Reinforcement Learning

AI 3000

Assignment No: 01

K. Surya Prakash

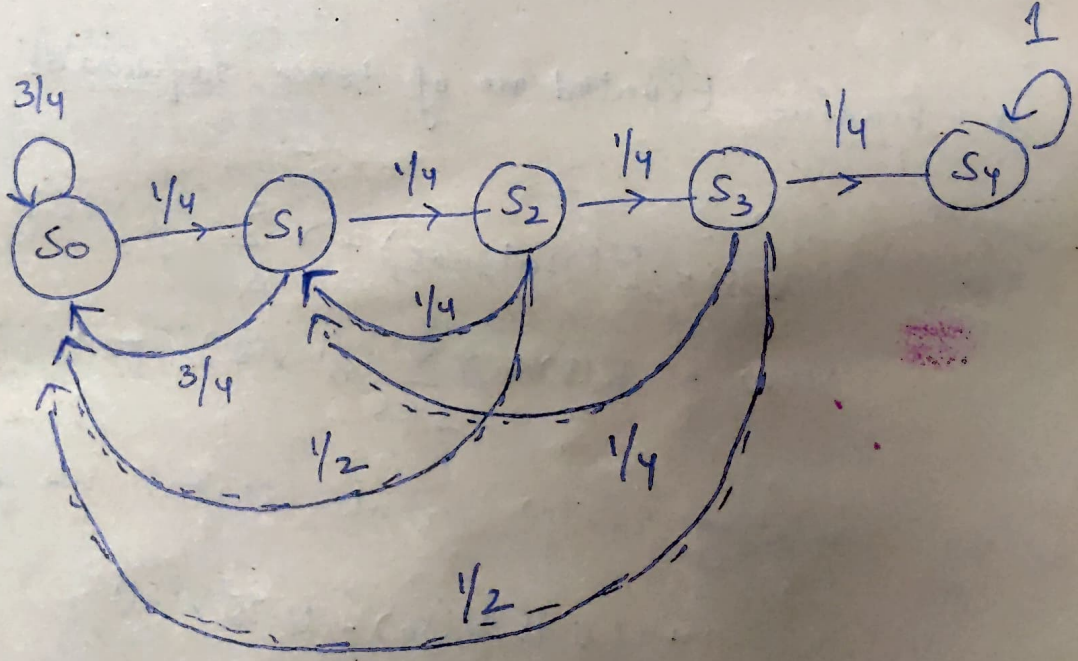
EE18BTECH11026

P1) Markov Reward Process:

@ States: $< S_0, S_1, S_2, S_3, S_4 >$

S_0: Start state
S_1: When a 1 appears on the die
S_2: When 2 appears after 1
S_3: When 3 appears after 2
S_4: When 4 appear after 3.

State transition Diagram:



$$P = \begin{array}{c} \\ S_0 \\ S_1 \\ S_2 \\ S_3 \\ S_4 \end{array} \begin{array}{c} \begin{array}{ccccc} S_0 & S_1 & S_2 & S_3 & S_4 \end{array} \\ \begin{bmatrix} 0.75 & 0.25 & 0 & 0 & 0 \\ 0.5 & 0.25 & 0.25 & 0 & 0 \\ 0.5 & 0.25 & 0 & 0.25 & 0 \\ 0.5 & 0.25 & 0 & 0 & 0.25 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

S_4 is a terminal state.

(b) Reward func: (R)

$$R(S) = -1 \quad \forall\, S \in \{S_0, S_1, S_2, S_3\}$$
$$= 0 \qquad S = S_4.$$

Discount factor: (γ)

$$\gamma = 1.$$

Finding Expected no. of tosses required.

Belmann Equ:

$$V(S) = E(r_{t+1} | s_t = s) + \gamma \sum_{s' \in S} P_{ss'} V(s')$$

$$V(S) = R(S) + \gamma \sum_{s' \in S} P_{ss'} V(s') \longrightarrow (1)$$

By calculating Equ (1) for all states, we get

$$V(S_0) = -1 + (0.75\, V(S_0) + 0.25\, V(S_1))$$

$$V(S_1) = -1 + (0.5\, V(S_0) + 0.25\, V(S_1) + 0.25\, V(S_2))$$

$$V(S_2) = -1 + (0.5\, V(S_0) + 0.25\, V(S_1) + 0.25\, V(S_3))$$

$$V(S_3) = -1 + (0.5\,V(S_0) + 0.25\,V(S_1) + 0.25\,V(S_4))$$

$$V(S_4) = 0$$

By solving the above eqns.

we get

$$V(S_0) = -256$$

$$V(S_1) = -252$$

$$V(S_2) = -240$$

$$V(S_3) = -192$$

$$V(S_4) = 0.$$

The value of S_0 is

Since we choose $R(s) = -1, \forall s \in \{S_0 \ldots S_3\}$, it represents that a step is counted whenever we move from one state, other than at S_4, which is a terminal state

$\therefore$ $[V(S_0)]$ should determine the average no. of steps taken to reach the pattern.

$\therefore$

"Average" no. of steps = 256//

P2 Finite Horizon MDP.

Reward : $3x^2+5$.

$S \in \{1, 2, 3, 4\}$ denoted as

$A \in \{\text{Continue, Quit}\} = \{C, Q\}$

$\gamma = 1$;

(a) $V_N(s)$ ..?

Since at 'N'; we need to quit anyways, we'll be left with the reward that we get

$\therefore \boxed{V_N(s) = 3s^2+5}$

$V_N(1) = 8$
$V_N(2) = 17$
$V_N(3) = 32$
$V_N(4) = 53$

(b) $Q^{N-1}(s,a)$ ..?

$Q^{N-1}(s, Q) = R(s) = 3s^2+5 \rightarrow (1)$

$Q^{N-1}(s, C) = 0 + \underset{s}{E}(V^N(s)) = \frac{1}{4}\left(\sum_{s=1}^{4} 3s^2+5\right) \rightarrow (2)$

$\Rightarrow Q^{N-1}(1, Q) = 8$
$Q^{N-1}(2, Q) = 17$
$Q^{N-1}(3, Q) = 32$
$Q^{N-1}(4, Q) = 53$

$Q^{N-1}(1, C)$
$Q^{N-1}(2, C)$
$Q^{N-1}(3, C)$
$Q^{N-1}(4, C)$

$\frac{1}{4}(8+17+32+53) = 27.5$

(c) $V^{N-1}(s) = \max_{a \in A} Q^{N-1}(s,a)$

$= \max\left(Q^{N-1}(s,Q), Q^{N-1}(s,C) \right)$

$V^{N-1}(1) = \max(8, 27.5) \quad = 27.5$

$V^{N-2}(2) = \max(17, 27.5) \quad = 27.5$

$V^{N-2}(3) = \max(32, 27.5) \quad = 32$

$V^{N-2}(4) = \max(53, 27.5) \quad = 53$

(d) for $2 < n \leq N$

$V^{n-1}(s) = \max_{a \in A} Q^{n-1}(s,a)$

$= \max\left(Q^{n-1}(s,Q), Q^{n-1}(s,C) \right)$

$= \max\left(3s^2 + 5, E(V^n(s)) \right)$

(e) for $2 < n \leq N$

$Q^{n-1}(s,C) = \underset{\text{no reward if contd.}}{0} + E_s(V^n(s))$

$= E_s\left(\max_{a \in A} (Q^n(s,a)) \right)$

$= E_s\left(\max\left(Q^n(s,Q), Q^n(s,C) \right) \right)$

$= E_s\left(\max\left(3s^2+5, Q^n(s,C) \right) \right)$

$$\Phi^{n-1}(s,c) = \underset{s}{E}\left(\max\left(3s^2+5 , \Phi^{n}(s,'c') \right) \right)$$

(b) Optimal policy:

$$\pi^{*n}(s) = \begin{cases} Q: \text{quit} & ; \text{ if } \Phi^{n}(s,Q) \geq \Phi^{n}(s,C) \\ C: \text{continue} & ; \text{ o.w} \end{cases}$$

$\forall \; 2 \leq n \leq N$

(c) Upon calculating $\Phi^{n}(s,c)$: $\forall \; s \in \{1,2,3,4\}$

$$\Phi^{N}(s,c) = 0$$

$$\Phi^{N-1}(s,c) = 27.5$$

$$\Phi^{N-2}(s,c) = 35$$

$$\Phi^{N-3}(s,c) = 39.5$$

$$\Phi^{n}(s,c) \in [35, 53) \quad \forall \; n \in [1, N-2]$$

$$\therefore \; \Phi^{n}(s,c) < 53$$

∴ The policy will be:

$\forall\ n \in [1, N-2]$, $\pi_n^*(s) = C$; $s \in \{1, 2, 3\}$

$\pi_n^*(s) = Q$; $s \in \{4\}$

$n = N-1$; $\pi_n^*(s) = C$; $s \in \{1, 2\}$

$\pi_n^*(s) = Q$; $s \in \{3, 4\}$

$n \geq N$; $\pi(s) = Q$; $s \in \{1, 2, 3, 4\}$

(g). ∴ $\pi^*(s)$ depends upon 'n'.

→ The optimal policy is non-stationary.

Q3. Value Iteration

$M \langle S, A, P, R, \gamma \rangle$

$|S| < \infty$; $|A| < \infty$, $\gamma \in [0, 1)$

$\hat{M} \langle S, A, P, \hat{R}, \gamma \rangle$

$$|R(s,a,s') - \hat{R}(s,a,s')| \leq \varepsilon$$

* Given a policy π, V^π, $\hat{V}^\pi$ are policies

a) Expression relating V^π, $\hat{V}^\pi$,

$$V^\pi(s) = \sum_a \pi(a|s) \cdot \sum_{s'} P^a_{ss'} \left(R^a_{ss'} + \gamma V^\pi(s') \right)$$

$$\hat{V}^\pi(s) = \sum_a \pi(a|s) \cdot \sum_{s'} P^a_{ss'} \left(\hat{R}^a_{ss'} + \gamma \hat{V}^\pi(s') \right)$$

Consider

$$V^\pi(s) - \hat{V}^\pi(s) = \sum_a \pi(a|s) \sum_{s'} P^a_{ss'} \left(R^a_{ss'} - \hat{R}^a_{ss'} \right) + \gamma \sum_a \pi(a|s) \cdot \sum_{s'} P^a_{ss'} \left(V^\pi(s') - \hat{V}^\pi(s') \right)$$

absolute value on both sides & $|a+b| \leq |a| + |b|$ on RHS

$$|V^\pi(s) - \hat{V}^\pi(s)| \leq \left| \sum_a \pi(a|s) \sum_{s'} P^a_{ss'} \left(R^a_{ss'} - \hat{R}^a_{ss'} \right) \right| + \left| \gamma \sum_a \pi(a|s) \sum_{s'} P^a_{ss'} \left(V^\pi(s') - \hat{V}^\pi(s') \right) \right|$$

$$\leq \sum \pi(a/s) \sum_{s'} P^a_{ss'} \left| R^a_{ss'} - \hat{R}^a_{ss'} \right|$$

$$+ \gamma \sum_a \pi(a/s) \sum_s P^a_{ss'} \left| V^\pi(s') - V^\pi(s') \right|$$

($\because$ π & P are ~~pos~~ non-negative)

$$\leq \sum \pi(a/s) \sum_{s'} P^a_{ss'} (\varepsilon)$$

$$+ \gamma \sum_a \pi(a/s) \cdot \sum_s P^a_{ss'} \left| V^\pi(s') - V^\pi(s') \right|$$

($\because \sum_a \pi(a/s) \sum_{s'} P^a_{ss'}(\varepsilon) = \sum_a \pi(a/s)\, \varepsilon = \varepsilon$)

$\because \sum_a \pi(a/s) \leq 0$

$$\left| \hat{V}(s) - \hat{V}^\pi(s) \right| \leq \varepsilon + \gamma \sum_a \pi(a/s) \sum_s P^a_{ss'} \left| V^\pi(s') - \hat{V}^\pi(s') \right|$$

(b) $$V_\alpha = \max_a \left[\sum_{s' \in S} P^a_{ss'} \left(R^a_{ss'} + \gamma V_\alpha(s') \right) \right]$$

$$\hat{V}_\alpha = \max_a \left[\sum_{s' \in S} P^a_{ss'} \left(\hat{R}^a_{ss'} + \gamma \hat{V}_\alpha(s') \right) \right]$$

$$V_\alpha - \hat{V}_\alpha = \max_a \left(\sum_{s' \in S} P^a_{ss'} \left(R^a_{ss'} + \gamma V_\alpha(s') \right) \right.$$

$$\left. - \max_a \left(\sum_{s' \in S} P^a_{ss'} \left(\hat{R}^a_{ss'} + \gamma \hat{V}_\alpha(s') \right) \right) \right.$$

$$\max(a, b) \geq \max(a) - \max(b)$$

$$\leq \max_a \left(\sum_{s' \in S} P^a_{ss'} \left((R^a_{ss'} - \hat{R}^a_{ss'}) + \gamma (V_\alpha(s') - \hat{V}_\alpha(s')) \right) \right.$$

$$\therefore R^a_{ss'} - \hat{R}^a_{ss'} \leq \varepsilon$$

$$\leq \max_a \left(\sum_{s' \in S} P^a_{ss'} (\varepsilon) + \gamma (V_\alpha(s') - \hat{V}_\alpha(s')) \right)$$

$$V_\alpha - \hat{V}_\alpha \leq \max_a \left(\varepsilon + \gamma \sum_{s' \in S} P^a_{ss'} (V_\alpha(s') - \hat{V}_\alpha(s')) \right)$$

$$V_\pi - \hat{V}_\pi \leq \varepsilon + \max_a \left(\gamma \sum_{s' \in S} P^a_{ss'} \left(V_\pi(s') - \hat{V}_\pi(s') \right) \right)$$

(c) Yes, M & M' have the same optimal policy. Because,

$V_\pi - \hat{V}_\pi$ are far by $\varepsilon \geq 0$

∴ The π which achieves V_π also achieves $\hat{V}_\pi$.

⇒ thus both have same optimal policy

P4. Effect of Noise & Discounting.

Need to map each given path to possible settings of η & γ

$\Rightarrow$ $\eta = \{0, 0.5\}$; $\gamma \in \{0.1, 0.9\}$

a) Close Exit (state with reward +1) but risk of cliff (dashed path to +1)

$\longrightarrow$ $\boxed{\gamma = 0.1,\ \eta = 0}$

Reason: with 'high' discounting factor discourages the agent to travel long paths as the final reward will be scaled down exponentially by γ.

→ Since ($\eta = 0$): there is no randomness, hence, risk of falling into cliff is less.

(b) Distant exit (state with +10) ; risk the cliff (dashed path +10)

$\longrightarrow$ $\boxed{\gamma = 0.9.\ \ \eta = 0}$

Reason:

With $\gamma \uparrow \Rightarrow$ agent will prefer the distant route if the net reward it gets at

the end is high. Since higher γ will not significantly scale down the reward.

→ $\eta = 0$ ⇒ No noise ⇒ Reduces risk of falling into a cliff; can pick the path with cliff.

(c) Close exit (reward +1); avoid cliff (solid path +1)

→ $\boxed{\gamma = 0.1 \quad \eta = 0.5}$

$\gamma \downarrow$ ⇒ prefers shortest path although the reward at final state is low;

$\eta \uparrow$ ⇒ more randomness ⇒ more risky to fall into cliff ⇒ discourages encourages the agent to avoid cliff.

(d) Distant exit (reward +10); avoiding cliff (solid path +10)

→ $\boxed{\gamma = 0.9 \; ; \; \eta = 0.5}$

→ if $\gamma \uparrow$; encourages to take distant path if net reward is greater.

→ $\eta \uparrow$ ⇒ risk of falling into cliff ⇒ encourages to avoid path with -ve rewards

P5. On Value Iteration Algo.

$M := \langle S, A, P, R, \gamma \rangle$; $|S| < \infty$; $|A| \leq \infty$, $\gamma \in [0, 1)$

→ Given : policy π ; need to evaluate $V^{\pi}(s)$ using iterative update.

$$V_{k+1} \leftarrow \sum_{a} \pi(a/s) \sum_{s'} P^{a}_{ss'} [R^{a}_{ss'} + \gamma V^{\pi}_{k}(s')]$$

$$\boxed{V^{\pi}_{k+1} = R + \gamma P V^{\pi}_{k}} \quad \text{(following notation from cls)}$$

→ (1)

while,

V^{π}_{k} Converge to $V^{\pi} = (I - \gamma P)^{-1} R$

$$\Rightarrow \boxed{V^{\pi} = R + \gamma P V^{\pi}} \rightarrow (2)$$

⇒ Given, algo will terminate

$$\| V_{k+1} - V_k \|_{\infty} \leq \varepsilon \quad ; \varepsilon > 0. \rightarrow (3)$$

Q1) Need to prove : $\| V_{k+1} - V^{\pi} \|_{\infty} \leq \frac{\varepsilon \gamma}{1 - \gamma}$

Consider LHS:

$$\| V_{k+1} - V^{\pi} \|_{\infty} = \| R + \gamma P V_k - (R + \gamma P V^{\pi}) \|_{\infty}$$

$$= \| \gamma P (V_k - V^\pi) \|_\infty$$

$$\leq \gamma \| P \|_\infty \| V_k - V^\pi \|_\infty$$

$$\leq \gamma \| V_k - V^\pi \|_\infty$$

$$\leq \gamma \| (V_k - V_{k+1}) + (-V^\pi + V_{k+1}) \|_\infty$$

$$\leq \gamma \left(\| V_k - V_{k+1} \|_\infty + \| V_{k+1} - V^\pi \|_\infty \right)$$

(Δk inequality)

$$\| V_{k+1} - V^\pi \|_\infty \leq \gamma \left(\| V_{k+1} - V_k \|_\infty + \| V_{k+1} - V^\pi \|_\infty \right)$$

$$(1-\gamma) \| V_{k+1} - V^\pi \|_\infty \leq \gamma \| V_{k+1} - V_k \|_\infty$$

$$\leq \gamma \varepsilon \quad (\because \text{From Eq (3)})$$

$$\boxed{\| V_{k+1} - V^\pi \|_\infty \leq \frac{\gamma \varepsilon}{1-\gamma}} \quad //$$

(b) Need to prove

$$\| V_{k+1} - V^\pi \|_\infty \leq \gamma^k \| V_1 - V_\infty^\pi \|.$$

Substituting using Eq (1) & (2),

Consider LHS:

$$\|V_{k+1} - V^{\pi}\|_\infty = \|(R+\gamma P V_k) - (R+\gamma P V^{\pi})\|_\infty$$

$$= \|\gamma P (V_k - V^{\pi})\|_\infty$$

$$\leq \gamma \|P\|_\infty \|V_k - V^{\pi}\|_\infty$$

$$\leq \gamma \|V_k - V^{\pi}\|_\infty \quad \text{(proved in class)}$$

$$\therefore \|V_{k+1} - V^{\pi}\|_\infty \leq \gamma \|V_k - V^{\pi}\|_\infty \rightarrow (a)$$

Similarly

$$\Rightarrow \|V_k - V^{\pi}\|_\infty \leq \gamma \|V_{k-1} - V^{\pi}\|_\infty$$

$$\|V_{k-1} - V^{\pi}\|_\infty \leq \gamma \|V_{k-2} - V^{\pi}\|_\infty$$

$$\vdots$$

$$\|V_2 - V^{\pi}\|_\infty \leq \gamma \|V_1 - V^{\pi}\|_\infty$$

} k-1 inequalities

$$\Rightarrow \|V_k - V^{\pi}\|_\infty \leq \gamma^{k-1} \|V_1 - V^{\pi}\|_\infty$$

Substituting in eq (a)

$$\|V_{k+1} - V^{\pi}\|_\infty \leq \gamma \left(\gamma^{k-1} \|V_1 - V^{\pi}\|_\infty\right)$$

$$\leq \gamma^k \|V_1 - V^{\pi}\|_\infty$$

Hence proved

(c) v : denote the value func.

Lets 'S' → define the states

$S \in \mathcal{S}$: $\mathcal{S}$: state space

$|\mathcal{S}| = d$: d states

then v is represented by a vector of d-dimension

$v : R^d \to R^d$

(∵ $\mathcal{L}(v) = \max_{a \in A} [R^a + \gamma P^a v]$

$\mathcal{L} : \mathcal{V} \to \mathcal{V}$: $\mathcal{L}$ also maps $R^d \to R^d$)

⇒ Need to prove:

$\mathcal{L}(u) \leq \mathcal{L}(v)$; if $u \leq v$

$u \leq v \Rightarrow u(i) \leq v(i) \; \forall \; i \in \{1, \ldots d\}$

⇒ Consider the following:

$v \geq u$. (as defined)

Consider an arbitrary action 'a' $\in A$; s.t.

$P^a_v \geq P^a(u)$ (holds true since P^a contains elements ranging from $[0,1]$)

$\gamma P^a_v > \gamma P^a u$ ($\because \gamma \in [0,1)$)

$$R^a + \gamma P^a_v \geq R^a + \gamma P^a_u$$ (adding R^a for both sides)

↳ (1)

⇒ Now consider the following

$$\max_{a \in A} (R^a + \gamma P^a_v) \geq R^a + \gamma P^a_v$$ (defn of maximum)

↳ ∀ a ∈ A.

$$\geq R^a + \gamma P^a_u \quad \text{(From eq. (1))}$$

$$\geq \max_{a \in A} (R^a + \gamma P^a_u) \quad (\because \text{By defn of max operator})$$

$$\therefore \max_{a \in A} (R^a + \gamma P^a_v) \geq \max_{a \in A} (R^a + \gamma P^a_u)$$

$$L(v) \geq L(u)$$

$$\Rightarrow L(w) \geq L(v) \text{ ; if } u \leq v$$

Hence, Bellman optimality operator is monotonic.

Pb. On Contractions

@ P, Q contractions on normed vector space $\langle V, \|\cdot\| \rangle$.

ie; $P: V \to V$; $Q: V \to V$

$\forall\ u, v \in V$; $r_p, r_q \in [0,1)$

$\|P(v) - P(u)\| \leq r_p \|v - u\| \to (1)$

$\|Q(v) - Q(u)\| \leq r_q \|v - u\| \to (2)$

By the defn of contraction

* Prooving $P \circ Q$, $Q \circ P$ are contractions on the same vector space.

⇒ Consider vector space $\langle V, \|\cdot\| \rangle$.

$\boxed{P \circ Q : V \to V}$ (since, range of Q is V and domain of P is V)

$\therefore P \circ Q : V \to V$

$\forall\ u, v \in V$

$\|P(Q(v)) - P(Q(u))\| \leq r_p \|Q(v) - Q(u)\|$

$\leq r_p \cdot r_q \|v - u\|$

$= r_p r_q \|v - u\|$ → using eq(2)

$\therefore \|P(Q(v)) - P(Q(u))\| \leq r_p r_q \|v - u\|$

Since $r_p, r_q \in [0,1)$, $\eta_{pq} = r_p r_q \in [0,1)$

$\therefore P \circ Q$ is a contraction, with η_{pq} as Lipsitz coefficient, & $\eta_{pq} \in [0,1)$

Consider $Q \circ P$

$\boxed{Q \circ P : \mathcal{V} \to \mathcal{V}}$ ($\because$ Range of P is $\mathcal{U}$ & Domain of Q is $\mathcal{U}$, Range of Q is $\mathcal{V}$)

consider $v, u \in \mathcal{V}$

$$\| Q(P(v)) - Q(P(u)) \| \leq r_q \| P(v) - P(u) \|$$

$$\leq r_q (r_p \| v - u \|)$$

$$= r_p r_q \| v - u \| \quad (\text{From eq-(1)})$$

$$\therefore \| Q(P(v) - Q(P(u)) \| \leq r_q r_p \| v - u \|$$

$\because r_p, r_q \in [0,1) \Rightarrow$ $\eta = r_p r_q \in [0,1)$

$$\Rightarrow \| Q(P(v) - Q(P(u)) \| \leq \eta_{qp} \| v - u \| \; ; \; \eta_{qp} \in [0,1)$$

Hence, $Q \circ P$ is also a contraction on the same normal vector space $\mathcal{V}$. with coefficient $\eta_{qp} \in [0,1)$

(b) As calculated earlier,

Lipschitz coefficient for $P \cdot \Phi = \Phi \circ P = \gamma_P \gamma_Q$

where γ_P is the coeff for P:

γ_Q " " for Φ.

(c) given $B := F \circ L$

$\Rightarrow$ L is the Bellman optimality operator.

It has been proved that, L is a contraction. (in class).

$\therefore$ L is a contraction, and is defined on a normal vector space denoted by $\langle V, \|\cdot\| \rangle$

$$L : V \to V$$

* Conditions for B to converge.

a) F should be defined on the same normal vector space as L i.e; 'V'

b) F should be a contraction.

c) The initial guess in the value iteration algo (lets say) $v \in V$

Reason:

→ Banach Fixed point theorem states that; if for normal vector space $\langle V, \|\cdot\| \rangle$, a contraction $B : V \to V$; . Then the iterative application of B converges to a fixed point in V (thus resulting in a unique soln).

∴ For $B = F \circ L$ should be a contraction on the same vector space of L i.e., 'V'

→ For that to happen, we proved in (6a) that if F, L defined on the same normed vector space $\langle V, \|\cdot\| \rangle$ and are contractions,

then $B = F \circ L$ is also defined on $\langle V, \|\cdot\| \rangle$ and is also a contraction.

⇒ Thus B satisfies Banach Fixed Point thm. & thus converges when the above conditions are satisfied

⟶