CS6700 - RL - ASSIGNMENT-1

BY: S. VISHAL
CH18B020

DECLARATION.

I pledge that I have not copied or given any unauthorised assistance on this assignment.

ACKNOWLEDGEMENT:

- Bertsekas DPOC vol 1 chap 1
- Class notes (notes taken by me during class & notes used for teaching)
- Classmates:
  i) Aniswar S (CS18B080), Abishek (EE18B001)
  - discussed Q3
  ii) Narendhiran (CH18B045)
  - checked Q4 transition probabilities, corrected & gave hints for Q5

BY: S. VISHAL
CH18B020

CS6700 - RL - ASSIGNMENT-1

(1)

a) Action space : { forage patch 1, forage patch 2, forage patch 3 }
(call them as a_1, a_2, a_3 respectively)

State Space : { Animal is dead, 1 units of energy in reserve, 2 units of energy in reserve, 3 units of energy in reserve, 4 units of energy in reserve }

We want to maximise probability of survival at the end of 3 time periods.

$$g_3(x_3) = \begin{cases} 1 & \text{if } x_3 \in \{2, 3, 4\} \\ 0 & \text{if } x_3 = \text{dead}. \end{cases}$$

↓
Terminal cost

So we give a terminal cost of 1 if the animal has survived & all other transition costs to be 0. (& we want to maximise π)

i.e. $g_k(i, a, j) = 0 \quad \forall \{k, i, a, j\}$

So by maximising J_1

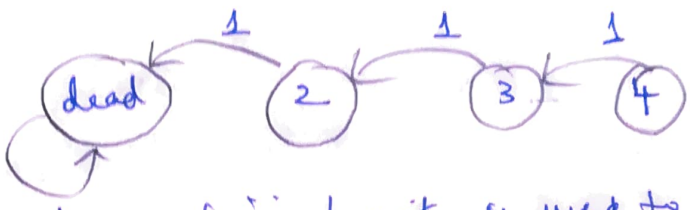
$$J = \max \mathbb{E}\left(g_3(x_3) + \sum_{k=0}^{k=2} g_k(x_k, u_k(x_k), x_{k+1})\right)$$

$$= \max \mathbb{E}\left(g_3(x_3)\right)$$

we in fact maximise the probability of survival $(= \mathbb{E}(g_3(x_3)))$

Let us note the transition probabilities for each action

a_1: Forage patch 1



($\because$ 1 unit is used to forage)

$P_{dd} = 1$

$P_{2d} = 1$ $\quad P_{ij} = 0$ for all others

$P_{32} = 1$

$P_{43} = 1$

a_2: Forage patch 2

Assume: Risk of predation independent of finding food.

$\therefore$ Risk of predation = 0.004, there is a 0.004 probability that the animal dies

In the 0.996 chance the animal survives, there is 0.3 chance to get food of 2 energy units

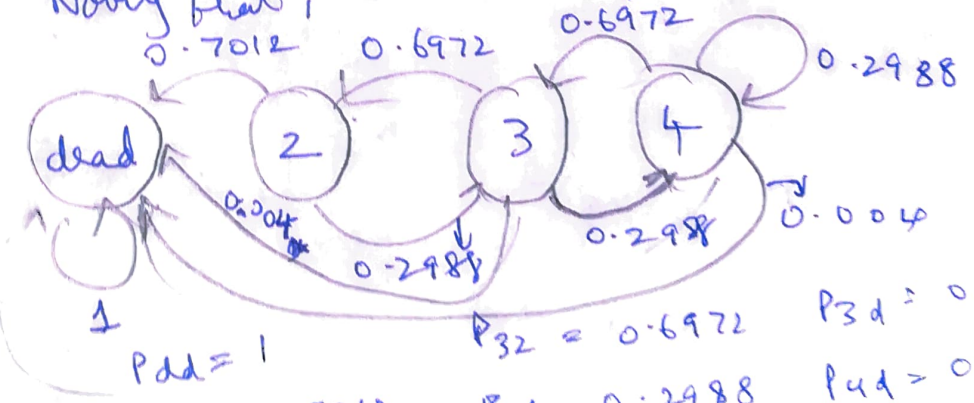
$P(\text{food} \mid \text{survival}) = 0.3 \Rightarrow P(\text{food} \cap \text{survival}) = 0.3 \times P_{\text{survival}}$

($\because$ gaining 1 energy unit as a result - 1 unit will be used to forage)

& 0.7 chance it doesn't get food

($\Rightarrow$ loses 1 unit as a result of foraging)

Noting that, $0.996 \times 0.3 = 0.2988$ & $0.996 \times 0.7 = 0.6972$



$P_{dd} = 1$

$P_{2d} = 0.7012$ $\quad P_{32} = 0.6972$ $\quad P_{3d} = 0.004$

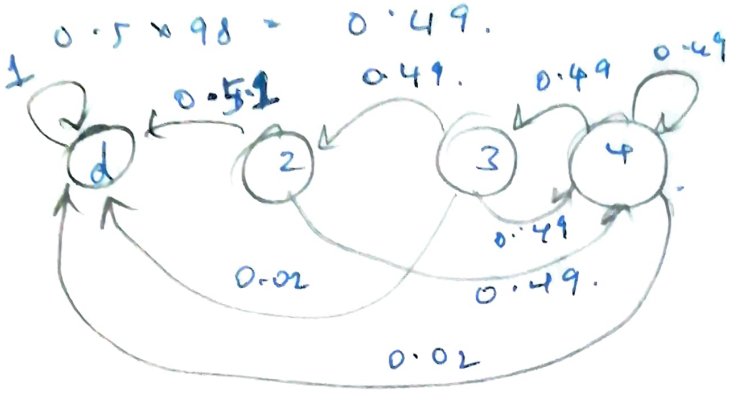
$P_{23} = 0.2988$ $\quad P_{34} = 0.2988$ $\quad P_{4d} = 0.004$

$P_{43} = 0.6972$ $\quad P_{ij} = 0$ for all other i,j

a3: Forage patch 3

Here the energy gain 3 units (= net gain of 2 units)

Proceeding similar to the case of a_2



$P_{dd} = 1$

$P_{2d} = 0.51$

$P_{24} = 0.49$

$P_{32} = 0.49$

$P_{3d} = 0.02$

$P_{34} = 0.49$

$P_{43} = 0.49$ $P_{44} = 0.49$

$P_{4d} = 0.02$ $P_{ij} = 0 \ \forall$ other $i \& j$

b) $$J_3^*(x_3) = g_3(x_3)$$

$$= \begin{cases} 1 & \forall\ x_3 \in \{2,3,4\} \\ 0 & \text{if } x_3 = d. \end{cases}$$

DP Algorithm:

$$J_2^*(x_2) = \max_a E_{x_3}\left(g_2(x_2, a, x_3) + J_3^*(x_3) \right)$$

$$= \max_a \sum_{\forall j \in S} P_{ij}(a) \left(g_2(x_2, a, j) + J_3^*(j) \right)$$

but $g(i, a, j) = 0$ always

$$\Rightarrow J_2^*(x_2) = \max_a \sum P_{ij}(a)\, J_3^*(j)$$

$x_2 = d$ ⇒ action doesn't matter the animal is already dead

action 1: forage patch 1

x_2	$\sum J^* P_{ij}$
d	0
2	$1 + 0 = 0$
3	$P_{32} J_3^*(2) = 1$
4	$P_{43} J_3^*(3) = 1$

$(\because P_{3j} = 0 \; \forall \; P_{4j} \neq 2)$

action 2: forage patch 2

x_2	$\sum J^* P_{ij}$
d	0
3	$0.2988 \, J_3^*(3) = 0.2988$
1	$0.2988 \, J_3^*(4) = 0.2988$
3	$+ 0.6972 \, J_3^*(2)$
4	$0.2988 \, J_3^*(4) = 0.996$
	$+ 0.6972 \, J_3^*(2)$

action 3: forage patch 3

x_2	$\sum J^* P_{ij}$
d	0

x_2 $\sum P_{ij} J_3^*(j)$

2 $P_{24} J^*(4) = 0.49.$

3 $0.49 * J^*(2) + 0.49 J^*(4) = 0.98$

4 $0.49 J^*(3) + 0.49 J^*(4) = 0.98$

choosing the actions that maximise the probability we land $J_2(x_2)$

x_2	$\max_a \sum_j P_{ij}(a) J_2^*(x_2)$	a.
1	0	(doesn't matter)
2	0.49.	a_3 (forage patch 3)
3	1	a_1 (forage patch 1)
4	1	a_1 (forage patch 1)

Now, $J_1^*(x_1) = \max_a E\left(g(x_1, a, j) + J_2^*(j) \right)$

$= \max_a E\left(J_2^*(j) \right)$ over j

If we choose action a_1,

$$E(J_2^*(j)) = \sum P_{ij}(a) J_2^*(j)$$

We make a table & compare $J_1(x_1)$ values for each action and then consider the optimal value

x_1	a_1	a_2	a_3
d	0	0	0
2	0	$0.2988 J_2^*(3)$ $= 0.2988$	$0.49 J_2^*(4)$ $= 0.49$
3	$1 \times J_2^*(2)$ $= 0.49$	$0.2988 J_2^*(4)$ $+ 0.6972 J_2^*(2)$ $= 0.6404$	$0.49 J_2^*(4)$ $+ 0.49 J_2^*(2)$ $= 0.7301$
4	$1 \times J_2^*(3)$ $= 1$	0.996	$0.49 J_2^*(4)$ $+ 0.49 J_2^*(3)$ $= 0.98$

choosing the action that maximises ~~probability of~~ return the reward,

x_1	$J_1^*(x_1)$	Action
d	0	– (does can't take any action – deceased)
2	0.49	Forage patch 3 (a_3)
3	0.7301	Forage patch 3 (a_3)
4	1	Forage patch 1 (a_1)

Similarly, $J_0^*(x_0) = \max_a \in (J_1^*(x_1))$

$= \max_a \sum_{\forall j} P_{ij}(a) \left(J_1^*(x_1)\right)$

We once again make a table of action & rewards

x_0	a_1	a_2	a_3
d	0	0	0
2	0	$0.2988(0.73)$ $= 0.218$	0.49
3	0.49	$0.2988(1) + 0.6972(0.49)$ $= 0.6504$	$0.49 + (0.49)^2$ $= 0.7301$
4	0.7301	$(0.2988) + (0.6972)(0.73)$ $= 0.8078$	$0.49 + 0.49 \times 0.7301$ $= 0.8477$

Choosing the action that optimises the reward,

x_0	$J_0^*(x_0)$	Action
d	0	– (dead)
2	0.49	Forage patch 3 (a_3)
3	0.7301	Forage patch 3 (a_3)
4	0.8477	Forage patch 3 (a_3)

Optimal policy for foraging obtained from the above analysis!

$$\pi^* = \{ \mu_0(x), \mu_1(x), \mu_2(x) \}$$

$$\mu_0(x) = \begin{cases} \text{can't perform action} & x = d \\ a_3 & x = 2 \quad \text{(forage patch 3)} \\ a_3 & x = 3 \quad \text{(forage patch 3)} \\ a_3 & x = 4 \quad \text{(forage patch 3)} \end{cases}$$

$$\mu_1(x) = \begin{cases} \text{cant perfor actin} & x = d \\ a_3 & x = 2 \text{ (Forage patch 3)} \\ a_3 & x = 3 \text{ (Forage patch 3)} \\ a_1 & x = 4 \text{ (Forage patch 1)} \end{cases}$$

$$\mu_2(x) = \begin{cases} \text{cant perfor actin} & x = d \\ a_3 & x = 2 \text{ (forage patch 3)} \\ a_1 & x = 3 \text{ (forage patch 1)} \\ a_1 & x = 4 \text{ (forage patch 1)} \end{cases}$$

CS6700 RL ASSIGNMENT - 1

BY: S. VISHAL
CH18B020

(2) a) State Space (x_k) : Let the state at the kth stage represent the crop production at that stage.

Action Space (u_k) : The action we perform u_k, is the fraction of ~~crops invested~~ production in the kth stage invested for production in $(k+1)$th stage

State evolution : $x_{k+1} = x_k + w_k u_k x_k$

where $\{w_k\}$ is i.i.d with $E(w_k) = \bar{w} > 0$

Here we want to optimize maximise the total crop production. So,

let cost $g(x_k, u_k, x_{k+1}) = (1-u_k) x_k$ (fraction of crops stored)

& the terminal cost $g_N(x_N) = x_N$

$$\Rightarrow J^* = \max_{u_0, u_1, u_2, \ldots, u_{N-1}} E\left(x_N + \sum_{k=0}^{N-1} (1-u_k) x_k\right)$$

Here the decision variables are u_k, $k \in \{0, \ldots, N-1\}$. and the expectation is over the sequence of random variables $\{w_k\}$

Thus by using this MDP, we have set the appropriate objective as desired in the problem as well as obtained an appropriate description.

b) We can use the DP algorithm to solve this problem

$$J_k^*(x_k) = \max_{u_k} E\left(J_{k+1}^*(x_{k+1}) + (1-u_k)x_k \right)$$

(transition cost)

& $J_N^*(x_N) = g_N(x_N) = x_N$ expectation over w_k or x

this procedure can be used computationally to arrive at a solution but numerically since we don't know the structure of $J_{k+1}^*(x_{k+1})$, I would like to evaluate the first 2 steps & take a guess and then proceed by induction to generalize the proof. (transition cost $g(\cdot)$)

$$J_{N-1}^*(x_{N-1}) = \max_{u_{N-1}} E\left((1-u_{N-1})x_{N-1} + J_N^*(x_N) \right)$$

x_N

$$= \max_{u_{N-1}} E_{w_{N-1}}\Big((1-u_{N-1})(x_{N-1}) + [x_{N-1} + w_{N-1}u_{N-1}x_{N-1}] \Big)$$

using the state evolution equation

$$= \max_{u_{N-1}} \Big[(1-u_{N-1})(x_{N-1}) + x_{N-1} + \bar{w}\,u_{N-1}x_{N-1} \Big]$$

$(\because E(w_k) = \bar{w})$

Notice the eqn is linear in u_{N-1} so the extrema lie at the ends of the ~~bounds~~ allowed values of u_{N-1}

Sub $u_{N-1} = 0$ & 1

$$\Rightarrow T^*_{N-1}(x_{N-1}) = \max_{u_{N-1}} \left\{ 2x_{N-1}, (1+\bar{\omega}) x_{N-1} \right\}$$

$2x_{N-1} > (1+\bar{\omega}) x_{N-1} \Rightarrow \bar{\omega} > 1 \quad (\because x \text{ is +ve})$

So, if $\bar{\omega} > 1$, $u^*_{N-1} = 1$

if $\bar{\omega} \leq 1$, $u^*_{N-1} = 0$

& the $T^*_{N-1}(x_{N-1})$ are $(1+\bar{\omega}) x_{N-1}$ & $2x_{N-1}$ corresponding

→ transition cost $g(\cdot)$

$$J^*_{N-2} = \min_{u_{N-2}} \left((1 - u_{N-2}) x_{N-2} + J^*_{N-1}(x_{N-1}) \right)$$

If $\bar{\omega} > 1$,

$$J^*_{N-2}(x_{N-2}) = \mathbb{E}\left((1 - u_{N-2})(x_{N-2}) + (1+\bar{\omega}) x_{N-1} \right)$$

$$= \mathbb{E}\left((1 - u_{N-2})(x_{N-2}) + (1+\bar{\omega})(x_{N-2} + \omega_{N-2} u_{N-2} x_{N-2}) \right)$$

$$= (1 - u_{N-2})(x_{N-2}) + (1+\bar{\omega})(x_{N-2} + u_{N-2} + \bar{\omega} x_{N-2})$$

$$= \max\left((2+\bar{\omega})(x_{N-2}), (x_{N-2})(1+\bar{\omega})^2 \right)$$

Since $\bar{\omega} > 1$,

$$J^*_{N-2} = (1+\bar{\omega})^2 x_{N-2} \quad \& \quad u^*_{N-2} = 1$$

(claim: If $\bar{\omega} > 1$, $J^*_{N-k}(x_{N-k}) = (1+\bar{\omega})^k x_{N-k}$

We have shown this for $k = 1$ & $k = 2$.

Now we can try showing it for $k+1$ if k is true.

$$J^*_{N-k-1} = \max_{u_{N-k-1}} E\left((1-u)_{N-k-1} x_{N-k-1} + J^*_{N-k}(x_{N-k}) \right)$$

$$= \max_{u_{N-k-1}} \left(1 - u_{N-k-1}\right) x_{N-k-1} + (1+\bar{\omega})^k \left(x_{N-k-1} + x_{N-k-1}\bar{\omega}\, u_{N-k-1}\right)$$

$$= \max \left\{ x_{N-k-1}\left(1 + (1+\bar{\omega})^k\right),\ x_{N-k-1}\left(1 + \bar{\omega}(1+\bar{\omega})^k\right) \right\}$$

$$= \max \left\{ x_{N-k-1}\left(1 + (1+\bar{\omega})^k\right),\ (x_{N-k-1})(1+\bar{\omega})^{k+1} \right\}$$

$k = 1 \Rightarrow (1+\bar{\omega})^k > 1 \Rightarrow 2(1+\bar{\omega})^k > 1 + (1+\bar{\omega})^k$

$\Rightarrow (1+\bar{\omega})^{k+1} > 1 + (1+\bar{\omega})^k$

$\rightarrow \because 1 + \bar{\omega} > 1 + 1 = 2$

$$= x_{N-k-1}(1+\bar{\omega})^{k+1}$$

& $u^*_{N-k-1} = 1$

Thus the proposition holds for all k. (by induction)

If $u^*_{N-k} = 1$ if $\bar{\omega} > 1$

$\left(\& \; J^*_{N-k}(x_{N-k}) = (1+\bar{\omega})^k x_{N-k} \right)$

If $\bar{\omega} < 1$

$$J^*_{N-2}(x_{N-2}) = \max_{u_{N-2}} E\Big((1-u_{N-2})(x_{N-2}) + J^*_N(x_{N-1}) \Big)$$

$$= \max_{u_{N-2}} E\Big((1-u_{N-2})(x_{N-2}) + 2x_{N-1} \Big)$$

$$= \max_{u_{N-2}} E\Big((1-u_{N-2})(x_{N-2}) + 2\big(x_{N-2} + u_{N-2} w_{N-2}^{x_{N-2}} \big) \Big)$$

$$= \max \{ 3x_{N-2}, (2+\bar{\omega}) x_{N-2} \}$$

$\because \bar{\omega} < 1, \quad 3 > 2+\bar{\omega}$

$$\Rightarrow J^*_{N-2}(x_{N-2}) = 2x_{N-2}.$$

Claim: $J^*_{N-k}(x_{N-k}) = k x_{N-k}$ if $\bar{\omega} < 1$

& $u^*_{N-k} > 0$

$$J^*_{N-k-1} = \max_{u_{N-k-1}} E\Big((1-u_{N-k-1}) x_{N-k-1} + k\big(x_{N-k-1} + (u_{N-k-1}) w_{N-k-1}^{x_{N-k-1}} \big) \Big)$$

$$= \max_{u_{N-k-1}} \Big[(1-u_{N-k-1}) x_{N-k-1} + k(x_{N-k-1} + u_{N-k-1} \bar{\omega}^{x_{N-k-1}}) \Big]$$

$$= \max\left\{(k+1)\bar{\omega}\, x_{N-k-1};\ x_{N-k-1}(k)(1+\bar{\omega})\right\}$$

$$(k+1)\, x_{N-k-1} \geq k\, x_{N-k-1}(1+\bar{\omega})$$

$$\Rightarrow (1+\bar{\omega}) < \frac{k+1}{k} \Rightarrow \bar{\omega} < \frac{1}{k}.$$

So if $\bar{\omega} \leq \frac{1}{N}$, we can complete the induction & the claim modified claim:

$$V^*_{N-k}(x_{N-k}) = k x_{N-k}, \ \& \ u_{N-k} = 0 \quad \forall k$$

if $\bar{\omega} \leq \frac{1}{N}$ holds

However if $\bar{\omega} < 1$ but $\bar{\omega} > \frac{1}{k}$,

new claim: choose $u^*_{N-j} = 0 \quad \forall j \in \{1, \dots, k\}$

$u^*_{N-j} = 1 \quad \forall j > k$

where k is such that

$$\frac{1}{k+1} \leq \bar{\omega} \leq \frac{1}{k}$$

From the previous proof we can infer that

as long as $\bar{\omega} < \frac{1}{k}$, $u^*_{N-k} = 0$.

We still need to show the claim not: $u^*_{N-k} = 1$ if $\bar{\omega} > \frac{1}{j}$

So consider a point after $N-k-1$

say $N-k-j$, $j>0$

we have shown that $u^*_{N-k-1} = 1$

(base case)

Now for if we can by induction

Assume $J^*_{N-k-j} = (k)(1+\bar{\omega})^j \, x_{N-k-j}$

(∵ k $u=0$ k times followed by $u=1$ j times)

$$J^*_{N-k-j-1} = \min_{u \in \{u_{N-k-j-1}\}} \left((1-u_{N-k-j-1}) \, x_{N-k-j-1} + J^{\alpha}_{N-k-j} \right)$$

$$= \min_{u_{N-k-j-1}} \left((1-u_{N-k-j-1}) \, x_{N-k-j-1} + k(1+\bar{\omega})^j \left(x_{N-k-j-1} + \omega_{N-k-j-1} u_{N-k-j-1} \right) \right)$$

$$= \min_{u_{N-k-j-1}} \left[(1-u_{N-k-j-1}) + k(1+\bar{\omega})^j \left(x_{N-k-j-1} + \bar{\omega} \, u_{N-k-j-1} \right) \right]$$

$$= \min \left\{ x_{N-k-j} \left(1 + k(1+\bar{\omega})^j \right), \; x_{N-k-j} \left| 1+\bar{\omega} \right|^{j+1} k \right\}$$

$u=0$ ↓ $\qquad$ $u=1$ ↓

$(1+\bar{\omega})^j k + 1$ $\qquad$ vs $\qquad$ $1 + k(1+\bar{\omega})^j$

$$= \max\left(\left[(1+\omega)^j k\right]\{1+\bar{\omega}\}, \left[(1+\bar{\omega})^j k\right]\left(\frac{1}{k(1+\omega)}+1\right)\right)$$

removing common factors

$$\max\left((1+\bar{\omega}), \frac{1}{k(1+\bar{\omega})^{j}}+1 \right)$$

$$\bar{\omega} > \frac{1}{k} \Rightarrow \bar{\omega} > \frac{1}{k+(1+\bar{\omega})}$$

↳ common factor >1

$$\therefore J^{+}_{N-k-j-1} = (1+\omega)^{j+1} k \{.$$

$\Rightarrow u = 1$ better.
$N-k-j-1$

$\therefore$ By principle of mathematical induction

$\therefore$ our claim is true

Optimal policy

i) If $\bar{\omega} > 1$, $u^*_0 = u^*_1 = \cdots = u^*_{N-1} = 1$

ii) If $0 < \bar{\omega} < \frac{1}{N}$, $u^*_0 = u^*_1 = \cdots = u^*_{N-1} = 0$

iii) If $\frac{1}{N} < \bar{\omega} < 1$, $u^*_0 = u^*_1 = \cdots = u^*_{N-k-1} = 1$

$$u^+_{Nk} = u^+_{N-k+1} = \cdots = u^+_{N-1} = 0$$

where k is such that $\frac{1}{k+1} < \bar{\omega} \leq \frac{1}{k}$

CS6700 – RL – ASSIGNMENT - 1

BY: S. VRSHAL
CH18B020

③ a) For an optimal policy π^* having cost J^*

$$T_{\pi^*} J^* = TJ^* \Rightarrow J^* = TJ^*$$

(prop 2)

and J^* is the unique solution to the above equation ——— (1)

additionally, J^* is the unique solution to

$$T_{\pi^*} J^* = J \quad \text{——} (2)$$

Upon introducing a new action a' in a state s', we get a new MDP. Let the optimal policy of such an MDP be π^*_{new} and cost be J^*_{new}.

From (2), since we don't have a change in transition probabilities & $\pi^*(i)$ is also unchanged

$$T_{\pi^*} J^* = J^* \text{ still holds} \quad \text{——} (3)$$

We still need to check if

$$J^* = TJ^*$$ to know whether π^* is indeed π^*_{new}.

[Assuming minimisation problem for part a) – although it really doesn't matter here]

$\therefore \pi^*$ was optimal for the older MDP

$(T U^*)(i) = J^*(i)$ is older MDP

$$\Rightarrow J^*(i) = \min_a \sum_{\forall j} \left(P_{ij}(a) \left(g(i,a,j) + J^*(j) \right) \right)$$

In the new MDP,

$$U^{**}(i) = \min_a \sum \left(P_{ij}(a) \left(g(i,a,j) + J^*(j) \right) \right)$$

holds for all $i \neq s'$ ——— (4)

(because nothing has changed
— no new actions have been added)

claim①: If $J^*(s') = \min_a \sum_{\forall j} \left(P_{s'j}(\pi^*(s')) \left(g(s', \pi^*(s'), j) + J^*(j) \right) \right)$

$$= \min_a \sum \left(P_{s'j}(a) \left(g(s', a, j) + J^*(j) \right) \right)$$

then, $\pi^{**} = \pi^*$ optimal & $J^{**} = J^*_{\text{optimal}}$

Proof: Procedure to test:

Assume that compare $J^*(s')$ with

$$\sum_{\forall j} P_{s'j}(a') \left(g(s', a', j) + J'(j) \right)$$

if $J^*(s') \neq \sum_{\forall j} P_{s'j}(a') \left(g(s', a', j) + J'(j) \right)$

then $\pi^*(s')$ is still the solution to the eqn because substituting $\pi^*(s')$ gives us a lower cost than a''

$\Rightarrow$ $\pi^*(s')$ is still optimal

& hence the claim is satisfied.

$\Rightarrow \pi^* = \pi^*_{new}$ & $J^* = J^*_{new}$

proof for the claim①:

If $J^*(s) = \min_a \sum P_{s'j}(a)\,(g(s',a,j) + J^*(j))$

then combining with eqn ④, we have

(for new MDP),

$J^* = J^*$ $\Rightarrow$ J^* is still the optimal cost

& we already showed that expected cost of policy π^* is J^*

$\Rightarrow J^*_{new} = J^*$ & $\pi^*_{new} = \pi^*$.

Summary: For the s' state equation, check if a' produces a minimum in RHS, if it doesn't the old policy remains the optimal policy

claim ②: If a' gives a lower value in T^*J^* (note that J^* is the optimum in old MDP, it is not so in the new MDP) for s' state; then for that state s', the action a' is optimal.

proof:

Suppose a' is not the optimal action, then it is as good as dealing with an MDP without that action

$\Rightarrow$ this is the old MDP

$\Rightarrow$ π^* & J^* are optimal in the new MDP then.

$\Rightarrow$ $J^*(s') = \min_a \sum_j P_{s'j}\,(g(s',a,j) + J(j))$

has solution $a = \pi^*(a)$ instead of a'

this is a contradiction because $J^*(s') \neq \sum_j P_{s'j}(\cdot)$

to find new policy $\Rightarrow$ the claim is true.

From claim ② we can infer that a' is the optimal action at s' if $\pi^* \neq \pi^+$ new

Let $\pi_0(i) = \begin{cases} \pi^*(i) & \forall i \neq s' \\ a' & i = s' \end{cases}$

Using this policy we can perform policy iteration & find the new optimal policy

[note that $\pi_0(i)$ is proper since

$J^* \leq T_{\pi_0} J^*$ — (if a' is optimal)

as shown in that eqn in claim ①]

Since we already found one optimal action, policy iteration might converge faster

b) Assuming it is a maximization problem we are given that the cumulative reward $J_{\pi^*}(s^t)$ decreases by Δ

$J_{\pi^*}(s^t) = E\left(g(s^t, \pi^*(s^t), j) + J_{\pi^*}(j) \right)$

Since no other information is given, I would like to interpret it as $J_{\pi^*}(s^+)$ has changed because of a Δ decrease in average expected transition cost from s^+

that $\hat{E}(g_{new}(s^+, \pi^*(s^+), j))$

$= E(g_{new}(s^+, \pi^*(s^+), j)) - \Delta$

Because of this change in transition cost from s^+, there are 2 implications

i) J_{π^*} might not be solution to $J = T_{\pi^*} J$ i.e. it may not be the expected cost of policy π^*

ii) π^* may not be optimal policy (because returns from s^+ has reduced, intuitively there is a possibility that a policy that reduces has lesser number of transitions to s^+ may be the new optimal policy, π^*_{new})

Checking if π^* is optimal

Step ①: Perform policy evaluation

$$J(i) = \sum_j P_{ij}(\pi_k(i)) \, g(i, \pi_k(i), j) + J(j)$$

where $g(i, \pi_k(i), j) = \begin{cases} g_{original}(i, \pi_k(i), j) & \forall i \neq s^+ \\ g_{original}(s^+, a^+, j) - & \end{cases}$

$$J(i) = E\left(g_{original}(i, \pi_k(i), j) + J(j) \right)$$

& for $i = s^+$

$$J(s^+) = E\left(g_{original}(s^+, \pi_k(s^+), j) + J(s^+) \right) \quad \text{— } \Delta$$

where $g_{original}$ is the transition cost of original MDP. Let soln to these eqns be J_{π^*}

Step ②: Perform policy improvement.

$$T_{\pi_{new}} J_{\pi^*} = T J_{\pi^*}$$

$$\Rightarrow (T_{\pi_{new}} J_{\pi^*})(i) = \max_a E\left(g(i, a, j) + J_{\pi^*}(j) \right)$$

Get $J_{\pi_{new}}$

Get π_{new} & evaluate $J_{\pi_{new}}$.

if $J_{\pi_{new}} = J_{\pi^*}$ then π^* is the optimal policy

If $J_{new} < J_{\pi^+}$, π^+ is not optimal.

In order to find the new optimal policy we can do policy iteration with J_{new} as the initial proper policy.

The expected cost will also be obtained via policy iteration in policy evaluation step.

CS6700 : RL - ASSIGNMENT-1

BY: S. VISHAL
CH18B020

(4) For convenience, I would like to label Kevin as player A, Karan as player B, Kannan as player C.

To model the problem as an MDP, where 'A' is the decision maker we need to know the decisions (in fact, optimal) of 'B' & 'C'.

2 - Game scenarios.

Consider the game has reduced to A & B and it is A's turn to fire. Let us analyse this from the point of view of B:

- B can't miss intentionally - has to shoot at A.
- Rule permits A to shoot into the air. But that will increase chance of B winning - because B will take more turns to fire & hence the probability of reaching the target (which will be a geometric distribution) increases

So, to obtain a lower bound of probability of B winning we can as well assume A always shoots & doesn't miss intentionally

$P_{B\text{ wins in }k\text{ turns}} = P(A \text{ misses } k \text{ turns despite shooting}) \times P(B \text{ misses } k-1 \text{ turns consecutively before firing in last turn})$

$$= (1-\alpha)^k (1-\beta)^{k-1} \beta.$$

$$\therefore P(B \text{ wins}) = \sum_{k=1}^{\infty} (1-\alpha)^k (1-\beta)^{k-1} \beta.$$

$$= \frac{(1-\alpha)}{1-(1-\alpha)(1-\beta)} \times \beta$$

Proceeding similarly

$$P(B \text{ wins in } B \text{ vs } C) = \frac{(1-\gamma)}{1-(1-\gamma)(1-\beta)} \times \beta$$

But $(1-\gamma) < (1-\alpha)$ $(\because \alpha > \gamma)$

$$\left(\Rightarrow 1-(1-\alpha)(1-\beta) < 1-(1-\gamma)(1-\beta)\right)$$

So we conclude

$$P(B \text{ wins } A \text{ vs } B) > P(B \text{ wins } B \text{ vs } C)$$

Similarly for C, arguing in a similar fashion,

$$P(C \text{ wins } C \text{ vs } A) > P(C \text{ wins } C \text{ vs } B)$$

From this analysis, we understand that B and C would seek to eliminate each other first because they prefer dueling A in a 2 vs 1 game scenario to maximise their winning chance.

So, when all of ABC are alive,

* B shoots C in his turn
* C shoots B in his turn

Now that we have decided the possible actions of B & C we can build an MDP.

- Note that a single transition in the MDP incorporates all changes in the position of the game between 2 turns of A (eg. if B & C are alive, one transition implies A, B, and C have all taken their turns)

State Space : { A ①, A and B ② are alive, A and B C ③ are alive, A, B and C ④ are alive, A ⑤ only alive (A has won!) }

Action space : {

state ① : no action - A already out of the game (terminal)

state ② : shoot at B, miss intentionally.

state ③ : shoot at C, miss intentionally

state ④ : shoot at B, shoot at C, miss intentionally

state ⑤ : A has won! - no actions (terminal)

Costs/Rewards :

∴ We want to maximise probability of winning,

$g(i,a,j) = 0 \quad \forall \{i,a,j\}$

$g_N(x_N) = \begin{cases} 0 & \text{if } x \neq x_5 \\ 1 & \text{if } x = x_5 \end{cases}$

Note that the A last 4 A wins are terminal states

& so objective $\max_a E\left(\alpha + J_N^*(x_N) - \sum_{k=1}^{N} g_k(x_k, u_k, x_{k+1}) \right)$

$= \max_a E\left(J_N^*(x_N) \right)$

$= \max_a E\left(g_N'(x_N) \right)$

which is eqvt to maximizing probability of winning.

→ (∵ probability = expectation of indicator variable for the event)

Transition probabilities

Let a_1: A shoots B a_2: A shoots C a_3: A shoots air.

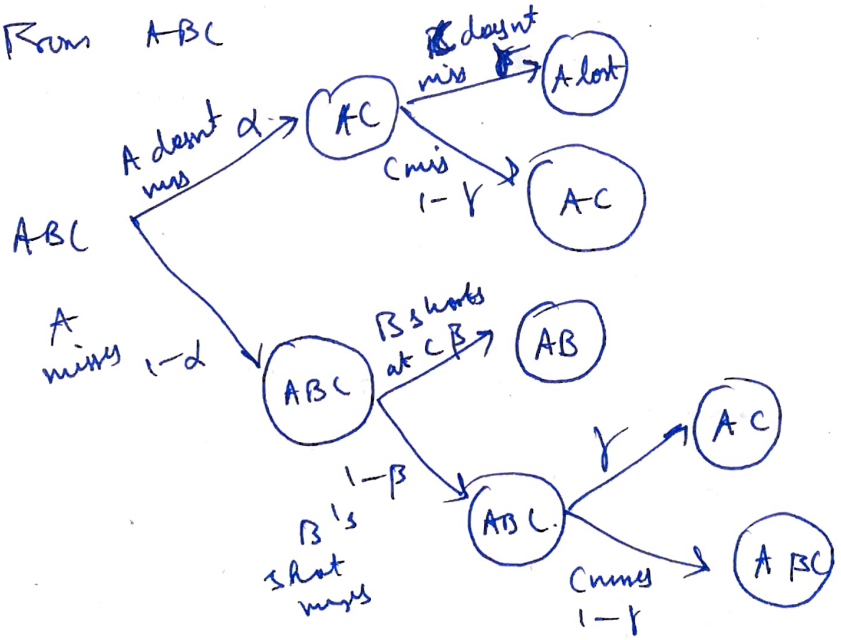
a) Action a_1:

From AB: A wins if A shoots properly : α

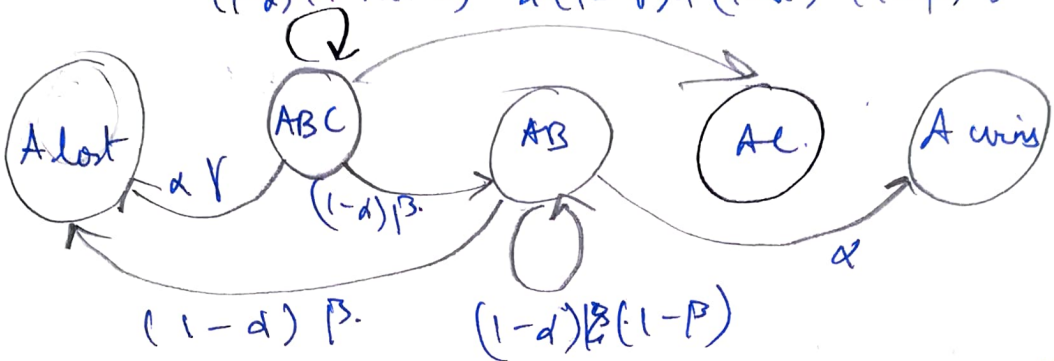
A loses if A misses & B shoots properly $(1-\alpha)\beta$

Back to same state if A & B both miss $(1-\alpha)(1-\beta)$

From ABC



Using these transition probabilities we get the markov chain as:



No transitions from AC because A can't shoot B in their turn.

$P_{21} = \alpha\gamma$ $P_{22} = (1-\alpha)(1-\beta)(1-\gamma)$ $P_{24} = \alpha(1-\gamma) + (1-\alpha)(1-\beta)\gamma$

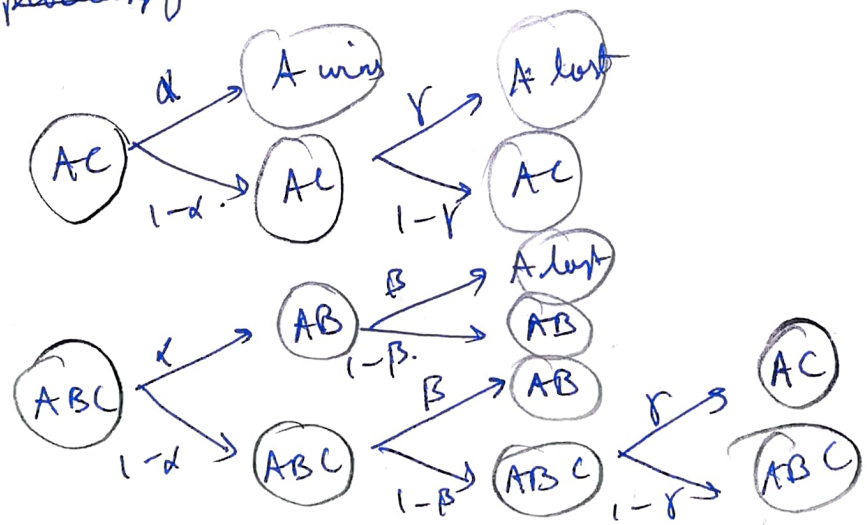
$P_{23} = (1-\alpha)\beta$

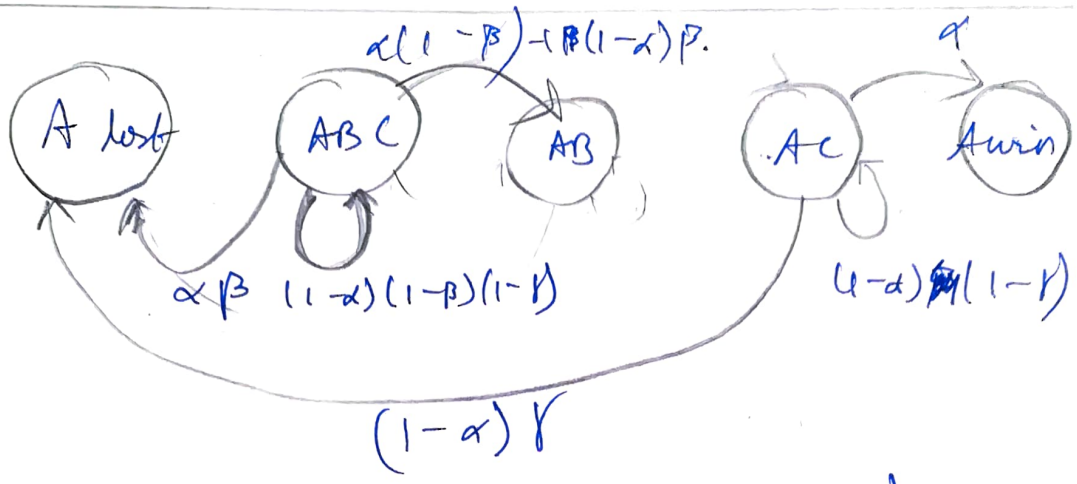
$P_{31} = (1-\alpha)\beta$ $P_{33} = (1-\alpha)(1-\beta)$

$P_{34} = \alpha$; & other $P_{ij} = 0$

ii) Action as: A shoots C

~~proceeding~~ Proceeding similarly:



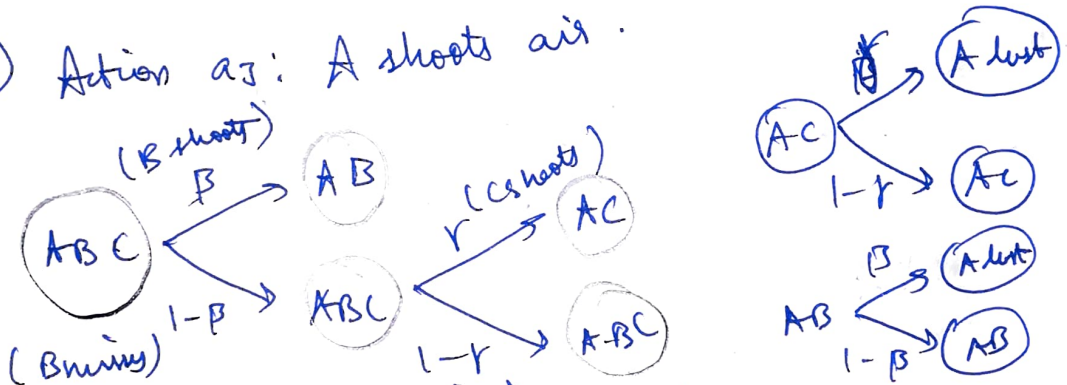


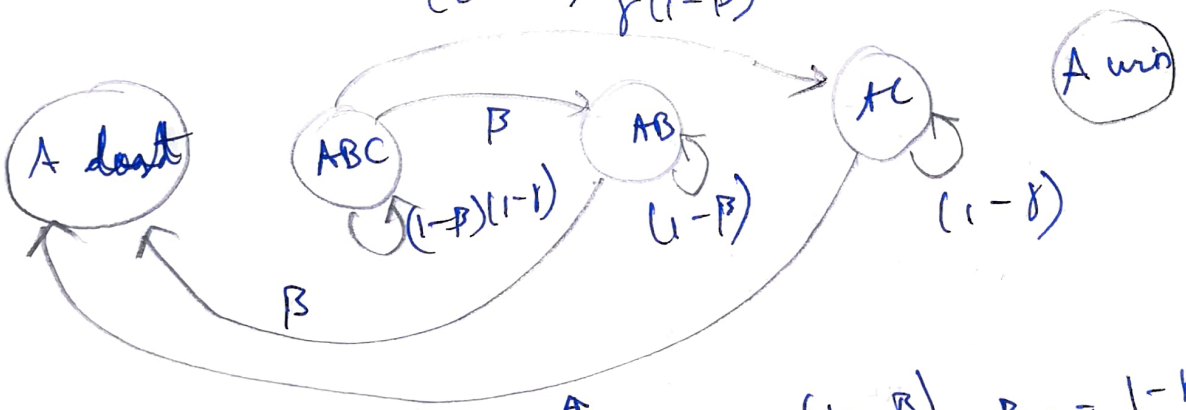
$P_{21} = \alpha\beta \qquad P_{22} = (1-\alpha)(1-\beta)(1-\gamma)$

$P_{23} = \alpha(1-\beta) + (1-\alpha)\beta \; ; P_{41} = (1-\alpha)\gamma$

$P_{44} = (1-\alpha)(1-\gamma) \; ; \; P_{45} = q \; ; \text{other } P_{ij} = 0$

iii) Action a_3: A shoots air.





$P_{22} = (1-\beta)(1-\gamma) \qquad \gamma \uparrow \qquad P_{33} = (1-\beta) \qquad P_{44} = 1-\gamma$

$P_{23} = \beta \qquad P_{24} = (1-\beta)\gamma \qquad P_{31} = \beta \qquad P_{41} = \gamma$

other $P_{ij} = 0$

b) Bellman optimality equations

$$T J^{+}_{N} = J^{+} \; ; \quad J^{+}(\text{Aim}) = 1$$

$$\Rightarrow J^{+}(i) = \max_{a \in A(i)} E\left(g(i,a,j) + J^{+}(j) \right)$$

($\because$ maximizing reward)

$$= \max_{a \in A(i)} E\left(J^{+}(j) \right)$$

~~$J(AB) = \max_{\alpha}$~~

~~At AB state~~ AB, we have 2 actions allowed

- shoot B or shoot at

$$\Rightarrow J(AB) = \max \left\{ \begin{array}{l} (1-\beta) J(AB), \\ \quad \text{A shoots air} \\ (1-\alpha)(1-\beta) J(AB) + \alpha \times J(\text{Aim}) \\ \quad \text{A shoots B} \end{array} \right\}$$

$$= \max \left\{ \begin{array}{l} (1-\beta) J(AB), \\ (1-\alpha)(1-\beta) J(AB) + \alpha \end{array} \right\}$$

——— (1)

Similarly

$$J(AC) = \max \left\{ \begin{array}{l} (1-\gamma) J(AC), \quad \longrightarrow \text{A shoots air} \\ (1-\alpha)(1-\gamma) J(AC) + \alpha \quad \longrightarrow \text{A shoots C} \end{array} \right\}$$

(2 actions : shoot air or shoot C)

——— (2)

At state ABC, all 3 actions are allowed, so we take max over the 3 actions

$$J(ABC) = \max \Big\{ \big[(1-\alpha)(1-\beta)(1-\gamma) J(ABC) + (\alpha(1-\gamma) + (1-\alpha)(1-\beta)\gamma) J(AC) + (1-\alpha)\beta J(AB) \big], \big[(1-\alpha)(1-\beta)(1-\gamma) J(ABC) + (\alpha(1-\beta) + (1-\alpha)\beta) J(AB) + (1-\alpha)(1-\beta)\gamma J(AC) \big], \big[(1-\beta)(1-\gamma) J(ABC) + \beta J(AB) + \gamma(1-\beta) J(AC) \big] \Big\}$$

——— (3)

c) Given $\alpha = 0.3$, $\beta = 0.5$, $\gamma = 0.6$. Substituting,

eqn ① : $J_{AB} = \max(0.5 J_{AB}, 0.5 * 0.7 + 0.3)$
$= 0.35 J_{AB} + 0.3$

part i) gives $J_{AB} = 0$ — not correct — there must probably be some error

part ii) using $J_{AB}^* = \frac{0.3}{0.65} = \boxed{0.461}$ — short B prepaid

(i.e. $0.5(0.461) < 0.35(0.461) + 0.3$)

eqn ② : $J_{AC}^* = \max(0.4 J_{AC}, 0.28 J(AC) + 0.3)$

using a similar argument,

$J_{AC}^* = \frac{0.3}{0.72} = \boxed{0.416}$ — short C prepaid

eqn ③

(i) term :

$$T^{+}_{ABC} = (0.7)(0.5)(0.4)\,T_{ABC} + [(0.3)(0.4) + (0.7)(0.5)(0.6)]\,T_{AC} + (0.7)(0.5\,T_{AB}$$

$$\Rightarrow \quad T^{+}_{ABC} = \frac{0.299}{0.86} = 0.3477$$

ii) term :

$$T^{+}_{ABC} = 0.14\,T_{ABC} + [(0.3)(0.5) + (0.7)(0.5)]\,T_{AB} + (0.5)(0.7)(0.6)\,T_{AC}$$

$$\Rightarrow T^{+}_{ABC} = \frac{0.3183}{0.86} \approx 0.37$$

iii) term :

$$T^{+}_{ABC} = 0.2\,T_{AB} + 0.5\,T_{AB} + 0.3\,T_{AC}$$

$$\Rightarrow T^{+}_{ABC} = \frac{0.35}{0.8} = \boxed{0.45}$$

Using $T^{*}_{ABC} = 0.444$,

(iii) term > (i) & (ii)

$\Rightarrow$ eqn is solved. — closest in all preferred.

State	$J^+(x)$	$\mu^+(x)$ (optimal action)
A lost	0	— (A can't do anything)
AB	0.4615	Shoot B
AC	0.4167	Shoot C
ABC	0.445	Shoot in air

BY: S. VISHAL
CH18B020

CS6700 - RL - ASSIGNMENT - 1

(5) Before I answer the question, I would like to show $\|a+b\|_{\xi} < \|a\|_{\xi} + \|b\|_{\xi}$

$$\|a+b\|_{\xi} = \max_i \frac{|a(i)+b(i)|}{\xi(i)}$$

[triangle inequality] $$< \max_i \left(\frac{|a(i)|+|b(i)|}{\xi(i)} \right)$$

$$\because \max((a+b)(i)) < \max(a(i)) + \max(b(i))$$

$$\leq \max_i \frac{|a(i)|}{\xi(i)} + \max_j \frac{|b(j)|}{\xi(j)}$$

$$\leq \|a\|_{\xi} + \|b\|_{\xi}$$

$$\Rightarrow \|a+b\|_{\xi} < \|a\|_{\xi} + \|b\|_{\xi} \quad \text{---(1)}$$

a) Termination criterion: $\|J_{m+1} - J_m\|_{\xi} < \varepsilon\left(\frac{1-\beta}{2\beta}\right)$ ---(2)

T is a contraction operator

$$\Rightarrow \|TJ - TJ'\|_{\xi} \leq \beta \|J - J'\|_{\xi} \quad \text{---(3)}$$

Apply operator T in eqn (2)

$$\Rightarrow \|TJ_{m+1} - TJ_m\| < \beta\varepsilon\left(\frac{1-\beta}{2\beta}\right)$$ (∵ T is a contraction - eqn (3))

$$\Rightarrow \| T_{m+2} - T_{m+1} \|_\xi < \rho \left[\varepsilon \left(\frac{1-\rho}{2\rho} \right) \right]$$

Recursively applying T,

$$\| T_{m+3} - T_{m+2} \|_\xi < \rho^2 \left[\varepsilon \left(\frac{1-\rho}{2\rho} \right) \right]$$

$$\vdots$$

$$\| T_{m+n} - T_{m+n-1} \|_\xi < \rho^{n-1} \left[\varepsilon \left(\frac{1-\rho}{2\rho} \right) \right]$$

Add all these eqns.

$$\Rightarrow \sum_{k=2} \| T_{m+k} - T_{m+k-1} \|_\xi < \sum_{k=1}^{n-1} \rho^k \left[\varepsilon \left(\frac{1-\rho}{2\rho} \right) \right] \quad (4)$$

Now we use $\| a+b \|_\xi < \| a \|_\xi + \| b \|_\xi$

and push the sum inside the modulus

i.e. $\| T_{m+n} - T_{m+n-1} + T_{m+n-1} - T_{m+n-2} + \cdots + T_{m+3} - T_{m+2} + T_{m+2} - T_{m+1} \|_\xi$

$$< \sum_{k=2}^{n} \| T_{m+k} - T_{m+k-1} \|_\xi$$

④ $\Rightarrow \| J_{m+n} - J_{m+1} \|_\infty < \sum_{k=1}^{n-1} \beta^k \left(\varepsilon \left(\frac{1-\beta}{2\beta} \right) \right)$

* - take $\lim_{n \to \infty}$

$$\lim_{n \to \infty} \| J_{m+n} - J_{m+1} \| < \lim_{n \to \infty} \sum_{k=1}^{n-1} \beta^k \left(\varepsilon \left(\frac{1-\beta}{2\beta} \right) \right)$$

$$\xi = \frac{\beta}{1-\beta} \varepsilon \left(\frac{1-\beta}{2\beta} \right)$$

$$\Rightarrow \lim_{n \to \infty} \| J_{m+n} - J_{m+1} \|_\infty < \frac{\varepsilon}{2} < \frac{\varepsilon}{2}$$

Also, we know that value iteration converges

$$\left(\lim_{k \to \infty} T^k J_0 = J^* \right)$$

$$\Rightarrow \| J_m^* - J_{m+1} \|_\infty < \varepsilon / 2$$

$$\Rightarrow \| J_{m+1} - J_m^* \|_\infty < \frac{\varepsilon}{2} \quad \text{---(5)}$$

Hence proved.

b) From part (a) we had

$$\| J_{m+1} - J_m^* \|_\infty < \frac{\varepsilon}{2}$$

Applying T operator & using T is a contraction

$$\Rightarrow \| T J_{m+1} - J^* \| < \frac{\beta\varepsilon}{2} \quad (\because T J^* = J^*)$$

But $T_{\pi\varepsilon} J_{m+1} = T J_{m+1}$ $\quad (\because T J^* = J^*)$

$$\Rightarrow \| T_{\pi\varepsilon} J_{m+1} - J^* \| < \frac{\beta\varepsilon}{2} \quad \text{---} (6)$$

In showing each policy iteration indeed gives us a policy improvement, (which is what we had also done)

we showed (in class)

$$J_{\pi\varepsilon} \leq J_{m+1} \Rightarrow T_{\pi\varepsilon} J_{\pi\varepsilon} \leq T_{\pi\varepsilon} J_{m+1}$$

($\because T_{\pi\varepsilon}$ is a monotone operator.)

$$\Rightarrow T_{\pi\varepsilon} J_{\pi\varepsilon} - J^* \leq T_{\pi\varepsilon} J_{m+1} - J^*$$

But $T_{\pi k} J_{\pi k} = J_{\pi k} \Rightarrow J_{\pi\varepsilon} - J^* \leq T_{\pi\varepsilon} J_{m+1} - J^*$

$$\Rightarrow \| J_{\pi\varepsilon} - J^* \| \leq \| T_{\pi\varepsilon} J_{m+1} - J^* \| \quad \text{---} (7)$$

plug (7) in (6)

$$\Rightarrow \| J_{\pi\varepsilon} - J^* \|_\xi \leq \| T_{\pi\varepsilon} J_{m+1} - J^* \|_\xi < \frac{\beta\varepsilon}{2}$$

$$\Rightarrow \| J_{\pi\varepsilon} - J^* \|_\xi < \frac{\beta\varepsilon}{2} ; \text{ But } 0 < \beta < 1 \Rightarrow \frac{\beta\varepsilon}{2} < \varepsilon$$

$$\Rightarrow \| J_{\pi\varepsilon} - J^* \|_\xi < \varepsilon$$

Hence proved