Qs ① Limiting ODE of the TD(0) algorithm with linear function approximation:

$$\dot{\omega}(t) = b - A\omega(t)$$

where,

$$b = \phi^T D R$$

$$A = \phi^T D [I - \gamma P] \phi$$

Global Lyapunov function for the limiting ODE

$$V(x) = \| \omega - \omega_* \|_p$$

Here, $\omega_* = A^{-1} b$

Verification:

① To show: $V(x) \geq 0 \quad \forall x$ with equality iff $x = x_*$

Here, $V(\omega) = \| \omega - \omega_* \|_p$

$$= \left(\sum_{i=1}^{d} |\omega_i - \omega_*|^p \right)^{1/p}$$

w.k.t $\omega_i \geq \omega_{*i}$.

$\therefore V(\omega) \geq 0$

with $V(\omega) = 0$ iff $\omega_i = \omega_{*i} \quad \forall i$

$\Rightarrow \omega = \omega_*$

(2) To show: $\lim_{\|w\| \to \infty} V(w) = +\infty$

$$\lim_{\|w\| \to \infty} \|w - w_*\|_p = \lim_{\|w\| \to \infty} \left(\sum_{i=1}^{d} |w_i - w_{*_i}|^p \right)^{1/p}$$

$$\leq \lim_{\|w\| \to \infty} \left(\sum_{i=1}^{d} \|w_i\|^p \|w_*\|^p \right)^{1/p}$$ [using cauchy swarchz Inequality]

$$\leq \lim_{\|w\| \to \infty} \left(\|w\|_p - \|w_*\|_p \right)$$

$$\leq +\infty$$

$$\Rightarrow \lim_{\|w\| \to \infty} \|w - w_*\|_p \leq +\infty$$

$$\Rightarrow \lim_{\|w\| \to \infty} \|w - w_*\|_p = +\infty$$

Qs (2) Given: $\phi(1) = 1$ $\phi(2) = 2$

$$\Rightarrow \phi = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Target policy $\cancel{\eta}$ μ and Behavior policy η

* $\eta(right|1) = 0.5$

$\therefore \eta(left|1) = 1 - \eta(right|1) = 1 - 0.5 = 0.5$

* $\eta(right|2) = 0.5$

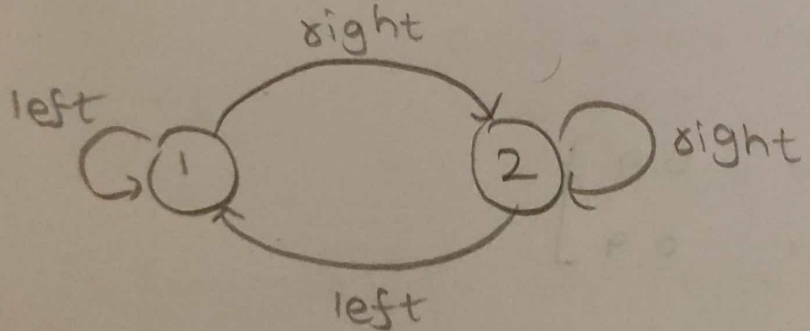
$\Rightarrow \eta(left|2) = 1 - \eta(right|2) = 1 - 0.5 = 0.5$

* $\mu(right|1) = 0.9$

$\Rightarrow \mu(left|1) = 1 - \mu(right|1) = 1 - 0.9 = 0.1$

* $\mu(right|2) = 0.9$

$\Rightarrow \mu(\cancel{right}\ \text{left}|2) = 1 - \mu(right|2) = 1 - 0.9 = 0.1$



Qs (2) (a) To find: $\mathbb{P}^{\mu}$

$$\mathbb{P}^{\mu}(s_{t+1}=1|s_t=1) = \mu(\text{right}|s_t=1)\,\mathbb{P}^{\mu}(s_{t+1}=1|s_t=1, a_t=\text{right})$$
$$+ \mu(\text{left}|s_t=1)\,\mathbb{P}^{\mu}(s_{t+1}=1|s_t=1, a_t=\text{left})$$

$$= 0.9 * 0 + 0.1 * 1 = 0.1$$

$$\mathbb{P}^{\mu}(s_{t+1}=2|s_t=1) = \mu(\text{right}|1)\,\mathbb{P}^{\mu}(2|1,\text{right}) +$$
$$\mu(\text{left}|1)\,\mathbb{P}^{\mu}(2|1,\text{left})$$

$$= 0.9 * 1 + 0.1 * 0 = 0.9$$

$$\mathbb{P}^{\mu}(s_{t+1}=1|s_t=2) = \mu(\text{right}|2)\,\mathbb{P}^{\mu}(1|2,\text{right}) +$$
$$\mu(\text{left}|2)\,\mathbb{P}^{\mu}(1|2,\text{left})$$

$$= 0.9 * 0 + 0.1 * 1 = 0.1$$

$$\mathbb{P}^{\mu}(s_{t+1}=2|s_t=2) = \mu(\text{right}|2)\,\mathbb{P}^{\mu}(2|2,\text{right}) +$$
$$\mu(\text{left}|2)\,\mathbb{P}^{\mu}(2|2,\text{left})$$

$$= 0.9 * 1 + 0.1 * 0 = 0.9$$

$$\therefore\ \mathbb{P}^{\mu} = \begin{bmatrix} 0.1 & 0.9 \\ 0.1 & 0.9 \end{bmatrix}$$

To find: d^{μ}

we know that $\quad d^{\mu^T}\mathbb{P}^{\mu} = d^{\mu^T}$

$$\Rightarrow [d^{\mu}(1)\ \ d^{\mu}(2)]\begin{bmatrix} 0.1 & 0.9 \\ 0.1 & 0.9 \end{bmatrix} = [d^{\mu}(1)\ \ d^{\mu}(2)]$$

$$\Rightarrow 0.1\,d^{\mu}(1) + 0.1\,d^{\mu}(2) = d^{\mu}(1)$$

$\Rightarrow 0.1\, d^{\mu}(2) = 0.9\, d^{\mu}(1)$

$\Rightarrow d^{\mu}(2) = 9\, d^{\mu}(1) \longrightarrow (1)$

We also know that $\quad d^{\mu}(1) + d^{\mu}(2) = 1$

Subs Eq (1)

$\Rightarrow d^{\mu}(1) + 9\, d^{\mu}(1) = 1$

$\Rightarrow 10\, d^{\mu}(1) = 1$

$\Rightarrow d^{\mu}(1) = 0.1$

From (1) $\quad d^{\mu}(2) = 9 * 0.1 = 0.9$

$\therefore d^{\mu} = \begin{bmatrix} 0.1 \\ 0.9 \end{bmatrix}$

To find : A

w.k.t $\quad A = \Phi^T D (I - \gamma P^{\mu}) \Phi$

$= [1 \;\; 2] \begin{bmatrix} 0.1 & 0 \\ 0 & 0.9 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \gamma \begin{bmatrix} 0.1 & 0.9 \\ 0.1 & 0.9 \end{bmatrix} \right) \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$= [0.1 \;\; 1.8] \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0.09 & 0.81 \\ 0.09 & 0.81 \end{bmatrix} \right) \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$= [0.1 \;\; 1.8] \begin{bmatrix} 0.91 & -0.81 \\ -0.09 & 0.19 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$= [-0.071 \;\; 0.261] \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$= [-0.071 + 0.522]$

$= [0.451]$

As the entry in A. is positive, here, A is a positive definite matrix.

$\therefore$ The limiting ODE associated with the on-policy TD(0) has a globally asymptotically stable equilibrium.

Qs 2 b To find: $\mathbb{P}^{\eta}$

$$\mathbb{P}^{\eta} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

To find: d^{η}

w.k.t $\quad d^{\eta T} \mathbb{P}^{\eta} = d^{\eta T}$

$$\Rightarrow [d^{\eta}(1) \;\; d^{\eta}(2)] \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix} = [d^{\eta}(1) \;\; d^{\eta}(2)]$$

$$\Rightarrow 0.5 d^{\eta}(1) + 0.5 d^{\eta}(2) = d^{\eta}(1)$$

$$\Rightarrow d^{\eta}(1) = d^{\eta}(2) \longrightarrow (2)$$

Also, $d^{\eta}(1) + d^{\eta}(2) = 1$

Subs Eq (2)

$$\Rightarrow d^{\eta}(1) + d^{\eta}(1) = 1$$

$$\Rightarrow d^{\eta}(1) = d^{\eta}(2) = 0.5$$

$$\therefore d^{\eta} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$$

To find : A'

w.k.t $A' = \phi^T D^{?} (I - \gamma P^{H}) \phi$

$$= [1 \ \ 2] \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \left[\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 0.9 \begin{bmatrix} 0.1 & 0.9 \\ 0.1 & 0.9 \end{bmatrix} \right] \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$= [0.5 \ \ 1] \begin{bmatrix} 0.91 & -0.81 \\ -0.09 & 0.19 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$= [0.365 \ \ -0.215] \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$= [0.365 - 0.43]$$

$$= [-0.065]$$

The entry in A' is negative.

$\therefore$ Here, A' is not a positive definite matrix

$\therefore$ The limiting ODE associated with the naive off-policy algorithm ~~here~~ does not have a globally asymptotically stable equilibrium.

Qs 2 c Given:

$$f(s) = d_\eta(s) \lim_{t \to \infty} \mathbb{E}_\eta[F_t \mid S_t = s]$$

Here, $F_t = \rho_{t-1} \gamma K_{t-1} + 1$

$$\Rightarrow f(s) = d_\eta(s) + \gamma \sum_{s'} \cancel{d_\pi} f(s') P^{\mu}_{s',s}$$

[Derived in Lecture 9]

$$\Rightarrow f(s) = d_\eta(s) + \gamma \cancel{d} f^T P_{\cdot,s}$$

$$\Rightarrow f^T (I - \gamma P^{\mu}) = d_\eta^T$$

$$\Rightarrow [f(1)\ f(2)] \left[\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 0.9 \begin{bmatrix} 0.1 & 0.9 \\ 0.1 & 0.9 \end{bmatrix} \right] = [0.5\ 0.5]$$

$$\Rightarrow [f(1)\ f(2)] \begin{bmatrix} 0.91 & -0.81 \\ -0.09 & 0.19 \end{bmatrix} = [0.5\ 0.5]$$

$$\Rightarrow 0.91 f(1) - 0.09 f(2) = 0.5 \longrightarrow (3)$$

$$-0.81 f(1) + 0.19 f(2) = 0.5 \longrightarrow (4)$$

Multiplying (3) by 0.19 and (4) by 0.09

$$\Rightarrow 0.1729 f(1) - 0.0171 f(2) = 0.095$$

$$\Rightarrow -0.0729 f(1) + 0.0171 f(2) = 0.045$$

Adding the two equations

$$\Rightarrow 0.1 \cancel{d} f(1) = 0.14$$

$$\Rightarrow f(1) = 1.4$$

Subs in (3)

$0.91 + 1.4 - 0.09 f(2) = 0.5$

$\Rightarrow 1.274 - 0.09 f(2) = 0.5$

$\Rightarrow -0.09 f(2) = -0.774$

$\Rightarrow f(2) = 8.6$

$\therefore f(1) = 1.4 \qquad f(2) = 8.6$

To find : A''

w.k.t $A'' = \phi^T D^e (I - \gamma P^\mu) \phi$

$= \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 1.4 & 0 \\ 0 & 8.6 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 0.9 \begin{bmatrix} 0.1 & 0.9 \\ 0.1 & 0.9 \end{bmatrix} \right) \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$= \begin{bmatrix} 1.4 & 17.2 \end{bmatrix} \begin{bmatrix} 0.91 & -0.81 \\ -0.09 & 0.19 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$= \begin{bmatrix} -0.274 & 2.134 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$= \begin{bmatrix} -0.274 + 4.268 \end{bmatrix}$

$A'' = \begin{bmatrix} 3.994 \end{bmatrix}$

As the entry in A'' is positive. Here, A'' is a positive definite matrix.

$\therefore$ The limiting ODE associated with emphatic TD(0) has a globaly asymptotically stable equilibrium.

Qs 3 To find: ~~f~~ $\sum_s f(s)$

w.k.t $f(s) = d_\eta(s) + \gamma f^T \mathbb{P}_{\cdot,s}$

$$\therefore \sum_s f(s) = \sum_s d_\eta(s) + \gamma f^T \mathbb{P}_{\cdot,s}$$

$$= \sum_s d_\eta(s) + \gamma \sum_s f^T \mathbb{P}_{\cdot,s}$$

$$= 1 + \gamma \sum_s f^T \mathbb{P}_{\cdot,s} \qquad [\because \sum_s d_\eta(s) = 1]$$

$$= 1 + \gamma \sum_s \sum_{s'} f(s') \mathbb{P}_{s',s}$$

$$= 1 + \gamma \sum_s f(s) \sum_{s'} \mathbb{P}_{s,s'}$$

$$= 1 + \gamma \sum_s f(s) . 1$$ [$\because$ sum of row vector of $\mathbb{P}^\mu$ is 1]

$$\Rightarrow \sum_s f(s) - \gamma \sum_s f(s) = 1$$

$$\Rightarrow \sum_s f(s) [1 - \gamma] = 1$$

$$\Rightarrow \sum_s f(s) = \frac{1}{1-\gamma}$$

Yes, we can identify $\sum_s f(s)$, which is

$$\sum_s f(s) = \frac{1}{1-\gamma}$$

Qs (4) Optimization Problem : $\min \mathcal{E}(w)$

$$\mathcal{E}(w) = \frac{1}{2} \| Q^{\mu}(s,a) - Q^{*}(s,a) \|_D^2$$

Intuition: The optimization problem tries to reduce the difference between the true (or target) Q value and the current estimate of the Q-value, which is what we want. we want our estimate of Q-value to be as close as to value of true Q-value.

* But, the problem (or difficulty) in doing this is, we don't know the $Q^{*}(s,a)$ value

∴ we try to approximate the $Q^{*}(s,a)$ value by

$$Q^{*}(s,a) \simeq \phi^{T}(s,a).w$$

where,

ϕ - apriori known feature matrix

w - parameter.

$$\therefore \min \mathcal{E}(w) = \min \frac{1}{2} \| Q^{\mu}(s,a) - \phi^{T}(s,a) w \|_D^2$$

Convex Optimization:

∴ Differentiate $\mathcal{E}(w)$ w.r.t w

Here, $\mathcal{E}(w) = \frac{1}{2} \sum_{a,s} d_{\mu}(s) \mu(a) \left(Q^{\mu}(s,a) - \phi^{T}(s,a) w\right)^2$

$$\therefore \nabla \mathcal{E}(\omega) = -\mathbb{E}\left[\left(Q^{\mu}(s_t,a_t) - \phi^T(s_t,a_t)\,\omega\right)\phi(s_t,a_t)\right] \quad \text{①}$$

where, $s_t \sim d_\mu$

$a_t \sim \mu$

From Bellman Equation, we know that

$$Q^{\mu}(s,a) = \mathbb{E}\left[r(s_t,a_t,s_{t+1}) + \gamma\, Q^{\mu}(s_{t+1},a_{t+1}) \mid s_t = s,\ a_t = a\right]$$

$$= \sum_{a',s'} \mathbb{P}\{a_{t+1}=a', s_{t+1}=s' \mid s_t = s, a_t = a\} \left[r(s,a,s') + \gamma\, Q^{\mu}(s',a')\right]$$

Subs value of $Q^{\mu}(s,a)$ in ①

$$\nabla \mathcal{E}(\omega) = -\sum_{a,s} \mathbb{P}\{s_t = s\}\,\mu(a_t = a)\left[Q^{\mu}(s,a) - \phi^T(s,a)\,\omega\right]\phi(s,a)$$

$$= -\sum_{a,s} \mathbb{P}\{s\}\,\mu(a) \sum_{a',s'} \mathbb{P}\{s',a' \mid s,a\}\left(r(s,a,s') + \gamma\, Q^{\mu}(s',a') - \phi^T(s,a)\,\omega\right)\phi(s,a)$$

$$= -\sum_{\substack{a,s,\\ s',a'}} \mathbb{P}\{s_t = s, a_t = a, s_{t+1} = s', a_{t+1} = a'\}\left(r(s,a,s') + \gamma\, Q^{\mu}(s',a') - \phi^T(s,a)\,\omega\right)\phi(s,a)$$

$$= -\mathbb{E}\left[\left(r(s,a,s') + \gamma\, Q^{\mu}(s',a') - \phi^T(s,a)\,\omega\right)\phi(s,a)\right]$$

where, $s_t \sim d_\mu$ $\quad s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_t)$

$a_t \sim \mu(\cdot \mid s_t)$ $\quad a_{t+1} \sim \mu(\cdot \mid s_{t+1})$

Hence, the SGD algorithm would be:

$$w_{t+1} = w_t + \alpha_t [\, r(s,a,s') + \gamma Q^{M}(s',a') - \phi^T(s,a)\, w \,]\, \phi(s,a)$$

where, $\alpha_t \rightarrow$ step size