IE 708
Markov Decision Processes

Ref: M. Puterman
Markov decision process
grading: class test 1 – 15%
(before midterm)
class test 2 - 20%
midterm - 30%
final exam/presentation - 35%
pre req: James Norris - chapter 1 read

29th July:

There is a notion of time, at a given time we have to make a decision, take them s.t we maximise something or some function.

Eg: Runig competition, speed fast or less at start compared to end, control speed at each time, finish race first

very broadly about sequential decision making under uncertainty

Eg: first shop, we want to keep stock, goal is to maximise profit, every morning order something s.t we meet demand and also less stock for next day doesn't want to order too many, minimise waistage, so how much to buy?

Uaing objectives:

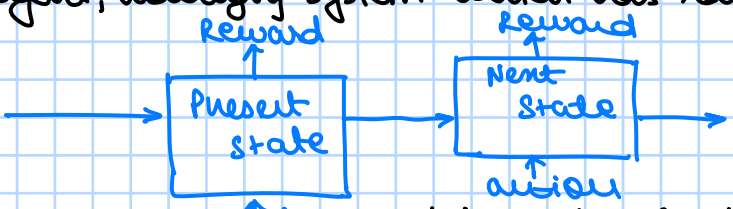
1) Model sequential decision problems using MDPs. (Markov decision problems)

2) Existance and structure of optimal strategies

3) Computing an optimal strategy

framework:

There is a single agent, underlying system which has randomness, at a given time:



← this action can be chosen in many ways and Single agent we assume agent knows everything about past

← time →

The decision maker wants to optimise action to maximise some form of reward reward can be MA or total average

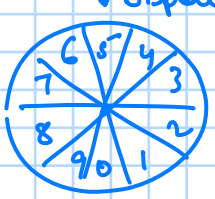
← Moving average / over period of time

Eg: For vegitable vendor:
State = stock available
action = how much to order
reward = revenue

Two sequencial problems:

who's counting:

↓ Display some numbers




team 1


team 2

First let's have only one team, after every spin we place num at some location, goal is to max out the num.

In optimal strategy, current state does not matter, only location matters

Obs Num	spin 1	spin 2	Spin 3	Spin 2	spin 1
0	5	4	3	2	1
1	5	4	3	2	1
2	5	4	3	2	1
3	4	3	3	2	1
4	3	3	2	2	1
5	3	2	2	1	1
6	2	2	1	1	1
7	1	1	1	1	1
8	1	1	1	1	1
9	1	1	1	1	1

Average score using this strategy: 78,734.12
Average score using random assignment: 49,999.5

K-hop tandem network: (Resource allocation problem)

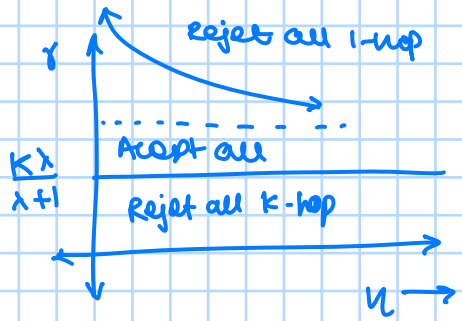
} 2 types of request

each link has a unit capacity
reward of 1-hop request: 1
reward of K-hop request: r

← 1 hop request

On an average there are λ arrivals (for one unit of time)
K hop request arrive at η
← per unit time 1-hop req

parameters: λ, η, r, K
← reward of K-hop per unit time K-hop req



Model description & Notations:

1) Decision epoch:
T → set of epoch, can be continuous/discrete time / random time
(mostly this two)

Random time is when the period is random

for finite discrete time, i.e $T = \{1, 2 \ldots N\}$ we call it finite horizon problem

for infite discret time, i.e $T = \{1, 2 \ldots\}$ we call it infinite horizon problem

2) State and actions:

S: set of states

mostly discrete { finite / countable

A_s : set of allowable actions for present state $s \in \mathcal{S}$

$A = \bigcup_{s \in \mathcal{S}} A_s$: set of all actions

mostly focus on discrete

Note: $P(A_s)$ = set of probability distributions on A_s

$q(\cdot) \in P(A_s)$ a random action, and so

$\therefore q(a)$ for $a \in A_s$ is a random taken action

Eg: $A_s = \{1, 2, \dots a_s\}$
↑ set of possible actions

$P(A_s) = \{q \mid \sum_{i=1}^{a_s} q_i = 1, q_i \geq 0\}$ ← vector of dim a_s

$q(\cdot) \in P(A_s)$
← A vector of probability and so

$q(a)$ for $a \in A_s$ is a random thing from A_s set

3) Reward and state evolution:
if in state $s \in \mathcal{S}$ and we take action $a \in A_s$
then reward we get is $r_t(s, a)$

$r_t(s, t)$: reward at time t, when system is at $s \in \mathcal{S}$ and we take action $a \in A_s$

$P_t(j \mid s, a)$: the prob that system is at $j \in \mathcal{S}$ (← this is at $t+1$) given prev is $s \in \mathcal{S}$ and we take $a \in A_s$ at given time t.

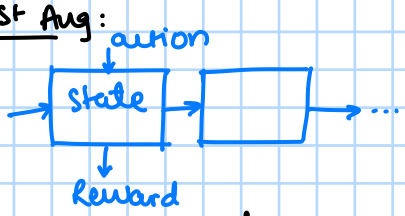
Note: $r_t(s, a, j)$ is reward if we are at $s \in \mathcal{S}$, take $a \in A_s$ and we go to $j \in \mathcal{S}$

in this case:

$$r_t(s, a) = \sum_{j \in \mathcal{S}} P_t(j \mid s, a) \times r_t(s, a, j)$$

Def: $(T, \mathcal{S}, A_s, s \in \mathcal{S}, r_t(s, a), P_t(j \mid s, a))$ is the markov decision process

1st Aug:



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

$T = \{1, 2, 3, \ldots N\}$ random horizon (stop when we find candidate)
↳ one of the cases (N is a random variable)

decision epochs are random is one more case of decision epochs

Stochastic inventory control:

S_t ← stock available at time t

new order D_t = demand in month t

$$P(D_t = j) = p_j$$

M = total capacity of warehouse so, $a_t + S_t \leq M$

$T = \{1, 2, 3, \ldots N\}$ $N < \infty$ for finite horizon
$N = \infty$ for infite horizon

$S = \{0, 1, \ldots M\}$ → state / items in warehouse

$A_S = \{0, 1, \ldots M - S\}$ $s \in S$
↳ maximum we can order

$$S_{t+1} = \max\{0, S_t + a_t - D_t\}$$

$$S_{t+1} = (S_t + a_t - D_t)^+$$ ← one way of writting

$h(u)$ = holding u units of stock per month
$O(u)$ = ordeing cost u units

$$O(u) = \begin{cases} 0 \,;\, u = 0 \\ c(u) + K \,;\, u > 0 \end{cases}$$

$c(u)$ ↓ cost of u ← some fixed cost

$f(j)$ = amount we get if we sell j units

if current state S_t and we order a_t units then
holding cost = $h(S_t + a_t)$
ordeing cost = $O(a_t)$
revenue cost = $f(\max\{D_t, S_t + a_t\})$

$$r_t(S_t, a_t) = f(\max\{D_t, S_t + a_t\}) - O(a_t) - h(S_t + a_t)$$

$S_t + a_t - S_{t+1}$

$$r_t(S_t, a_t, S_{t+1}) = f(S_t + a_t - S_{t+1}) - O(a_t) - h(S_t + a_t)$$

← Avg profit if we start with u units

$$F(u) = \sum_{j=0}^{u} f(j) P(j) + f(u)\left[\sum_{j>u} P(j)\right]$$ → expected profit for given u units

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

or $r_t(s,a) = F(s+a) - h(s+a) - o(a)$

$F(u) = \mathbb{E}[f(\min(D_t, u))]$ D_t is a random variable

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

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

Exe: so who's couting game MDP formulation ? → doubt

Decision rules:

prescribes an action at decision epoch t

$d_t : S \to A$

so $d_t(s) \in A_s$ action when current state is s

also called Markovian Deterministic (MD)

$(A_s \subseteq A$ so $d_t(s) \in A_s \subseteq A)$

$h_t = (s_1, a_1, s_2, a_2, \ldots s_{t-1}, a_{t-1}, s_t)$ is called history upto time t

$h_t \in \underbrace{S \times A \times \cdots \times}_{1} \underbrace{S \times A \times S}_{t-1 \; t}$

$d_t : H_t \to A$

$d_t(h_t) \in A_{s_t}$ History dependent deterministic (HD)

M. Randomised $d_t : S \to P(A)$

H. Randomised $d_t : H_t \to P(A)$

D_t^K : set of all decision rules of type K

$K \in \{MD, HD, MR, HR\}$

policy:

Specifies the decision rule at each decision epoch

$\pi = (d_1, d_2, \ldots)$

$\pi = (d, d, \ldots) \to$ this is called a stationary policy

Π^K = set of all policies for $K \in \{SD, SR, MD, MR, HD, HR\}$

$\Pi^{SD} \subseteq \Pi^{MD} \subseteq \Pi^{MR} \subseteq \Pi^{HR}$

$\Pi^{SD} \subseteq \Pi^{MD} \subseteq \Pi^{HD} \subseteq \Pi^{HR}$

$\Pi^{SR} \times \Pi^{MD}$

$\Pi^{HD} \times \Pi^{MR}$

Eg: $d(s) = \begin{cases} 0 & \text{if } s \geq \sigma \\ M-s & s < \sigma \end{cases}$

σ ↓ fixed threshold

$d : \mathcal{S} \to A$ and as $\pi = (d, d, \ldots\ldots)$ we have

$\pi \in \Pi^{SD}$

if

$d_t(s) = \begin{cases} 0 & ; s \geq \sigma_t \\ M-s & ; s < \sigma_t \end{cases}$

then $d_t : \mathcal{S} \to A$

and $\pi = (d_1, d_2 \ldots\ldots)$

$\Rightarrow \pi \in \Pi^{MS}$

5th Aug:

$\Omega = \underbrace{S}_{1} \times A \times \underbrace{S}_{2} \times A \times S \times A \times \cdots \times \underbrace{S}_{N} \times A \times S$ Finite horizon ($N<\infty$) ← canonical space

also assume finite states S^N as S^N actions

$B(\Omega)$ = set of all subsets of Ω (this will be a σ-field as finite Ω)

$\omega \in \Omega$

$\omega = (s_1, a_1, s_2, a_2, \ldots, s_{N-1}, a_{N-1}, s_N)$

$X_t(\omega) = s_t$ state at time t

$Y_t(\omega) = a_t$ action at time t

$Z_t(\omega) = (s_1, a_1, s_2, a_2, \ldots s_t)$ history upto time t

now if $\pi \in \Pi^{HR}$ (the biggest policy space)

$\pi = (d_1, \ldots)$

s.t $d_t : Z_w(t) \longrightarrow P(A)$

this induces a probability distribution P^π

let distn$(X_1) = P_1$, so under P^π:

$P^\pi(X_1 = s) = P_1(s)$ ← probability that at $t=1$, state is s

$P^\pi(Y_t = a \mid Z_t = h_t)$ ← given π, what is prob we take action a at time t given history at time t $Z_t = h_t$

$= q_{d_t(h_t)}(a)$

here $d_t : H_t \longrightarrow P(A)$

so $d_t(h_t) \in P(A)$

& then $q(a)$ is a probability

now, $P^\pi(X_{t+1} = s_{t+1} \mid Z_t = h_t, Y_t = a) = P_t(s_{t+1} \mid s_t, a)$

→ probability of next state is s_{t+1} given history and action

$P^\pi((s_1, a_1, s_2, a_2, \ldots s_{N-1}, a_{N-1}, s_N)) = P_1(s_1) \times q_{d_1(s_1)}(a_1) \times P_1(s_2 \mid s_1, a_1) \times q_{d_2(h_2)}(a_2)$

$\times P_2(s_3 \mid s_2, a_2) \times \cdots \times P_{N-1}(s_N \mid s_{N-1}, a_{N-1})$

so, $(\Omega, B(\Omega), P^\pi)$ describes the MDP

Note: (X_t, Y_t) will not be markovian here as actions will depend on history

P^π for $\pi \in \Pi^{MR}$:

$P^\pi(X_{t+1} = s_{t+1} \mid Z_t = h_t)$

$= \sum_{a \in A_{s_t}} P^\pi(X_{t+1} = s_{t+1}, Y_t = a \mid Z_t = h_t)$ (law of total prob)

$= \sum_{a \in A_{s_t}} P^\pi(Y_t = a \mid Z_t = h_t) \times P^\pi(X_{t+1} = s_{t+1} \mid Y_t = a, Z_t = h_t)$ (conditional prob)

$= \sum_{a \in A_{s_t}} q_{d_t(s_t)}(a) P_t(s_{t+1} \mid s_t, a)$

this is s_t and not h_t as $\pi \in \Pi^{MR}$

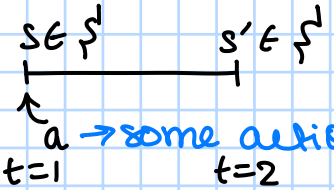
so, $\{X_t\}$ is a discrete time Markov chain (DTMC) by defn

Exe: (X_t) is DTMC, then is $(f(X_t))$ a markov chain? $f: S \to S'$
Ans: if f is one-one it is DTMC but if not one-one then not a markov chain

Note: $(X_t, r_t(X_t, Y_t))$ is called a markov reward process, not necessary a MC

one period markov decision problem:

$N=2$



At time 2, we get a reward $v(s')$ for some $v: S \to \mathbb{R}$

If we took action a' at $t=1$, then we get reward $r_1(s,a')$

$$\sum_{j \in S'} v(j) P_1(j|s,a') = \text{Average reward} = \mathbb{E}_s^{\pi}[v(X_2)|X_1=s] = \mathbb{E}_s^{\pi}[V(X_2)]$$

total reward $= r_1(s,a') + \sum v(j) P_1(j|s,a')$
so given s state at $t=1$, the action to take is
(action to take at $t=1$)

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

MR case:
$q_1, q \in P(A_s)$

$$\max_q \sum_a q(a)\left[r_1(s,a) + \sum_j v(j) P(j|s,a)\right] = \text{maximum reward}$$

Note: $\max_{a' \in A_s} \{r_1(s,a') + \sum v(j) P_1(j|s,a')\} = \max_{q \in P(A_s)} \left\{\sum_a q(a)\left[r_1(s,a) + \sum_j v(j) P(j|s,a)\right]\right\}$
so in this case randomisation does not help

so if we start at s:

$$\mathbb{E}_s^{\pi}[r_1(X_1,Y_1) + v(X_2)]$$ ← fix state, this is the avg reward we get if we are using π policy

Note: For finite case we can find an action to maximise reward, but for infinite case not always true

Sequence of rewards:

$N<\infty$, finite horizon case $R=(R_1, \ldots R_N)$ sequence of rewards
we can compare random variables by:
U, V, U stochastically dominate V if
$P(U>t) \geq P(V>t)$ $\forall t$

now if $U=(U_1 \ldots U_N)$
$V=(V_1 \ldots V_N)$ then we can compare U, V by:

$\mathbb{E}[f(U_1 \dots U_N)] \geq \mathbb{E}[f(V_1 \dots V_N)]$ for any f s.t $U \leq V \Rightarrow f(U) \leq f(V)$ (this is coordinate wise)

(this is a very strong condition)

Eg: π_1, π_2 we want to compare two policies, $N=2$

$P^{\pi_1}(R=(0,0)) = P^{\pi_1}(R=(1,0)) = P^{\pi_1}(R=(0,1)) = P^{\pi_1}(R=(1,1)) = 1/4$

$P^{\pi_2}(R=(0,0)) = P^{\pi_2}(R=(1,1)) = 1/2$

$f(U,V) = U+V$ then $\mathbb{E}^{\pi_1}[f] = (2)\times\frac{1}{4} + 1\times\frac{1}{4} + 1\times\frac{1}{4} = 1$

$\mathbb{E}^{\pi_2}[f] = 2\times\frac{1}{2} = 1$

by this fusion π_1 and π_2 are same 2

$f(U,V) = \max\{U,V\}$

$\mathbb{E}^{\pi_1}[f] = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4}$

$\mathbb{E}^{\pi_2}[f] = \frac{1}{2}$ so π_1 is better than π_2

similarly for $f = \min\{U,V\}$ π_2 is better than π_1

Note: The above example is why we cannot use the method of fusions to compare

Expected total reward $\pi \in \Pi^{HR}$

$r_N(s)$: reward at time N

at time N we get

under policy π $\rightarrow$ $\mathbb{E}_s^{\pi}\left[\sum_{t=1}^{N-1} r_t(X_t, Y_t) + r_N(X_N)\right]$ = reward by starting at s and using policy π

s: initial state s (start from s) = total expected reward under π, given $X_1 = s$

$= V_N^{\pi}(s)$

goal of decision maker is to maximise $V_N^{\pi}(s)$

or $\sup_{\pi \in \Pi^{HR}} V_N^{\pi}(s)$ $\rightarrow$ goal of policy maker to pick $\pi \in \Pi^{HD}$ with $\max V_N^{\pi}(s)$

8th Aug:

Quiz1 - Aug 29

$X_1 = s$, given $\pi \in \Pi^{HR}$ policy π (Random), we get $(\Omega, \mathcal{B}(\Omega), P^\pi)$ where $N < \infty$ (finite horizon problem)

we get $r_t(X_t, Y_t)$

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

(final reward depends upon final state)

$\mathbb{E}_s^\pi$: conditioned to start at s

$V_N^\pi(s)$: initial state, $\pi \in \Pi^{HR}$: if fixed

Expected total reward starting at $X_1 = s$ under π

S, A are discrete

if $\sup_{s,a} |r_t(s,a)| < \infty$, $\sup |r_N(s)| < \infty$ then $V_N^\pi(s)$ is always finite, $V_N^*(s) = \sup_{\pi \in \Pi^{HR}} V_N^\pi(s)$ (so $\exists$ supremum)

Defn: (optimal policy) a policy $\pi^* \in \Pi^{HR}$ is optimal if

$$V_N^*(s) = V_N^{\pi^*}(s) \geq V_N^\pi(s) \quad \forall s \in S, \ \forall \pi \in \Pi^{HR}$$

(i.e $V_N^{\pi^*}(s) = \sup V_N^\pi(s)$)

(so optimal policy does not always exist)

Defn: (ε optimal) π is epsilon optimal if

$$V_N^\pi(s) + \varepsilon \geq V_N^*(s) \quad \forall s \in S$$ (this will always exist)

Note: From defn of supremum, there will always be an Epsilon optimal strategy

$\mathbb{E}[V_N^\pi(X_1)]$ for some strategies also get maximised by π^*

now, $\mathbb{E}_s^\pi \left[\sum_{t=1}^{N-1} \lambda^{t-1} r_t(X_t, Y_t) + \lambda^{N-1} r_N(t_N) \right]$ expected total discount reward

so we want to give more reward to first few time (since use of γ rate or λ) ($\lambda < 1$, but not comp)

Note: Expected total discounted reward and normal reward will be maximised similarly

$V_N^\pi(s)$ for given π ($\pi \in \Pi^{HD}$):

Defn: (reward-to-go function) $U_t^\pi : H_t \to \mathbb{R}$

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

($h_t = (s_1, a_1, \ldots, s_{t-1}, a_{t-1}, s_t)$)

Expected reward from t to N given history till t h_t

Now at $T = t$ action taken is $d_t(h_t)$ ($d_t(h_t) = a_t \in A_{s_t}$)

and given s_t so

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

now $\mathbb{E}[f(X) | X] = f(X)$

$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$ → tower property

$\mathcal{F} \supseteq \mathcal{G}$ 2 σ-fields (h_{t+1}, h_t)

$$\mathbb{E}\left[\mathbb{E}[X|\mathcal{F}]\,|\,\mathcal{G}\right] = \mathbb{E}[X|\mathcal{G}]$$

$$X = \sum_{n=t+1}^{N-1} r_n(X_n, Y_n) + r_N(X_N)$$

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

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

only thing that is random

Now, set $t = N$, $U_N^\pi(h_N) = r_N(s_N)$ $\forall h_N$ ($h_N = (h_{N-1}, a_{N-1}, s_N)$)

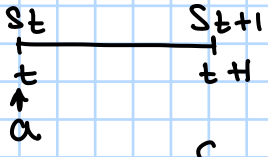
Substitute $t-1$ in place of t to evaluate (1) to compute $U_{t-1}(h_{t-1})$ $\forall h_{t-1} \in H_{t-1}$

and then stop when $t=1$, $U_1^\pi(s)$ $\forall s \in S$

and so, $V_N^\pi(s) = U_1^\pi(s)$

Note: The above algorithm to calculate $V_N^\pi(s)$ is called policy evaluation algorithm

Optimal reward-to-go:



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

↑ not for given π like before

↑ choose action that maximizes the above

$$U_N(h_N) = r_N(s_N)$$

Note: The above equation is optimality equation / Bellman equation

Now, let $U_t^*(h_t) = \sup_{\pi \in \Pi^{HR}} U_t^\pi(h_t)$

But we still have to show U_t^* will satisfy the Bellman eqn

Theorem: Let U be a solution to optimality equation, then $U_t(h_t) = U_t^*(h_t)$ $\forall h_t \in H_t$, in particular $U_1(s) = V_N^*(s)$ $\forall s \in S$

proof:

we will show that $U_n(h_n) \geq U_n^*(h_n)$ $\forall h_n$ and then we will show that $\forall \varepsilon > 0$, $\exists \pi \in \Pi^{HP}$ s.t

$$U_n^\pi(h_n) + (N-n)\varepsilon \geq U_n(h_n)$$

then, $U_n^\pi(h_n) + (N-n)\varepsilon \geq U_n(h_n) \geq U_n^*(h_n)$

but by defn of sup: $U_n^*(h_n) \geq U_n^\pi(h_n)$

$$\Rightarrow U_n^*(h_n) + (N-n)\varepsilon \geq U_n(h_n) \geq U_n^*(h_n)$$

and we let $\varepsilon \to 0$

12th Aug:

$$u_t(h_t) = \sup_{a \in A_{s_t}} \left\{ r_t(s_t, a) + \sum_{j \in S} p_t(j \mid s_t, a) \cdot u_{t+1}(h_t, a, j) \right\}, \quad h_t \in H_t$$

$$u_N(h_N) = r_N(s_N) \leftarrow \text{terminal reward at } t = N$$

} Bellman equation

$$u_t^*(h_t) = \sup_{\pi \in \Pi^{HR}} u_t^{\pi}(h_t), \quad h_t \in H_t$$

Theorem: Let u_t be a solution to bellman equation then (Bellman equation / optimality eqn)

$$u_t(h_t) = u_t^*(h_t) \quad \forall h_t \in H_t, \forall t$$

in particular $u_1(s) = u_1^*(s) = v_N^*(s) = \sup_{\pi} v_N^{\pi}(s)$

↙ roll to eqn ↙ sup value

proof:

1) $u_n(h_n) \geq u_n^*(h_n) \quad \forall h_n, n \in \mathbb{N}$
2) $\forall \varepsilon > 0, \exists \pi \in \Pi^{HD}$ s.t. $u_n^{\pi}(h_n) + \varepsilon(N-n) \geq u_n(h_n)$

} we have to show this two to prove theorem

from 1, 2 we get: $u_n^*(h_n) + \varepsilon(N-n) \geq u_n^{\pi}(h_n) + \varepsilon(N-n) \geq u_n(h_n) \geq u_n^*(h_n)$

as $\varepsilon \to 0$

$\Rightarrow u_n^*(h_n) = u_n(h_n)$ from sandwich theorm

now let's show 1:

at $t = N$, $u_N(h_N) = r_N(s_N) = u_N^{\pi}(h_N) \quad \forall \pi$

$u_N(h_N) = u_N^*(h_N) \quad \forall h_N$

if 1 holds for $n+1, \ldots, N$ then for n:

$$u_n(h_n) = \sup_a \left\{ r_n(s_n, a) + \sum_{j \in S} p_n(j \mid s_n, a) \times u_{n+1}(h_n, a, j) \right\}$$

$$\geq \sup_a \left\{ r_n(s_n, a) + \sum_{j \in S} p_n(j \mid s_n, a) \times u_{n+1}^*(h_n, a, j) \right\}$$

$$\geq \sup_a \left\{ r_n(s_n, a) + \sum_{j \in S} p_n(j \mid s_n, a) \times u_{n+1}^{\pi}(h_n, a, j) \right\} \forall \pi \in \Pi^{HD}$$

$$\geq \sum_a q_{d_n(h_n)}(a) \left(r_n(s_n, a) + \sum_{j \in S} p_n(j \mid s_n, a) \times u_{n+1}^{\pi}(h_n, a, j) \right\}$$

$$= u_n^{\pi}(h_n) \quad \forall \pi \in \Pi^{HD}$$

from defn $\forall \pi \in \Pi^{HD}$

$\Rightarrow u_n(h_n) \geq u_n^{\pi}(h_n) \quad \forall \pi \in \Pi^{HD}$

$\Rightarrow \sup_{\pi} u_n(h_n) \geq \sup_{\pi} (u_n^{\pi}(h_n))$

$\Rightarrow u_n(h_n) \geq u_n^*(h_n)$

now let's show 2:

for $t = N$

$u_N^{\pi}(h_N) = u_N(h_N)$, so true $\forall \varepsilon > 0$

now let $\varepsilon > 0$

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

$r_n(s_n, d_n(h_n)) + \sum p_n(j \mid s_n, d_n(h_n)) \, u_{n+1}(h_n, d_n(h_n), j) + \varepsilon$

$\geq u_n(h_n)$

↑ this is from $u_t(h_t) = \sup\{\}$ def of sup

↑ By defn sup over action

$\forall n, h_n$, choose $d_n(h_n)$

↑ $\exists$ action $= d_n(h_n)$ s.t. $(\,) + \varepsilon \geq (\,)$

So, given that we can find action

now this is true for N as

$u_N^{\pi}(h_N) = r_N(s_N) = u_N(h_N)$

if true for $n+1, \ldots, N$ then for n:

$$U_n^\pi(h_n) = r_n(s_n, d_n(h_n)) + \sum_{j\in S} P_n(j|s_n, d_n(h_n))\, U_{n+1}^\pi(h_n, d_n(h_n), j)$$

$$\geq r_n(s_n, d_n(h_n)) + \sum_{j\in S} P_n(j|s_n, d_n(h_n))\, U_{n+1}(h_n d_n(h_n), j) - \varepsilon(N-(n+1))$$

$$\geq \underbrace{U_n(h_n)}_{\text{from depth of } \pi \text{ we took}} - \varepsilon - \varepsilon(N-(n+1))$$

$$= U_n(h_n) - \varepsilon[1+N-n+1]$$

$$\Rightarrow U_n^\pi(h_n) + \varepsilon(N-n) \geq U_n(h_n)$$

so, $\forall \varepsilon > 0$, $\exists \pi \in \Pi^{HD}$ s.t. above is true

Note: If "sup" are attained in $U_t(h_t) = \sup_{a\in A_{s_t}} \{r_t(s_t,a) + \sum P_t(j|s_t,a) U_{t+1}(h_t,a,j)\}$ $\forall h_t, \forall t$ then there is an HD optimal policy

Theorem: Let U_t^* be solution to Bellman equation, then:

1) $\forall t$, $U_t^*(h_t)$ depends on h_t only through s_t

2) $\forall \varepsilon > 0$, $\exists \pi \in \Pi^{MD}$ that is ε optimal

3) If all sup are attained, then $\exists$ an optimal MD policy

proof: At N, $U_N^*(h_N) = r_N(s_N)$ $\forall h_N \in H_N$

so 1 holds for N

now assume it holds for $n+1, \cdots N$

$$U_n^*(h_n) = \sup_a \{r_n(s_n,a) + \sum_j P_n(j|s_n,a) U_{n+1}^*(j)\}$$

so $U_n^*(h_n)$ depends just on s_n, 2,3 follows trivially from previous theorm and backward induction

Note: From previous 2 theorms, we get:

$$\sup_{\pi\in\Pi^{HR}} V_N^\pi(s) = \sup_{\pi\in\Pi^{MD}} V_N^\pi(s), \quad s\in S$$

Dynamic programming principle / Backward induction algo:

we are going to assume that all sup are attained

step 1: Set $t=N$ and $U_N^*(s_N) = r_N(s_N)$ $\forall s_N \in S$

Step 2: Substitute $t-1$ in place of t and compute $U_t^*(s)$ from bellman equation and $A_{s,t}^* \in \operatorname{argmax}_{a\in A_{s_t}}$ (Bellman eqn)

step 3: stop at $t=1$ and so:

$d_t(s_t) = A_{s_t,t}^*$, $s_t \in S$, $t<N$ is optimal MD policy

Note: $K^2 L^{|A|}(N-1)$ = total multiplication in above algorithm

(K^2: $|S|$; $L^{|A|}$)

$(L^K)^{N-1}$ = total number of MD policies $d_i: S \to A$ (K, L)

Eg: Sell a stock, if a stock price crosses time we sell as one of the policies so problem is when to stop the policy

Optimal stopping problems:

$T = \{1, 2, \dots, N\}$, $N < \infty$, at everytime we can either continue or stop

$S = S' \cup \{\Delta\}$ ← absorbing state

S' = state space of DTMC

$$A_s = \begin{cases} \{c\} & s = \Delta \\ \{c, Q\} & s \in S' \end{cases}$$

$$r_t(s,a) = \begin{cases} -f_t(s) & ; s \in S', a = c \\ g_t(s) & ; s \in S', a = Q \\ 0 & ; s = \Delta, a = c \end{cases}$$

$r_N(s) = h(s)$ → terminal reward

$$P_t(j|s,a) = \begin{cases} P_t(j|a) & ; s \in S', a = c \\ 1 & ; s \in S', a = Q, j = \Delta \\ 1 & ; s = \Delta, j = \Delta, a = c \\ 0 & ; \text{otherwise} \end{cases}$$

$t < N$

19th Aug:

Optimal stopping problem:

$S = S' \cup \{\Delta\}$

$A_S = \{C, Q\}$ $s \in S'$

$A_\Delta = \{C\}$ for Δ state

$$r_t(s,a) = \begin{cases} -f_t(s) & ; s \in S', a = C \\ g_t(s) & ; s \in S', a = Q \\ 0 & ; s = \Delta, a = C \end{cases}$$

$-f_t(s)$: price of continuing; $g_t(s)$: reward of quitting

$$P_t(j|s,a) = \begin{cases} P_t(j|s) & ; j \in S', s \in S', a = C \\ 1 & ; j = \Delta, s \in S', a = Q \\ 0 & ; \text{otherwise} \end{cases}$$

The secretary problem:

N candidates for job offer, $\{1,2 \ldots N\}$, company decides to choose the candidate or reject

objective is to maximize the prob of choosing the best candidate

$T = \{1,2 \ldots N\}$ $S' = \{0,1\}$ (1: Best so far; 0: not the best candidate compared to last 1)

$A_S = \{C, Q\}$ (C: give offer to next person; Q: reject)

$f_t(S) = 0$ (continuing No cost)

$g_t(0) = 0$

$g_t(1) = \text{Prob}\{\text{Best candidate is in first } t\}$

$= \dfrac{\text{Number of sets of size } t \text{ which include the best}}{\text{NO. of sets of subsets } t}$

$= \dfrac{{}^{n-1}C_{t-1}}{{}^{n}C_{t}} = \dfrac{\frac{(n-1)!}{(n-t)!(t-1)!}}{\frac{n!}{(n-t)!t!}} = \dfrac{t}{n}$

$\Rightarrow g_t(1) = \dfrac{t}{N}$

the terminal reward: $h(1) = 1$, $h(0) = 0$

$P_t(1|0) = \dfrac{1}{t+1}$ ← uniformly picking one guy from $t+1$

$P_t(1|1) = \dfrac{1}{t+1}$ ← so $P_t(1,s) = \dfrac{1}{1+t}$ for $s \in \{0,1\}$

$P_t(0|S) = \frac{t}{1+t}$ → optimal prob for $S \in \{0,1\}$

other probability same as optimal stopping problem

optimal solution for secretary problem:

let $U_t^*(1)$ = maximum probability of choosing the best candidate given that you are at the best so far (at time t)

$U_t^*(0)$ = maximum probability of choosing the best candidate given we are not at best so far (at time t)

terminal $t=N$:

$U_N^*(1) = h(1) = 1$

$U_N^*(0) = h(0) = 0$

$U_N^*(\Delta) = 0$

for $t<N$: $U_t^*(0) = \max_{Q,C} \{V^Q, V^C\}$ ← value of choosing Q or C

$= \max\{ g_t(0) + U_{t+1}^*(\Delta), -f_t(0) + P_t(1|0)U_{t+1}^*(1) + P_t(0|0)U_{t+1}^*(0)\}$

cost to go function if we quit/continue given $S_t = 0$

$= \max\{0, \frac{1}{t+1}U_{t+1}^*(1) + \frac{t}{t+1}U_{t+1}^*(0)\} = \frac{1}{t+1}U_{t+1}^*(1) + \frac{t}{1+t}U_{t+1}^*(0)$

$U_t^*(1) = \max\{g_t(1) + U_{t+1}^*(\Delta), -f_t(1) + P_t(1|1)U_{t+1}^*(1) + P_t(0|1)U_{t+1}^*(0)\}$

cost to go function if we quit/cont given $S_t = 1$

$= \max\{\frac{t}{N}, \frac{1}{t+1}U_{t+1}^*(1) + \frac{t}{1+t}U_{t+1}^*(0)\}$

$= \max\{\frac{t}{N}, U_t^*(0)\}$

$U_t^*(\Delta) = 0 + U_{t+1}^*(\Delta) \Rightarrow U_t^*(\Delta) = 0 \quad \forall t \in \{1, \ldots N\}$

prob of reaching Δ is 1

so in case of $S_t = 0$, we always continue as it is maximising $U_t^*(0)$

for case of $S_t = 1$, we have 2 choices of continuing or quitting

for some $\tau \in \{1,2 \ldots N\}$ if we iterate $t < \tau$ we continue always after that for smallest $t > \tau$, s.t $S_t = 1$, we do action Q

formally, suppose it is optimal to some $\tau \in \{1,2 \ldots N\}$ let the optimal be C

$U_\tau^*(1) > \tau/N$ (as we are continuing we pick $U_t^*(1) = U_t^*(0)$)

$\Rightarrow U_\tau^*(0) = U_\tau^*(1)$

then for $t = \tau - 1$

$U_{\tau-1}^*(0) = \frac{1}{\tau}U_\tau^*(1) + \frac{\tau-1}{\tau}U_\tau^*(0)$

$= U_\tau^*(1) > \tau/N > \tau-1/N$

this is just to show that if $\exists$ such τ then all values less is continue

$\Rightarrow U_{\tau-1}^*(0) = U_{\tau-1}^*(1) > (\tau-1)/N$ so we continue

So, optimal policy: $\exists \tau \in \{1,2 \dots N\}$ s.t when $t \le \tau$, the optimal action is C

when $t > \tau$, if $S_t = 0 \rightarrow$ continue
$S_t = 1 \rightarrow$ Quit

Lemma: If $N > 2$, then $\tau \ge 1$

proof: If Not then $\tau < 1$, or $\forall t$, $U_t^*(1) = \frac{t}{N}$ we quit once we see $S_t = 1$

$$U_t^*(0) = \frac{1}{t+1}\cdot\frac{t+1}{N} + \frac{t}{t+1} U_{t+1}^*(0)$$

$$\Rightarrow U_t^*(0) = \frac{1}{N} + \frac{t}{t+1} U_{t+1}^*(0)$$

$$U_N^*(0) = 0, \quad U_{N-1}^*(0) = \frac{1}{N}$$

$$U_{N-2}^*(0) = \frac{1}{N} + \frac{N-2}{N-1}\left(\frac{1}{N}\right) = \frac{N-2}{N}\left[\frac{1}{N-1} + \frac{1}{N-2}\right]$$

$$\vdots$$

$$U_t^*(0) = \frac{t}{N}\left[\frac{1}{t} + \frac{1}{t+1} + \cdots + \frac{1}{N-1}\right]$$

$$U_1^*(0) = \frac{1}{N}\left[1 + \frac{1}{2} + \cdots \frac{1}{N-1}\right] > \frac{1}{N} = U_1^*(1) \ge U_1^*(0) \quad (\because \text{max value})$$

$\Rightarrow U_1^*(0) > U_1^*(0)$ this is a contradiction and so

$\tau \ge 1$

Note: So for $N > 2$, $\tau \ge 1$ i.e $U_1^*(0) = U_2^*(0) = \cdots = U_\tau^*(0) \rightarrow$ continuing $t < \tau$

and $U_1^*(0) = U_1^*(1)$

$\vdots \qquad \vdots$

$U_\tau^*(0) = U_\tau^*(1)$

for $t > \tau$: $U_t^*(1) = \frac{t}{N}$

$t > \tau$: $U_t^*(0) = \frac{t}{N}\left[\frac{1}{t} + \frac{1}{t+1} + \cdots + \frac{1}{N-1}\right]$

$$\frac{t}{N}\left[\frac{1}{t} + \cdots + \frac{1}{N-1}\right] > \frac{t}{N}$$

$\Rightarrow \frac{1}{t} + \cdots + \frac{1}{N-1} > 1$ for all this t we continue no matter what

for $\frac{1}{t} + \cdots + \frac{1}{N-1} < 1$ we Quit if $S_t = 1$

Note: $\max\left\{t \in \{1, \dots N\} \,\middle|\, \left[\frac{1}{t} + \frac{1}{t+1} + \cdots + \frac{1}{N-1}\right] > 1\right\} = \tau$

$$1 \approx \left[\frac{1}{t} + \cdots + \frac{1}{N-1}\right] = \int_{\tau(N)}^{N-1} \frac{1}{x}\,dx = 1$$

$$\Rightarrow \log(N-1) - \log(\tau(N)) = 1$$

$$\Rightarrow \log(N-1) - \log(e) = \log(\pi(N))$$

$$\Rightarrow \frac{N-1}{e} = \pi(N)$$

as $N \to \infty$ $\lim_{N\to\infty} \frac{\pi(N)}{N} = e^{-1}$

for N large it will be $\frac{N}{e}$

22nd Aug:

Infinite horizon:

$0<\lambda<1$, $\lim_{N\to\infty} \mathbb{E}_s^{\pi}\left[\sum_{t=1}^{N} \lambda^{t-1} r_t(X_t, Y_t)\right]$ ← Discounted reward case → we want to maximise this

$\lim_{N\to\infty} \frac{1}{N} \mathbb{E}_s^{\pi}\left[\sum_{t=1}^{N} r_t(X_t, Y_t)\right]$ ← Avg reward case → we want to maximise this

Discounted case:

Expected total discounted reward $= \lim_{N\to\infty} \mathbb{E}_s^{\pi}\left[\sum_{t=1}^{N} \lambda^{t-1} r_t(X_t, Y_t)\right]$

first assumption: $0 \le \lambda < 1$ is discount factor, we assume S, A are discrete (atmost countable)

second assumption: time homogenious rewards & t.p. functions i.e $r(s,a) = r_t(s,a)$ ($r_t(S_t, A_t)$) (transition prob functions) $P(j|s,a) = P_t(j|s,a)$ ← third assumption

final assumption: $\sup_{s\in S} \sup_{a\in A_s} |r(s,a)| < \infty$ ← ∞ limit always exist

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

we get λ as discount factor as if Horizon $\sim \nu \sim$ geom$(1-\lambda)$ (motivation only why λ)

ν $P(\nu = k) = \lambda^{k-1}(1-\lambda)$ $k \ge 1$

$$\begin{aligned}
&\mathbb{E}_s^{\pi}\left[\mathbb{E}_\nu\left(\sum_{t=1}^{\nu} r(X_t, Y_t)\right)\right] \\
&= \mathbb{E}_s^{\pi}\left[\sum_{n=1}^{\infty} \lambda^{n-1}(1-\lambda) \sum_{t=1}^{n} r(X_t, Y_t)\right] \\
&= \mathbb{E}_s^{\pi}\left[\sum_{t=0}^{\infty} \left(\sum_{n=t}^{\infty} \lambda^{n-1}(1-\lambda)\right) r(X_t, Y_t)\right] \quad \text{telescopic} \\
&= \mathbb{E}_s^{\pi}\left[\sum_{t=0}^{\infty} \lambda^{t-1} r(X_t, Y_t)\right]
\end{aligned}$$

Defn: (optimal policy in case of $\lambda, N \to \infty$) π^* is suitable optimal if $\sup_{\pi\in\pi^{HR}} v_\lambda^{\pi}(s) = v_\lambda^{\pi^*}(s)$

Defn: $v_\lambda^*(s)$ is value of MDP starting from s

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

$= \sum_{t=1}^{\infty} \lambda^{t-1} \underbrace{\mathbb{E}_s^{\pi}[r(X_t, Y_t)]}_{\text{only depends of prob dist of } X_t, Y_t}$

P^{π} is joint dist of all X_t, Y_t if we can construct (X_t', Y_t') s.t $(X_t, Y_t) \overset{d}{=} (X_t', Y_t')$ then $V_\lambda^{\pi}(s)$ will not change even if (X_t', Y_t') MR

Theorem: let $\pi = (d_1, d_2, \dots) \in \pi^{HR}$, For every $s \in S$, $\exists\, \pi' = (d_1', d_2', \dots) \in \pi^{MR}$ s.t

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

consequently $V_\lambda^{\pi}(s) = V_\lambda^{\pi'}(s)$

proof: $d_t' : \mathcal{S}' \longrightarrow P(A)$

$q_{d_t'(j)}(a) = \sum_{(x_1 x \cdots x x_{t-1}) \in H_{t-1} \text{ s.t } (x_1)=s} p^\pi(Y_t = a \mid X_t = j, (x_1 x y_1 x \cdots x x_{t-1}) \in H_{t-1}, X_1 = s)$ (we use induction)

$\Rightarrow q_{d_t'(j)}(a) = p^\pi(Y_t = a \mid X_t = j, X_1 = s)$ (this is our defn)

then eqn true for $n=1$ (trivial), if true till $n-1$ then for n:

$$p^{\pi'}(X_n = j, Y_n = a \mid X_1 = s) = p^{\pi'}(X_n = j \mid X_1 = s)\, p^{\pi'}(Y_n = a \mid X_n = j, X_1 = s)$$

$$= p^{\pi'}(X_n = j \mid X_1 = s)\, p^{\pi}(\cdots \mid \cdots) \leftarrow \text{from defn}$$

$$\overset{?}{=} p^{\pi}(X_n = j \mid X_1 = s)\, p^{\pi}(\cdots \mid \cdots)$$

$$= p^{\pi}(X_n = j, Y_n = a \mid X_1 = s)$$

so if we know $p^\pi(X_n = j \mid X_1 = s) = p^{\pi'}(X_n = j \mid X_1 = s)$ then we are done

$$p^\pi(X_n = j \mid X_1 = s) = \sum_{i,a} p^\pi(X_n = j, X_{n-1} = i, Y_{n-1} = a \mid X_1 = s)$$

$$= \sum_{i,a} p^\pi(X_{n-1} = i, Y_{n-1} = a \mid X_1 = s)\, p^\pi(X_n = j \mid X_{n-1} = i, Y_n = a, X_1 = s)$$

$$= \sum_{i,a} p^{\pi'}(\cdots)\, p^{\pi'}(\cdots) \leftarrow P(j \mid i, a) \text{ same as } P_t(j \mid i, a) = P(j \mid i, a)$$

$$= p^{\pi'}(X_n = j \mid X_1 = s)$$

(from induction)

so π' depends on s

Exe: if $X_1 \overset{d}{=} \nu$, then $\exists \pi'$ (depending on ν) s.t conclusion of theorem holds (does not depend on s) → done done

Note: $\sup_{\pi \in \Pi^{HR}} v_\lambda^\pi(s) = \sup_{\pi \in \Pi^{MR}} v_\lambda^\pi(s)$

Policy evaluation:

we want to compute $V_\lambda^\pi(s)$ $\quad |\mathcal{S}| = d$

$V_\lambda^\pi = (V_\lambda^\pi(s), s \in \mathcal{S}')$ ← vector of dim d

$d \in D^{MD}$ $\quad r(s, d(s)) = r_d(s)$

decision rule one time

$r_d = (r(s, d(s)), s \in \mathcal{S}')$

$r_d = \left(\sum_a r(s,a)\, q_{d(s)}(a), s \in \mathcal{S}'\right)$ $d \in D^{MD}$ ← Expected value of reward

matrix

$(P_d)_{s,j} = P(j \mid s, d(s))$ matrix $d \times d$ as $|\mathcal{S}'| = d$

v $\quad \|v\| = \sup_{s \in S} |v(s)|$ (our defn)

$v^\pi : \mathcal{S}' \longrightarrow \mathbb{R}$ and let V: set of all bounded fn from $\mathcal{S}' \longrightarrow \mathbb{R}$

norm for V is the sup norm

now $(V, \|\ \|)$ is Banach space and $P: V \to V$ is a linear map (← cauchy conv ⇒ conv)

($V_\lambda^\pi \in V$ as $V_\lambda^\pi : S \to \mathbb{R}$ and $r_d \in V$ as $r_d : S \to \mathbb{R}$)

$$\|P\| = \sup\left\{ \|Pv\|, \|v\| \le 1 \right\}$$ (matrix)

then let $BL(V)$ = space with matrix norm s.t all maps are bounded i.e $\sup < \infty$

both V & $BL(V)$ are vector spaces and both are Banach space

now, $\pi = (d_1, \dots)$

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

$$= \underbrace{\sum_a r(s,a) q_{d_1(s)}(a)}_{r_{d_1}(s)} + \underbrace{\lambda \mathbb{E}_s^\pi[r(X_2, Y_2)]}_{\lambda \sum_j r(j, d_2(j)) P(j \mid s, d_1(s))} + \dots$$

$$= \lambda \sum_j r_{d_2}(j)(P_{d_1})_{s,j}$$ ← right multiplication

$$= \lambda (P_{d_1} \cdot r_{d_2})_s$$

the third term: $\lambda^2 (P_{d_1} P_{d_2} r_{d_3})_s$

so, $V_\lambda^\pi = r_{d_1} + \lambda P_{d_1} r_{d_2} + \lambda^2 P_{d_1} P_{d_2} r_{d_3} + \dots$

$$= r_{d_1} + \lambda P_{d_1}\left[r_{d_2} + \lambda P_{d_2} r_{d_3} + \dots \right]$$

$$\Rightarrow V_\lambda^\pi = r_{d_1} + \lambda P_{d_1} \cdot V_\lambda^{\pi'}$$

where $\pi' = (d_2, d_3, \dots)$

Note: If $\pi \in \pi^{SR}$ i.e $\pi = (d, d, \dots)$

$$\Rightarrow V_\lambda^\pi = r_d + \lambda P_d V_\lambda^\pi$$

Exe: If $X_1 \overset{d}{=} \nu$, then $\exists \pi'$ (depending on ν) s.t conclusion of theorem holds

Ans:

Now as $X_1 \sim \nu$ i.e it is a random variable instead of $X_1 = s$, we make it an $\mathbb{E}$ value

so as $V_\lambda^\pi(s)$ known for $X_1 = s$

$\Rightarrow \mathbb{E}(V_\lambda^\pi(X_1))$ is what we want to maximise, in all cases where we see s, so theorem still will hold

26th Aug:

$V_\lambda^\pi(s) = \mathbb{E}_s^\pi\left[\sum_{t=1}^{\infty} \lambda^{t-1} r(X_t, Y_t)\right]$ ← only depends on marginal so $V_\lambda^\pi(s) = V_\lambda^{\pi'}(s)$ for some $\pi' \in \Pi^{MR}$ (proved previously)

$\pi = (d, d, \dots)$ then $V_\lambda^\pi = r_d + \lambda P_d V_\lambda^\pi$ $(\pi \in \Pi^{SR})$ (vector wise)

Note: let L_d be a map s.t

$$L_d : \mathcal{V} \to \mathcal{V} \quad (V_\lambda^\pi \in \mathcal{V})$$

$$\text{s.t } L_d v = r_d + \lambda P_d v$$

$r_d(s) = \sum r(s,a) q_{d(s)}(a)$ as $\sup_{s \in S} |r_d(s)| \le M$ because of bounded rewards

$\Rightarrow r_d \in \mathcal{V}$ (Bounded reward assumption),

now, $\|P_d v\| \le \|P_d\| \cdot \|v\|$ ($\because \|P\| = \sup\{\|Pv\|; \|v\| \le 1\}$ from our defn $\Rightarrow \|P\| \ge \frac{\|Pv\|}{\|v\|}$ to make $\|v\| \le 1$ i.e $\frac{v}{\|v\|}$)

and $\|P_d\| = \sup_{s \in S} \sum_j |P(j|s)| = 1$ ← this is equal to our defn of $\|P\| = \sup\{\|Pv\|; \|v\| \le 1\}$ then $\|P\| = \sup_{s \in S} \sum_j |P(j|s)|$

$\Rightarrow \|P_d v\| \le \|v\|$

$\Rightarrow P_d v \in \mathcal{V}$ as $\|v\|$ is bounded so $P_d v$ is also bounded

now as $r_d \in \mathcal{V}$, $P_d v \in \mathcal{V}$ we have $r_d + \lambda P_d v \in \mathcal{V}$ (as $\mathcal{V}$ is a vector space)

so $L_d v = r_d + \lambda P_d v \quad L_d : \mathcal{V} \to \mathcal{V}$

and now $L_d V_\lambda^\pi = V_\lambda^\pi$

Note: V_λ^π is a fixed point of L_d

Theorem: let $d \in D^{MR}$, then V_λ^π is the unique fixed point of L_d, moreover $V_\lambda^\pi = (I - \lambda P_d)^{-1} r_d$

proof:

for some $v \in \mathcal{V}$ if $L_d v = r_d + \lambda P_d v = v$

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

$\Rightarrow (I - \lambda P_d) v = r_d \quad \lambda < 1$

now as P_d is a stochastic matrix $\Rightarrow \max|\mu| = 1$ (μ = eigenvals of P_d)

as $\|P_d\| = \max(|\mu|)$ by defn of our eigenvalues and norm

$\|P_d\| = 1$

$\Rightarrow 1 = \max(|\mu|)$

and so $\max|\lambda\eta| < 1$ as $\lambda < 1$

& so 0 cannot be eigenvalue of $I - \lambda P_d$

as 0 is not an eigenvalue of $I - \lambda P_d \Rightarrow I - \lambda P_d$ is invtible

as $\det(I - \lambda P_d) = \lambda_1 \lambda_2 \dots \lambda_n \neq 0$ → all eigvalues multiplied

so $I - \lambda P_d$ is invertible

and so $v = (I - \lambda P_d)^{-1} r_d$ ↓ unique

Defn: (spectral radius) let $T : \mathcal{V} \to \mathcal{V}$, then $\sigma(T) = \sup\{|\mu| \mid (\mu I - T)$ is not invtible$\}$

Note: $\sigma(T) = \lim_{n \to \infty} \|T^n\|^{\frac{1}{n}}$

Note: if $\sigma(T) < 1$ then $I - T$ is invertible and $(I-T)^{-1} = \sum_{n=0}^{\infty} T^n$ } alternate way to show $(I - \lambda P_d)$ is invertible

we can compute $(I-T)^{-1}$ using same

now, $\sigma(\lambda P_d) = \lambda < 1$ so $(I - \lambda P_d)$ is invertible

$\left(\because \sigma(\lambda P_d) = \lim_{n \to \infty} \lambda^{n} {}^{1}/_{n} = \lambda \right)$

Now, $\sup_{a \in A_s} \left\{ r(s,a) + \lambda \sum_{j \in S} p(j|s,a) v(j) \right\} = v(s)$, is the bellman equation for infinite horizon problem also called optimality equation

as infinite horizon

now, $L : V \to V$ be the bellman operator

s.t $Lv = \sup_{d \in D^{MD}} \{ (r_d + \lambda P_d v) \}$

s.t $Lv(s) = \sup_{d \in D^{MD}} \sup_{a \in A_s} ((r_d + \lambda P_d v)(s))$

Note: $Lv = \sup_{d \in D^{MR}} \{ r_d + \lambda P_d v \}$ from previous lemma (as sup over not matter)

suppose $v \geqslant Lv$ componentwise

then we want to show $v \geqslant v_\lambda^*$ ← optimal reward $v_\lambda^* = \sup_{\pi} v_\lambda^{\pi}$

as $v \geqslant Lv \geqslant r_{d_1} + \lambda P_{d_1} v$ from defn of L

$\Rightarrow v \geqslant r_{d_1} + \lambda P_{d_1} (r_{d_2} + \lambda P_{d_2} \cdot v)$

$\vdots$

$\geqslant r_{d_1} + \lambda P_{d_1} r_{d_2} + \cdots + \lambda^{n-1} P_{d_1} \cdots P_{d_{n-1}} r_{d_n} + \lambda^n P_{d_1} \cdots P_{d_n} v$

now $\pi = (d_1, d_2 \ldots) \in \Pi^{MR}$

then $v_\lambda^\pi = r_{d_1} + \lambda P_1 r_{d_2} + \cdots$

$\forall \pi \in \Pi^{MR}$

then $v \geqslant v_\lambda^\pi - \sum_{k=n}^{\infty} \lambda^k P_{d_1} \cdots P_{d_{k-1}} r_{d_{k+1}} + \lambda^n P_{d_1} \cdots P_{d_n} v$

$\Rightarrow v - v_\lambda^\pi \geqslant \lambda^n P_{d_1} \cdots P_{d_n} v - \sum_{k=n}^{\infty} \lambda^k P_{d_1} \cdots P_{d_k} r_{d_{k+1}}$

this is true $\forall n$ ← as $P_{d_1} \times P_{d_2}$ is also stochastic matrix

now $|\lambda^n P_{d_1} \cdots P_{d_n} v| \leqslant \lambda^n \times 1 \times 1 \times \cdots \times 1 \times \|v\|$ as $n \to \infty$ $\lambda^n \to 0$ so given $\varepsilon > 0$, $\exists N$ s.t $-\varepsilon e \leqslant \lambda^n \leqslant \varepsilon e$

→ all one vector $e = (1, 1 \ldots)$

$\sum \lambda^k P_{d_1} \cdots P_{d_k} r_{d_{k+1}} \leqslant M \sum \lambda^k \underbrace{P_{d_1} \cdots P_{d_k} \cdot e}_{e \text{ as stochastic matrix}}$

$= Me \frac{\lambda^n}{1-\lambda}$

$\exists N' \geqslant N$ s.t $\frac{\lambda^n M}{1-\lambda} < \varepsilon$ as $\lambda^n \to 0$ as $n \to \infty$

so for sufficient large n we get:

$v - v_\lambda^\pi \geqslant -2\varepsilon e \quad \forall \varepsilon > 0$

then as $\varepsilon \to 0$

$\Rightarrow v - v_\lambda^\pi \geqslant 0$

$\Rightarrow v \geqslant v_\lambda^\pi \quad \forall \pi \in \Pi^{MR}$

$\Rightarrow v \geq \sup_{\pi \in \Pi^{MR}} v_\lambda^\pi = v_\lambda^*$

$\Rightarrow v \geq v_\lambda^*$

so if $v \geq Lv \Rightarrow v \geq v_\lambda^*$

now, we want to find $v \leq Lv$ then does $v \leq v_\lambda^*$, for this:

let $\varepsilon > 0$, then

$v \leq r_d + \lambda P_d v + \varepsilon . e$ for some d (from defn of supremum)

$\Rightarrow (I - \lambda P_d) v \leq r_d + \varepsilon e$

$\Rightarrow v \leq (I - \lambda P_d)^{-1} (r_d + \varepsilon e)$ ($\because (I - \lambda P_d)^{-1} = \sum_{n=0}^{\infty} (\lambda P_d)^n$ all positive so positive and can go on right)

as $(I - \lambda P_d)^{-1}(r_d + \varepsilon e) = v_\lambda^{(d,d,\ldots)} + \frac{\varepsilon . e}{1-\lambda}$ (first term is straight from prev theorm, last term from $\sum_{n=0}^{\infty} (\lambda^n) \underbrace{P_d^n}_{e \text{ as } P_d \text{ is stochastic}} x e = \frac{e}{1-\lambda}$)

$\Rightarrow v \leq \sup_{\pi \in \Pi^{MR}} v_\lambda^\pi + \frac{\varepsilon . e}{1-\lambda}$ as ε is arbitary

$\Rightarrow v \leq v_\lambda^*$

Theorm: (a) if $v \geq Lv$ then $v \geq v_\lambda^*$
(b) if $v \leq Lv$ then $v \leq v_\lambda^*$
(c) if $v = Lv$ then $v = v_\lambda^*$

proof: The proof follows trivially from calculations done above

Note: (c) part does not imply L has a fixed point, as we know if $Lv = v$ then $v = v_\lambda^*$ (we know if L has a fixed point then it is v_λ^*)

Defn: $T : V \to V$ is said to be a contraction if $\exists \lambda < 1$ s.t $\|Tu - Tv\| \leq \lambda \|u - v\|$ $\forall u, v \in V$

Theorm: (Banach fixed point theorm) Let $T : V \to V$ be a contraction then:
(a) $\exists$ a unique v^* s.t $Tv^* = v^*$
(b) Starting with any $v^0 \in V$, $v^{n+1} = Tv^n = T^2 v^{n-1} \ldots = T^{n+1} v^0$ for $n \geq 1$ converges to v^*

2nd sept:

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

$(d, d, \ldots)$

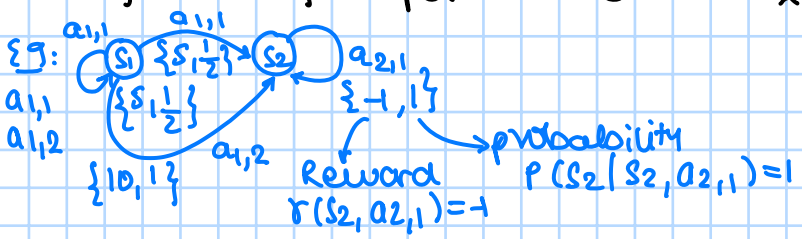
$L_d v = r_d + \lambda P_d v$

& we get $v = \sup_{d \in D^{MD}} \{r_d + \lambda P_d v\}$ Bellman equation

Lv

Lv s.t $v = Lv \Rightarrow v = v_\lambda^*$ ← from theorem done

so if v is a fixed point then $v = v_\lambda^*$



$d(s_1) = a_{1,1}$

$\pi = (d, d, \ldots)$ then we want to find $v_\lambda^\pi = v$

$d(s_2) = a_{2,1}$ → only one possible action

$v(s_2) = -1 + \lambda \begin{pmatrix} 0 \\ 1 \end{pmatrix} (V(s_1), V(s_2))$

$V(s_2) = -1 + \lambda V(s_2)$

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

now, $v(s_1) = 5 + \lambda \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix} (V(s_1) \;\; V(s_2))$

$= 5 + \lambda \left[\frac{V(s_1)}{2} + \frac{1}{2} V(s_2)\right]$

$\Rightarrow V(s_1) = \frac{10 - 11\lambda}{(2-\lambda)(1-\lambda)}$

$v_\lambda^\pi = (I - \lambda P_d)^{-1} r_d$ can also be used

Note: in above eg for optimality equation as only $a_{2,1}$

we get $V_\lambda^*(s_2) = \frac{-1}{1-\lambda}$

and for s_1: $v(s_1) = \max\left\{5 + \lambda\left[\frac{1}{2} v(s_1) + \frac{v(s_2)}{2}\right], 10 + \lambda v(s_2)\right\}$

if $\lambda = 0$ then $v(s_1) = \max\{5, 10\} = 10$

so $v(s_2) = -1$

$v(s_1) = 10$ optimal action is $a_{1,2}$

$\lambda = \frac{1}{2}$: $V(s_2) = -2$

$v(s_1) = \max\left\{5 - \frac{1}{2} + \frac{1}{4} v(s_1), 9\right\}$

$v(s_1) = 9$ as $\frac{3}{4} v(s_1) = \frac{13}{2} \Rightarrow v(s_1) = 6$ in $a_{1,1}$ case

so for $\lambda = 1/2$ $a_{1,2}$ is optimal action

Exe: Find an optimal action at s_1 as a function of λ → done down

now if $v = Lv$ then $v = v_\lambda^*$ (theorem done previously)

Defn: $T: V \to V$ is said to be a contraction if $\exists \lambda < 1$ s.t $\|Tu - Tv\| \leq \lambda \|u - v\|$ $\forall u, v \in V$

Theorem: (Banach fixed point theorem) Let $T: V \to V$ be a contraction then: where V is a banach space

(a) $\exists$ a unique v^* s.t $Tv^* = v^*$

(b) Starting with any $v^0 \in V$, $v^{n+1} = Tv^n = T^2 v^{n-1} \dots = T^{n+1} v^0$ for $n \geq 1$ converges to v^*

$\|v^n - v^*\| \to 0$ as $n \to \infty$

proof: we fix m, n ←—|—|—→ we want to compare

n $\quad$ $m+n$

$$\|v^{n+m} - v^n\| \leq \sum_{k=0}^{m-1} \|v^{n+k+1} - v^{n+k}\| \quad \text{by triagle identity}$$

$$= \sum_{k=0}^{m-1} \|T^{n+k} v^1 - T^{n+k} v^0\|$$

$$\leq \sum_{k=0}^{m-1} \lambda^{n+k} \|v^1 - v^0\| \quad (\because \text{By contraction property})$$

$$= \lambda^n \|v^1 - v^0\| \sum_{k=0}^{m-1} \lambda^k$$

$$= \lambda^n \|v^1 - v^0\| \left[\frac{(1-\lambda^m)}{(1-\lambda)} \right]$$

$$\Rightarrow \|v^{n+m} - v^n\| \leq \lambda^n \|v^1 - v^0\| \left[\frac{(1-\lambda^m)}{(1-\lambda)} \right]$$

as $\lambda < 1$, $\forall \varepsilon > 0, \exists N$ s.t $\|v^{n+m} - v^n\| < \varepsilon$ $\forall m \geq 1, n \geq N$

so $\{v^n\}$ is a cauchy sequence and as V is a banach space (every cauchy conv) so $\{v^n\}$ converges, let the conv be v^* ← this is from completness

so $\|v^n - v^*\| \to 0$ as $n \to \infty$

and now we want to show $Tv^* = v^*$

$$\|Tv^* - v^*\|$$
$$= \|Tv^* - v^n + v^n - v^*\|$$
$$\leq \|Tv^* - v^n\| + \|v^n - v^*\|$$
$$= \|Tv^* - Tv^{n-1}\| + \|v^n - v^*\|$$
$$= \lambda \|v^* - v^{n-1}\| + \|v^n - v^*\|$$

as $n \to \infty$ $\|Tv^* - v^*\| \to 0$

$\Rightarrow Tv^* = v^*$

also v^* is unique as if not then $\exists u^*$ s.t $Tu^* = u^*$, $Tv^* = v^*$ & $u^* \neq v^*$

$$\|Tu^* - Tv^*\| \leq \lambda \|u^* - v^*\| < \|u^* - v^*\|$$

$\Rightarrow \|u^* - v^*\| < \|u^* - v^*\|$ this is a contradiction

so, $u^* = v^*$

& so v^* is a unique fixed point

Note: $L : \mathcal{V} \longrightarrow \mathcal{V}$ has to be a contraction for us to have a fixed point

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

Theorm: Suppose all sup are attained, then L is a contraction. By all sup we mean for all $V \in \mathcal{V}$, sup is attained

proof: we want to show that $\|LU - LV\| \leq \lambda \|U - V\|$ (this is same λ as our objective function)

let's fix a state s, for some $V, U \in \mathcal{V}$, let

$LV(s) \geq LU(s)$, s fixed (WLOG)

let $a_s^* \in \underset{a \in A_s}{\operatorname{argmax}} \{ r(s,a) + \lambda \sum_{j \in S} P(j|s,a) V(j) \}$

$$LU(s) \geq \{ r(s, a_s^*) + \lambda \sum_{j \in S} P(j|s,a_s^*) U(j) \} \quad (\because \text{plugging } a_s^* \text{ for } U(s))$$

where $LV(s) - LU(s) \geq 0$

$$LV(s) - LU(s) \leq \{ r(s, a_s^*) + \lambda \sum P(j|s,a_s^*) V(j) \} - \{ r(s, a_s^*) + \lambda \sum P(j|s,a_s^*) U(j) \}$$

$$\Rightarrow LV(s) - LU(s) \leq \lambda \sum P(j|s,a_s^*)[V(j) - U(j)]$$

$$\leq \lambda \left(\sum P(j|s,a_s^*) \right) \|V - U\| \quad (\text{max of } j \in S \text{ as } \mathcal{V} \text{ is equipped with sup norm})$$

$$= \lambda \|V - U\|$$

so, $|LV(s) - LU(s)| \leq \lambda \|V - U\| \quad \forall s \in S$

taking sup we get $\|LV - LU\|$ from defn of $\mathcal{V}$

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

$\Rightarrow L$ is a contraction

Ex: show that when sup need not be attained, still L is a contraction → done down

Note: $L_d V = r_d + \lambda P_d V$, we can show that L_d is also a contraction and from banach fixed point theorm we get $\exists$ unique fixed point

now, we showed that $L : \mathcal{V} \longrightarrow \mathcal{V}$, then $\exists$ a unique fixed point

$LV^* = V^*$ & $V^* = V_\lambda^*$ (from prev theorm)

where $V_\lambda^* = \sup_\pi V_\lambda^\pi$

Theorm: A policy $\pi^* \in \Pi^{HR}$, then it is optimal iff $V_\lambda^{\pi^*}$ is a solution to the optimality equations

proof: ($\Rightarrow$) as π^* is optimal, $V_\lambda^{\pi^*} = V_\lambda^* = L V_\lambda^*$ (By theorm done before)

$$\Rightarrow V_\lambda^{\pi^*} = L V_\lambda^{\pi^*}$$

$\Rightarrow V_\lambda^{\pi^*}$ is solution to optimality eqn

($\Leftarrow$) As L has unique fixed point

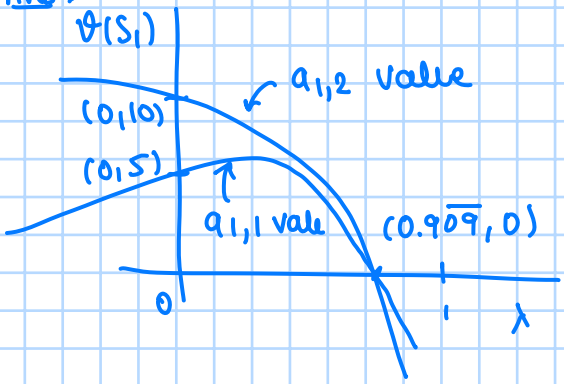
if $V_\lambda^{\pi^*} = L V_\lambda^{\pi^*}$

$\Rightarrow V_\lambda^{\pi^*} = V_\lambda^*$ i.e π^* is optimal

as from prev theorm V_λ^* is only fixed point

Exe: Find an optimal action at s_1 as a function of λ

Ans:



for $\lambda > 0.90\overline{9}$ $a_{1,1}$ is optimal action
or else $a_{1,2}$ is optimal

Exe: Show that when sup need not be attained, still L is a contraction

Ans: similar to when sup is attained
for some $U, V \in \mathcal{V}$ and some $s \in S$
wlog $LV(s) \geq LU(s)$ then

and as sup need not be attained $\forall \varepsilon > 0 \ \exists a^*$ s.t

$$r(s,a^*) + \lambda \sum_{j \in S} p(j|s,a^*)\, v(j) + \varepsilon \geq \sup_{a \in A_s} \{ r(s,a) + \lambda \textstyle\sum p(j|s,a)\, v(j) \}$$

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

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

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

($\|\cdot\|$: sup norm)

$$\Rightarrow |LV(s) - LU(s)| \leq \lambda \|V - U\| + \varepsilon$$

true for all s

so

$$\|LV - LU\| \leq \lambda \|V - U\| + \varepsilon$$

($\|\cdot\|$: sup norm) and so for $\varepsilon \to 0$
we get: $\|LV - LU\| \leq \lambda \|V - U\|$ so L is a contraction

9th Sept:

so far, $L: V \to V$ is a bellman operator, showed that L is a contraction and then from banach fixed point theorem:

(we cannot comment on uniqueness of policy π, only $v_\lambda^* \in V$)

$\exists$ a unique fixed point $v^* \in V$ for L

s.t $v^* = v_\lambda^*$

$L v_\lambda^* = v_\lambda^*$

where $v_\lambda^* = \underbrace{\sup_d \{r_d + \lambda P_d v_\lambda^*\}}_{\text{Bellman}}$

Def: $d^* \in D^{MD}$ is said to be conserving if $d^* \in \underset{d \in D^{MD}}{\operatorname{argmax}} \{r_d + \lambda P_d v_\lambda^*\}$

where if we fix s, $a_s^* \in \underset{a \in A_s}{\operatorname{argmax}} \{r(s,a) + \lambda \sum_j p(j|s,a) v_\lambda^*(j)\}$

i.e $d^*(s) = a_s^*$

Theorem: Suppose all suprema are attained, then ($L: V \to V$ all sups attained)

a) $\exists$ a conserving rule d^*

b) If d^* is conserving, then $\pi = (d^*, d^*, \dots) \in \Pi^{SD}$ is optimal

proof: a. is trivial as all suprema are attained, so d^* exist

b. It is suffices to show $v_\lambda^\pi = v_\lambda^*$, for this

as $v_\lambda^* = L v_\lambda^* = \sup_d \{r_d + \lambda P_d v_\lambda^*\}$

For stationary policy $= r_{d^*} + \lambda P_{d^*} v_\lambda^* = L_{d^*}(v_\lambda^*)$

$L_d: V \to V$ s.t $L_d(v) = r_d + \lambda P_d v$

so, $L_{d^*}(v_\lambda^*) = v_\lambda^*$

$\Rightarrow v_\lambda^*$ is a fixed point for L_{d^*}

we had shown that L_d has a unique fixed point v_λ^π for $\pi = (d, d, \dots)$

so $v_\lambda^* = v_\lambda^\pi$ for $\pi = (d^*, d^*, \dots)$

so as $v_\lambda^\pi = v_\lambda^* \Rightarrow \pi = (d^*, d^*, \dots)$ is optimal and $\pi \in \Pi^{SD}$

Note: In particular, it means that if all sups are attained in $L: V \to V$ then:

$\sup_{\pi \in \Pi^{HR}} v_\lambda^\pi(s) = \sup_{\pi \in \Pi^{SD}} v_\lambda^\pi(s)$ (as shown $\sup_{\pi \in \Pi^{HR}} v_\lambda^\pi(s) = v_\lambda^*(s)$)

Now lets see when all sups are not attained

Theorem: For any $\varepsilon > 0$, $\exists$ an ε optimal SD policy

proof: we have $L v_\lambda^* = v_\lambda^*$ as unique fixed point of $L: V \to V$

$L v = \sup_{d \in D^{MD}} \{r_d + \lambda P_d v\}$ (depn of Bellman operator)

given $\varepsilon > 0$, $\exists d_\varepsilon \in D^{MD}$ s.t

$r_{d_\varepsilon} + \lambda P_{d_\varepsilon} v_\lambda^* \geq L v_\lambda^* - \underbrace{\varepsilon e(1-\lambda)}_{e = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \text{ all one vector}}$ (From sup property)

consulting ε as $\varepsilon(1-\lambda)$

$\Rightarrow r_{d_\varepsilon} + \lambda P_{d_\varepsilon} v_\lambda^* \geq v_\lambda^* - \varepsilon e(1-\lambda)$ ($\because L v_\lambda^* = v_\lambda^*$)

$\Rightarrow r_d + \varepsilon e(1-\lambda) \geq (I - \lambda P_d) v_\lambda^*$

as $(I-\lambda P_d)^{-1}$ multiplied previous inequality (∵ done before)

$\Rightarrow v_\lambda^* \leq \underbrace{(I-\lambda P_d)^{-1} r_d} + \varepsilon(1-\lambda)(I-\lambda P_d)^{-1} e$

v_λ^π when $\pi = (d_\varepsilon, d_\varepsilon, \ldots) \in \Pi^{SD}$

and $(I-\lambda P_{d_\varepsilon})^{-1} e = \sum_{n=0}^{\infty} (\lambda P_{d_\varepsilon})^n e$

$= (1 + \lambda + \lambda^2 + \cdots) e$

$= \frac{1}{1-\lambda} e$

so, $v_\lambda^* \leq v_\lambda^\pi + \varepsilon . e$

this means, $(d_\varepsilon, d_\varepsilon, \ldots) \in \Pi^{SD}$ is an ε optimal policy

Note: $\sup_{\pi \in \Pi^{HR}} v_\lambda^\pi(s) = \sup_{\pi \in \Pi^{SD}} v_\lambda^\pi(s)$ in all cases (∵ $\varepsilon \to 0$, we get this)

Value iteration algorithm:

we want to algorithmically find an optimal policy given r_d, P_d, λ

we will assume all suprema are attained

Algorithm:

1. pick $v^0 \in V$, fix $\varepsilon > 0$, set $n = 0$
2. while $(\| v^n - L v^n \| \geq \varepsilon \left(\frac{1-\lambda}{2\lambda}\right))$:
   $v^{n+1} = L v^n$
   $n = n+1$

$v^{n+1}(s) = \max_{a \in A_s} \{ r(s,a) + \lambda \sum_j P(j|s,a) v^n(j) \} \quad \forall s \in S$

3. as $\| v^n - L v^n \| < \varepsilon \left(\frac{1-\lambda}{2\lambda}\right)$ now, $v^{n+1} = L v^n$

   choose $d_\varepsilon \in \operatorname{argmax}_d \{ r_d + \lambda P_d v^{n+1} \}$
4. return (v^{n+1}, d_ε)

we now have to show ① v^{n+1} is close to v_λ^*

② d_ε is ε optimal

Theorem: For the above algorithm, given $\{v^n\}$ by $v^{n+1} = L v^n$

a) $v^n \to v_\lambda^*$ as $n \to \infty$

b) $\exists N$ s.t $\| v^{n+1} - v^n \| < \varepsilon \left(\frac{1-\lambda}{2\lambda}\right)$ holds $\forall n \geq N$

c) $\pi^\varepsilon = (d_\varepsilon, d_\varepsilon, \ldots) \in \Pi^{SD}$ is ε-optimal

d) $\| v^{n+1} - v_\lambda^* \| < \frac{\varepsilon}{2}$ whenever $\| v^{n+1} - v^n \| < \varepsilon \left(\frac{1-\lambda}{2\lambda}\right)$ holds

Proof: a) is trivial from banach fixed point theorem

b) from banach fixed point, as given $\{v^n\}$ is a cauchy sequence, as V is a banach space, from defn $\forall \varepsilon > 0$, $\exists$ such N

for c) and d), $\|v_\lambda^{\pi_\varepsilon} - v_\lambda^*\|$

$$= \|v_\lambda^{\pi_\varepsilon} + v^{n+1} - v^{n+1} - v_\lambda^*\|$$

$$\leq \|v_\lambda^{\pi_\varepsilon} - v^{n+1}\| + \|v^{n+1} - v_\lambda^*\| \quad \text{——— ①}$$

($\because$ triagle inequality)

where $v_\lambda^{\pi_\varepsilon} = L_{d_\varepsilon} V_\lambda^{\pi_\varepsilon}$ from L_{d_ε} constructing

$L v^{n+1} = L_{d_\varepsilon} v^{n+1}$ from defn of L and calculation of d_ε

→ showing this $< \varepsilon/2$ will prove part d

$$\|v_\lambda^{\pi_\varepsilon} - v^{n+1}\| = \|L_{d_\varepsilon} v_\lambda^{\pi_\varepsilon} - v^{n+1} + Lv^{n+1} - Lv^{n+1}\|$$

$$\leq \|L_{d_\varepsilon} v_\lambda^{\pi_\varepsilon} - \underbrace{Lv^{n+1}}_{L_{d_\varepsilon} v^{n+1}}\| + \underbrace{\|Lv^{n+1} - Lv^n\|}_{\leq \lambda \|v^{n+1} - v^n\|}$$

$$\leq \lambda \|v_\lambda^{\pi_\varepsilon} - v^{n+1}\| + \lambda \|v^{n+1} - v^n\|$$

from contraction property of L_{d_ε} — from contraction property of L

$$\Rightarrow (1-\lambda)\|v_\lambda^{\pi_\varepsilon} - v^{n+1}\| < \lambda(\varepsilon)\left(\frac{1-\lambda}{2\lambda}\right) \quad (\because \text{from algorithm } \|v^{n+1} - v^n\| < \varepsilon\left(\frac{1-\lambda}{2\lambda}\right))$$

$$\Rightarrow \|v_\lambda^{\pi_\varepsilon} - v^{n+1}\| < \frac{\varepsilon}{2} \quad \text{——— ②}$$

so, now from ②, ① becomes: $\|v_\lambda^{\pi_\varepsilon} - v_\lambda^*\| < \frac{\varepsilon}{2} + \|v^{n+1} - v_\lambda^*\|$

now, $\|v^{n+1} - v_\lambda^*\| = \|v^{n+1} - Lv_\lambda^*\| \quad (\because Lv_\lambda^* = v_\lambda^*)$

$$\leq \|\underbrace{v^{n+1}}_{Lv^n} - Lv^{n+1}\| + \|Lv^{n+1} - Lv_\lambda^*\|$$

$$\leq \lambda \|v^n - v^{n+1}\| + \lambda \|v^{n+1} - v_\lambda^*\|$$

$$\Rightarrow \|v^{n+1} - v_\lambda^*\| \leq \frac{\lambda}{1-\lambda}\|v^n - v^{n+1}\|$$

$$< \frac{\varepsilon}{2} \quad (\because \text{stopping condition}) \quad \text{——— ③}$$

so from ②, ③, ① becomes:

$$\|v_\lambda^{\pi_\varepsilon} - v_\lambda^*\| < \varepsilon \text{ and } \|v^{n+1} - v_\lambda^*\| < \frac{\varepsilon}{2}$$

12th Sept:

Value Iteration:

$$v^{n+1} = L v^n$$
$$\| v^{n+1} - v^n \| < \frac{\varepsilon}{d\lambda}(1-\lambda)$$
$$d_\varepsilon \in \underset{d}{\operatorname{argmax}} \{ r_d + \lambda P_d v^{n+1} \}$$

Theorem: (a) If $u, v \in V$, $v \geq u$ (for all components) then $Lv \geq Lu$

(b) If $Lv^N \geq v^N$ for some N, then $v^{N+m+1} \geq v^{N+m}$ $\forall m \geq 0$
and similarly
$Lv^N \leq v^N$ for some N, then $v^{N+m+1} \leq v^{N+m}$ $\forall m \geq 0$

proof: (a) let $\delta \in \underset{d}{\operatorname{argmax}} \{ r_d + \lambda P_d u \}$, then

$$Lu = r_\delta + \lambda P_\delta u$$

also as P_δ was non negative entries (transition matrix)
$P_\delta u \leq P_\delta v$

$$\Rightarrow Lu = r_\delta + \lambda P_\delta u \leq r_\delta + \lambda P_\delta v \leq \max_d \{ r_d + \lambda P_d v \} = Lv$$

$$\Rightarrow Lu \leq Lv$$

(b) $v^{N+m+1} = L^{m+1} v^N$
$= L^m (Lv^N)$
as $Lv^N \geq v^N$ for some N

$\Rightarrow L(Lv^N) \geq L(v^N)$ (by part (a))
$\vdots$

$\Rightarrow L^m (Lv^N) \geq L^m (v^N)$

$\Rightarrow v^{m+N+1} \geq v^{N+m}$ $\forall m \geq 0$

Exe: Show the $\leq$ part of (b) → done down

let $r(s,a) \geq 0$, $Lv = \max_d \{ r_d + \lambda P_d v \}$, then we would make $v^0 = 0$ as

$$Lv^0 = \max_d \{ r_d + 0 \} \geq 0$$
$$\Rightarrow Lv^0 \geq v^0$$
$$\Rightarrow v^1 \geq v^0$$

so, $v^n \uparrow v_\lambda^*$ v^n is monotonically inc to v_λ^*

similarly if $r_d \leq 0$ then $v^0 = 0$ gives
$v^n \downarrow v_\lambda^*$ v^n is monotonically dec to v_λ^*

Rates of convergence:

$\{v^n\}$ is a seq of vectors, then $v^n \longrightarrow v^*$ if $\| v^n - v^* \| \longrightarrow 0$ as $n \longrightarrow \infty$ (globally)

Defn: $\{v^n\}$ conv at order (at least α) if $\alpha > 0$ and $\| v^{n+1} - v^* \| \leq K \| v^n - v^* \|^\alpha$ $\forall n$ for some constant $K > 0$ (globally)

Defn: (rate of conv) Smallest such K is called rate of convergence

globally

Note: $\alpha = 1$: linear convergence (order 1)

Defn: (Asymptotic rate of convergence) $\overline{\lim_{n\to\infty}} \left(\frac{\|v^n - v^*\|}{\|v^0 - v^*\|} \right)^{\frac{1}{n}}$ gives order = 1

we call it AARC

order

$\{v^n\}$ converges $O(f(n))$ if $\overline{\lim_{n\to\infty}} \frac{\|v^n - v^*\|}{f(n)} < \infty$

and we say $v^n = v^* + O(f(n))$

Note: If $f(n) = \beta^n$, $\beta < 1$, then we say geometric convergence at rate β

Theorem: $\{v^n\}$ from value iteration, then the following global conv. property holds:

(i) conv. is linear with rate λ
(ii) AARC is λ
(iii) convergence is $O(\lambda^n)$

proof:

(i) $\|v^{n+1} - v_\lambda^*\| = \|Lv^n - Lv_\lambda^*\|$

$\leq \lambda \|v^n - v_\lambda^*\|$, $\forall n$

so, conv is atleast linear, if $v^0 = v_\lambda^* + k \cdot e$, $k \neq 0$ vector can be any

then $v^1 = \max_d \{r_d + \lambda P_d (v_\lambda^* + k \cdot e)\}$

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

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

$v^1 = \lambda k \cdot e + v_\lambda^*$

$\Rightarrow v^1 - v_\lambda^* = \lambda (v^0 - v_\lambda^*)$ $\forall n$ with above choice of v^0

so, $\alpha = 1$ and rate $= \lambda$

(ii) $\|v^n - v_\lambda^*\| \leq \lambda^n \|v^0 - v_\lambda^*\|$ ——— ①

so, $\overline{\lim_{n\to\infty}} \left(\frac{\|v^n - v^*\|}{\|v^0 - v_\lambda^*\|} \right)^{\frac{1}{n}} \leq \lambda$ but from (a) $v^0 = v_\lambda^* + k.e$ $k \neq 0$ makes AARC $\not< \lambda$

(AARC)

so, AARC $= \lambda$

(iii) $\overline{\lim_{n\to\infty}} \frac{\|v^n - v^*\|}{\lambda^n} = \|v^0 - v^*\| < \infty$ (from ①) so, converges $O(\lambda^n)$

gauss-seidel value iteration:

$\{s_1, \ldots s_N\}$ states

algorithm:

① Fix $\varepsilon > 0$, $v^0 \in V$, $n > 0$

② $v^{n+1}(s_j^\circ) = \max_{a \in A_{s_j^\circ}} \{ r(s_j, a) + \lambda \left[\sum_{i<j}^{\circ} P(s_i | s_j, a) v^{n+1}(i^\circ) + \sum_{i \geq j}^{\circ} P(s_i^\circ | s_j^\circ, a) v^n(i) \right] \}$

sequentially for $j = 1, 2, \ldots$

③ If $\|v^{n+1} - v^n\| < \frac{\varepsilon}{2\lambda}(1-\lambda)$ then stop and return v^{n+1} and d_ε, otherwise continue to step 2

let d attain max, then

$P_d = P_d^L + P_d^U$ ← upper diagonal including diagonals
← lower triagular

Step 2 says: $v^{n+1} = (I - \lambda P_d^L)^{-1} r_d + (I - \lambda P_d^L)^{-1} \lambda P_d^U . v^n$

Exe: show $v^{n+1} = (I - \lambda P_d^L)^{-1} r_d + (I - \lambda P_d^L)^{-1} \lambda P_d^U . v^n$

As: as d attain max $v^{n+1} = r_d + \lambda P_d^L v^{n+1} + \lambda P_d^U v^n$, then follows

so, $I - \lambda P_d = Q_d - R_d$, and $Q_d^{-1} \geq 0$ and $R_d \geq 0$ then this is called regular splitting

in value iteration (V.I) $Q_d = I$ $R_d = \lambda P_d$
in Gauss seidel V.I $Q_d = I - \lambda P_d^L$
$R_d = \lambda P_d^U$

$$v^{n+1} = Q_d^{-1} r_d + Q_d^{-1} R_d v^n \quad (\text{G.S V.I})$$

Theorem: let (Q_d, R_d) be a regular splitting of $I - \lambda P_d$, $\forall d \in D^{MD}$, assume

$$\alpha = \sup_d \|Q_d^{-1} R_d\| < 1$$

a) For any $v^0 \in V$, the iteration scheme

$v^{n+1} = \max_d \{Q_d^{-1} r_d + Q_d^{-1} R_d v^n\} = T v^n$, converges to v_λ^*

b) v_λ^* is unique fixed point of T

c) linear convergence, with rate $\leq \alpha$
AARC $\leq \alpha$
converges $O(\beta^n)$ with $\beta \leq \alpha$

proof: T is a contraction is already seen in problem sheet
$\|Tu - Tv\| \leq \alpha \|u - v\|$ is also seen in problem sheet
$\Rightarrow \exists$ unique v^* s.t $Tv^* = v^*$ (Banach fixed point theorem)
sufficient to show $v^* = v_\lambda^*$

$v^* = T v^*$
as $v^* \geq Q_d^{-1} r_d + Q_d^{-1} R_d v^*$, $\forall d$ (By defn of max)

$\Rightarrow (I - Q_d^{-1} R_d) v^* \geq Q_d^{-1} r_d$
$\Rightarrow v^* \geq (I - Q_d^{-1} R_d)^{-1} Q_d^{-1} r_d$
$\Rightarrow v^* \geq \underbrace{((Q_d)(I - Q_d^{-1} R_d))^{-1}}_{Q_d - R_d} r_d$ ($\because \sup_d \|Q_d^{-1} R_d\| < 1$ so $(I - Q_d^{-1} R_d)^{-1}$ is all positive and in addition)

$A^{-1} B^{-1} = (AB)^{-1}$

$\Rightarrow v^* \geq (I - \lambda P_d)^{-1} r_d$

$\Rightarrow v^* \geq v_\lambda^\pi$ for $\pi = (d, d, \ldots) \in \pi^{SD}$

$\Rightarrow v^* \geq \sup_{\pi \in \pi^{SD}} v_\lambda^\pi = v_\lambda^*$

$\Rightarrow v^* \geq v_\lambda^*$, now as $\exists d^*$ s.t $v^* = r_{d^*} + \lambda P_{d^*} v^* = L_{d^*} v^*$
$L_{d^*} v^* = v^* \Rightarrow v^* = v_\lambda^{(d,d,\ldots)} d^* v^*$

as L_{d^*} has a unique fixed point

now, $\exists d^*$ s.t $V^* = \vartheta_\lambda^{(d^*, d^*, \dots)}$

$\Rightarrow \vartheta^* = \vartheta_\lambda^*$

part (c) follows from previous theorem

Note: (Q_1, R_1) and (Q_2, R_2) two regular splittings of $I - \lambda P$ if $R_2 \leq R_1 \leq \lambda P$ then $\|Q_2^{-1} R_2\| \leq \|Q_1^{-1} R_1\|$

then $R_1 = \lambda P_d$

$R_2 = \lambda P_{d^\nu}$

then $R_2 \leq R_1$ holds as we are just removing some values,

$$\Rightarrow \| (I - \lambda P_{d^\nu})^{-1} \lambda P_{d^\nu} \| \leq \| \lambda P_d \| = \lambda < 1 \quad \forall d$$

$$\Rightarrow \sup_d \| (I - \lambda P_{d^\nu})^{-1} \lambda P_{d^\nu} \| \leq \| \lambda P_d \| = \lambda < 1$$

$$\Rightarrow \alpha \leq \lambda < 1$$

Note: If $P_d \neq P_{d^\nu}$ then $\alpha < \lambda$

so, to recap, we saw $\vartheta_\lambda^\pi(s) = \mathbb{E}_s^\pi \left[\sum_{t=1}^{\infty} \lambda^{t-1} r(X_t, Y_t) \right]$ our objective function

① suffices to look at markovian randomised policies

② $L_d \vartheta = r_d + \lambda P_d \vartheta$, we saw L_d has unique fixed point $\vartheta_\lambda^{(d,d,\dots)}$

③ $L\vartheta = \vartheta$, if $\vartheta \geq L\vartheta$ then $\vartheta \geq \vartheta_\lambda^*$

$\vartheta \leq L\vartheta$ then $\vartheta \leq \vartheta_\lambda^*$

$\vartheta = L\vartheta$ then $\vartheta = \vartheta_\lambda^*$

④ L is a contraction, by banach fixed point, $\exists$ unique v^* s.t $Lv^* = v^*$

⑤ observing: $d^* \in \operatorname{argmax} \{ r_d + \lambda P_d \vartheta_\lambda^* \}$

⑥ V.I

Exe: Show the $\leq$ part of (b)

Ans: $L\vartheta^N = \vartheta^{N+1} \leq \vartheta^N$

$\Rightarrow L(\vartheta^{N+1}) \leq L(\vartheta^N)$

$\vdots$

$\Rightarrow L^m(\vartheta^{N+1}) \leq L^m(\vartheta^N)$

$\Rightarrow \vartheta^{N+m+1} \leq \vartheta^{N+m} \quad \forall m \geq 0$

26th Sept:

S, A, r(S,A), λ considering infinite horizon $X_1 \to Y_1 \to X_2 \to Y_2 \cdots$ realisation

X_t state → Y_t action

R_t = random reward

$R_t = r(X_t, Y_t)$

Note: $R_t = r(X_t, Y_t, X_{t+1})$ ← depends on future state

$$r(S,A) = \sum_{S' \in S} P(S' \mid S, A)\, r(S, A, S')$$

curely we look at: $\mathbb{E}_s^{\pi}\left[\sum \lambda^{t-1} R_t\right]$ ← called risk sensitive mdp

$d_S(a)$ → MD ← choose a particular action for every S

d is for every S, ∃ one action

$r_d = (r_d(S))_S = (r(S, d(S)) \ \forall S) \to$ vector

Similarly P_d is a matrix

if $\pi = (d, d, \ldots)$ we call it SMD or $d^\infty = \pi$

Note: we have already seen that it's same to work with SMD

Policy evaluation:

$\pi = (d_1, d_2, \ldots)$ v^π = vector s.t

$$v^\pi = (v_S^\pi, \forall S)$$

$$v^\pi = r_{d_1} + \lambda P_{d_1} v^{\tilde{\pi}} \qquad \tilde{\pi} = (d_2, d_3, \ldots)$$

if $\pi = d^\infty$:

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

$$\Rightarrow v^d = (I - \lambda P_d)^{-1} r_d$$

also, $v^* = \max_d (r_d + \lambda P_d v^*)$ we assume max as everything else is finite

Note: To solve a very long N finite horizon MDP, we can count that to $N \to \infty$ and solve, given λ is given

d is a decision rule, d^∞ is a policy (SMD), value of this:

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

$$\Rightarrow v^d = (I - \lambda P_d)^{-1} r_d \to \text{value under } d^\infty$$

we want to improve v^d:

$A^d \triangleq \arg\max_{\gamma} (r_\gamma + \lambda P_\gamma v^d)$ i.e if $d_1 \in A^d$ then

set of all decision rules

$d \in A^d$ $\quad$ $d \notin A^d$

$$d_1(S) \in \arg\max_{a_S \in A_S} \left(r(S, a_S) + \lambda \sum_{S'} P(S' \mid S, a_S)\, v^d(S')\right)$$

case I: $d \in \pi^d$ then $v^d = v^*$ ← optimal of objective function

case II: $d \notin \pi^d$, then d is not optimal

Let's pick $d_1 \in \pi^d$ s.t
if $d(s) \in \pi^d(s)$ for some s
then $d_1(s) = d(s)$ else we pick

now we have d_1, we get v^{d_1} under d_1^∞
so, $v^{d_1} > v^d$ → improvement step
(atleast one component strictly better)

Now, as $\pi^{d_1} \triangleq \underset{a}{\text{argmax}}(r_a + \lambda P_a v^{d_1})$

$d_2 \in \pi^{d_1}$ if $d_1 \notin \pi^{d_1}$

$\vdots$

$d_{k+1} \in \pi^{d_k}$ if $d_k \notin \pi^{d_k}$

algorithm will stop in finite its as finite d (SMD) and v cannot repeate as improving, never get back discarded dp

Lemma: $v^{d_1} > v^d$

proof:

$d_k \to v^{d_k}$

$d_{k+1} \in \pi^k$

$v^{d_{k+1}} \geq v^{d_k}$ → we want to show this

By defn of d_{k+1}:

$$r_{d_{k+1}} + \lambda P_{d_{k+1}} v^{d_k} \geq r_d + \lambda P_d v^{d_k} \quad \forall d$$

for $d = d_k$

$$\Rightarrow r_{d_{k+1}} + \lambda P_{d_{k+1}} v^{d_k} \geq r_{d_k} + \lambda P_{d_k} v^{d_k} = v^{d_k}$$

so we have to show, $r_{d_{k+1}} + \lambda P_{d_{k+1}} v^{d_k} \leq v^{d_{k+1}}$ and

so, $v^{d_k} \leq v^{d_{k+1}}$

now, $r_{d_{k+1}} + \lambda P_{d_{k+1}} v^{d_k} \geq v^{d_k}$

$$\Rightarrow r_{d_{k+1}} \geq (I - \lambda P_{d_{k+1}}) v^{d_k}$$

$$\Rightarrow (I - \lambda P_{d_{k+1}}) v^{d_{k+1}} \geq (I - \lambda P_{d_{k+1}}) v^{d_k}$$

as $r_{d_{k+1}} = (I - \lambda P_{d_{k+1}}) v^{d_{k+1}}$

now if A is positive matrix, $Ax \geq Ax'$ (true as $\lambda > 0 \ \forall \lambda$ of A) ← $Av \geq Av' \Rightarrow v \geq v'$

so, $v_{k+1} \geq v_k$

NOW, $v^d = r_d + \lambda P_d v^d$ is known to us, so let's consider any $v \in V$

s.t $v \geq r_d + \lambda P_d v \ \forall d$ ← component wise version

ie $v(s) \geq r(s,a) + \lambda \sum_{s'} P(s'|s,a) v(s') \quad \forall s, a$

now if $\nabla = \{v \mid v(s) \geq r(s,a) + \lambda \sum_{s'} P(s'|s,a) v(s'), \forall s,a\}$

then minimizer of $\nabla = v$ s.t v is optimal

and $v(s) = r(s,a) + \lambda \sum_{s'} P(s'|s,a) v(s') \quad \forall s,a$

this will always happen from $v_{k+1} \geq v_k$ argument ($\because$ finite many)

Note: our problem becomes: given r, P, λ:

for all $\alpha(s) > 0, \forall s, \quad \sum \alpha(s) = 1$

$\min_{v} \sum_{s} \alpha(s) v(s)$ is our solution s.t

v is: $v(s) \geq r(s,a) + \lambda \sum_{s'} P(s'|s,a) v(s') \quad \forall a, s$ } primal LP

This becomes LP as $(I - \lambda P_d) v \geq r_d$ is what we are solving, i.e

$(1 - \lambda \sum P(s'|s,a)) v(s) \geq r(s,a) \quad \forall s,a$

$\max_{(x(s,a))_{s,a}} \sum_{s,a} r(s,a) x(s,a)$ where $x(s,a) \geq 0 \quad \forall s,a$

$\sum_{a} x(s',a) - \lambda \sum_{s,a} P(s|s,a) x(s,a) = \alpha(s'), \forall s'$ } Dual of LP

now if v^* is solution for primal, then there will be a dual solution x^*

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

$x^* \rightarrow d^*(s,a) \triangleq \frac{x^*(s,a)}{\sum_{a'} x^*(s,a')} \quad \forall a, \forall s$

↖ if we choose basis solutions, then $d^{*\infty} \in \Delta^{MD}$

we claim $d^{*\infty}$ is optimal

$v^* \rightarrow$ value v of MDP

$(d^*)^\infty \rightarrow$ optimal policy for MDP

Note: $\mathbb{E}_{\alpha}^{\pi}[\sum \lambda^{t-1} R_t]$

↖ starting state probability of states

$X_1 \sim \alpha$

now let, $J_{\alpha,\lambda}^{\pi} = \mathbb{E}_{\alpha}^{\pi}[\sum_t \lambda^{t-1} R_t]$ where $v_{\alpha,\lambda} \triangleq \sup_{\pi} J_{\alpha,\lambda}^{\pi}$ ← unconstrained MDP problem

let, $C_{\alpha,\lambda}^{\pi} \triangleq \mathbb{E}_{\alpha}^{\pi}[\sum_t \lambda^{t-1} c_t]$, constrait optimisation problem:

↖ constrait fuction

$c_t = c(x_t, y_t)$

$v_{\alpha,C} \triangleq \sup_{\pi} J_{\alpha}^{\pi}$, s.t $C_{\alpha,\lambda}^{\pi} \leq \beta$ is a constrait

this can be solved by using a constrait from dual

$\sum c(s,a) x(s,a) \leq \beta$

we get this from $d_{\underline{x}}(s,a) = \frac{x(s,a)}{\sum x(s,a)} \rightarrow d_{\underline{x}}^{\infty}$

$J_{\alpha,\lambda}^{d_{\underline{x}}^{\infty}} = \sum_{s,a} r(s,a) x(s,a)$ is a Lemma for $\underline{x} \in$ dual is a policy for MDP

$C_{\alpha,\lambda}^{d_{\underline{x}}^{\infty}} = \sum c(s,a) x(s,a)$