Solutions: Assignment 2 & 3

MAINAK BISWAS
IISc, PMRF

08 September 2025 19:42

THEORY

1) : – (6)

terminal state

$\mathcal{S} = \{s_1, s_2, s_3, s_4, s_0\}$

Given trajectories:

T1: $\{s_2, -1, s_3, 2, s_2, 5, s_1, 0, s_1, 3, s_4, 7, s_0\}$

T2: $\{s_4, 3, s_2, -3, s_1, 2, s_1, 5, s_0\}$

T3: $\{s_3, 2, s_3, 1, s_4, 4, s_0\}$

T4: $\{s_1, 1, s_2, 0, s_3, 2, s_4, 2, s_0\}$

T5: $\{s_4, 6, s_1, -1, s_2, -2, s_2, 2, s_3, 6, s_4, 3, s_0\}$

(Let $\gamma = 1$) – as all reach terminal state – simply add rewards.

(any $\gamma \in (0,1]$ will be accepted)

(a) First Visit Method:

$V(s_0) = 0$ (terminal state)

Collecting trajectory rewards :–

→ (sum of reward of a trajectory)

	s_1	s_2	s_3	s_4
T_1	10 (0, 3, 7)	16 (−1, 2, 5, 0, 3, 7)	17 (2, 5, 0, 3, 7)	7 (7)
T_2	7 (2, 5)	4 (−3, 2, 5)	–	7 (3, −3, 2, 5)
T_3	–	–	7 (2, 1, 4)	4 (4)
T_4	5 (1, 0, 2, 2)	4 (0, 2, 2)	4 (2, 2)	2 (2)
T_5	8 (−1, −2, 2, 6, 3)	9 (−2, 2, 6, 3)	9 (6, 3)	14 (6, −1, −2, 2, 6, 3)
$\hat{V}$ =	(30)	33	37	34

$\hat{V}$ = $\left(\frac{30}{4}\right)$ $\frac{33}{4}$ $\frac{37}{4}$ $\frac{34}{5}$

7.5 8.25 9.25 6.8

$\hat{V}^{(1)}$ $\hat{V}^{(2)}$ $\hat{V}^{(3)}$ $\hat{V}^{(4)}$

(b) Every visit Method:

	S_1	S_2	S_3	S_4
T_1	(0, 3, 7) 10	(-1, 2, 5, 0, 3, 7) 16	(2, 5, 0, 3, 7) 17	(7) 7
	(3, 7) 10	(5, 0, 3, 7) 15		
T_2	(2, 5) 7	(-3, 2, 5) 4	-	(3, -3, 2, 5) 7
				(5) 5
T_3	-	-	(2, 1, 4) 7	(4) 4
			(1, 4) 5	
T_4	(1, 0, 2, 2) 5	(0, 2, 2) 4	(2, 2) 4	(2) 2
T_5	(-1, -2, 2, 6, 3) 8	(-2, 2, 6, 3) 9	(6, 3) 9	(6, -1, -2, 2, 6, 3) 14
		(2, 6, 3) 11		(3) 3

$\hat{V}$: 40 59 42 42

(# occurrences) → 5 6 5 7

8 9.83 8.4 6

11 11 11 11

$\hat{V}(1) = 8$ $\quad \hat{V}(2) = 9.83$ $\quad \hat{V}(3) = 8.4$ $\quad \hat{V}(4) = 6$

2) :— (9)

(a) $q(s,a) = \mathbb{E}\left[\sum_{i=0}^{T} r(s_i, a_i, s_{i+1}) \mid S_0 = s, A_0 = a\right]$ (stochastic end time)

$$= \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[r(s,a,s') + \sum_{i=1}^{T} r(s_i, a_i, s_{i+1}) \,\middle|\, \begin{matrix} S_0 = s \\ A_0 = a \end{matrix}\right]$$

(Controlled MC)

$$= \mathbb{E}\left[r(s,a,s_1) + \mathbb{E}\left[\sum_{i=1}^{T} r(s_i, a_i, s_{i+1}) \,\middle|\, S_1 = s'\right] \,\middle|\, \begin{matrix} S_0 = s \\ A_0 = a \end{matrix}\right]$$

$$q(s,a) = \mathbb{E}\left[r(s,a,s') + V(s') \mid S_0 = s, A_0 = a\right]$$

Let V^* and q^* be the optimal value functions —

Hence,

$$q^*(s,a) = \mathbb{E}\left[r(s,a,s') + V^*(s') \mid S_0 = s, A_0 = a\right]$$

$$\therefore V^*(s) = \max_{a' \in A(s')} q^*(s', a')$$

Therefore, the Bellman optimality Eqn:

$$q^*(s,a) = \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[r(s,a,s') + \max_{a' \in A(s')} q^*(s', a') \,\middle|\, \begin{matrix} S_0 = s \\ A_0 = a \end{matrix}\right]$$

Thus, the Bellman optimality operator's (as discussed in class):

Then; the Bellman optimality operator's (as discussed in class);

natural choice would be:

$\mathcal{H} : \mathbb{R}^{|S| \times |A|} \to \mathbb{R}^{|S| \times |A|}$

$$\mathcal{H} Q(s,a) = \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[r(s,a,s') + \max_{a' \in A(s')} Q(s',a') \,\middle|\, s_0 = s, A_0 = a \right]$$

$$= \sum_{s' \in S} P(s'|s,a) \left[r(s,a,s') + \max_{a' \in A(s')} Q(s',a') \right]$$

[QED]

(b) To show: for proper policy $\exists\, Z \in \mathbb{R}^{|S|}$; $Z(s) > 0$

s.t. $\sum_{s' \in S} P_\pi(s'|s)\, Z(s') \leq \beta Z(s)$ for some $\beta \in (0,1)$

Let $V_{\pi_*}(s) = \max_a Q_{\pi_*}(s,a)$

$$= \max_a \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[r(s,a,s') + V_{\pi_*}(s') \,\middle|\, s,a \right]$$

$$= \max_{a \in A(s)} \sum_{s' \in S} P(s'|s,a) \left[r(s,a,s') + V_\pi(s') \right]$$

(Is valid for any choice of r) — safely assume $r = 1$

$$= 1 + \max_{a \in A(s)} \sum_{s' \in S} P(s'|s,a)\, V_\pi(s')$$

$$\geq \sum_{s' \in S} P(s'|s,a)\, V_\pi(s')$$

$a \sim \pi(\cdot|s)$ (for all policies by definition)

$$\geq 1 + \sum_{a \sim \pi(s)} \pi(a|s) \sum_{s' \in S} P(s'|s,a)\, V_\pi(s')$$

$$= 1 + \sum_{s'} \underbrace{\left(\sum_{a'} p(s'|s,a)\, \pi(a|s) \right)}_{P_\pi(s'|s)} v_{\pi_*}(s')$$

$$v_{\pi_*}(s) \geq 1 + \sum_{s'} P_\pi(s'|s)\, v_{\pi_*}(s')$$

Let; $Z(s) = v_\pi(s)$,

$$\Rightarrow \sum_{s' \in S} P_\pi(s'|s)\, Z(s') \leq Z(s) - 1$$

Let $\beta = \max_s \left(\frac{Z(s) - 1}{Z(s)} \right) \in (0,1)$

$$\Rightarrow \sum_{s' \in S} P_\pi(s'|s)\, Z(s') \leq \beta\, Z(s)$$

$\therefore (\exists \beta$ for $Z)$

(Q E D)

(c) To show $\mathcal{B}$ is a contraction under a Z norm :-

Let $\mathcal{B}$ be a vector of dimension $|S| \times |A|$ such that:

$$\mathcal{B} = \left(\mathcal{B}(s,a) ; \forall\, s \in S,\ a \in A(s) \right)^T$$

Consider two Q-value functions $\mathcal{B}$, $\bar{\mathcal{B}}$

from (a),

$$\therefore H\mathcal{B}(s,a) = \sum_{s' \in S} p(s'|s,a) \left[R(s,a,s') + \max_{a' \in A(s')} \mathcal{B}(s',a') \right]$$

$$\Rightarrow H\mathcal{B}(s,a) - H\bar{\mathcal{B}}(s,a) = \sum_{s' \in S} p(s'|s,a) \left[\max_{a' \in A(s')} \mathcal{B}(s',a') - \max_{a' \in A(s')} \bar{\mathcal{B}}(s',a') \right]$$

(from lemma*) $\leq$ (absolute value) $\sum_{s' \in S} p(s'|s,a) \max_{a \in A(s')} \left| \mathcal{B}(s',a') - \bar{\mathcal{B}}(s',a') \right|$

(from above) $\leq \sum_{s' \in S} P(s'|s,a) \max_{a' \in A(s')} \left| Q(s',a') - \bar{Q}(s',a') \right|$ — (†)

Lemma * $\max_x f(x) - \max_x g(x) \leq \max_x |f(x) - g(x)|$

Proof:

$f(x) - g(x) \leq \max_x (f(x) - g(x))$

$\leq$

$f(x) - \max_x g(x)$

$\Rightarrow$

$\Rightarrow f(x) - \max_x g(x) \leq \max_x (f(x) - g(x))$ (true for $\forall x$)

$\therefore \max_x f(x) - \max_x g(x) \leq \max_x (f(x) - g(x))$

$\leq \max_x |f(x) - g(x)|$ (definition of absolute value)

Hence

$$HQ(s,a) - H\bar{Q}(s,a) \leq \sum_{s' \in S} P(s'|s,a) \max_{a' \in A(s')} \left| Q(s',a') - \bar{Q}(s',a') \right|$$

Note:

(We can put 1.1 on LHS as if we replace Q by $\bar{Q}$; the RHS of eqn† shall remains the same)

multiply and divide by $Z(s')$

$$\therefore \left| HQ(s,a) - H\bar{Q}(s,a) \right| \leq \sum_{s' \in S} Z(s') P(s'|s,a) \max_{a' \in A(s')} \frac{\left| Q(s',a') - \bar{Q}(s',a') \right|}{Z(s')}$$

$$\leq \sum_{s' \in S} P(s'|s,a) Z(s') \max_{a' \in A(s')} \frac{\left| Q(s',a') - \bar{Q}(s,a) \right|}{Z(s')}$$

$s' \in S$ (added)

$$= \|Q - \bar{Q}\|_2 \sum_{s'\in S} P(s'|s,a)\, z(s')$$

from theorem in (b), $\exists\ \beta \in [0,1)$ s.t.

$$|HQ(s,a) - H\bar{Q}(s,a)| \le \|Q - \bar{Q}\|\, \beta\, z(s)$$

$$\Rightarrow \max_{s,a} \frac{|HQ(s,a) - H\bar{Q}(s,a)|}{z(s)} \le \beta \|Q - \bar{Q}\|$$

(true for all a, s)

$$\Rightarrow \|HQ - H\bar{Q}\|_z \le \beta \|Q - \bar{Q}\|_z.$$

$\therefore$ (H is a contraction under z norm)

$(Q = D)$

(d) Now; since H is a contraction and.

$HQ_* = Q_*$ (Bellman optimality eqn).

By Banach's fixed pt theorem, there exists a unique fixed pt for H. and Q^* is the fixed point.

$\therefore$ Q-learning algorithm using H converges to Q^*

Thus; we proved convergence of the Q-learning algorithm.

(tabular and model based)

3) : − (20)

(a) Definition:

$$V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1}) \,\middle|\, s_t = s\right]$$

$$V_\pi(s) = \mathop{\mathbb{E}}_{\substack{a\sim\pi(\cdot|s)\\ s'\sim P(\cdot|s,a)}}\left[\sum_{i=0}^{\infty}\gamma^i R(S_{t+i}, a_{t+i}, S_{t+i+1}) \,\middle|\, S_t = s\right]$$

$$= \mathbb{E}\Big[R(S_t, a_t, S_{t+1}) + \gamma R(S_{t+1}, a_{t+1}, S_{t+2}) + \sum_{t=2}^{\infty}\gamma^i R(S_{t+i}, a_{t+i}, S_{t+i+1}) \,\Big|\, S_t = s\Big]$$

$$= \mathbb{E}\Big[R(S_t, a_t, S_{t+1}) + \gamma R(S_{t+1}, a_{t+1}, S_{t+2}) + \gamma^2 \underset{\pi}{\mathbb{E}}\Big[\sum_{t=0}^{\infty}\gamma^i (S_{t+i+2}, a_{t+i+2}, S_{t+i+3}) \,\Big|\, S_{t+2}\Big] \,\Big|\, S_t = s\Big]$$

(iterated Expectation)

$$= \mathbb{E}\Big[R(S_t, a_t, S_{t+1}) + \gamma R(S_{t+1}, a_{t+1}, S_{t+2}) + \gamma^2 V_\pi(S_{t+2}) \,\Big|\, S_t = s\Big]$$

$a_t \sim \pi(\cdot|s_t)$

$S_{t+1} \sim P(\cdot|S_t, a_t)$ (2step reward function)

$a_{t+1} \sim \pi(\cdot|S_{t+1})$

$S_{t+2} \sim P(\cdot|S_{t+1}, a_{t+1})$ [Hence proved]

(b)

Objective - linear

$$d(\omega) = \frac{1}{2}\sum_{s\in S} d_\pi(s)\left(V_\pi(s) - \Phi^T(s)\omega\right)^2.$$

Gradient descent based update:

$$\frac{\partial L}{\partial \omega} = -\sum_{s\in S} d_\pi(s)\left(V_\pi(s) - \Phi^T(s)\omega\right)\Phi(s)$$

$$= -\mathop{\mathbb{E}}_{s\sim d_\pi}\left[\left(V_\pi(s) - \Phi^T(s)\omega\right)\Phi(s)\right]$$

From part(a),

$$= \mathop{\mathbb{E}}_{s\sim d_\pi}\Big[\mathbb{E}\left[R(S_t, a_t, S_{t+1}) + \gamma R(S_{t+1}, a_{t+1}, S_{t+2}) + \gamma^2 V_\pi(S_{t+2}) - \Phi^T(S_t)\omega\right]\Phi(s)\Big]$$

$S \sim d_\pi$ $\quad + \gamma^2 V_\pi(S_{t+2}) - \hat{\Phi}^T(S_t)w\big]$
$\hat{\Phi}(S_t)\Big]$

$$= \mathbb{E}_{\substack{S_t \sim d_\pi(\cdot) \\ a_t \sim \pi(\cdot|S_t) \\ S_{t+1} \sim P(\cdot|S_t, a_t) \\ a_{t+1} \sim \pi(\cdot|S_{t+1}) \\ S_{t+2} \sim P(\cdot|S_{t+1}, a_{t+1})}} \Big[\Big(r(S_t, a_t, S_{t+1}) + \gamma r(S_{t+1}, a_{t+1}, S_{t+2}) + \gamma^2 V_\pi(S_{t+2}) - \hat{\Phi}^T(S_t)w \Big) \hat{\Phi}(S_t) \Big]$$

$\therefore$ SGD update: replace the Expectation and update per sample.

$$\therefore w_{t+1} = w_t + \alpha_t \Big(r(S_t, a_t, S_{t+1}) + \gamma r(S_{t+1}, a_{t+1}, S_{t+2}) + \gamma^2 V_\pi(S_{t+2}) - \hat{\Phi}^T(S_t)w_t \Big] \hat{\Phi}(S_t)$$

$V_\pi(S_{t+2})$ is unknown: (Sutton's Idea)

$\therefore$ let $V_\pi(S_{t+2}) \approx \hat{\Phi}^T(S_{t+2})\, w_t$

Algorithm:

$$\Rightarrow w_{t+1} = w_t + \alpha_t \Big(r(S_t, a_t, S_{t+1}) + \gamma r(S_{t+1}, a_{t+1}, S_{t+2}) + \gamma \hat{\Phi}^T(S_{t+2})w_t - \hat{\Phi}^T(S_t)w_t \Big) \hat{\Phi}(S_t)$$

Iterate until convergence - with sampling S_{t+1}'s based on $P(\cdot|S_t, a_t)$ and a_t's based on π.

(c) limiting ODE:

let:

$$f(w_t, S_t, a_t, S_{t+1}, a_{t+1}, S_{t+2}) = \Big(r(S_t, a_t, S_{t+1}) + \gamma r(S_{t+1}, a_{t+1}, S_{t+2}) + \gamma^2 \hat{\Phi}^T(S_{t+2})w_t - \hat{\Phi}^T(S_t)w_t \Big) \hat{\Phi}(S_t)$$

Let $\mathcal{F}_t = \sigma(w_0, S_0, a_0, S_1, w_1, S_1, a_1, \ldots\ldots w_t, S_t)$ be a

. Collection of observable for these r.v.s

$$\therefore f(w_t, s_t) = \mathbb{E}\left[f(w_t, s_t, a_t, s_{t+1}, a_{t+1}, s_{t+2}) \,\middle|\, \mathcal{F}_t \right]$$

$$= \mathbb{E}\left[\Big(\underbrace{r(s_t, a_t, s_{t+1})}_{\text{expected reward.}} + \gamma r(s_{t+1}, a_{t+1}, s_{t+2}) + \gamma^2 \bar{\phi}^T(s_{t+2}) w_t - \bar{\phi}^T(s_t) w_t \Big) \bar{\phi}(s_t) \,\middle|\, w_t, s_t \right]$$

($\bar{\phi}^T(s_t) w_t$: known)

$$= r(s_t)\bar{\phi}(s_t) - \bar{\phi}(s_t)\bar{\phi}^T(s_t) w_t + \gamma \mathbb{E}\left[\underbrace{r(s_{t+1}, a_{t+1}, s_{t+2}) + \gamma \bar{\phi}^T(s_{t+2}) w_t}_{\text{III}} \,\middle|\, \underset{s_t}{w_t,} \right] \bar{\phi}(s_t) \quad \text{---(C.1)}$$

$$\text{III} = \mathbb{E}\Big[\gamma \bar{\phi}(s_t) \mathbb{E}\Big[\Big(r(s_{t+1}, a_{t+1}, s_{t+2}) + \gamma \bar{\phi}^T(s_{t+2}) w_t \Big) \Big| s_{t+1} \Big] \Big| s_t, w_t \Big]$$

(Iterated Exp)

(should be a function of s_t, w_t only)

$s_t \xrightarrow{\pi} a_t \xrightarrow{P} s_{t+1} \xrightarrow{\pi} a_{t+1}$

$$= \gamma \bar{\phi}(s_t) \sum_{a_t, s_{t+1}} \pi(a_t|s_t) P(s_{t+1}|s_t, a_t) \mathbb{E}\left[r(s_{t+1}, a_{t+1}, s_{t+2}) + \gamma \bar{\phi}^T(s_{t+2}) w_t \,\middle|\, s_{t+1} \right]$$

$$= \gamma \bar{\phi}(s_t) \sum_{a_t, s_{t+1}} \pi(a_t|s_t) P(s_{t+1}|s_t, a_t) \left[\sum_{a_{t+1}, s_{t+2}} \pi(a_{t+1}|s_{t+1}) P(s_{t+2}|s_{t+1}, a_{t+1}) \Big(\underbrace{r(s_{t+1}, a_{t+1}, s_{t+2})}_{r(s_{t+1})} + \gamma \bar{\phi}^T(s_{t+2}) w_t \Big) \right]$$

$$= \gamma \bar{\phi}(s_t) \sum_{s_{t+1}} \underbrace{P_\pi(s_{t+1}|s_t) r(s_{t+1})}_{} + \gamma^2 \bar{\phi}(s_t) \sum_{s_{t+1}} \underbrace{P_\pi(s_{t+1}|s_t) \sum_{s_{t+2}} P_\pi(s_{t+2}|s_{t+1}) \bar{\phi}^T(s_{t+2}) w_t}$$

$$= \Bigg[\sum_{s_{t+2}} \sum_{s_{t+1}} P(s_{t+1}|s_t) P(s_{t+2}|s_t) \Bigg] \bar{\phi}^T(s_{t+2}) w_t$$

$$\left[S_{t+2} \mid S_{t+1} \right] \qquad \mathbb{P}^2(S_{t+2} \mid S_t)$$

$$= \gamma \Phi(S_t) \sum_{S_{t+1}} P_\pi(S_{t+1} \mid S_t) R(S_{t+1}) + \gamma^2 \Phi(S_t) \sum_{S_{t+2}} \mathbb{P}^2_\pi(S_{t+2} \mid S_t) \Phi^T(S_{t+2}) w_t$$

∴ Putting in (·1):

$$F(w_t, S_t)$$

$$= \Phi(S_t) R(S_t) + \gamma \Phi(S_t) \sum_{S_{t+1}} P_\pi(S_{t+1} \mid S_t) R(S_{t+1}) + \gamma^2 \Phi(S_t) \sum_{S_{t+2}} \mathbb{P}^2_\pi(S_{t+2} \mid S_t) \Phi^T(S_{t+2}) w_t$$

$$- \Phi(S_t) \Phi^T(S_t) w_t \qquad \text{—(C·2)}$$

Now,

$$w_{t+1} = w_t + \alpha_t f(w_t, S_t, a_t, S_{t+1}, a_{t+1}, S_{t+2})$$

$$= w_t + \alpha_t \left[F(w_t, S_t) + M_{t+1} \right]$$

where: $M_{t+1} = f(w_t, S_t, a_t, S_{t+1}, a_{t+1}, S_{t+2}) - F(w_t, S_t)$

Show M_{t+1} is a martingale:

$$\mathbb{E}[M_{t+1} \mid \mathcal{F}_t] = \mathbb{E}[f \mid \mathcal{F}_t] - \mathbb{E}[F(w_t, S_t) \mid w_t, S_t]$$

(constant given w_t, S_t)

$$= F(w_t, S_t) - F(w_t, S_t)$$

(from C·2)

$$= 0$$

(Hence M_{t+1} is a valid Martingale)

(d) As seen in class (class of stochastic approximation algorithms); the solution of the SDE is the following ODE:

(d) As seen in class (class of stochastic approximation algorithms), the solution of the SDE is the following ODE:

$$\dot{w}(t) = \sum_{s_t} d_\pi(s_t)\, \pi F(w_t, s_t)$$

(Puterman C.2)

$$= \sum_{s_t} d_\pi(s_t)\, \Phi(s_t)\, r(s_t) + \gamma \sum_{s_t} d_\pi(s_t)\, \Phi(s_t) \sum_{s_{t+1}} P_\pi(s_{t+1}|s_t)\, r(s_{t+1})$$

$\Phi^T D_\pi$; $P_\pi r_{S\times 1}$

$$+ \gamma^2 \sum_{s_t} d_\pi(s_t)\, \Phi(s_t) \sum_{s_{t+2}} P^2(s_{t+2}|s_t)\, \Phi^T(s_{t+2})\, w_t$$

(III)

$$- \sum_{s_t} d_\pi(s_t)\, \Phi(s_t)\, \Phi^T(s_t)\, w_t$$

$\Phi^T D_\pi \Phi\, w_t$ (quadratic)

$\Phi^T_{d\times S}$ $D_{\pi\,(S\times S)}$ $r_{(S\times 1)}$

(matrix from for all w's)

(III)

$$\gamma^2 \sum_{s_t} d_\pi(s_t)\, \Phi(s_t) \sum_{s_{t+2}} P^2(s_{t+2}|s_t)\, \Phi^T(s_{t+2})\, w_t$$

$$\gamma^2\, \Phi^T D_\pi P_\pi^2 \Phi\, w_t$$

$S\times S$; $S\times d$; $d\times 1$

$$\therefore \dot{w}(t) = \Phi^T D_\pi r + \gamma \Phi^T D_\pi P_\pi r + \gamma^2 \Phi^T D_\pi P_\pi^2 \Phi\, w_t - \Phi^T D_\pi \Phi\, w_t.$$

$$= \underbrace{\Phi^T D_\pi (I + \gamma P_\pi) r}_{\tilde{b}} - \underbrace{\Phi^T D_\pi (I - \gamma^2 P_\pi^2) \Phi}_{\tilde{A}}\, w_t.$$

$$= \tilde{b} - \tilde{A}\, w_t.$$

$$\therefore \tilde{b} = \Phi^T D_\pi (I + \gamma P_\pi) r$$

$$\tilde{A} = \Phi^T D_\pi (I - \gamma P_\pi^2) \Phi$$

(e) To show that $\tilde{A}$ is invertible:

$$\tilde{A} = \Phi^T D_\pi \Phi - \gamma \Phi^T \underbrace{D_\pi P_\pi^2}_{M} \Phi$$

compute. (diagonal)

$$|x^T M x| = |x^T D_\pi P_\pi^2 x| = |x^T D_\pi^{1/2} D_\pi^{1/2} P_\pi^2 x|$$

(Cauchy Schwartz)
$$\leq \| D_\pi^{1/2} x \|_2 \, \| D_\pi^{1/2} P_\pi^2 x \|_2$$

$$= \|x\|_{D_\pi} \boxed{\| P_\pi^2 x \|_{D_\pi}}$$

$$\| P_\pi^2 x \|_{D_\pi}^2 = \sum_s d_\pi(s) \Big[\sum_{s'} \underbrace{P^2(s'|s)}_{\text{(prob. distribut. - 2 step)}} x(s') \Big]^2$$

(Jensen's inequality)

Convex: $f(E(x)) \leq E f(x)$

$$\leq \sum_s d_\pi(s) \sum_{s'} P^2(s'|s) \, x(s')^2$$

$$= \sum_{s'} x(s')^2 \sum_s d_\pi(s) \, P^2(s'|s)$$

$$= \sum_{s'} x(s')^2 \, d_\pi(s')$$

$$= \boxed{\|x\|_{D_\pi}^2} \quad \Rightarrow \| P_\pi^2 x \|_D \leq \|x\|_D.$$

$$|x^T M x| \leq \|x\|_{D_\pi} \|x\|_{D_\pi} = \|x\|_{D_\pi}^2$$

$$|x^T M x| \leq \|x\|_{D_\pi} \|x\|_{D_\pi} = \|x\|_{D_\pi}^2 \quad \text{——} (e.1)$$

Now,

$$|x^T \hat{A} x| = \underbrace{x^T \Phi^T}_{z} D_\pi (I - \gamma^2 P_\pi^2) \underbrace{\Phi x}_{z}$$

(let Φ be full column rank: $\Phi x \neq 0$ if $w \neq 0$)

$$= z^T D_\pi (I - \gamma^2 P_\pi^2) z$$

$$= \|z\|_D^2 - \gamma^2 z^T \underbrace{D_\pi P_\pi^2}_{M} z$$

(from e.1)

$$\geq \|z\|_D^2 - \gamma^2 \|z\|_D^2$$

$$= (1-\gamma^2) \|z\|_D^2 > 0 \quad \text{(Hence positive definite)}$$

$\therefore$ ($\hat{A}^{-1}$ Exists)

$$\therefore w_*' = \underbrace{[\Phi^T D_\pi (I - \gamma^2 P_\pi^2) \Phi]^{-1}}_{\hat{A}^{-1}} \underbrace{\Phi D_\pi (I + \gamma P_\pi) r}_{\hat{b}}$$

[CASE]

$$\Phi w_\# = \Phi^T (\Phi D \Phi)^{-1} \Phi D v_\pi \leftarrow \text{(optimal)}$$

(f) New Bound: Use the following lemmas from class.

Lemma 1: $T_\pi v = r + \gamma P_\pi v$ is a contraction

Lemma 2: Let $\|V - \Pi V\|_D \leq \min_w \|V - \Phi w\|_D$.

then $\|\Pi V\|_D \leq \|V\|_D$

Lemma 3: ΠT_π is a contraction w.r.t $\|\cdot\|_D$.

we show

Lemma 4 $T_\pi^2 V_\pi = T_\pi (T_\pi V)$ (Let T_π^2 is a contraction too)

$\|T^\pi (T_\pi V_1 - T_\pi V_2)\|_D$

(contraction on the)

$\leq \|T_\pi (\gamma P_\pi V_1 - \gamma P_\pi V_2)\|$

$= \gamma \|P_\pi (T_\pi V_1 - T_\pi V_2)\|_D$

$\leq \gamma \|T_\pi V_1 - T_\pi V_2\|_D$

(as seen in class: $\|P_\pi z\|_D \leq \|z\|_D$)

(lemma 1) $\leq \gamma^2 \|V_1 - V_2\|_D$

(T_π^2 is a γ^2 - contraction)

Now: $\|\Pi T_\pi^2 V_1 - \Pi T_\pi^2 V_2\|_D$

$= \|\Pi (T_\pi^2 V_1 - T_\pi^2 V_2)\|_D \overset{\text{(lemma 2)}}{\leq} \leq \|T_\pi^2 V_1 - T_\pi^2 V_2\|_D$

$\leq \gamma^2 \|V_1 - V_2\|_D$.

$\therefore \Pi T_\pi^2$ is also a γ^2 - contraction.

Now what is πT_π^2 ?

$$\pi T_\pi^2 V = \pi T_\pi (T_\pi V)$$

$$= \pi T_\pi (r + \gamma P_\pi V)$$

$$= \pi \left(r + \gamma P_\pi (r + \gamma P_\pi V) \right)$$

$$= \pi \left[(I + \gamma P_\pi) r + \gamma^2 P_\pi^2 V \right]$$

showing that $\Phi w_*'$ is the fixed pt for πT_π^2.

$$\pi T_\pi^2 [\Phi w_*']$$

$$= \pi \left[(I + \gamma P_\pi) r + \gamma^2 P_\pi^2 \Phi w_*' \right]$$

$$= \pi (I + \gamma P_\pi) r + \gamma^2 \pi P_\pi^2 \Phi w_* \qquad (\text{f.1})$$

[we have seen in class: $\pi = \Phi (\Phi^T D \Phi)^{-1} \Phi^T D$]

$$\therefore \pi (I + \gamma P_\pi) r = \Phi (\Phi^T D_\pi \Phi)^{-1} \Phi^T D_\pi (I + \gamma P_\pi) r \qquad (\text{f.2})$$

Now form the sol^n : $A w_*' = b$

we know: (part c-d)

$$\Rightarrow \Phi^T D (I - \gamma^2 P_\pi^2) \Phi w_*' = \Phi^T D (I + \gamma P) r$$

$$\Rightarrow \Phi^T D (I - \gamma^2 P_\pi^2) \Phi w_* = \Phi^T D (I + \gamma P) r$$

∴ (LHS becomes)

$$\Pi (I + \gamma P_\pi) r = \Phi (\Phi^T D_\pi \Phi)^{-1} (\Phi^T D_\pi \Phi) w_* - \gamma^2 \Phi (\Phi^T D_\pi \Phi)^{-1} \Phi^T D P_\pi^2 \Phi w_*$$

$$= \Phi w_* - \gamma^2 \Pi P_\pi^2 \Phi w_*$$

Putting in LHS, $\Pi T_\pi^2 \Phi w_* =$

$$\Pi \left((I + \gamma P_\pi) r + \gamma^2 P_\pi^2 \Phi w_* \right)$$

$$= \Phi w_* - \gamma^2 \Pi P_\pi \Phi w_* + \gamma^2 \Pi P_\pi^2 \Phi w_*$$

$$= \Phi w_*$$

∴ Φw_* is the fixed pt of ΠT_π^2 (operator of interest)

Now — Bound

$$\therefore \| \Phi w_* - v_\pi \|$$

(fixed pt)

$$= \| \Pi T_\pi^2 \Phi w_* - \Phi w_\# + \Phi w_\# - v_\pi \|_D$$

($w_\#$ optimal)

Triangle inequality

$\leq$

$$\| \Pi T_\pi^2 \Phi w_* - \| \Phi w_\# - v_\pi \|$$

$\leq$

$$\| \Pi T_\pi^2 \Phi w_* - \Phi w_\# \|_D + \| \Phi w_\# - v_\pi \|_D .$$

$\Phi w_\#$: Πv_π (ophial sol^n)

$\Pi v_\pi = T_\pi v_\pi$

$$T_\pi^2 v_\pi = T_\pi (T_\pi v) = v_\pi .$$

$$= \| \Pi T_\pi^2 \Phi w_* - \Pi T_\pi^2 v_\pi \|_D + \| \Phi w_\# - v_\pi \|_D .$$

(It is a γ^2 contraction - lemma 4)

$$\leq \gamma^2 \| \Phi w_* - v_\pi \|_D + \| \Phi w_\# - v_\pi \|_D .$$

$$\Rightarrow (1-\gamma^2) \| \Phi w_* - v_\pi \|_D \leq \| \Phi w_\# - v_\pi \|_D .$$

$$\boxed{\| \Phi w_* - v_\pi \|_D \leq \frac{1}{1-\gamma^2} \| \Phi w_\# - v_\pi \|_D .}$$

Old bound $\| \Phi w_* - v_\pi \|_D \leq \frac{1}{1-\gamma} \| \Phi w_\# - v_\pi \|_D$

(1 step reward)

The new algorithm is better for a tighter upper bound.

$\gamma^2 < \gamma \quad * \quad \gamma \in (0,1)$

$1-\gamma^2 > 1-\gamma$.

$\Rightarrow \frac{1}{1-\gamma^2} < \frac{1}{1-\gamma}$ (Tighter upperbound — better algorithm)

Coding = (25 pts) — see Assignment2-3.ipynb code.

8.1 : Model-free TD(λ) algorithm -

Update :

$\forall s \in S$;

$$V_{k+1}(s) = V_k(s) + \left[\gamma \sum_{m=k}^{L} \alpha_k \lambda^{m-k} \bar{d}_m \right] \quad \alpha_k = \frac{1}{\#\text{ visits}}$$

(L ← End of episode)

where : $\bar{d}_m = r(s_m, a_m, s_{m+1}) + V_k(s_{m+1}) - V_k(s_m)$

where $(s_k = s)$

(γ is the discount factor) ; { λ^{m-k} is the weight for future errors $\bar{d}_m$.

Algorithm :

Overall idea :

1) Store : $[s_0, s_1, s_2 \ldots\ldots s_L] = S$

$[d_0\ d_2 \ldots\ldots\ldots d_L] = D$

$[d_0 \; d_2 \; \cdots \; d_L] = D$ (temporal differences)

(And at the end of the trajectory)

$[\; d_{L-1} + \lambda d_L \;\; d_L \;]$

Do dp:

for $\theta \in \{L-1, L-2, \cdots 0\}$

$\text{cumuD}[\theta] = \lambda \, \text{cumu}[\theta+1] + D[\theta]$.

$\Downarrow$

$V^{k+1}(S[i]) := \text{cumuD}[i] \quad \forall \; i \in \{0, 1, 2 \cdots\}$

Q.2:

Bellman Equation:

$$Q_\pi(s,a) = \sum_{s,r'} P(s'|s,a)\left[r(s,a,s') + \gamma \max_{a' \in A(s')} Q_\pi(s',a')\right]$$

Q.3:

$$w_{t+1} = w_t + \alpha_t \left[r(s_t, a_t, s_{t+1}) + \gamma \Phi^T(s_{t+1}) w_t - \Phi^T(s_t) w_t \right] \Phi(s_t)$$