IE 708 tutorial

Tutorial-1 :

1. $P_t(s'|s)$, $r_t(s)$ if system occupies state s at time t
either don't do anything $= \phi$
or choose from A_s
s.t $P_t(s'|s,a)$ reward $r_t(s,a)$

(a) $(T, S', A'_s, s \in S', r'_t(s,a), P'_t(j|s,a))$ is MDP
then let $T = T$ (given discrete-time MDP) ($N=\infty$ or $N<\infty$)
(i.e $T = \{1,2 \ldots N\}$)

now $S' =$ given set of all states S

now $A'_s = A_s \cup \{\phi\}$, as A_s given and we have one more action $\forall s \in S'$
ϕ s.t we don't do anything

now, $r'_t(s,a) = \begin{cases} r_t(s) & ; a = \phi \leftarrow \text{this is given} \\ r_t(s,a) & ; a \neq \phi \leftarrow \text{this is also given} \end{cases}$

$P'_t(j|s,a) = \begin{cases} P_t(j|s) & ; a = \phi \\ P_t(j|s,a) & ; a \neq \phi \end{cases}$

(b) $T=2$, $v(j)$ = terminal reward then

$$MTR(s_1) = \max_{a' \in A_s \cup \{\phi\}} \left(r'_t(s_1,a') + \sum P'_t(j|s_1,a')\, v(j) \right)$$

now this is for given s_1

now, $MTR(s_1) = \max_{a' \in A_s} \{ r_t(s_1,a') + \sum_{\text{case } a' \in A_s} P_t(j|s_1,a')\, v(j),\ r_1(s_1) + \sum_{\text{case } \phi \text{ case}} P_t(j|s_1)\, v(j) \}$

we try to pick $a' \in A_s \cup \{\phi\}$ s.t it maximises the above
so let:

$$a_1^*(s) = \underset{a' \in A_s \cup \{\phi\}}{\operatorname{argmax}} \{ r'_t(s,a') + \sum P'_t(j|s,a')\, v(j) \}$$

then $a_1^*(s)$ is function s.t for state $s \in S'$ what
action to do at $T=1$ to maximise reward

2. $T = \{1\}$
$S' = \{s_1, s_2\}$ $A_{s_1} = \{a_{11}, a_{12}\}$
$A_{s_2} = \{a_{21}, a_{22}\}$
$r_1(s_1, a_{11}) = 5$ $r_1(s_2, a_{21}) = -1$
$r_1(s_1, a_{12}) = 10$ $r_1(s_2, a_{22}) =$

	s_1	s_2
(s_1, a_{11})	0.5	0.5
(s_1, a_{12})	0	1
(s_2, a_{21})	0.8	0.2
(s_2, a_{22})	0.1	0.9

(a) $v(s) = 0$ $\forall s \in S'$ then

$$a^* = \underset{a \in A_s}{\operatorname{argmax}} \{ r_1(s,a) + \sum \underbrace{v(j)}_{0} P_1(j|s,a) \}$$

$\Rightarrow a^* = \underset{a \in A_s}{\text{argmax}}\ r_1(s,a)$

now if at s_1 at $t=1$ then

$a^*(s_1) = \underset{a \in \{a_{11}, a_{12}\}}{\text{argmax}}\ r_1(s_1, a)$

$a^*(s_1) = a_{12}$ as $r_1(s_1, a_{12}) \geq r_1(s_1, a)\ \forall\ a \in A_{s_1}$

now $a^*(s_2) = \underset{a \in \{a_{21}, a_{22}\}}{\text{argmax}}\ r_1(s_2, a) = a_{22}$

so, $a^* = \begin{cases} a_{12} ; \text{at } t=1 \text{ at } s_1 \\ a_{22} ; \text{at } t=1 \text{ at } s_2 \end{cases}$

(b) now total reward $= r_1(s, a') + \sum v(j) P_1(j|s, a')$

now, if at s_1: $TR = r_1(s, a') + v(s_1) P_1(s_1 | s_1, a) + v(s_2) P_1(s_2 | s_1, a)$ then

$a^*(s_1) = \underset{a \in \{a_{11}, a_{12}\}}{\text{argmax}} \left(5 + d \times (0.5) + e \times 0.5 , 10 + d \times (0) + e \times 1 \right)$

$= \underset{a \in \{a_{11}, a_{12}\}}{\text{argmax}} \left(5 + \frac{d+e}{2}, 10 + e \right)$

$a^*(s_2) = \underset{a \in \{a_{21}, a_{22}\}}{\text{argmax}} \left(-1 + d(0.8) + e(0.2), 2 + d(0.1) + e(0.9) \right)$

$= \underset{a \in \{a_{21}, a_{22}\}}{\text{argmax}} \left(-1 + \frac{4d+e}{5}, 2 + \frac{d+9e}{10} \right)$

now, $a^*(s_1) = a_{11}$ if $5 + \frac{d+e}{2} > 10 + e$

$\Rightarrow \frac{d-e}{2} > 5 \Rightarrow d - e > 10$

if $d - e < 10 \Rightarrow a^*(s_1) = a_{12}$

$a^*(s_2) = a_{21}$ if $-1 + \frac{4d+e}{5} > 2 + \frac{d+9e}{10}$

$\Rightarrow \frac{8d+2e}{10} > 3 + \frac{d+9e}{10}$

$\Rightarrow 8d + 2e > 30 + d + 9e$

$\Rightarrow 7d > 30 + 7e$

$\Rightarrow d - e > \frac{30}{7}$

$a^*(s_2) = a_{22}$ if $d - e < \frac{30}{7}$

(c) $r_1(s, d(s)) = \begin{cases} P(A=a_{12}) r_1(s_1, a_{12}) + P(A=a_{11}) r_1(s_1, a_{11}) ; s = s_1 \\ P(A=a_{21}) r_1(s_2, a_{21}) + P(A=a_{22}) r_1(s_2, a_{22}) ; s = s_2 \end{cases}$

$= \begin{cases} q[5] + (1-q)[10] ; s = s_1 \\ 0[2] + 1[-1] ; s = s_2 \end{cases} = \begin{cases} 10 - 5q ; s = s_1 \\ -1 ; s = s_2 \end{cases}$

$$P_1(j|s,d_1(s)) = \begin{cases} q \times [0.5] + (1-q)0 & ; j=s_1,\ s=s_1 \\ q\times[0.5] + (1-q)1 & ; j=s_2,\ s=s_1 \\ 1\times 0.8 & ; j=s_1,\ s=s_2 \\ 1\times 0.2 & ; j=s_2,\ s=s_2 \end{cases}$$

$$\Rightarrow P_1(j|s,d_1(s)) = \begin{cases} 0.5q & ; j=s_1,\ s=s_1 \\ 1-0.5q & ; j=s_2,\ s=s_1 \\ 0.8 & ; j=s_1,\ s=s_2 \\ 0.2 & ; j=s_2,\ s=s_2 \end{cases}$$

(d) let $d: S \to P(A)$ be a random policy s.t

$$d(s_1) = \begin{bmatrix} p \\ 1-p \end{bmatrix} \begin{matrix} a_{11} \\ a_{12} \end{matrix}$$

$$d(s_2) = \begin{bmatrix} q \\ 1-q \end{bmatrix} \begin{matrix} a_{21} \\ a_{22} \end{matrix}$$

then now to maximise:

$$TR(s_1) = p[5 + (5)(0.5) + (-5)(0.5)] + (1-p)[10 + (5)(0) + (-5)(1)]$$

$$TR(s_2) = q[-1 + (5)(0.8) + (-5)(0.2)] + (1-q)[2 + (5)(0.1) + (-5)(0.9)]$$

$$TR(s_1)(p) = p[5] + (1-p)(5) = 5$$

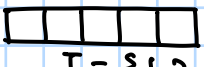
$$TR(s_2)(q) = q[-1+4-1] + (1-q)[2-4] = q[2] + (1-q)(-2) = 2q - 2 + 2q = 4q - 2$$

↑ maximum when $q=1$

so, $d(s_1) = \begin{bmatrix} p \\ 1-p \end{bmatrix}$ does not depend on p

$$d(s_2) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

3.



$T = \{1,2,3,4,5\}$ $N = 5$

at $\forall t \in T$

(a) $X_t \sim \text{Unif}\{0,1,2,3 \ldots 9\}$

↓ chosen digit

(b) choose one of empty positions

let positions be 1s, 10s, 100s, 1000s, 10000s

then Actions $= \{1, 10, 100, 1000, 10000\}$ → choose any 1

S = state of current number, total places filled

$S_1 = (0, [00000])$

↳ NO places filled

$r_1(S_1, a) = X_1 \times a$ ← reward is what we add to the number as total reward will optimise avg of number

where $A_{S_t} = \{10^{i-1} \mid i^{th}$ place is zero for digit for $S_t\}$

so if $S_1 = (0, [00000])$

then

$A_{S_1} = \{10^{1-1}, 10^{2-1}, \ldots 10^{5-1}\}$

$A_{S_1} = \{1, 10, 100, \ldots 10000\}$

if $S_2 = (1, [00001])$

then $A_{S_2} = \{10, 100, \ldots 10000\}$

now, to formulate MDP:

$(T, S, A_s, s \in S, r_t(s,a), P_t(j|s,a))$

$T = \{1, 2 \ldots 5\}$

S = all states $= \left\{ (x, [a_1 a_2 a_3 a_4 a_5]) \;\middle|\; x = \sum_{i=1}^{5} \alpha_i a_i,\ \alpha_i \in \{0, 1 \ldots 9\},\ a_i \in \{0, 1\} \right\}$

A_s for $s \in S$ is

$$A_s = \left\{ 10^{i-1} \;\middle|\; s = (x, [a_1 a_2 \ldots a_5]),\ i \text{ is s.t } a_i = 0 \right\}$$

$s = (x, [a_1 \ldots a_5])$

$r_t(s, a) = \mathbb{E}[X_t] \times a$

$$\mathbb{E}[X_t] = \frac{0 + 1 + \ldots 9}{10} = \frac{9 \times 10}{2 \times 10} = 4.5$$

$\Rightarrow r_t(s,a) = 4.5 \times a$ → as total reward will optimise avg of final number formed

now, $P_t(j|s,a)$ now if $s = (x, [a_1 \ldots a_5])$

$j = (y, [b_1 \ldots b_5])$

now if action a is chosen then

j will have $b_i = 1$ s.t $a = 10^{i-1}$

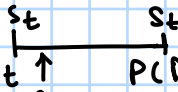
and $y = x + X_t a$

$$\text{so, } P_t(j|s,a) = P_t(j|s, 10^{i-1}) = \begin{cases} \frac{1}{10};\ j = (x + r10^{i-1}, [b_1 \ldots b_5]) \\ \quad \text{s.t } r \in \{0, 1 \ldots 9\},\ b_i = 1, \\ \quad b_j = a_j \quad j \neq i \\ 0;\ \text{otherwise} \end{cases}$$

($a = 10^{i-1}$)

4.

(a)



$P(D_t = j) = P_j$ demand

M = total capacity

$a_t + s_t \leq M$

$T = \{1 \ldots N\}$ $\quad M = 3$

$S = \{0, 1 \ldots M\}$ $\quad N = 3$ so $t=1$, $t=2$ $t=3 \rightarrow r_3(S_3) = 0$ No action

$A_S = \{0, 1 \ldots M - s\}$

$s_{t+1} = \max\{0, s_t + a_t - D_t\}$

$h(u)$ = holding cost
$o(u)$ = ordering cost

$$o(u) = \begin{cases} 0 & ; u = 0 \\ (u) + k & ; u > 0 \end{cases}$$

((u): cost per u; k: fixed cost)

$$o(u) = \begin{cases} 0 & ; u = 0 \\ 4 + 2u & ; u \neq 0 \end{cases}$$

$f(j)$ = amount we get if we sell j units $\quad f(u) = 8u$

$$r_t(s_t, a_t) = f(\min\{D_t, s_t + a_t\}) - o(a_t) - h(s_t + a_t) \quad h(u) = u$$

($\min\{D_t, s_t + a_t\} = s_t + a_t - s_{t+1}$)

$$F(u) = \sum_{j=0}^{u-1} f(j) p(j) + f(u)\left[\sum_{j \geq u} p(j)\right] \rightarrow \text{expected } f \text{ value if we start with } u \text{ units}$$

$$r_t(s_t, a_t) = F(s_t + a_t) - h(s_t + a_t) - o(a_t)$$

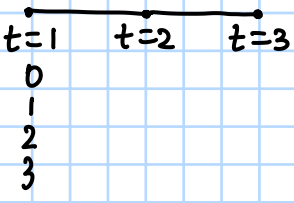
$$p_t(j \mid s, a) = \begin{cases} 0 & ; j > s + a \\ P(D_t = s + a - j) & ; 0 < j < s + a \\ P(D_t \geq s + a) & ; j = 0 \end{cases}$$

$$= \begin{cases} 0 & ; j > s + a \\ p_{s+a-j} & ; 0 < j < s + a \\ \sum_{i \geq s+a} p_i & ; j = 0 \end{cases}$$

$p_0 = \frac{1}{4} \quad p_1 = \frac{1}{2} \quad p_2 = \frac{1}{4} \quad p_i = 0$ for $i \neq 0, 1, 2$

$S = \{0, 1, 2, 3\}$

now,



$A_0 = \{0, 1, 2, 3\}$
$A_1 = \{0, 1, 2\}$
$A_2 = \{0, 1\}$
$A_3 = \{0\}$

$F(0) = 0$

$F(1) = p_0 \times f(0) + p_1 \times f(1) + p_2 \times f(1) = \left(\frac{1}{2} + \frac{1}{4}\right)(8) = 4 + 2 = 6$

↑ start with 1

$F(2) = p_0 \times f(0) + p_1 \times f(1) + p_2 \times f(2) = \frac{1}{2} \times 8 + \frac{1}{4} \cdot 8 \times 2 = 8$

$F(3) = p_0 \times f(0) + p_1 f(1) + p_2 f(2) = 8$

↑ start with 3

now, $h(u) = u \quad o(u) = \begin{cases} 0 & ; u = 0 \\ 4 + 2u & ; u \neq 0 \end{cases}$

$$r_t(s_t, a_t) = F(s_t + a_t) - h(a_t + s_t) - o(a_t)$$

$r_3(s_3) = 0 \quad \forall s_3, a_3$ possible as terminal case

$r_1(0,0) = F(0) - h(0) - u(0) = 0 - 0 - 0 = 0 = r_2(0,0)$
$r_1(0,1) = F(1) - h(1) - u(1) = 6 - 1 - 6 = -1 = r_2(0,1)$
$r_1(0,2) = F(2) - h(2) - u(2) = 8 - 2 - 4 - 4 = -2 = r_2(0,2)$
$r_1(0,3) = F(3) - h(3) - u(3) = 8 - 6 - 4 - 6 = -8 = r_2(0,3)$

$r_1(1,0) = F(1) - h(1) - u(0) = 6 - 1 - 0 = 5 = r_2(1,0)$
$r_1(1,1) = F(2) - h(2) - u(1) = 8 - 2 - 4 - 2 = 0 = r_2(1,1)$
$r_1(1,2) = F(3) - h(3) - u(2) = 8 - 3 - 4 - 4 = -3 = r_2(1,2)$

$r_1(2,0) = F(2) - h(2) - u(0) = 8 - 2 - 0 = 6 = r_2(2,0)$
$r_1(2,1) = F(3) - h(3) - u(1) = 8 - 3 - 4 - 2 = -1 = r_2(2,1)$

$r_1(3,0) = F(3) - h(3) - u(0) = 8 - 3 = 5 = r_2(3,0)$

now, pf at $t=2$ at state $s=0$

$$a_2^*(0) = \underset{a \in \{0,1,2,3\}}{\operatorname{argmax}} \; (r_2(0,a) + 0)$$

$= 0$ so, $a_2^*(0) = 0$

$a_2^*(1) = 0$

$a_2^*(2) = 0$

$a_2^*(3) = 0$

so at $t=2$ we don't buy anything

now at $t=1$ $s_1 = 0$

$$a_1^*(0) = \operatorname{argmax} \{ r_1(0,a) + P(S_2 = a)\, r_2(a,0) + P(S_2 = a-1)\, r_2(a-1,0) + P(S_2 = a-2)\, r_2(a-2,0) \}$$

$a_1^*(0)$: $a=0$: $r_1(0,0) + 1 \times r_2(0,0) = 0$

$a=1$: $r_1(0,1) + \frac{1}{4} r_2(1,0) + \frac{3}{4} r_2(0,0) = -1 + \frac{1}{4}(5) + 0 = \frac{1}{4}$

$a=2$: $r_1(0,2) + \frac{1}{4} r_2(2,0) + \frac{1}{2} r_2(1,0) + \frac{1}{4} r_2(0,0)$

$= -2 + \frac{1}{4}(6) + \frac{1}{2}(5) + \frac{1}{4}(0)$

$= -2 + 4 = 2$

$a=3$: $r_1(0,3) + \frac{1}{4} r_2(3,0) + \frac{1}{2} r_2(2,0) + \frac{1}{4} r_2(1,0)$

$= -8 + \frac{1}{4}(5) + \frac{1}{2}(6) + \frac{1}{4}(5)$

$= -8 + \frac{5}{2} + \frac{6}{2} = -8 + \frac{11}{2}$

$a_1^*(0) = 2$

$a_1^*(1)$: $a=0$: $r_1(1,0) + \frac{1}{4} r_2(1,0) + \frac{3}{4} r_2(0,0) = 5 + \frac{1}{4}(5) + \frac{3}{4}(0) = 6.25$

$a=1$: $r_1(1,1) + \frac{1}{4} r_2(2,0) + \frac{1}{2} r_2(1,0) + \frac{1}{4} r_2(0,0) = 0 + 2.5 = 2.5$

$a=2$: $r_1(1,2) + \frac{1}{4} r_2(3,0) + \frac{1}{2} r_2(2,0) + \frac{1}{4} r_2(1,0) = -3 + 3 + 1.25 = 1.25$

$a_1^*(1) = 0$

$a_1^*(2)$: $a=0$: $r_1(2,0) + \frac{1}{4} r_2(2,0) + \frac{1}{2} r_2(1,0) + \frac{1}{4} r_2(0,0)$

$= 6 + 1.5 + 2.5 + 0 = 10$

$a=1$: $r_1(2,1) + \frac{1}{4} r_2(3,0) + \frac{1}{2} r_2(2,0) + \frac{1}{4} r_2(1,0)$

$= -1 + 1.25 + 3 + 1.25$

$a_1^*(2) = 0$

$a_1^*(3) = 0 \rightarrow$ trivial

so at all states $a=0$, but if $S_1 = 0$ then $a_1 = 2$

(b) $b(u)$ = backlogging cost, if applied at start

i.e $r(S_t, a_t) = f((S_t + a_t - D_t)^+) - o(a_t) - h(S_t + a_t) - b((D_t - S_t - a_t)^+)$

for D_t

then

$r(S_t, a_t) = \mathbb{E}[f(S_t + a_t - D_t)^+] - o(a_t) - h(S_t + a_t) - \mathbb{E} b(D_t - S_t - a_t)^+$

let $\mathbb{E}[f(u - D_t)^+] = F(u)$

$$= \sum_{D_t=0}^{u-1} f(D_t) P(D_t) + \sum_{D_t=u}^{\infty} f(u) P(D_t)$$

$$\mathbb{E}[r(D_t - u)^+] = R(u)$$

$$= \sum_{D_t=0}^{u} r(0) P(D_t) + \sum_{D_t=u+1}^{\infty} r(D_t - u) P(D_t)$$

then $r_t(S_t, a_t) = F(S_t + a_t) - o(a_t) - h(S_t + a_t) - r(S_t + a_t)$

$P(D_t = 0) = 1/4 \quad P(D_t = 1) = 1/2 \quad P(D_t = 2) = 1/4$

and $u_t^*(S_t) = \max_{a_t \in A_{S_t}} \left(r(S_t, a_t) + \sum_{j \in S} P(j \mid S_t, a_t) u_{t+1}^*(j) \right)$

Reward-to-go function

$a_t^*(S_t) = \operatorname{argmax}_{a_t \in A_{S_t}} \left(r(S_t, a_t) + \sum_{j \in S} P(j \mid S_t, a_t) u_{t+1}^*(j) \right)$

Optimal quantity to order at t, given S_t

now $R(u) = \sum_{D_t=0}^{u} r(0) P(D_t) + \sum_{D_t=u+1}^{\infty} r(D_t - u) P(D_t)$

for $u = 0$

$$R(0) = \sum_{D_t=1}^{2} r(D_t) P(D_t)$$

$$= \frac{1}{2} \times (3) + \frac{1}{4} \times (3)(2)$$

$$R(0) = 3$$

$R(1) = \frac{1}{4} \times 3 \times 1 = \frac{3}{4}$

$R(2) = 0$

$R(3) = 0$, everything else functions same as (a)

(c) code on website

5. S, A_S are countable

N = Horizon length ($N<\infty$)

(a) π be HR policy i.e $\pi=(d_1, d_2 \dots d_{N-1})$ s.t

$d_t : H_t \longrightarrow P(A)$

↑ takes in History till t and outputs probability of action

$H_t = S_1 \times A_1 \times \cdots S_{t-1} \times A_{t-1} \times S_t$

$V_N^\pi(s) = \mathbb{E}_s^\pi \left[\sum_{t=1}^{N-1} r_t(X_t, Y_t) + \gamma_N(X_N) \right]$

↑ start at s

$U_t^\pi : H_t \longrightarrow \mathbb{R}$ be reward-to-go function

$U_t^\pi(h_t) = \mathbb{E}_{h_t}^\pi \left[\sum_{n=t}^{N-1} \gamma_n(X_n, Y_n) + \gamma_N(X_N) \right]$

$= r_t(s_t, d_t(h_t)) + \mathbb{E}_{h_t}^\pi \left[\sum_{n=t+1}^{N-1} \gamma_n(X_n, Y_n) + \gamma_N(X_N) \right]$

$U_t^\pi(h_t) = r_t(s_t, d_t(h_t))$

$+ \mathbb{E}_{h_t}^\pi \left[\mathbb{E}_{h_{t+1}}^\pi \left(\sum_{n=t+1}^{N-1} \gamma_n(X_n, Y_n) + \gamma_N(X_N) \right) \right]$

by tower property

$= r_t(s_t, d_t(h_t)) + \mathbb{E}_{h_t}^\pi \left[U_{t+1}^\pi(h_t, d_t(h_t), X_{t+1}) \right]$

↑ only thing that is random

now algorithm: set $t=N$, $U_N^\pi(h_N) = \gamma_N(s_N) \ \forall h_N \in H_N$

then put $t=N-1$:

$U_{N-1}^\pi(h_{N-1}) = \gamma_{N-1}(s_{N-1}, d_{N-1}(h_{N-1})) \quad \forall h_{N-1} \in H_{N-1}$

$+ \mathbb{E}_{h_{N-1}}^\pi \left[U_N^\pi(h_{N-1}, d_{N-1}(h_{N-1}), X_N) \right]$

iterate $t = N-2, N-3, \dots 1$

then $U_1^\pi(s) = V_N^\pi(s) \ \forall s \in S$

(b) Now $U_t(h_t) = \sup_{a \in A_{s_t}} \left\{ r_t(s_t, a) + \sum_{j \in S} p_t(j | s_t, a) U_{t+1}(h_t, a, j) \right\}$

$U_N(h_t) = r_N(s_N)$ is the bellman equation where

U_t^* be the solution to bellman equation, it is also the optimal strategy to follow

now, $U_t^*(h_t)$ only depends on h_t through s_t as

for $t=N$: $U_N^*(h_N) = \gamma_N(s_N)$ (true for $t=N$)

if true for $t \in \{n+1 \dots N\}$ then for n:

$U_n^*(h_n) = \sup_{a \in A_{s_n}} \left\{ r_n(s_n, a) + \sum_{j \in S} p_t(j | s_n, a) U_{n+1}^*(j) \right\}$

$\Rightarrow$ true for n

now, as solution to bellman equation depends on S_t and not n_t

$$U_n^*(S_t) = \sup_{a_t \in A_{S_t}} \{ r_t(S_t, a) + \sum_{j \in S} P_t(j | S_t, a) U_{t+1}^*(j) \}$$

by defn of supremum, $\exists a \in A_{S_t}$ s.t for given ε let call selecting that action given S_t π:

$$\varepsilon + U_n^{\pi}(S_t) \geq U_n^*(S_t)$$

then action a depends on S_t for given t and so similarly for every time same

$$\varepsilon + U_1^{\pi}(S_1) \geq U_1^*(S_1)$$

$$\pi = (d_1, d_2 \ldots d_{N-1})$$

s.t $d_i : S_i \to A_{S_i}$ this $d_i^0(S_i) = a_i$ s.t ε-optimal

(c) MR optimal policy $\pi = (d_1 \ldots d_{N-1})$ given

s.t $d_i : S_i \to P(A_{S_i})$ ← this is MD from (b)

optimal policy is s.t $U_t^*(S_t) = \sup_{\pi \in MR} U_t^{\pi}(S_t)$

and the supremum is for given $\pi \in MR$

with π is given optimal MD policy, now

then

MD policy is $\pi' = (d_1' \ldots d_{N-1}')$

s.t

$d_i' : S_i \to A_{S_i}$

optimal policy should satisfy bellman equation, we can find

$U_t^{\pi}(S_t)$ using (a) algorithm and so

let $a_t^*(S_t) = \underset{a \in A_{S_t}}{\operatorname{argmax}} \left(r_t(S_t, a) + \sum_{i \in S} P(i | S_t, a) U_{t+1}^{\pi}(i) \right)$

as same as $U_{t+1}^*(i)$ and computed by our algorithm

then $a_t^*(S_t)$ is optimal action to take at time t given S_t

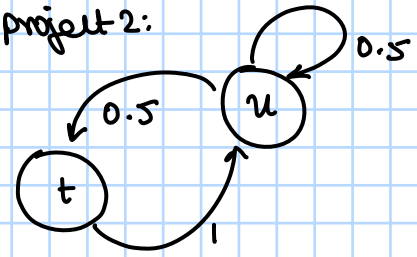
or $d_i' : S \to A_S$

$$d_i'(S) = a_t^*(S)$$

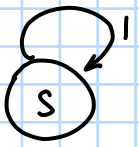
$$d_i'(S) = \underset{a \in A_S}{\operatorname{argmax}} \left(r_t(S, a) + \sum P(i | s, a) U_{t+1}^{\pi}(i) \right)$$

$\pi' = (d_1', d_2' \ldots, d_{N-1}')$ is optimal MD policy

6. project 2:



project 1:



now, $t=1$ 2 cases (s,t), (s,u)
$t=2$ 2 cases (s,t), (s,u)
$t=3$ 2 cases (s,t), (s,u)

so, total states $= \{(s,t),(s,u)\} = S$

actions: $A_{(s,t)} = \{1,2\}$

$A_{(s,u)} = \{1,2\}$ ← what project to pick

$r((s,t),1) = 1 \qquad r((s,t),2) = 0$

$r((s,u),1) = 1 \qquad r((s,u),2) = 2$

$r_3(s) = 0 \quad \forall s \in S$

now, $p((s,t)\,|\,(s,t),1) = 1$
$p((s,t)\,|\,(s,u),1) = 0$
$p((s,u)\,|\,(s,t),1) = 0$
$p((s,u)\,|\,(s,u),1) = 1$
$p((s,t)\,|\,(s,t),2) = 0$
$p((s,u)\,|\,(s,t),2) = 1$
$p((s,t)\,|\,(s,u),2) = 1/2$
$p((s,u)\,|\,(s,u),2) = 1/2$

now, for $N=3$:

$u_3^*((s,t)) = 0$

$u_3^*((s,u)) = 0$

$N=2$:

$u_2^*(s,t) = \max\{1+\Sigma(\ )(0),\ 0+\Sigma(\ )(0)\}$

$= \max\{1,0\}$

$u_2^*(s,t) = 1$ & $a_2^*(s,t) = 1$

$u_2^*(s,u) = \max\{1+\Sigma(\ )(0),\ 2+\Sigma(\ )(0)\}$
$= 2$

$u_2^*(s,u) = 2, \quad a_2^*(s,u) = 2$

now, $u_1^*(s,t) = \max\{1+1\times 1,\ 0+1\times 2\}$

$= 2 \quad a_1^*(s,t) = \{1,2\}$ does not matter

$$u_1^*(s,u) = \max\{1+1\times 2,\ 2+\tfrac{1}{2}\times 2+\tfrac{1}{2}\times 1\}$$

$$= \max\{3, 3.5\}$$

$$= 3.5$$

$$q_1^*(s,u) = 2$$

so,

State	action at $t=1$	action at $t=2$
(s,u)	2	2
(s,t)	1 or 2	1

Tutorial-2:

1. $v_\nu^\pi(s) = \mathbb{E}_s^\pi\left[\mathbb{E}_\nu\left[\sum_{t=1}^{\nu} r(X_t, Y_t)\right]\right]$

$\nu \sim$ Hypergeometric

$$P(\nu = n) = (n-1)(1-\lambda)^2 \lambda^{n-2} \quad n = 2, 3, \ldots$$

To prove: $v_\nu^\pi(s) = v_\lambda^\pi(s) + (1-\lambda)\dfrac{\partial v_\lambda^\pi(s)}{\partial \lambda}$

proof: $v_\lambda^\pi(s) = \lim_{N\to\infty} \mathbb{E}_s^\pi\left[\sum_{t=1}^{N} \lambda^{t-1} r(X_t, Y_t)\right]$

$$\frac{\partial v_\lambda^\pi(s)}{\partial \lambda} = \lim_{N\to\infty} \mathbb{E}_s^\pi\left[\sum_{t=1}^{N} (t-1)\lambda^{t-2} r(X_t, Y_t)\right]$$

$$\Rightarrow (1-\lambda)\frac{\partial v_\lambda^\pi(s)}{\partial \lambda} = \lim_{N\to\infty} \mathbb{E}_s^\pi\left[\sum_{t=1}^{N} (1-\lambda)(t-1)\lambda^{t-2} r(X_t, Y_t)\right]$$

$$\Rightarrow v_\lambda^\pi(s) + (1-\lambda)\frac{\partial v_\lambda^\pi(s)}{\partial \lambda} = \lim_{N\to\infty} \mathbb{E}_s^\pi\left[\sum_{t=1}^{N} \lambda^{t-1} r(X_t, Y_t) + \sum_{t=1}^{N} (1-\lambda)(t-1)\lambda^{t-2} r(X_t, Y_t)\right]$$

$$= \lim_{N\to\infty} \mathbb{E}_s^\pi\left[\sum_{t=1}^{N} \left(\lambda^{t-1} + (1-\lambda)(t-1)\lambda^{t-2}\right) r(X_t, Y_t)\right] \quad \text{—— ①}$$

now, $v_\nu^\pi(s) = \mathbb{E}_s^\pi\left[\mathbb{E}_\nu\left[\sum_{t=1}^{\nu} r(X_t, Y_t)\right]\right]$

$$= \lim_{N\to\infty} \mathbb{E}_s^\pi\left[\sum_{n=1}^{N} P(\nu = n)\left(\sum_{t=1}^{n} r(X_t, Y_t)\right)\right]$$

$$= \lim_{N\to\infty} \mathbb{E}_s^\pi\left[\sum_{n=1}^{N} (n-1)(1-\lambda)^2 \lambda^{n-2}\left(\sum_{t=1}^{n} r(X_t, Y_t)\right)\right]$$

$$= \mathbb{E}_s^\pi\left[\sum_{n=1}^{\infty} (n-1)(1-\lambda)^2 \lambda^{n-2}\left(\sum_{t=1}^{n} r(X_t, Y_t)\right)\right]$$

$$= \mathbb{E}_s^\pi\left[\sum_{t=1}^{\infty}\left(\sum_{n=t}^{\infty} \underbrace{(n-1)(1-\lambda)^2\lambda^{n-2}}_{②}\right) r(X_t, Y_t)\right] \quad \text{—— ③}$$

$(\because$ telescopic sum$)$

now for an AGP of type: $ab + (a+d)br + (a+2d)br^2 + \ldots\ldots = S$

given $r<1$ $\quad rS = abr + (a+d)br^2 + \ldots\ldots$

$$S - rS = ab + dbr + dbr^2 + dbr^3 + \ldots$$

$$(1-r)S = ab + dbr\left(\frac{1}{1-r}\right)$$

$$S = \frac{ab}{1-r} + \frac{dbr}{(1-r)^2}$$

so, ② is s.t $\begin{cases} a = t-1 \\ b = (1-\lambda)^2\lambda^{t-2} \\ d = 1 \\ r = \lambda \end{cases}$ $\quad S = (t-1)(1-\lambda)\lambda^{t-2} + \lambda^{t-1}$

so, $v_\nu^\pi(s) = \mathbb{E}_s^\pi\left[\sum_{t=1}^{\infty}((t-1)(1-\lambda)\lambda^{t-2} + \lambda^{t-1})\, r(X_t, Y_t)\right]$

$$= \lim_{N\to\infty} \mathbb{E}_s^\pi\left[\sum_{t=1}^{N} \lambda^{t-1} + (t-1)(1-\lambda)\lambda^{t-2})\, r(X_t, Y_t)\right]$$

$$v_\nu^\pi(s) = v_\lambda^\pi(s) + (1-\lambda)\frac{\partial v_\lambda^\pi(s)}{\partial \lambda} \quad \text{from } ①$$

2. $S = \{1, 2, \ldots\}$ $A_s = \{a\}$ $P(s+1 \mid s, a) = 1$, $r(s, a) = 0$ $\forall s, a$

(a) $v_1 = (0, 0, \ldots)$

$v_2 = (1, \frac{1}{\lambda}, \frac{1}{\lambda^2}, \ldots)$

$d(s) = a \ \forall s \Rightarrow r_d = (r(s, d(s)) = (0, 0, 0, \ldots)$

$(P_d)_{s,j} = P(j \mid s, d(s)) = \begin{cases} 1 ; j = s+1 \\ 0 ; \text{otherwise} \end{cases}$

$$P_d = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 1 & 0 & \cdots \\ 0 & 0 & 0 & \ddots & \ddots & \ddots \\ 0 & \cdots & & & & \end{bmatrix}$$

$$Lv = \sup_a \{r_d + \lambda P_d v\}$$

$$Lv_1 = \sup_a \left\{ v_1 + \lambda \begin{bmatrix} 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & 0 \cdots \\ 0 & 0 & 0 & 1 \cdots \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \end{bmatrix} \right\}$$

$v_1 = (0, 0, \ldots) = r_d$

$= (0, 0, 0 \ldots 0)$ $\qquad Lv_1 = v_1$

$$Lv_2 = \lambda \begin{bmatrix} 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & 0 \cdots \\ \cdot & \cdot & \cdot & \end{bmatrix} \begin{bmatrix} 1 \\ 1/\lambda \\ 1/\lambda^2 \\ \vdots \end{bmatrix} = \lambda \left[\frac{1}{\lambda} \ \frac{1}{\lambda^2} \ \frac{1}{\lambda^3} \ldots\right] = v_2$$

so, $Lv_1 = v_1$, $Lv_2 = v_2$

(b) let $V = \{v \mid \|v\| = \sup_{s \in \{1,2\ldots\}} v(s) < \infty\}$

then as we define $L : V \to V$
as d is any one

$$Lv = r_d + \lambda P_d v$$
$$= \lambda P_d v$$

$$L(v^1, v^2, \ldots) = \lambda(v^2, v^3, \ldots)$$

now for any $v, u \in V$

$$\|L(v^1, v^2, \ldots) - L(u^1, u^2 \ldots)\| = \lambda \|(v^2, v^3, \ldots) - (u^2, u^3, \ldots)\|$$

$$= \lambda \sup_{s \in \{2,3,\ldots\}} \|v^s - u^s\|$$

but if v, u s.t $\|u - v\| = \|u' - v'\|$ and u', v' s.t

$\|u^1 - v^1\| > \|u^s - v^s\| \ \forall\ s \in \{2, \dots\}$

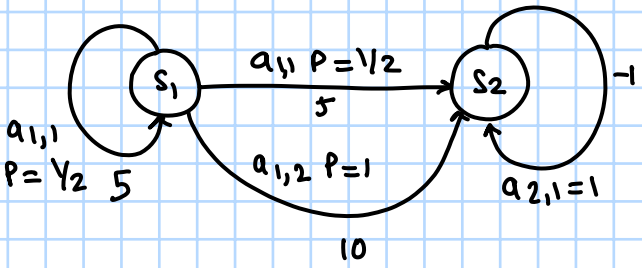
then

$\|Lu - Lv\| > \lambda \|v - u\|$ so L is not a contraction map

and we cannot apply banach fixed point theorm

(if T is banach, if T is not banach then also same)

3.



we want to maximise reward to go

$$v(s_2) = r(s_2, a_{2,1}) + \lambda P(s_2 | s_2, a_{2,1}) v(s_2)$$

$v(s_2)$: reward to go for s_2 at time t; $v(s_2)$: reward to go at $t+1$

$$\Rightarrow (1-\lambda) v(s_2) = -1$$
$$\Rightarrow v(s_2) = \frac{-1}{1-\lambda}$$

now for s_1: $\underbrace{v(s_1)}_{\text{Reward to go at } t} = \max_{a \in A_{s_1}} r(s_1, a) + \lambda \left(P(s_2 | s_1, a) v(s_2) + P(s_1 | s_1, a) v(s_1) \right)$

$$= \max \Big\{ r(s_1, a_{1,1}) + \lambda \left(P(s_2 | s_1, a_{1,1}) v(s_2) + P(s_1 | s_1, a_{1,1}) v(s_1) \right),$$
$$r(s_1, a_{1,2}) + \lambda P(s_2 | s_1, a_{1,2}) v(s_2) \Big\}$$

$$= \max \Big\{ \underbrace{5 + \lambda \left(\frac{1}{2} v(s_2) + \frac{1}{2} v(s_1) \right)}_{a_{1,1}}, \underbrace{10 + \lambda v(s_2)}_{a_{1,2}} \Big\}$$

$$v(s_2) = \frac{-1}{1-\lambda}$$

$$v(s_1) = \max \Big\{ \underbrace{5 + \frac{\lambda}{2} \left(v(s_1) - \frac{1}{1-\lambda} \right)}_{a_{1,1}}, \underbrace{10 - \frac{\lambda}{1-\lambda}}_{a_{1,2}} \Big\}$$

now $a^*(s_1) = a_{1,1}$ if

① $5 + \frac{\lambda}{2} \left(v(s_1) - \frac{1}{1-\lambda} \right) > 10 - \frac{\lambda}{1-\lambda}$

② $v(s_1) = 5 + \frac{\lambda}{2} \left(v(s_1) - \frac{1}{1-\lambda} \right) \Rightarrow \left(1 - \frac{\lambda}{2}\right) v(s_1) = 5 - \frac{\lambda}{2(1-\lambda)}$

$$\Rightarrow v(s_1) = \frac{5 - \frac{\lambda}{2(1-\lambda)}}{1 - \frac{\lambda}{2}}$$

$a^*(s_1) = a_{1,2}$ if

① $5 + \frac{\lambda}{2} \left(v(s_1) - \frac{1}{1-\lambda} \right) < 10 - \frac{\lambda}{1-\lambda}$

② $v(s_1) = 10 - \frac{\lambda}{1-\lambda}$

so, for $a^*(s_1) = a_{1,1}$ we get:

$$5 + \frac{\lambda}{2}\left(\frac{5 - \frac{\lambda}{2}}{\frac{2(1-\lambda)}{2-\lambda}} - \frac{1}{1-\lambda}\right) > 10 - \frac{\lambda}{1-\lambda}$$

$$\frac{\lambda}{2}\left(\frac{10}{2-\lambda} - \frac{2\lambda}{2(1-\lambda)(2-\lambda)} - \frac{1}{1-\lambda}\right) > 5 - \frac{\lambda}{1-\lambda}$$

$$\Rightarrow \frac{\lambda}{2}\left(\frac{10(1-\lambda) - (2-\lambda)}{(2-\lambda)(1-\lambda)} - \frac{\lambda}{(1-\lambda)(2-\lambda)}\right) > \frac{5(1-\lambda) - \lambda}{1-\lambda}$$

$$\Rightarrow \frac{\lambda}{2}\left(\frac{10 - 10\lambda - 2 + \lambda - \lambda}{2-\lambda}\right) > 5 - 5\lambda - \lambda$$

$$\Rightarrow \lambda(5 - 5\lambda - 1) > 5 - 6\lambda$$

$$\Rightarrow \lambda\left(\frac{4 - 5\lambda}{2-\lambda}\right) > 5 - 6\lambda$$

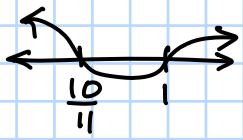
$$\Rightarrow 4\lambda - 5\lambda^2 > 10 - 12\lambda - 5\lambda + 6\lambda^2$$

$$\Rightarrow 11\lambda^2 - 21\lambda + 10 < 0$$

$$\lambda = \frac{21 \pm \sqrt{(21)^2 - 4(11)(10)}}{2 \times 11}$$

$$= \frac{21 \pm 1}{22}$$

so, $(\lambda - 1)(\lambda - \frac{10}{11}) < 0$



so $\lambda \in \left(\frac{10}{11}, 1\right)$ for $a^*(s_1) = a_{1,1}$

now, $5 + \frac{\lambda}{2}\left(10 - \frac{\lambda}{1-\lambda} - \frac{1}{1-\lambda}\right) < 10 - \frac{\lambda}{1-\lambda}$

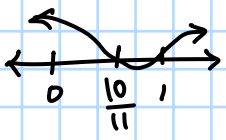
$$\Rightarrow \frac{\lambda}{2}\left(\frac{10 - 10\lambda - \lambda - 1}{1-\lambda}\right) < \frac{5 - 5\lambda - \lambda}{1-\lambda}$$

$$\Rightarrow \frac{\lambda}{2}(9 - 11\lambda) < 5 - 6\lambda$$

$$\Rightarrow 9\lambda - 11\lambda^2 < 10 - 12\lambda$$

$$\Rightarrow 0 < 11\lambda^2 - 21\lambda + 10$$

$$\Rightarrow (\lambda - 1)(\lambda - \tfrac{10}{11}) > 0$$



$\lambda \in \left[0, \frac{10}{11}\right]$

for $a^*(s_1) = a_{1,2}$

4. $L : \mathcal{V} \longrightarrow \mathcal{V}$

To prove: L is a contraction map

proof: let $U, V \in \mathcal{V}$

and let $LU = \sup_{d \in DMD} \{r_d + \lambda P_d U\}$ by defn

wlog: $LV(s) \geq LU(s)$ for some $s \in S$

$$LV(s) = \sup_{a \in A_s} \{ r(s,a) + \lambda \sum_{j \in S} P(j|s,a) V(j) \}$$

by defn of sup, $\forall \varepsilon > 0$, $\exists a^*$ s.t.

$$r(s,a^*) + \lambda \sum_{j \in S} P(j|s,a^*) V(j) + \varepsilon \geq LV(s)$$

and $$LU(s) \geq r(s,a^*) + \lambda \sum_{j \in S} P(j|s,a^*) U(j)$$

$$LV(s) - LU(s) \leq \varepsilon + \lambda \sum_{j \in S} P(j|s,a^*) V(j) - \lambda \sum_{j \in S} P(j|s,a^*) U(j)$$

$$\Rightarrow |LV(s) - LU(s)| \leq \varepsilon + \lambda \| \sum_{j \in S} P(j|s,a^*) \| \, \| V - U \|$$

$$\Rightarrow \|LV - LU\| \leq \varepsilon + \lambda \|V - U\| \quad (\because \text{true } \forall s \in S)$$

as $\varepsilon \to 0$

$$\Rightarrow \|LV - LU\| \leq \lambda \|V - U\|$$

so, $\forall U, V \in \mathcal{V}$, $\|LU - LV\| \leq \lambda \|U - V\|$

so L is a contraction map

5. To prove: $\sup_{d \in D^{MD}} \{ r_d + \lambda P_d V \} = \sup_{d \in D^{MR}} \{ r_d + \lambda P_d V \}$

proof:

let $FV(s) = \sup_{a \in A_s} \{ r(s,a) + \lambda \sum_{j \in S} P(j|s,a) V(j) \}$

↑ we define this function

$$\tilde{F}V(s) = \sup_{P(A_s)} \sum_{a \in A_s} \left(P(a) \left(r(s,a) + \lambda \sum_{j \in S} P(j|s,a) V(j) \right) \right)$$

now to maximise weighted average, we have to assign all weights to one index that is maximum

so from this

$$\sup_{a \in A_s} \{ r(s,a) + \lambda \sum_{j \in S} P(j|s,a) V(j) \} = \sup_{P(A_s)} \sum_{a \in A_s} P(a) \left(r(s,a) + \lambda \sum_{j \in S} P(j|s,a) V(j) \right)$$

$$\Rightarrow F(V(s)) = \tilde{F}(V(s)) \quad \forall s \in S$$

$$\Rightarrow \sup_{d \in D^{MD}} \{ r_d + \lambda P_d V \} = \sup_{d \in D^{MR}} \{ r_d + \lambda P_d V \}$$

6. To prove: If $V \in \mathcal{V}$ s.t. $LV = V$ then

$$V = \sup_{\pi \in \Pi^{SR}} V_\lambda^\pi$$

proof: If $V \geq LV$ then (componentwise)

now as $LV \geq r_d + \lambda P_d V$ by defn (componentwise)

$$\Rightarrow V \geq r_d + \lambda P_d V$$

and $V_\lambda^\pi = r_d + \lambda P_d V_\lambda^\pi$ for $\pi \in \pi^{SR}$ ($\because$ done in class)

$(I - \lambda P_d) V \geq r_d$

as $(I - \lambda P_d)^{-1}$ is positive ($\because$ done in class)

$\Rightarrow V \geq \underbrace{(I - \lambda P_d)^{-1} r_d}_{V_\lambda^\pi \text{ for } \pi = (d, d, \ldots)}$

$\Rightarrow V \geq V_\lambda^\pi$ for $\pi \in \pi^{SR}$

as this is true $\forall\, d$, we get

$V \geq V_\lambda^\pi \;\forall\; \pi \in \pi^{SR}$

$\Rightarrow V \geq V_\lambda^* = \sup_{\pi \in \pi^{SR}} V_\lambda^\pi$ ——— ①

now, if $V \leq LV$ then from defn of sup
$\forall\, \varepsilon > 0, \exists\, d \in D^{MD}$ s.t

$V \leq r_d + \lambda P_d V + \varepsilon \cdot \mathbf{e}$

$\Rightarrow (I - \lambda P_d) V \leq r_d + \varepsilon \cdot \mathbf{e}$

$\Rightarrow V \leq \underbrace{(I - \lambda P_d)^{-1} r_d}_{V_\lambda^\pi \text{ for } \pi \in \pi^{SD}} + \underbrace{(I - \lambda P_d)^{-1} \varepsilon \cdot \mathbf{e}}_{\frac{\varepsilon}{1-\lambda} \cdot \mathbf{e}}$

$\Rightarrow V \leq V_\lambda^\pi + \frac{\varepsilon}{1-\lambda} \cdot \mathbf{e}$

taking sup

$\Rightarrow V \leq \sup_{\pi \in \pi^{SD}} V_\lambda^\pi + \frac{\varepsilon}{1-\lambda} \cdot \mathbf{e} \leq \sup_{\pi \in \pi^{SR}} V_\lambda^\pi + \frac{\varepsilon}{1-\lambda} \cdot \mathbf{e}$

as $\pi^{SD} \subseteq \pi^{SR}$

$\Rightarrow \sup_{\pi \in \pi^{SD}} V_\lambda^\pi \leq \sup_{\pi \in \pi^{SR}} V_\lambda^\pi$

$\Rightarrow V \leq \sup_{\pi \in \pi^{SR}} V_\lambda^\pi$ as $\varepsilon \to 0$ ——— ②

now if $V = LV$ then $V \leq LV \Rightarrow V \leq V_\lambda^*$ from ②
$V = LV$ then $V \geq LV \Rightarrow V \geq V_\lambda^*$ from ①

so as $V \leq V_\lambda^*$ & $V \geq V_\lambda^*$
$\Rightarrow V = V_\lambda^*$
$\Rightarrow V = \sup_{\pi \in \pi^{SR}} V_\lambda^\pi$

7. given $\exists\, \pi \in \pi^{HR}$ s.t

$$V_\lambda^\pi = V_\lambda^* = \sup_{\pi \in \pi^{HR}} V_\lambda^\pi$$

To prove: $\exists\, \tilde{\pi} \in \pi^{SD}$ s.t $V_\lambda^{\tilde{\pi}} = V_\lambda^\pi$

proof: let $\pi = (d_1, d_2, \ldots) \in \pi^{HR}$, then we know
$\exists\, \pi' \in (d_1', d_2', \ldots) \in \pi^{MR}$ s.t

$d_t' : S \longrightarrow P(A)$

$P^{\pi'}(X_n = j, Y_n = a \mid X_1 = s) = P^{\pi}(X_n = j, Y_n = a \mid X_1 = s)$ (theorem done in class)

i.e $V_\lambda^{\pi}(s) = V_\lambda^{\pi'}(s)$

$$\vartheta_\lambda^{\pi'}(s) = \sup_{P(A_j)} \sum_{a \in A_s} r(s,a)\, q_{d_1'(s)}(a) + \lambda \sum_{j \in S} \Big(\sum_{a \in A_s} P(j \mid s,a)\, q_{d_1'(s)}(a) \Big) \vartheta_\lambda^{(d_{2(j)}', d_{3(j)}', \ldots)}(j)$$

where $q_{d_t'(s)}(j) = P^{\pi}(Y_t = a \mid X_t = j, X_1 = s)$

as sup over $P(A_s)$, $\exists\, a \in A_s$, s.t $\{P(A_s = a) = 1\}$ $\in$ argmax of all $P(A_s)$

as average maximum, more all weight given to particular

so let this a be a_s^* then

$$\vartheta_\lambda^{\pi'}(s) = r(s, a_s^*) + \lambda \sum_{j \in S} P(j \mid s, a_s^*)\, \vartheta_\lambda^{(d_2'(j), \ldots)}(j)$$

let $d : S \longrightarrow A$ s.t

$d(s) = a_s^*$ as above can be done $\forall s$

now,

$$\vartheta_\lambda^{(d_2'(j), \ldots)}(j) \leq \vartheta_\lambda^{(d_1'(j), \ldots)}(j) = \vartheta_\lambda^{\pi}(j)$$

as optimal policy

$$\Rightarrow \vartheta_\lambda^{(d_2'(j), \ldots)}(j) \leq r(j, a_j^*) + \lambda \sum_{k \in S} P(k \mid j, a_j^*)\, \vartheta_\lambda^{(d_2'(k), \ldots)}(k)$$

$$\Rightarrow \vartheta_\lambda(j)^{(d_2'(j), \ldots)} \leq r(j, a_j^*) + \lambda \sum_{k \in S} P(k \mid j, a_j^*)$$

$$\Big(r(k, a_k^*) + \lambda \sum_{k_2 \in S} P(k_2 \mid k, a_k^*) \Big)$$

$$\Big(r(k_2, a_{k_2}^*) + \lambda \sum_{k_3 \in S} \cdots \Big) \Big)$$

recursively

$$\Rightarrow \vartheta_\lambda(j)^{(d_2'(j), \ldots)} \leq \sum_{k=0}^{\infty} \lambda^k P_d^k r_d(j) = \vartheta_\lambda^{(d,d,\ldots)}(j)$$

$$\Rightarrow \vartheta_\lambda(j)^{(d_2'(j), \ldots)} \leq \vartheta_\lambda^{(d,d,\ldots)}(j)$$

now, $\vartheta_\lambda^{\pi'}(s) \leq r(s, d(s)) + \lambda \sum P(j \mid s, d(s))\, \vartheta_\lambda^{(d, \ldots)}(j)$

$\vartheta_\lambda^{\pi'}(s) \leq \vartheta_\lambda^{(d,d,\ldots)}(s)$ $\forall i \in S$

but as $V_\lambda^{\pi'}(s) = \sup_{\pi \in \pi^{HR}} \vartheta_\lambda^{\pi} \leq \vartheta_\lambda^{(d, \ldots)}(s)$

$$\Rightarrow \vartheta_\lambda^{\pi'}(s) = \vartheta_\lambda^{(d, \ldots)}(s)$$

so, let $\tilde{\pi} = (d, d, \ldots)$ $\vartheta_\lambda^{\pi'}(s) = \vartheta_\lambda^{\tilde{\pi}}(s) = \vartheta_\lambda^{\pi}(s)$ $\forall s \in S$

so, $\vartheta_\lambda^{\tilde{\pi}}$ is optimal

8. $v \in V$ $d_v \in D^{MD}$ is v-improving if $d_v \in \underset{d \in D^{MD}}{\operatorname{argmax}} \{r_d + \lambda P_d v\}$

(a) $v_\lambda^{(d_v, d_v, \dots)} < v$ given $d_v \in \underset{d \in D^{MD}}{\operatorname{argmax}} \{r_d + \lambda P_d v\}$

let $v = (1000)$ only one state S
$r(s,a) = (0)$ only one action a
then $d_v(S) = a$ only one $d \in D^{MD}$

$$v_\lambda^{\pi} = (I - \lambda P_d)^{-1} r_d = (0) \Rightarrow v_\lambda^{\pi} = (0)$$

$v_\lambda^{\pi} < v$

(b) let S, a only one state and one action and let $r_d = (1-\lambda)$
then for $v = (1)$
we get

$$v = (1) \leq (1-\lambda) + \lambda(1) = (1)$$

and

$$v_\lambda^{(d_v, d_v \dots)} = (I - \lambda P_{d_v})^{-1} r_{d_v} = \frac{1}{1-\lambda}(1-\lambda) = (1$$

so, $r_{d_v} + \lambda P_{d_v} v \geq v$ and $v_\lambda^{(d_v, d_v, \dots)} = v$

(c) To prove: $v_\lambda^{(d_v, \dots)}(s'') < v(s'')$ for some $s'' \in \mathcal{S}$ if $r_{d_v}(s') + \lambda P_{d_v} v(s') \geq v(s')$ for some $s' \in \mathcal{S}$

proof: $\exists s' \in \mathcal{S}$ $r_{d_v}(s') + \lambda P_{d_v} v(s') \geq v(s')$

as $d_v \in \underset{d \in D^{MD}}{\operatorname{argmax}} \{r_d + \lambda P_d v\}$

$$r_d(s') + \lambda P_d v(s') \leq r_{d_v}(s') + \lambda P_{d_v}(s') \quad \forall d \in D^{MD}$$

if $\exists s''$ s.t. $r_{d_v}(s'') + \lambda P_{d_v}(s'') < v(s'')$

then, $r_d(s'') + \lambda P_d v(s'') < v(s'') \quad \forall d \in D^{MD}$

$$\Rightarrow (I - \lambda P_d)^{-1} r_d(s'') < v(s'') \quad \forall d \in D^{MD}$$

$$\Rightarrow (I - \lambda P_{d_v})^{-1} r_{d_v}(s'') < v(s'')$$

$$\Rightarrow v_\lambda^{(d_v, d_v \dots)}(s'') < v(s'')$$

if $\forall s \in \mathcal{S}$ $r_{d_v}(s) + \lambda P_{d_v}(s) v \geq v(s)$

$\Rightarrow r_{d_v} + \lambda P_{d_v} v \geq v$ (componet vise)

$\Rightarrow (I - \lambda P_{d_v})^{-1} r_{d_v} \geq v$

$\Rightarrow v_\lambda^{(d_v, d_v \dots)} \geq v$, so as in statement $\exists s'$, we get for all other $s'' \neq s'$ we get above, if all s', then not possible

9. finite state and actions $\{s_1, s_2 \dots s_N\}$ $N \geq 1$ $d \in D^{MD}$ let $P_d^1 + P_d^v = P_d$

$Q_d = (I - \lambda P_d^1)$

$R_d = \lambda P_d^v$

$$T: \mathcal{V} \longrightarrow \mathcal{V}$$

$$T\vartheta = \max_{d \in D^{MD}} \{ \vec{\theta}_d \vec{r}_d + \vec{Q}_d \vec{R}_d \vartheta \}$$

wlog $T\vartheta(s) \geq TU(s)$

Let $d^* \in \operatorname{argmax}_{d \in D^{MD}} \{ \vec{\theta}_d \vec{r}_d + \vec{Q}_d \vec{R}_d \vartheta \}$

$$T\vartheta(s) = \vec{\theta}_{d^*} \vec{r}_{d^*}(s) + \vec{Q}_{d^*} \vec{R}_{d^*} \vartheta(s)$$

And $TU(s) \geq \vec{\theta}_{d^*} \vec{r}_{d^*}(s) + \vec{Q}_{d^*} \vec{R}_{d^*} U(s)$

$$\Rightarrow T\vartheta(s) - TU(s) \leq \vec{\theta}_{d^*} \cancel{\vec{r}_{d^*}(s)} + \vec{\theta}_{d^*} \vec{R}_{d^*} V(s) \cancel{- \vec{\theta}_{d^*} \vec{r}_{d^*}(s)} - \vec{Q}_{d^*} \vec{R}_{d^*} U(s)$$

$$\Rightarrow TV(s) - TU(s) \leq \vec{Q}_{d^*} \vec{R}_{d^*} (V(s) - U(s))$$

$$\Rightarrow |TV(s) - TU(s)| \leq \| \vec{Q}_{d^*} \vec{R}_{d^*} \| \, |V(s) - U(s)| \quad \forall s \in \mathcal{S}$$

$$\Rightarrow \|TV - TU\| \leq \| \vec{Q}_{d^*} \vec{R}_{d^*} \| \|V - U\|$$

now seen in case $\| \vec{Q}_{d^*} \vec{R}_{d^*} \| < 1$, so T is a contraction map

10. on website, noting on monotonicity as $\vec{r}$ not ≥ 0 or ≤ 0