*)^{-1} - P^* = (I-P)^{\#}$

as $(I-P+P^*)P^* = P^* - PP^* + P^*P^* = P^*$

$$\Rightarrow P^* = (I-P+P^*)^{-1}P^*$$

and so, $(I-P+P^*)^{-1} - P^* = (I-P+P^*)^{-1}(I-P^*) = (I-P)^{\#}$

Note: we get $(I-P)^{\#} = (I-P+P^*)^{-1}(I-P^*)$

now, $(I-P)^{\#} r = (I-(P-P^*))^{-1}(I-P^*)r$ where $g = P^* r$ ← vector of gain

$$= \sum_{K=0}^{\infty} (P-P^*)^K (I-P^*) r \quad \left(\because (I-A)^{-1} = \sum_{K=0}^{\infty} A^K\right)$$

$$(P-P^*)^2 = (P-P^*)(P-P^*)$$

$$= P^2 - P^* \quad \text{similarly } (P-P^*)^K = P^K - P^*$$

then, $(I-P)^{\#} r = \sum_{k=0}^{\infty} (P^k - P^*)(I-P^*)\, r$

$$= \sum_{k=0}^{\infty} (P^k - P^*)\, r$$

as for $k=0$: $(P-P^*)^0 (I-P^*)\, r$
$= (I-P^*)\, r$

also $P^* = P^k P^*$ so,

$(I-P)^{\#} r = \sum_k (P^k r - P^k g)$ as $g = P^* r$

$P^k r \to \mathbb{E}[r(X_n, Y_n)]$; $P^k g \to \mathbb{E}[g(X_n)]$

so, $(I-P)^{\#} r = \sum_{k=0}^{\infty} \mathbb{E}[r(X_n, Y_n) - g(X_n)]$

→ Bias from before

Def^n: (Bias) Bias will be $(I-P)^{\#} r$

Note: for more than one state, we'll use cesàro and not lim

16th Oct :

$$(I-P+P^*) = W^{-1}\begin{bmatrix} I-Q & 0 \\ 0 & I \end{bmatrix} W$$

$$(I-P)^{\#} \triangleq W^{-1}\begin{bmatrix} (I-Q)^{-1} & 0 \\ 0 & 0 \end{bmatrix} W$$

$$(I-P+P^*)^{-1} = W^{-1}\begin{bmatrix} (I-Q)^{-1} & 0 \\ 0 & I \end{bmatrix} W$$

$$P^* = W^{-1}\begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} W$$

$$(I-P+P^*)^{-1} - P^* = (I-P)^{\#}$$

Note : we had $(I-P+P^*)^{-1}P^* = P^*$

So, $\underbrace{(I-P+P^*)^{-1}(I-P^*)}_{H_P} = (I-P)^{\#}$ ← Pseudo inverse of $(I-P)$

then we get, $\sum (P-P^*)^n (I-P^*) = H_P$ (from previous result)

$$\Rightarrow H_P = \sum (P^n - P^*)(I-P^*) \quad (\because P^*P = PP^* = P^*P^* = P^*)$$

$$= \sum_{n=0}^{\infty} (P^n - P^*)$$

from λ case: $(I-\lambda P)^{-1} r = v$, here we have $(I-P)^{\#} r$ we want to find what comes when $(I-P)^{\#} r$ is taken

now, $\lim_{N\to\infty} \sum_{n=0}^{N} (P^n - P^*) r = H_P r$

now, $\sum_{n=0}^{N} P^n r = \sum_{n=0}^{N} \mathbb{E}[r(x_n, y_n)]$

$$= v_N^{\pi}$$

So, $H_P r = \lim_{N\to\infty} (V_N^{\pi} - N g^*)$

let $h(s) = \lim_{N\to\infty} (V_N^{\pi}(s) - N g^*(s)) = v_N^{\pi}(s) - N g^*(s) + \sum_{n>N} (P^n - P^*) r$

So, $h(s) \approx v_N^{\pi}(s) - N g + O(\,)$

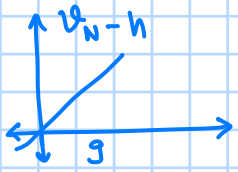
↑ If we start in stationarity

So, pseudo inverse gives us what happens when we start - start in stationary

$h(s) - h(j) = v_N^{\pi}(s) - v_N^{\pi}(j)$ ← Bias of starting states

$s, j \in$ same C_c

really same but fastness is in this argument



periodic:

$$H^{p}r \triangleq \lim_{N\to\infty} \underbrace{\frac{1}{N}\sum_{n\in N}\left(\sum_{n=0}^{N} P^{n} - P^{*}\right) r}_{\text{Cesaro sum or avg sum}}$$

$$\frac{1}{N}\sum_{n\in N} v_n(s) - N g(s) \quad \forall s,\ h(s) = \frac{1}{N}\sum v_n(s) - N g(s) + o(N)$$

$$\forall s, i,\ h(s) - h(j) = \frac{1}{N}\sum (v_n(s) - v_n(j))$$

Theorem: The gain $g(s) = \lim_{N\to\infty} \frac{1}{N} v_N(s)$ and $\underline{h} = H^{p} r$ and $\underline{g} = (g(s), \forall s)$ then:

(I) (i) $(I - P)g = 0$ and

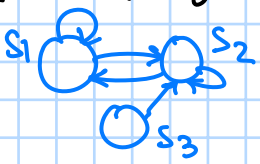
(ii) $g + (I - P)h = r$

(II) if $\underline{g}_1, \underline{h}_1$ satisfy (i) and (ii) then they are gain and bias vectors

we assume chain is aperiodic

$$\begin{bmatrix} p & 1-p & 0 \\ 1-p & p & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad \{s_1, s_2, s_3\} = S$$

top one is recurrent, bottom is transient

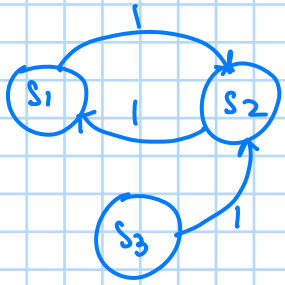


if $p = 0$: then not aperiodic

$p > 0$: then aperiodic and $P^{*} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \end{bmatrix}$ ← stationary distribution

$$P^{N} \approx P^{*} + o(N)$$

now if $p = 0$:



we will have $P^{*} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \end{bmatrix}$ ← Cesaro sum

but $P^{N} = \begin{cases} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}; & N \text{ even} \\ \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}; & N \text{ odd} \end{cases}$ ← periodic markov chain

$$C\lim_{N\to\infty} P^{N} = P^{*}$$

average and not individual case

so in both cases, $P^{N} \approx \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \end{bmatrix} + o(N)$

Theorem: The gain $g(s) = \lim_{N\to\infty} \frac{1}{N} v_N(s)$ and $\underline{h} = H^{p} r$ and $\underline{g} = (g(s), \forall s)$ then:

(I) (i) $(I - P)g = 0$ and

(ii) $g + (I - P)h = r$

(II) If g_1, h_1 satisfy (i) and (ii) then they are gain and bias vectors

proof:

(I) $g = P^* r$

$$(I-P)g = (I-P)P^* r = (P^* - PP^*)r = 0$$

now, $H_P r = (I - P + P^*)^{-1}(I - P^*) = \sum_{n=0}^{\infty}(P^n - P^*)$

$$(I-P)H_P = H_P(I-P) \quad (\because PP^* = P^*P = P^*P^* = P^*)$$

$$= \sum(P^n - P^*)(I-P)$$

$$= \sum P^n - P^* - P^{n+1} + P^*$$

$$= \sum P^n - P^{n+1}$$

$$= \lim_{N\to\infty} \sum_{n=0}^{N} P^n - \sum_{n=0}^{N} P^{n+1}$$

$$= \lim_{N\to\infty} I - P^{N+1} \quad (\because \text{aperodic})$$

← for periodic case we take $\lim_{N\to\infty}$ to get $\lim_{N\to\infty} P^N = P^*$

$$= I - P^*$$

so, $\lim_{N\to\infty}\left(\frac{1}{N}\sum_{n\in N}(I-P^{N+1})\right) = I - \lim_{N\to\infty} P^{N+1}$ ← periodic case

$$= I - P^*$$

$$(I-P)H_P = I - P^* \quad \text{——} (1)$$

so, $(I-P)H_P r = Ir - P^* r$

$$= r - g$$

$$\Rightarrow g + (I-P)H_P r = r$$

(II) g_1, h_1 that satisfy (i), (ii) given

$$(I-P)g_1 = 0$$

and $g_1 + (I-P)h_1 = r$

$$P^* g_1 + P^*(I-P)h_1 = P^* r = g$$

so, $(I-P)g_1 + g = g$

$$\Rightarrow (I-P)g_1 + P^* g_1 = g \ (\because \text{from (I)})$$

$$\Rightarrow (I - P + P^*)g_1 = g$$

$$\Rightarrow (I - P + P^*)g_1 = P^* r$$

$$\Rightarrow (I - P + P^*)g_1 = (I - P + P^*)P^* r$$

$$(\because (I - P + P^*)P^* = P^*)$$

$\Rightarrow g_1 = P^* r = g$

$\Rightarrow g_1 = g$

$g_1 + (I-P)h_1 = r$

$\Rightarrow P^* r + (I-P)h_1 = r$

$\Rightarrow (I-P)h_1 = (I-P^*)r$

$\Rightarrow (I-P)h_1 = (I-P)H_p r$ as $(I-P)H_p = I-P^*$ from ①

$\Rightarrow (I-P)(h_1-h) = 0$

$\Rightarrow h_1 - h \in N(I-P)$

$\Rightarrow h_1 = h + v$ for some $v \in N(I-P)$

Note: Numerically calculating h, we will get different h's but $h - \tilde{h} \in N(I-P)$

Now, $(I-P)g = 0$ and $g + (I-P)h = r$ and for λ case we had

$(I-\lambda P)v = r$

$v = \max\{r + \lambda P v\} \rightarrow$ DP equation

↖ gain + bias

were $g + h = r + Ph$

lexi $\rightarrow A_1 \rightarrow \max_x f(x)$

pick set A_1 and then do $\max_{x \in A_1} f_2(x)$

17th oct:

Theorem: if g and h are gain and bias then

(i) $(I-P)g=0$

(ii) $g+(I-P)h=r$

and if (g,h) satisfy (i) and (ii) then g = gain, h = bias

$\lambda \in [0,1]$ if $\lambda \to 1$ then we will have average cost version

$$\mathbb{E}[v] = \frac{1}{T}\sum_{t=1}^{T} \lambda^{t-1} r(s,a)$$

if $\lambda \to 1$ and $T \to \infty$ we want to find $\mathbb{E}[v]$

now, $v_\lambda = \underbrace{(I-\lambda P)^{-1}}_{\text{for } \lambda=1 \text{ not invertible, so we will take pseudoinverse}} r$

for $\lambda \in (0,1)$

$$\lambda = \frac{1}{1+\rho} \qquad \rho > 0$$

$$= (1+\rho)^{-1}$$

then, $(I-\lambda P)^{-1} = (I-(1+\rho)^{-1}P)^{-1} = (1+\rho)\underbrace{(I-P+\rho I)^{-1}}_{\text{Resolvent of } (1+\rho)}$

$$R_\rho = (I-P+\rho I)^{-1}$$

$$I-P = W^{-1}\begin{bmatrix} I-Q & 0 \\ 0 & 0 \end{bmatrix} W$$

$$I-P+\rho I = W^{-1}\begin{bmatrix} I-Q & 0 \\ 0 & 0 \end{bmatrix} W + W^{-1}\rho I W$$

$$= W^{-1}\begin{bmatrix} I-Q+\rho I & 0 \\ 0 & \rho I \end{bmatrix} W$$

let $I-Q = B$

$$\Rightarrow R_\rho^{-1} = (I-P+\rho I) = W^{-1}\begin{bmatrix} B+\rho I & 0 \\ 0 & \rho I \end{bmatrix} W$$

now

$$R_\rho = W^{-1}\begin{bmatrix} (B+\rho I)^{-1} & 0 \\ 0 & \rho^{-1} I \end{bmatrix} W$$

$$= W^{-1}\begin{bmatrix} (B+\rho I)^{-1} & 0 \\ 0 & 0 \end{bmatrix} W + \frac{1}{\rho} W^{-1}\begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} W$$

$$= W^{-1}\begin{bmatrix} (B+\rho I)^{-1} & 0 \\ 0 & 0 \end{bmatrix} W + \frac{1}{\rho} P^*$$

$$(B+\rho I)^{-1} = (I+\rho B^{-1})^{-1}B^{-1}$$
$$= \sum_{n=0}^{\infty} (-1)^n \rho^n B^{-1-n}$$

$$\text{so, } R_\rho = W^{-1}\begin{bmatrix} \sum_{n=0}^{\infty}(-1)^n\rho^n B^{n-1} & 0 \\ 0 & 0 \end{bmatrix} W + \frac{P^*}{\rho}$$

$$\Rightarrow V_\lambda = (I-\lambda P)^{-1}r = (1+\rho)R_\rho r = (1+\rho)\rho^{-1}P^* r$$
$$+ (1+\rho)W^{-1}\begin{bmatrix} \sum(-1)^n\rho^n(I-Q)^{-n-1} & 0 \\ 0 & 0 \end{bmatrix} Wr$$

$$H_\rho = W^{-1}\begin{bmatrix} (I-Q)^{-1} & 0 \\ 0 & 0 \end{bmatrix} W$$

$$\Rightarrow V_\lambda = (1+\rho)\rho^{-1}P^* r + (1+\rho)\sum_{n=0}^{\infty}(-1)^n\rho^n H_\rho^{n+1} r$$

$\lambda = (1+\rho)^{-1}$
$\lambda \to 1$ as $\rho \to 0$

$$\text{so, } V_\lambda = \left(1+\frac{1}{\rho}\right)P^* r + \underbrace{\sum(-1)^n\rho^n H_\rho^{n+1} r} + \sum(-1)^n\rho^{n+1}H_\rho^{n+1} r$$
$$= H_\rho r + \sum_{n=1}^{\infty}(-1)^n\rho^n H_\rho^{n+1} r$$

$$\Rightarrow V_\lambda = \left(1+\frac{1}{\rho}\right)P^* r + h + \rho\Big[\underbrace{\sum_{n=1}^{\infty}(-1)^n\rho^{n-1}H_\rho^{n+1} r + \sum_{n=0}^{\infty}(-1)^n\rho^n H_\rho^{n+1} r}_{\text{this is bounded}}\Big]$$

$f(\lambda)$ term
$f(\lambda) \to 0$ as $\lambda \to 1$

$$\Rightarrow V_\lambda = (1-\lambda)^{-1}g + h + f(\lambda)$$

as $\lambda \to 1$, $h \to 0$, $f(\lambda) \to 0$

$$\text{so, } g = \lim_{\lambda \to 1}(1-\lambda)V_\lambda$$

↖ this is just an heuristic to understand when $\lambda \to 1$ we get $g = V_\lambda$

Stopping time:

If $\tau \sim$ geom r.v

$$\mathbb{E}[r(X_\tau)] = \sum_n P(\tau = n)\,\mathbb{E}[r(X_n)]$$
$$= \sum_{n=0}^{\infty}(1-\lambda)\lambda^{n-1}\,\mathbb{E}[r(X_n)]$$

succeeds ↑ failing with prob λ

$$= (1-\lambda)\sum_{n=0}^{\infty}\lambda^n\,\mathbb{E}[r(X_n)]$$

$$\Rightarrow \mathbb{E}[r(X_t)] = (1-\lambda)v_\lambda$$

as $\lambda \to 1$, $t \to \infty$ almost surely

and so,

$$g = \lim_{\lambda \to 1} (1-\lambda) v_\lambda$$

one more way to look what happens when $\lambda \to 1$

Optimality equation:

$$\max\{r_d + (\lambda P_d - I) v_\lambda^*\} = 0$$

$$\text{or } v_\lambda^* = \max\{r_d + \lambda P_d v_\lambda^*\}$$

$$\text{now } v_\lambda = (1-\lambda)^{-1} g + h + f(\lambda)$$

$$\Rightarrow \max\{r_d + (\lambda P_d - I)[(1-\lambda)^{-1} g + h + f(\lambda)]\} = 0$$

if $\lambda \to 1$ then

Note: if we assume unichain, g becomes $g.e$, $g \in \mathbb{R}$

$$r_d + (\lambda P_d - I)[(1-\lambda)^{-1} g + h + f(\lambda)]$$

$g \in \mathbb{R}$ ↘ ← all one vector

$$= r_d + (\lambda P_d - I)(1-\lambda)^{-1} g.e + (\lambda P_d - I)h$$

$$= r_d + \lambda P_d . e\, g\,(1-\lambda)^{-1} - (1-\lambda)^{-1} g e + (\lambda P_d - I)h$$

$$= r_d + \lambda g.e\,(1-\lambda)^{-1} - (1-\lambda)^{-1} g e + (\lambda P_d - I)h$$

$$= r_d - g.e\,(\cancel{1-\lambda})\cancel{(1-\lambda)} + (\lambda P_d - I)h$$

$$= r_d - g.e + (\lambda P_d - I)h$$

$$0 = r_d - g.e + (P_d - I)h$$

So, optimality equation becomes:

$$\sup_d \{r_d - g.e + (P_d - I)h\} = 0$$

Theorem: If $\max_d \{r_d - g.e + (P_d - I)h\} \leq 0$ for some (g, h) then $g.e \geq g^*$

Now from above theorm, we can see if we take min of all such g's we will get g^*

i.e $\min g$

$$g + h(s) - \sum_j P(j|s,a) h(j) \geq r(s,a) \quad \forall s, a$$

} LP for average cost mdp

then dual becomes

$$\max_{x(s,a)} \sum_{s,a} r(s,a)\, x(s,a)$$

s.t

$$\sum_a x(j,a) - \sum_{s,a} P(j|s,a)\, x(s,a) = 0$$

$$\sum_{s,a} x(s,a) = 1$$

and $x(s,a) \geq 0 \quad \forall s,a$

} Dual of MDP

← Solving this gets us the optimal policy d^∞

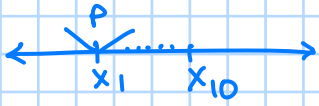
POMDP:

POMDP are Partially Observable MDPs, a particular case of POMDP is IOMDP or Intermittently Observable MDP

IOMDP:

t X_t is observable with probability ρ
is lost with $1-\rho$

$S = \{s_1 \dots s_n\}$



b_t is a vector

$$b_t = P(S_t | s_1 \dots s_{t-1}, a_1 \dots a_{t-1}, a_t)$$

only ones which are observed

$$= P(S_t | S_{t_k}, a_{t_k}, a_{t_{k+1}}, \dots a_t)$$

S_{t_k} ← last observed state

$a_{t_k}, a_{t_{k+1}}, \dots a_t$ ← All actions taken after last observed state as we have observed new state

then b_0 —ρ→ $e_i = (0,0 \dots 1, 0 \dots 0)$

b_0 —$1-\rho$→ $P_{a_1} b_0$

$P_{a_1} b_0$ —ρ→ e_i

$P_{a_1} b_0$ —$1-\rho$→ $P_{a_2} P_{a_1} b_0$

$P_{a_1} b_0$ ← we have not observed state

transition probabilities

$$Q(b_{t+1} | b_t, a_t) = \begin{cases} \rho\, [P_{a_t} b_t]_i & ; b_{t+1} = e_i \; ; i \in S \\ (1-\rho) & ; b_{t+1} = P_{a_t} b_t \\ 0 & ; \text{o.w} \end{cases}$$

← we have observed S_{t+1}

← we have not observed b_{t+1}

Note: b is a vector s.t $b = (b(s), \forall s)$

now, for IOMDP, the states are b_t's

$\forall b, R(b,a) = \sum_s b(s) r(s,a)$ ← expected value

↑ fution of s as b is a vector

Average reward IOMDP:

we want $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[R(b_t, a_t)]$ as $t \to \infty$

then we have an avg reward MDP with new state space which is NOT countable

Dual: $\max_{x(b,a)} x(b,a) R(b,a)$

s.t $\sum_a x(b',a) - \sum_{b,a} Q(b'|b,a)\, x(b,a) = 0 \quad \forall b'$

$\sum_{b,a} x(b,a) = 1$ and $x(b,a) \geq 0 \ \forall b,a$

as $b \in [0,1]$ then uncountable many states and we cannot solve MDP

but: ① if underlying MDP has finite states and action set

② underlying MDP is recurrent

under ①, ② we have strong duality and we can solve MDP

→ this means $b \in$ countable infinite space

21st oct:

we know $(I-P)g=0$, $r-g+(P-I)h=0$

we have $\pi = d^\infty$

$h = h_d$, $g = g_d$ } depends on d

If g and h satisfy the equations above and $P_d^* h = 0$ then $h = H_{P_d} r$ → Bias

Optimality:

$$g_+^\pi = \limsup_{N\to\infty} \frac{1}{N} v_N^\pi \qquad \pi \in \Pi^{HR}$$

↑ any history dependent random policy

$$g_-^\pi = \liminf_{N\to\infty} \frac{1}{N} v_N^\pi$$

Def: (optimality) π^* is optimal if $g_-^{\pi^*} \geq g_+^\pi \ \forall \pi$

Note: $g_-^\pi \leq g_+^\pi \ \forall \pi$ and equality if optimal

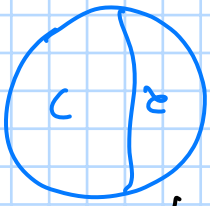
Def: $g_+^* = \sup_{\pi\in\Pi^{HR}} g_+^\pi$

$g_-^* = \sup_{\pi\in\Pi^{HR}} g_-^\pi$

Unichains:

$\forall d^\infty \in MR$ policy, the chain has almost one + the recurrent class

If $|S| < \infty$



$g^{d^\infty} = g \cdot e$ ($g \in \mathbb{R}$, ↑ all one vector)

also we don't care about periodic/aperiodic as we take $\lim$

let $g_d = g_d e$ (g_d ← scalar) $\quad e = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \ \forall d^\infty \in MR$

$(I-P_d)g_d = 0$

$\Rightarrow g_d (I - P_d)(e) = 0$

$e - e = 0$

this is meaningless

$r_d - g_d e + (P_d - I) h_d = 0$ ($\because$ second equation)

$v_\lambda^{d^\infty} = (1-\lambda)^{-1} g_d e + h + f_d(\lambda)$ ($\because$ from previous result)

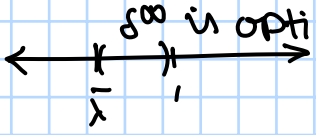
↑ converge differently for different d^∞

$\exists \bar{\lambda} < 1$ s.t. $v_\lambda^* = v^{\delta^\infty}, \forall \lambda \geq \bar{\lambda}$

so, optimal policy $= \delta^\infty, \forall \lambda \geq \bar{\lambda}$

we also know, $\forall d^\infty$, $\lim_{\lambda \to 1} v_\lambda^{d^\infty}(1-\lambda) = g^{d^\infty}$

the optimal policy δ^∞ remains same for every discount rate $\forall \lambda$ sufficiently big



↑ avg cost optimal

we did max to find

$$v_\lambda^* = \max_d \{ r_d + \lambda P_d v_\lambda^* \}$$

let $B(g,h) \triangleq \max_d \{ r_d - g\underline{e} + (P_d - I)h \}$

↑ scalar, ↖ vector

then our optimality equations says if $B(g,h)=0$ then g is optimal gain

Theorem: (i) If $\exists g$ scalar and vector h s.t $B(g,h) \leq 0$ then $g\underline{e} \geq g_+^*$

(ii) If $\exists g$ scalar and vector h s.t $B(g,h) \geq 0$, then $g\underline{e} \leq g_-^*$

(iii) If $\exists g, h$ s.t $B(g,h)=0$ then $g\underline{e} = g^* = g_-^* = g_+^*$
& $\exists d^\infty$ s.t $g = g_d = g^*$

proof: (i) given a vector and scalar (h,g) s.t $B(g,h) \leq 0$

let $\pi = (d_1, d_2, d_3, \ldots) \in \pi^{MR}$

$\forall d$, $r_d - g\underline{e} + (P_d - I)h \leq 0 \quad \forall d \in \pi^{MR}$

$\Rightarrow g\underline{e} \geq r_{d_t} + (P_{d_t} - I)h \quad \forall t = 1,2,\ldots$ (for π) — ①

we want to show $g\underline{e} \geq g_+^\pi \quad \forall \pi \in \pi^{MR}$ (it implies for all HR as we can map HR and MR)

as $\forall \pi \in \pi^{HR}$, $\exists \tilde{\pi} \in \pi^{MR}$ s.t $g_+^\pi = g_+^{\tilde{\pi}}$

$g_-^\pi = g_-^{\tilde{\pi}}$

now, $g\underline{e} = P_{d_1} g\underline{e} \geq P_{d_1}(r_{d_2} + (P_{d_2} - I)h)$ from ① as P_{d_1} is positive

similarly $P_{d_1} P_{d_2} g\underline{e} \geq P_{d_1} P_{d_2}(r_{d_3} + (P_{d_3} - I)h)$

$\forall t \quad P_{d_1} P_{d_2} \ldots P_{d_t} g\underline{e} \geq P_{d_1} \ldots P_{d_t}(r_{d_{t+1}} + (P_{d_{t+1}} - I)h$

summing this $\forall t \leq N$ ↙ telescopic

$$\Rightarrow N g\underline{e} \geq \sum_t P^{\pi,t} r_{d_{t+1}} + \sum_t (P^{\pi,t+1} - P^{\pi,t})h$$

now $P^{\pi,t} = P_{d_1} P_{d_2} \ldots P_{d_t}$

$$\Rightarrow N g\underline{e} \geq \sum_t P^{\pi,t} r_{d_{t+1}} + (P^{\pi,N+1} - I)h$$

$\Rightarrow g e \geq \frac{1}{N} \sum_t P^{\pi,t} r_{t+1} + \frac{1}{N}(P^{\pi,N+1} - I)h$ (↖ Bounded)

$\xrightarrow[\text{as } N \to \infty]{} 0$

$\Rightarrow g e \geq g_+^\pi \geq g_-^\pi \quad \forall \pi \in \Pi^{MR}$

$\Rightarrow g e \geq g_+^\pi \quad \forall \pi \in \Pi^{MR}$ and so $\forall \pi \in \Pi^{HR}$

(ii) given a vector and scalar (h, g), s.t $B(g,h) \geq 0$

$\Rightarrow B(g,h) = \max_a (\quad) = r_{d^*} - ge + (P_{d^*} - I)h \geq 0$

$\Rightarrow ge \leq r_{d^*} + (P_{d^*} - I)h$

then $P_{d^*} ge \leq P_{d^*}(r_{d^*} + (P_{d^*} - I))h$

$\vdots$

$\Rightarrow ge \leq g_-^{d^\infty *} e$ (from previous part)

$\Rightarrow ge \leq g_-^* e$ ($\because$ By defn of g^*)

given (i), (ii), part 3 is trivial and follows

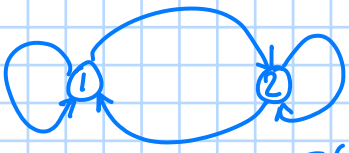
$g_+^* \leq ge \leq g^{d^\infty *} \leq g_-^* \leq g_+^*$

$\Rightarrow ge = g_+^* = g_-^* = g^{d^\infty *}$

Eg: $A_1 = \{a_{11}, a_{12}\}$ $A_2 = \{a_{21}, a_{22}\}$ $r(s,a)$

$\forall P(1|1,a) > 0 \quad \forall a$

$\forall P(1|2,a) > 0 \quad \forall a$



condition for no-self loop or unichain

$\inf_{a,s'} P(s'|s,a) > 0 \quad \forall s$

Queuing model:

λ (customer arrive)

queue

server

μ = capacity, let $\lambda \sim$ Poiss Arrival process (λ)

service times are exp(μ)

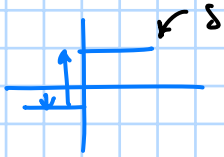
ghost server means too much congestion

$c(q)$ = congestion cost (↖ larger)

C_s = additional server cost, policy: $\exists \theta$ s.t $\forall s \leq \theta \quad a^* = 0$ else $a^* = 1$

Inventory control:

(n, Cs $c_0 = k + \tilde{c}_0(q)$ if $k > 0$ ss policy:



↖ ss policy is to refill

24th oct:

we have $B(g,h)=0$, $g \to$ scalar, $h \to$ vector

$$B(g,h) = \max_d \{r_d + (P_d - I)h - ge\}$$

we have shown that if $B(g,h) \leq 0 \Rightarrow g \geq g^*$
$B(g,h) \geq 0 \Rightarrow g \leq g^{+*}$
$B(g,h) = 0 \Rightarrow g = g^*$

Existance:

THEOREM: If $|S| < \infty$, then assuming unichain $\exists$ scalar g, vector h s.t $B(g,h)=0$

we have already seen, $\exists \bar{\lambda} < 1$ s.t $\forall \lambda \geq \bar{\lambda}$, the unique optimal policy $d^*_\lambda = \delta^\infty \ \forall \lambda \geq \bar{\lambda}$ for some

THEOREM: If $|S| < \infty$, then assuming unichain $\exists$ scalar g, vector h s.t $B(g,h)=0$

proof: consider $\lambda_n \uparrow 1$ let $v_{\lambda_n} \to$ value vector $\forall \lambda_n$

$$v_{\lambda_n} = \max_d \{r_d + \lambda_n P_d v_{\lambda_n}\}$$

$$\Rightarrow 0 = \max_d \{r_d + (\lambda_n P_d - I) v_{\lambda_n}\}$$

fixing a d^∞

$v_{\lambda_n} = (1-\lambda_n)^{-1} g + h + f(\lambda)$ where $v_{\lambda_n} = v_{\lambda_n}^{\delta^\infty}$
↑ comes from δ^∞, so does h

$$0 \geq r_d + (\lambda_n P_d - I)(1-\lambda_n)^{-1} g$$

$ge = g^{\delta^\infty}$, $h = H_p^{\delta^\infty} r_\delta$

$$0 \geq r_d + (\lambda_n P_d - I)((1-\lambda_n)^{-1} ge + h + f(\lambda_n))$$

$$\Rightarrow 0 \geq r_d + (P_d - I)h + (\lambda_n - 1)P_d h - ge + f(\lambda_n)(.)$$

as λ_n inc, h, g does not change as δ^∞ policy

as $\lambda_n \uparrow 1$

$$0 \geq r_d + (P_d - I)h - ge \ \forall a$$

for $d = \delta$ true $\geq$ covers into equality as it's the maximiser

so, $0 = r_\delta + (P_\delta - I)h^{\delta^\infty} - g^{\delta^\infty} e \Leftrightarrow B(g^{\delta^\infty}, h^{\delta^\infty}) = 0$

Two queues Age of Information:

we have 2 sources, when they update age = 0 if updated
so, we are assuming source has collision

age → age+1 if not updated

State space: $X = (x_1, x_2)$ age of source 1, age of source 2

now, if $x_{1_0} \neq x_{2_0}$ (starting) then

$$\mathcal{S} = \{ x = (x_1, x_2) \mid x_1 \neq x_2 \}$$

↑ as if we do not start at equal we can never get to it

$$= \mathcal{S}_1 \cup \mathcal{S}_2$$

$$\mathcal{S}_1 = \{ (x_1, x_2) \mid x_1 < x_2 \}$$

$$\mathcal{S}_2 = \{ (x_1, x_2) \mid x_2 < x_1 \}$$

we need a smaller state space if possible
(i) It is still markov
(ii) No loss of optimal

we can make $\mathcal{S}_1 = \{ (x_1, x_2) \mid x_1 < x_2 \text{ s.t } x_2 \leq L \}$
$x_1 \leq L-1$

$\mathcal{S}_2 = \{ (x_1, x_2) \mid x_2 < x_1 \text{ s.t } x_2 \leq L-1 \}$
$x_1 \leq L$

if we see SMD policy s.t
in $\mathcal{S}_1 \to a_1 = 1$, $\mathcal{S}_2 \to a_2 = 2$

then this policy is not-unichain
as 2 diff irreducible classes

Algorithms for unichain models :

v^0 $d_1 \leftarrow \arg\max_d (r_d + \lambda P_d v^0)$

$v^1 \leftarrow d_1$

$\vdots$

$d_n \leftarrow \arg\max_d (r_d + \lambda P_d v^{n-1})$

$v^n \leftarrow d_n$

above is for λ or discounted cost MDP, as $\lambda < 1$ we get contraction

$L_d v = r_d + \lambda P_d v$ $\forall d$ linear for every d

$\mathcal{L} v = \max_d L_d v$ non-linear

↓ Both are contractions

$$\| \mathcal{L} v - \mathcal{L} v' \| \leq \lambda \| v - v' \|$$

$v^{n+1} = \mathcal{L} v^n \to v^*$

↑ unique fixed point

Avg cost :

$0 = \max_d \{ r_d + (\lambda P_d - I) v_\lambda \}$

$0 = \max_d \{ r_d + (P_d - I) h - g \underline{e} \}$ ↖ do not want this

↑ we want v here

Defn: (span norm) Not a norm, but

$$sp(v) = \max_i v_i - \min_i v_i$$

If we take $sp(v)$, $sp(v + qe)$ ← this fixes our issue with qe vector

we have $sp(v) = sp(v + qe)$

now, $|f(v) - f(v')| \leq \lambda |v - v'|$ is what we want

$sp\,Tv \leq \alpha\, sp(v)$ for $\alpha < 1$

↑ linear operator

then we want to see if we will get solution

$$sp(T^n v) \leq \alpha^n sp(v) \to 0$$

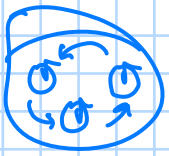
now, if we know d^*, then

$$T_{d^*}(h) = g_{d^*} + P_{d^*} h - ge$$

$$T_{d^*}(h - h') = g_{d^*} + P_{d^*}(h - h') - g_{d^*}$$

$$= P_{d^*}(h - h')$$

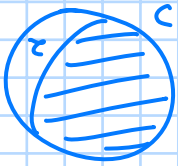
28th Oct:



unicain if periodic $\lim P^N \to P^*$

i.e $P^{dn+j} \to P_j^*$ as $n \to \infty$, d is the period

Multichain models:



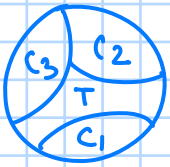
The unicain has only one recurrent class and so

$g_d = g_d e$ ← constant/same

g_d ↑ gain vector

$\frac{\sum_{n \le N} (P_d^\infty)^n_{ss}}{N} \to g$ as $N \to \infty$, given $|S'| < \infty$ or finite state space

Note: If $|S'| = \infty$, then we can also have cases more all is T



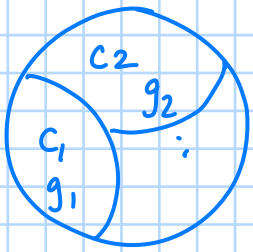
The markov chain will depend on d

Exe: $\exists$ unique recurrent class $\forall d^\infty$, $d \in \Pi^{SMD}$ → done down
iff $\exists$ unique recurrent class $\forall d^\infty$, $d \in \Pi^{SMR}$

Note: for unicain from above we can only take $\pi \in \Pi^{SMD}$

Defn: (Multichain) If $\exists d \in MD$, with more than one recurrent class, then it is multichain

Properties of g_d:



$\begin{pmatrix} g_1 \\ \vdots \\ g_1 \\ g_2 \\ \vdots \\ g_2 \\ g_3 \\ \vdots \\ g_3 \end{pmatrix} = g$, same for every s in same recurrent class

Policy evaluation Eqn:

given d^∞,

(i) $(P_d - I)g = 0$

(ii) $r_d + P_d h - h - g = 0$

} this works for both unicain and multichain

we have (gain, bias) vectors satisfy above, and if

some (g', h') satisfy above equations, then
$g' = g^*$ and $h' = h^* + \text{Null}(I - P_d)$

Note: if $P^* h = 0$ then $h' = h$, bias vector is h

Optimality equations:

we need $\max_{d \in MD} (P_d - I) g = 0$

and $\max_{d \in MD} (r_d + P_d h - h - g) = 0$ } modified optimality equation

if (h, g) satisfy above, but we will have an issue as many diff d's satisfy diff equation

so, $\max_d (P_d - I) g = 0$

$\max_{d \in E} (r_d + P_d h - h - g) = 0$ where } multichain optimality equation

$E = \{d : d(s) = a, \text{ where } (P_d - I) g = 0\}$

so, g should be s.t $(P_d - I) g \leq 0 \quad \forall d \in MD$

$$\sum_{j \in \{\}} P(j|s,a) g(j) - g(s) \leq 0 \quad \forall s, a \quad \text{——— ①}$$

and $\exists$ atleast one a' for fixed s s.t ① is satisfied with equality

$$\sum_{j \in \{\}} P(j|s,a') g(j) - g(s) = 0$$

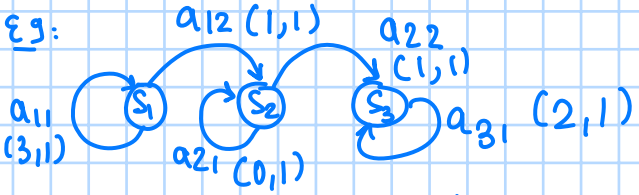
$\forall s$, let $B_s = \{a' \in A_s \mid \sum_{j \in \{\}} P(j|s,a') - g(s) = 0\}$ — ②

$$\max_{a' \in B_s} \left(r(s,a') + \sum_j P(j|s,a') h(j) - h(s) - g(s) \right) = 0, \quad \forall s \quad \text{——— ③}$$

so, $E = \{d \mid d(s) \in B_s \ \forall s\}$

Theorm: If (g^*, h^*) vectors satisfy multichain OE and are bounded, then $\exists M < \infty$ s.t $\exists (g^*, h^* + M g^*)$ satisfing modified multichain OE

Eg:



$P(s_1|s_1, a_{11}) = 1$ $\quad r(s_1, a_{11}) = 3$

$P(s_2|s_1, a_{12}) = 1$ $\quad r(s_1, a_{12}) = 1$

$P(s_2|s_2, a_{21}) = 1$ $\quad r(s_2, a_{21}) = 0$

$P(s_3|s_2, a_{22}) = 1$ $\quad r(s_2, a_{22}) = 1$

$P(s_3|s_3, a_{31}) = 1$

$r(s_3, a_{31}) = 2$

from unichain optimality equation, g is a scalar and

$$\max_d (r_d + P_d h - h - g\mathbb{1}) = 0$$

i.e $g + h_1 = \max\{3 + h_1, 1 + h_2\}$

$h_i = h(s_i)$

$g + h_2 = \max\{0 + h_2, 1 + h_3\}$

$g + h_3 = \max\{2 + h_3\} = 2 + h_3$

if we assume unichain

$g = 2$ from last

$2 + h_2 = \max\{h_2, 1 + h_3\}$

$2 + h_1 = \max\{3 + h_1, 1 + h_2\}$

$\not\geq 3 + h_1$

$\Rightarrow$ this is a contradiction

optimal policy: $g^*(s_1) = 3$ if on s_1 we stay on s_1

$g^*(s_2) = g^*(s_2) = 2$ if on s_2 we go to s_3 and stay

same for s_3

Theorem: If $\exists$ a pair of vectors (g, h) that simultaneously satisfy

$$\max_d (P_d - I)\, g \leq 0$$

and $\max_d (r_d + P_d h - h - g) \leq 0$

then (i) $g \geq g^* = \sup_{\pi \in \Pi^{HR}} \limsup_{N \to \infty} \frac{v_{N+1}^{\pi}}{N+1}$

if $\exists (g, h)$ for some d' s.t $(P_{d'} - I)g \geq 0$ and $r_{d'} + P_{d'} h - h - g \geq 0$

then (ii) $g \leq \underline{g}^{(d')^\infty} \leq \underline{g}^* = \sup_{\pi \in \Pi^{HR}} \liminf_{N \to \infty} \frac{v_{N+1}^{\pi}}{N+1}$

if $\exists (g, h)$ satisfying $\max_d (P_d - I)g = 0$ and $\max_d (r_d + P_d h - h - g) = 0$

then (iii) $g = g^*$

with below theorm, we have multichain and modified OE

Theorm: If (g^*, h^*) vectors satisfy multichain OE and are bounded, then $\exists\, M < \infty$ s.t $\exists (g^*, h^* + Mg^*)$ satisfing modified multichain OE (same)

proof: let (g^*, h^*) satisfy multichain optimality equation

say for some (s, a)

$$c_1 = r(s,a) + \sum_j P(j|s,a) h^*(j) - h^*(s) - g^*(s) > 0$$

$\Rightarrow a \notin B_s$ as g^*, h^* satisfy multichain OE

consider $h = h^* + Mg^*$ for some $M > 0$

$$r(s,a) + \sum_j P(j|s,a)h(s) - h(s) - g^*(s)$$
$$= r(s,a) + \sum_j P(j|s,a)h^*(s) - h^*(s) - g^*(s) + M\left(\sum P(j|s,a)g^*(j) - g^*(s)\right)$$

$$= c_1 + M\left(\sum P(j|s,a)g^*(j) - g^*(s)\right)$$

as $a \notin B_s$ $\sum P(j|s,a)g^*(j) - g^*(s) < 0$ (underbrace c_2)

true for $M = -\frac{c_1}{c_2}$

as $c_1 > 0$, $c_2 < 0 \Rightarrow M > 0$

s.t $c_1 + Mc_2 = 0$

or $(g^*, h^* + Mg^*)$ satisfy modified OE

Note: From above theorm, we have both modified OE and multichain OE

Exe: $\exists$ unique recurrent class $\forall d^\infty, d^\infty \in \pi^{SMD}$ iff $\exists$ unique recurrent class $\forall d^\infty, d^\infty \in \pi^{SMR}$

Ans:

if $\exists$ unique recurrent class, $\forall d^\infty \in \pi^{SMR}$

true as $\pi^{SMD} \subseteq \pi^{SMR}$, $\forall d^\infty \in \pi^{SMD}$

$\exists$ unique recurrent class

now, $\exists$ unique recurrent class $\forall d^\infty \in \pi^{SMD}$

true

let $d : S \longrightarrow P(A)$ be s.t $d^\infty \in \pi^{SMR}$

now if $\nexists$ unique recurrent class for $d^\infty \in \pi^{SMR}$

true let

$T, C_1, C_2, \ldots C_K$ be classes $K > 1 \Rightarrow K > 2$

true

$g_d \neq g_e$ $\forall d, \exists s_1, s_2$ s.t $g_d(s_1) \neq g_d(s_2)$

$g_d(s_1) = \sum q_{d(s_1,a)} g_d(s_1, a)$ } same

$= \sum q_{d(s_2,a)} g_d(s_2, a)$

$= g_d(s_2)$ this is a contradiction

31st Oct :

multichain MDPs :

$$\left.\begin{aligned}&\max_d (P_d - I)g = 0\\ &\max_{d\in E} r_d + (P_d - I)h - g = 0\end{aligned}\right\} \text{multichain OE}$$

$\Updownarrow$ Finite MDPs

$$\left.\begin{aligned}&\max_d (P_d - I)g = 0\\ &\max_d r_d + (P_d - I)h - g = 0\end{aligned}\right\} \text{modified OE}$$

$d \in MD$ ← if (g^*, h^*) satisfy multi OE then $(g^*, h^* + Mg^*)$ satisfy modi OE

where $B_s = \{a \in A_s \mid \sum P(j|s,a) g(j) - g(s) = 0\}\ \forall s \in I$

$E = \{d \mid d(s) \in B_s, \forall s\}$

Theorem: If $\exists$ a pair of vectors (g,h) that simultaneously satisfy

$$\max_d (P_d - I)g \le 0$$

$$\text{and } \max_d (r_d + P_d h - h - g) \le 0$$

then (i) $g \ge g^*_+ = \sup_{\pi \in \Pi^{HR}} \limsup_{N\to\infty} \frac{v^\pi_{N+1}}{N+1}$

if $\exists (g,h)$ for some d' s.t $(P_{d'} - I)g \ge 0$ and $r_{d'} + P_{d'}h - h - g \ge 0$

then (ii) $g \le g^{(d')^\infty}_- \le g^*_- = \sup_{\pi \in \Pi^{HR}} \liminf_{N\to\infty} \frac{v^\pi_{N+1}}{N+1}$

if $\exists (g,h)$ satisfying $\max_d (P_d - I)g = 0$ and $\max_d (r_d + P_d h - h - g) = 0$

then (iii) $g = g^*$

proof: (i) we have $\exists (g,h)$ s.t $\max_d (P_d - I)g \le 0$, $\max_d (r_d + (P_d - I)h - g) \le 0$

$\Rightarrow (P_d - I)g \le 0$ and $r_d + (P_d - I)h - g \le 0$ $\forall d \in MD$ —— ①

now $g^*_+ = \sup_{\pi \in \Pi^{HR}} \overline{\lim_{N\to\infty}} \frac{v^\pi_{N+1}}{N+1}$

take any $d \in MD$, then

$(P_d - I)g \le 0$, $r_d + (P_d - I)h - g \le 0$ → $r_{d_1} + (P_{d_1} - I)h \le g$

let $\pi \in HR$, let $\pi = (d_1, d_2, \ldots)$

$g \ge P_d g \Rightarrow g \ge P_{d_1} g$ $\forall i$ ($\because$ ①)

then $g \ge P_{d_2} g \ge P_{d_1}(r_{d_2} + (P_{d_2} - I)h)$

$\vdots$

$g \ge P_{d_1} \cdots P_{d_n} g \ge P_{d_1} \cdots P_{d_n}(r_{d_{n+1}} + P_{d_{n+1}} h - h)$

$\Rightarrow Ng \ge \sum_t P^{\pi,t} r_{d_{t+1}} + \sum_t (P^{\pi,t+1} - P^{\pi,t})h$ $\quad P^{\pi,t} = P_{d_1} P_{d_2} \cdots P_{d_t}$

$\ge \sum_t P^{\pi,t} r_{d_{t+1}} + (P^{\pi,N+1} - I)h$

$$\Rightarrow g \geq \frac{1}{N}\sum_t p^{\pi,t} r_{d_{t+1}} + \underbrace{\frac{1}{N}(p^{\pi,N+1} - I)h}_{\text{Bounded} \to 0 \text{ as } N \to \infty}$$

$$\Rightarrow g \geq g_+^{\pi} \quad \forall \pi$$

$$\Rightarrow g \geq \sup_{\pi} g_+^{\pi} = g_+^*$$

(ii) $r_{d^*} - g + (P_{d^*} - I)h \geq 0$ and $(P_{d^*} - I)g \geq 0$ for some (g, h, d^*)

$\Rightarrow g \leq r_{d^*} + (P_{d^*} - I)h$ and $g \leq P_{d^*} g$

$\Rightarrow P_{d^*} g \leq P_{d^*}(r_{d^*} + (P_{d^*} - I)h)$

$\Rightarrow g \leq P_{d^*}(r_{d^*} + (P_{d^*} - I)h)$

$\vdots$

then $g \leq \sum \frac{p^{\pi,t}}{N} r_{d_{t+1}} = g_-^{d^\infty_*}$

$\Rightarrow g \leq g_-^*$

(iii) from (i), (ii) this follows trivially

Note: In λ case, v_λ^* was unique but policy not unique, also MD policies are finite

we know that if $\lambda_n \to 1$, then we can take a subsequence s.t δ^∞ is optimal $\forall \lambda_n \to$ subsequence;

$\exists \lambda^* < 1$ s.t δ^∞ optimal $\forall \lambda \geq \lambda^*$, δ^∞ is optimiser for average cost MDP

Theorem: (Existence) If $|S| < \infty$, $|A_s| < \infty$, $\forall s \in S$, then $\exists$ a solution to multichain OE

proof: $g = g^{\delta^\infty}$, $h = h^{\delta^\infty}$ are the candidates

$$0 = r_\delta + (\lambda_n P_\delta - I) v_{\lambda_n} \geq r_d + (\lambda_n P_d - I) v_{\lambda_n} \quad \forall d \in MD$$

$$0 \geq r_d + (\lambda_n P_d - I)\left(\frac{g}{(1-\lambda_n)} + h + f(\lambda_n)\right) \quad (\because \text{optimality of } \lambda<1 \text{ case}) \quad \forall d \in MD$$

$$= r_d + \underbrace{(P_d - I)g}_{1-\lambda_n} - \underbrace{(\lambda_n - 1)P_d g}_{1-\lambda_n} + (P_d - I)h + (\lambda_n - 1)P_d h + \tilde{f}(\lambda_n)$$

$$= \underbrace{\frac{(P_d - I)}{1-\lambda_n} g}_{\to \infty \text{ as } \lambda_n \to 1 \text{ if } (P_d - I)g < 0 \text{ then not} \to \infty} + \underbrace{(r_d - P_d g + (P_d - I)h)}_{\lambda_n \to 1 \text{ this term remains same}} + \underbrace{(\lambda_n - 1)P_d h + \tilde{f}(\lambda_n)}_{\to 0 \text{ as } \lambda_n \to 1}$$

Case I: $(P_d - I)g > 0$ is not possible as after some λ_n we cannot satisfy the equation

Case II: $(P_d - I)g = 0$

then $0 \geq r_d - P_d g + (P_d - I)h \quad \forall d$

$\Rightarrow 0 \geq r_d - g + (P_d - I)h$ as $(P_d - I)g = 0$

Case III: $(P_d - I)g < 0$, this case is possible, but d does not matter as in multichain OE, we will only take d: $(P_d - I)g = 0$

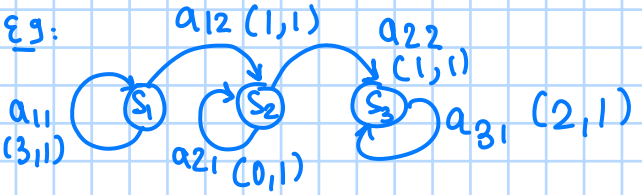
from case II and case III

$$\max_d (P_d - I)g = 0 \quad (\because d = \delta^\infty, \text{we have } P_{\delta^\infty} g = g)$$

$$\max_{d \in E} (r_d + (P_d - I)h - g) = 0 \quad (\because d = \delta^\infty \text{ we have } \max = 0)$$

Note: Above we have existance of multichain OE, we will have existance of modified OE by $h^* + Mg^*$ theorm

Eg:



$P(S_2|S_2, a_{21}) = 1$ $\quad r(S_2, a_{21}) = 0$ $\quad P(S_3|S_2, a_{22}) = 1$ $\quad r(S_2, a_{22}) = 1$

$P(S_3|S_3, a_{31}) = 1$

$P(S_1|S_1, a_{11}) = 1$ $\quad r(S_1, a_{11}) = 3$ $\quad P(S_2|S_1, a_{12}) = 1$ $\quad r(S_1, a_{12}) = 1$ $\quad r(S_3, a_{31}) = 2$ $\quad g_1, g_2, g_3$ $\quad h_1, h_2, h_3$

$$\left.\begin{aligned} &\max\{3 + h_1, 1 + h_2\} = h_1 + g_1 \\ &\max\{0 + h_2, 1 + h_3\} = h_2 + g_2 \\ &\text{and } 2 + h_3 = h_3 + g_3 \end{aligned}\right\} \text{second multichain OE}$$

$$\left.\begin{aligned} g_1 &= \max\{g_1, g_2\} \\ g_2 &= \max\{g_2, g_3\} \\ g_3 &= g_3 \end{aligned}\right\} \text{first multichain OE}$$

as $g_1 \geq 3 \Rightarrow g_1 > g_3$ and $g_1 \geq g_2$, $g_3 = 2$

If: $g_1 = g_2$

but as $\max\{0 + h_2, 1 + h_3\} = h_2 + g_2$

we pick a_{22} then $g_2 = g_3 \Rightarrow$ contradiction

so, $g_1 > g_2 \Rightarrow g_1 \geq 3$, $g_2 = 2 = g_3$ then $g_1 = 3$

asif $3 + h_1 < 1 + h_2$ then we pick a_{12}

but $g_1 \geq 3$ and not $g_1 = g_2 = 2$

so contradiction

so $3 + h_1 = g_1 + h_1 \Rightarrow g_1 = 3$

so $g^* = (3, 2, 2)^T$

lemma: If (g^*, h^*) satisfy multi OE and $d^* \in \underset{d}{\text{argmax}} \{r_d + P_d h^*\}$

then $(d^*)^\infty$ is optimal

lemma: If $(\tilde{g}, \tilde{h})$ satisfies modified OE and d^* s.t $P_{d^*}\tilde{g} = \tilde{g}$ and

$d^* \in \underset{d}{\text{argmax}} \{r_d + P_d \tilde{h}\}$

then $(d^*)^\infty$ is optimal

policy improvement algorithm:

for unichain: d_0

step n: find g_n, h_n

$$r_{d_n} + (P_{d_n} - I)h_n - g_n e = 0$$

improvement: $d_{n+1} \in \underset{d}{\text{Argmax}} (r_d + (P_d h_n))$

if $d_{n+1} = d_n$ we stop

first preference of d_{n+1} is d_n, otherwise as many s $d_n(s) = d_{n+1}(s)$ possible

for multichain: d_0

step n: $r_{d_n} + (P_d - I)h_n - g_n = 0$

$(P_{d_n} - I)g_n = 0$

improvement: $d_{n+1} \in \underset{d}{\text{argmax}} (P_d g_n)$