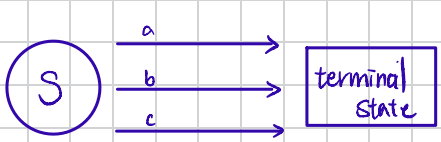
(a) What are the mean vector of $\hat{\nabla} V$ (denoted by $\mathbb{E}[\hat{\nabla} V]$) and the covariance matrix of $\hat{\nabla} V$ (i.e., $\mathbb{E}[(\hat{\nabla} V - \mathbb{E}[\hat{\nabla} V])(\hat{\nabla} V - \mathbb{E}[\hat{\nabla} V])^\intercal]$)?



$$\nabla_\theta V^{\pi_\theta}(M) = \mathbb{E}_{\tau \sim \mathbb{P}_M^{\pi_\theta}}\left[\sum_{t=0}^{\infty} \gamma^t \cdot Q^{\pi_\theta}(s_t, a_t)\, \nabla_\theta \log \pi_\theta(a_t|s_t)\right]$$

$Q^{\pi_\theta}(s,a) = 100$

$Q^{\pi_\theta}(s,b) = 98$

$Q^{\pi_\theta}(s,c) = 95$

$$\because \pi_\theta(\cdot|s) = \frac{\exp(\theta_\cdot)}{(\exp(\theta_a) + \exp(\theta_b) + \exp(\theta_c))} = \frac{\exp(\theta_\cdot)}{1+5+4}$$

$$\Rightarrow \pi_\theta(a|s) = \frac{1}{10} \quad \pi_\theta(b|s) = \frac{1}{2} \quad \pi_\theta(c|s) = \frac{2}{5}$$

$$\because \boxed{\begin{aligned} \frac{\partial \log \pi_\theta(s,a)}{\partial \theta(s,a)} &= 1 - \pi_\theta(s,a) \\ \frac{\partial \log \pi_\theta(s,a)}{\partial \theta(s,b)} &= -\pi_\theta(s,b) \end{aligned}}$$

$$\nabla_\theta \log \pi_\theta(s,a) = \left[\left(1-\tfrac{1}{10}\right), \tfrac{-1}{2}, \tfrac{-2}{5}\right] = \left[\tfrac{9}{10}, \tfrac{-1}{2}, \tfrac{-2}{5}\right]$$

$$\rightarrow \nabla_\theta \log \pi_\theta(s,b) = \left[\tfrac{-1}{10}, \left(1-\tfrac{1}{2}\right), \tfrac{-2}{5}\right] = \left[\tfrac{-1}{10}, \tfrac{1}{2}, \tfrac{-2}{5}\right]$$

$$\nabla_\theta \log \pi_\theta(s,c) = \left[\tfrac{-1}{10}, \tfrac{-1}{2}, \left[1-\tfrac{2}{5}\right]\right] = \left[\tfrac{-1}{10}, \tfrac{-1}{2}, \tfrac{3}{5}\right]$$

$$\mathbb{E}[\hat{\nabla} V] = \frac{1}{10} \cdot 100 \cdot \begin{pmatrix} 0.9 \\ -0.5 \\ -0.4 \end{pmatrix} + \frac{1}{2} \cdot 98 \cdot \begin{pmatrix} -0.1 \\ 0.5 \\ -0.4 \end{pmatrix} + \frac{2}{5} \cdot 95 \cdot \begin{pmatrix} -0.1 \\ -0.5 \\ 0.6 \end{pmatrix} = \begin{pmatrix} 9 + -4.9 - 3.8 \\ -5 + 24.5 - 19 \\ -4 - 19.6 + 22.8 \end{pmatrix} = \begin{pmatrix} 0.3 \\ 0.5 \\ -0.8 \end{pmatrix}$$

want: $Cov(\hat{\nabla} V) = \mathbb{E}\left[\left(\hat{\nabla} V - \mathbb{E}[\hat{\nabla} V]\right) \cdot \left(\hat{\nabla} V - \mathbb{E}[\hat{\nabla} V]\right)^\intercal\right]$

$$\left(\hat{\nabla} V - \mathbb{E}\hat{\nabla} V\right) = \begin{pmatrix} 90 - 0.3 \\ -50 - 0.5 \\ -40 + 0.8 \end{pmatrix} \cdot \begin{pmatrix} -9.8 - 0.3 \\ 49 - 0.5 \\ -39.2 + 0.8 \end{pmatrix} \cdot \begin{pmatrix} -9.5 - 0.3 \\ -47.5 - 0.5 \\ 57 + 0.8 \end{pmatrix} = \begin{pmatrix} 89.7 \\ -50.5 \\ -39.2 \end{pmatrix} \cdot \begin{pmatrix} -10.1 \\ 48.5 \\ -38.4 \end{pmatrix} \cdot \begin{pmatrix} -9.8 \\ -48 \\ 57.8 \end{pmatrix}$$

$$Cov(\hat{\nabla} V) = \frac{1}{10} \begin{pmatrix} 7921, & -4494.5, & -3488.8 \\ -4494.5, & 2550.25, & 1979.6 \\ -3488.8, & 1979.6, & 1536.64 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} 102.01, & -489.85, & 389.84 \\ -489.85, & 2352.25, & -1862.4 \\ 389.84, & -1862.4, & 1474.56 \end{pmatrix} + \frac{2}{5} \begin{pmatrix} 96.04, & 470.4, & -566.44 \\ 470.4, & 2304, & -2774.4 \\ -566.44, & -2774.4, & 3340.84 \end{pmatrix}$$

$$= \begin{pmatrix} 894.03, & -509.75, & -384.28 \\ -509.75, & 2352.75, & -1843 \\ -384.28, & -1843, & 2227.28 \end{pmatrix}$$

(b) Suppose we leverage the value function $V^{\pi_\theta}(s)$ as the baseline and denote by $\tilde{\nabla} V$ the corresponding estimated policy gradient. Then, what are the mean vector and the covariance matrix of $\tilde{\nabla} V$? (Note: $\tilde{\nabla} V$ is also a random vector)

$$V^{\pi_\theta}(s) = \sum \pi_\theta(a|s) \cdot Q^\pi(s,a) = \frac{1}{10} \cdot 100 + \frac{1}{2} \cdot 98 + \frac{2}{5} 95 = 10 + 49 + 38 = 97$$

$$\mathbb{E}[\tilde{\nabla} V] = \frac{1}{10} \cdot \underbrace{(3)}_{100-97} \cdot \begin{pmatrix} 0.9 \\ -0.5 \\ -0.4 \end{pmatrix} + \frac{1}{2} \cdot \underbrace{(1)}_{} \cdot \begin{pmatrix} -0.1 \\ 0.5 \\ -0.4 \end{pmatrix} + \frac{2}{5} \cdot \underbrace{(-2)}_{} \begin{pmatrix} -0.1 \\ -0.5 \\ 0.6 \end{pmatrix} = \begin{pmatrix} 0.3 \\ 0.5 \\ -0.8 \end{pmatrix}$$

$$\left(\tilde{\nabla} V - \mathbb{E}[\hat{\nabla} V]\right) = \begin{pmatrix} 2.7 & -0.3 \\ -1.5 & -0.5 \\ -1.2 & +0.8 \end{pmatrix} \cdot \begin{pmatrix} -0.1 - 0.3 \\ 0.5 - 0.5 \\ -0.4 + 0.8 \end{pmatrix} \cdot \begin{pmatrix} 0.2 - 0.3 \\ 1 \quad -0.5 \\ -1.2 + 0.8 \end{pmatrix} = \begin{pmatrix} 2.4 \\ -2 \\ -0.4 \end{pmatrix} \cdot \begin{pmatrix} -0.4 \\ 0 \\ 0.4 \end{pmatrix} \cdot \begin{pmatrix} -0.1 \\ 0.5 \\ -0.4 \end{pmatrix}$$

$$Cov(\tilde{\nabla} V) = \frac{1}{10} \begin{pmatrix} 5.76, & -4.8, & -0.96 \\ -4.8, & 4, & 0.8 \\ -0.96, & 08, & 0.16 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} 0.16 & 0 & -0.16 \\ 0 & 0 & 0 \\ -0.16 & 0 & 0.16 \end{pmatrix} + \frac{2}{5} \begin{pmatrix} 0.01, & -0.05, & 0.04 \\ -0.05, & 0.25, & -0.2 \\ 0.04, & -0.2, & 0.16 \end{pmatrix}$$

$$= \begin{pmatrix} 0.66 & -0.5 & -0.16 \\ -0.5 & 0.5 & 0 \\ -0.16 & 0 & 0.16 \end{pmatrix} \#$$

(c) Let $B(s)$ denote a baseline function and ∇V_B be the corresponding estimated policy gradient (∇V_B is again a random vector). Suppose we say that a baseline function $B(s)$ is *optimal* if it attains the minimum trace of the corresponding covariance matrix of ∇V_B among all possible state-dependent baselines. Please try to find one such optimal $B(s)$.

$$\mathbb{E}[\hat{\nabla V}] = \frac{1}{10}(100-B)\begin{pmatrix}0.9\\-0.5\\-0.4\end{pmatrix} + \frac{1}{2}(98-B)\begin{pmatrix}-0.1\\0.5\\-0.4\end{pmatrix} + \frac{2}{5}(95-B)\begin{pmatrix}-0.1\\-0.5\\0.6\end{pmatrix} = \begin{pmatrix}0.3\\0.5\\-0.8\end{pmatrix} \rightarrow \text{not change}$$

$$\nabla V_B - \mathbb{E}[\nabla V_B] = (100-B)\begin{pmatrix}0.9\\-0.5\\-0.4\end{pmatrix} - \begin{pmatrix}0.3\\0.5\\-0.8\end{pmatrix} - (98-B)\begin{pmatrix}-0.1\\0.5\\-0.4\end{pmatrix} - \begin{pmatrix}0.3\\0.5\\-0.8\end{pmatrix} - (95-B)\begin{pmatrix}-0.1\\-0.5\\0.6\end{pmatrix} - \begin{pmatrix}0.3\\0.5\\0.8\end{pmatrix}$$

$\because$ The trace of matrix = Inner product $\rightarrow (\nabla V_B - \mathbb{E}[\nabla V_B])^T \cdot (\nabla V_B - \mathbb{E}[\nabla V_B])$, let $k = 98 - B$

$$\frac{1}{10}\left((1k+2)\,0.9-0.3)^2 + \underbrace{(-(k+2)0.5-0.5}_{-0.5k-1.5})^2 + (-(k+2)0.4+0.8)^2\right) + \frac{1}{2}\left((-0.1k-0.3)^2 + (0.5k-0.5)^2 + (-0.4k+0.8)^2\right)$$
$$+ \frac{2}{5}\left((-0.1(k-3)-0.3)^2 + (-0.5(k-3)-0.5)^2 + (0.6(k-3)+0.8)^2\right)$$

$$= \left[(0.81+0.25+0.16)\frac{1}{10} + (0.01+0.25+0.16)\times\frac{1}{2} + (0.01+0.25+0.36)\times\frac{2}{5}\right]k^2$$
$$+ \frac{1}{10}(2.7+1.5+0) + \frac{1}{2}(0.06-0.5-0.64) + \frac{2}{5}(0-1-1.2)k + C$$

$$\Rightarrow (0.122+0.21+0.248)k^2 + (0.42-0.54-1.76)k + C$$

$$\Rightarrow 0.58k^2 - 1.88k \rightarrow \frac{\partial}{\partial k} \rightarrow 1.16k - 1.88 = 0 \text{ 時有 min.} \Rightarrow k \doteq 1.62 \rightarrow B = 96.38$$

Problem 2 (Non-Uniform Polyak-Lojacsiewicz Condition in RL) (8+8=16 points)

As described in Lecture 12, let us prove the fundamental Polyak-Lojacsiewicz condition in RL: Let π^* be an optimal policy and let $a^* := \arg\max_a \pi^*(a|s)$ (essentially, a^* is an optimal action). Under softmax policies, we have

$$\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 \geq \frac{1}{\sqrt{S}}\left\|\frac{d_\mu^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1} \cdot \min_{s\in S}\pi_\theta(a^*(s)|s) \cdot [V^*(\mu) - V^{\pi_\theta}(\mu)]. \tag{1}$$

To show this, you would also need the celebrated "Performance difference lemma" as follows: For any two policies π_1 and π_2, we always have

$$V^{\pi_2}(\mu) - V^{\pi_1}(\mu) = \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d_\mu^{\pi_2}}\mathbb{E}_{a'\sim\pi_2(\cdot|s')}\left[A^{\pi_1}(s',a')\right]. \tag{2}$$

(a) To begin with, show the following result:

$$\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 \geq \frac{1}{1-\gamma}\cdot\frac{1}{\sqrt{S}}\sum_s d_\mu^{\pi_\theta}(s)\cdot\pi_\theta(a^*(s)|s)\cdot\left|A^{\pi_\theta}(s,a^*(s))\right|. \tag{3}$$

$a^*(s) = \arg\max_a \pi^*(a|s)$ $(s\in S)$

(Hint: You would need to first apply Cauchy-Schwarz inequality and leverage the Policy Gradient expressio under softmax policies. This subproblem shall require about 5-8 lines of proof.)

By def. of norm-2 : $\left\| \frac{\partial V^{\pi_\theta}(M)}{\partial \theta} \right\|_2 = \sqrt{\left[\sum_{S,a} \left(\frac{\partial V^{\pi_\theta}(M)}{\partial \theta(s,a)} \right)^2 \right]} \geq \sqrt{\sum_S \left(\frac{\partial V^{\pi_\theta}(M)}{\partial \theta(s,a^*(s))} \right)^2}$

By Cauchy-Schwarz : $\|x\|_2 \cdot \|1\|_2 \geq (x_1 + x_2 + \dots x_n) = \|x\|_1$

$$\Rightarrow \sqrt{\sum_S \left(\frac{\partial V^{\pi_\theta}(M)}{\partial \theta(s,a^*(s))} \right)^2} \geq \frac{1}{\sqrt{S}} \sum_S \left| \frac{\partial V^{\pi_\theta}(M)}{\partial \theta(s,a^*(s))} \right|$$

$\because$ Softmax policy gradient w.r.t. θ is $\frac{\partial V^{\pi_\theta}(M)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} \cdot d_M^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot A^{\pi_\theta}(s,a)$, Prob > 0

$$\frac{1}{\sqrt{S}} \sum_S \left| \frac{\partial V^{\pi_\theta}(M)}{\partial \theta(s,a^*(s))} \right| = \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \sum_S d_M^{\pi_\theta}(s) \cdot \pi_\theta(a^*(s)|s) \cdot \left| A^{\pi_\theta}(s, a^*(s)) \right|$$

$$\Rightarrow \left\| \frac{\partial V^{\pi_\theta}(M)}{\partial \theta} \right\|_2 \geq \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \sum_S d_M^{\pi_\theta}(s) \cdot \pi_\theta(a^*(s)|s) \cdot \left| A^{\pi_\theta}(s, a^*(s)) \right|$$

(b) Next, please use the results in (a) and the Performance difference lemma to conclude that the PL condition in (2) indeed holds. (Hint: Try to handle each term in (3) separately. This subproblem shall require only about 5-8 lines of proof.)

by (3) $\Rightarrow \left\| \frac{\partial V^{\pi_\theta}(M)}{\partial \theta} \right\|_2 \geq \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \sum_S d_M^{\pi_\theta}(s) \cdot \pi_\theta(a^*(s)|s) \cdot \left| A^{\pi_\theta}(s,a) \right|$

Define the distribution mismatch coefficient as $\left\| \frac{d_M^{\pi^*}}{d_M^{\pi_\theta}} \right\|_\infty = \max_S \frac{d_M^{\pi^*}(s)}{d_M^{\pi_\theta}(s)}$

$$\Rightarrow \left\| \frac{\partial V^{\pi_\theta}(M)}{\partial \theta} \right\|_2 \geq \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \sum_S \frac{d_M^{\pi_\theta}(s)}{d_M^{\pi^*}(s)} \cdot d_M^{\pi^*}(s) \cdot \pi_\theta(a^*(s)|s) \cdot \left| A^{\pi_\theta}(s, a^*(s)) \right|$$

$$\geq \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \sum_S \left\| \frac{d_M^{\pi^*}}{d_M^{\pi_\theta}} \right\|_\infty^{-1} \cdot \min_S \pi_\theta(a^*(s)|s) \cdot \sum_S d_M^{\pi^*}(s) \cdot \left| A^{\pi_\theta}(s, a^*(s)) \right|$$

$$\geq \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \sum_S \left\| \frac{d_M^{\pi^*}}{d_M^{\pi_\theta}} \right\|_\infty^{-1} \cdot \min_S \pi_\theta(a^*(s)|s) \cdot \sum_S d_M^{\pi^*}(s) \cdot A^{\pi_\theta}(s, a^*(s))$$

$$= \frac{1}{\sqrt{S}} \sum_S \left\| \frac{d_M^{\pi^*}}{d_M^{\pi_\theta}} \right\|_\infty^{-1} \cdot \min_S \pi_\theta(a^*(s)|s) \cdot \left(\frac{1}{1-\gamma} \sum_S d_M^{\pi^*}(s) \cdot A^{\pi_\theta}(s, a^*(s)) \right) \quad \text{— ①}$$

$\because$ $V^{\pi_2}(M) - V^{\pi_1}(M) = \frac{1}{1-\gamma} \mathbb{E}_{s' \sim d_M^{\pi_2}} \mathbb{E}_{a' \sim \pi_2(\cdot|s')} \left[A^{\pi_1}(s', a') \right]$.

by def , $= \frac{1}{1-\gamma} \sum_{s'} d_M^{\pi_2}(s') \cdot \sum_{a'} \pi_2(a'|s') \cdot A^{\pi_1}(s', a')$

$\Rightarrow$ change $\pi_2 \to \pi^*$; $\pi_1 \to \pi_\theta$; $s' \to s$; $a' \to a^*(s)$

$$\Rightarrow \frac{1}{1-\gamma} \sum_S d_M^{\pi^*}(s) \boxed{\sum_{a^*(s)} \pi^*(a^*(s)|s)} \cdot A^{\pi_\theta}(s, a^*(s))$$

$\sum_{a^*(s)} \pi^*(a^*(s)|s) \to$ Prob (choose $a^*(s)$ in s) $= 1$

$$= \frac{1}{1-\gamma} \sum_S d_M^{\pi^*}(s) \cdot 1 \cdot A^{\pi_\theta}(s, a^*(s)) = ②$$

$$\Rightarrow ① = V^{\pi^*}(M) - V^{\pi_\theta}(M)$$

$$\Rightarrow \left\| \frac{\partial V^{\pi_\theta}(M)}{\partial \theta} \right\|_2 \geq \frac{1}{\sqrt{S}} \sum_S \left\| \frac{d_M^{\pi^*}}{d_M^{\pi_\theta}} \right\|_\infty^{-1} \cdot \min_S \pi_\theta(a^*(s)|s) \cdot \left(\frac{1}{1-\gamma} \sum_S d_M^{\pi^*}(s) \cdot A^{\pi_\theta}(s, a^*(s)) \right)$$

$$= \frac{1}{\sqrt{S}} \sum_S \left\| \frac{d_M^{\pi^*}}{d_M^{\pi_\theta}} \right\|_\infty^{-1} \cdot \min_S \pi_\theta(a^*(s)|s) \cdot \left(V^{\pi^*}(M) - V^{\pi_\theta}(M) \right) \quad *$$

Problem 3.

- Property 1: Show that the true value function at state S (denoted by $V(S)$) satisfies that

$$V(S) = \frac{P_S}{P_T} R_S + R_T.$$

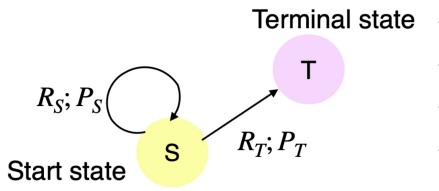


Figure 1: A simple 2-state MRP.

Consider all possible state transform senarios:

$T \Rightarrow S \to T$ | $S \to S \to T$ | $S \to S \to S \to T$ | ... $\Rightarrow V(s) = R_T P_T + (R_S + R_T) P_S P_T + (2R_S + R_T)(P_S)^2 P_T$

prob $\Rightarrow P_T$ | $P_S P_T$ | $P_S P_S P_T$

reward $\Rightarrow R_T$ | $R_S + R_T$ | $R_S + R_S + R_T$

$= R_T \cdot (P_T + P_S P_T + P_S^2 P_T \cdots)$

$+ R_S \cdot (P_S P_T + 2P_S^2 P_T \cdots)$

assuming $\gamma = 1$

$$\Rightarrow V(s) = R_T P_T \cdot \sum_{i=0}^{\infty} P_S^i + R_S P_T \cdot \sum_{i=1}^{\infty} i \cdot P_S^i$$

$$= R_T P_T \cdot \frac{(1 - P_S^{\infty})}{1 - P_S} = R_T P_T \cdot \frac{1}{1 - P_S}$$

$$k = R_S P_T (P_S + 2P_S^2 + 3P_S^3 \cdots)$$

$$-P_S k = R_S P_T (P_S^2 + 2P_S^3 \cdots)$$

$$(1 - P_S) k = R_S P_T \sum_{i=1}^{\infty} P_S^i = R_S P_T \cdot \frac{P_S (1 - P_S^{\infty})}{1 - P_S}$$

$$\Rightarrow k = R_S P_T \frac{P_S}{(1 - P_S)^2}$$

$$V(s) = R_T P_T \cdot \frac{1}{1 - P_S} + R_S P_T \frac{P_S}{(1 - P_S)^2}, \because P_S + P_T = 1 \Rightarrow V(s) = R_T + R_S \frac{P_S}{P_T} \#$$

- Property 2: Suppose we construct an every-visit MC estimate based on only 1 trajectory τ (denoted by $\hat{V}_{\text{MC}}(S; \tau)$). Then, please show that

$$\mathbb{E}_{\tau}[\hat{V}_{\text{MC}}(S; \tau)] = \frac{P_S}{2P_T} R_S + R_T. \tag{5}$$

(Hint: To begin with, you shall consider all possible trajectories and the corresponding probabilities. Accordingly, you would obtain that $\mathbb{E}_{\tau}[\hat{V}_{\text{MC}}(S; \tau)] = \sum_{k=0}^{\infty} P_T P_S^k \left(\frac{R_S + 2R_S + \cdots + kR_S + (k+1)R_T}{k+1} \right)$.)

τ | τ_0: $S \to T$ | τ_1: $S \to S \to T$ | τ_2: $S \to S \to S \to T$ | ... $\text{prob}(\tau_k) = P_T \cdot P_S^k$

prob $\Rightarrow$ | P_T | $P_S P_T$ | $P_S P_S P_T$

reward $\Rightarrow$ | R_T | $\frac{1}{2}(R_S + 2R_T)$ | $\frac{1}{3}(3R_S + 3R_T)$ | $\text{reward}(\tau_k) = \frac{1}{k+1}\left(R_S \left(\sum_{i=1}^{k} i\right) + (k+1) R_T\right)$

$$\mathbb{E}_{\tau}[\hat{V}_{\text{MC}}(S; \tau)] = \sum_{k=0}^{\infty} P_T \cdot P_S^k \cdot \frac{1}{k+1} \cdot \left(R_S \left(\sum_{i=1}^{k} i\right) + (k+1) R_T\right) = \sum_{k=0}^{\infty} P_T \cdot P_S^k \cdot \frac{1}{k+1} \cdot \left(\frac{k(k+1)}{2} R_S + (k+1) R_T\right)$$

$$= \sum_{k=0}^{\infty} P_T \cdot P_S^k \left(\frac{k}{2} R_S + R_T\right) = P_T R_T \sum P_S^k + \boxed{P_T \sum P_S^k \cdot \frac{k}{2} R_S}$$

$$= \left(P_T R_T \cdot \frac{1}{1 - P_S}\right) + \frac{1}{2} R_S \frac{P_S}{1 - P_S}$$

$$= R_T + \frac{P_S}{2P_T} R_S. \#$$

$$H = P_T \left(0 + P_S^1 \cdot \frac{1}{2} R_S + P_S^2 \cdot \frac{2}{2} R_S + P_S^3 \cdot \frac{3}{2} R_S \cdots\right)$$

$$-P_S H = P_T \left(\quad + P_S^2 \cdot \frac{1}{2} R_S + P_S^3 \cdot \frac{2}{2} R_S \cdots\right)$$

$$(1 - P_S) H = P_T \left(\frac{1}{2} \sum_{i=1}^{\infty} P_S^i \cdot R_S\right)$$

$$(\because 1 - P_S = P_T) = R_T \left(\frac{1}{2} R_S \frac{P_S}{1 - P_S}\right)$$

林庭寫

2024年4月19日

RL HW2

REINFORCEMENT and Function Approximation

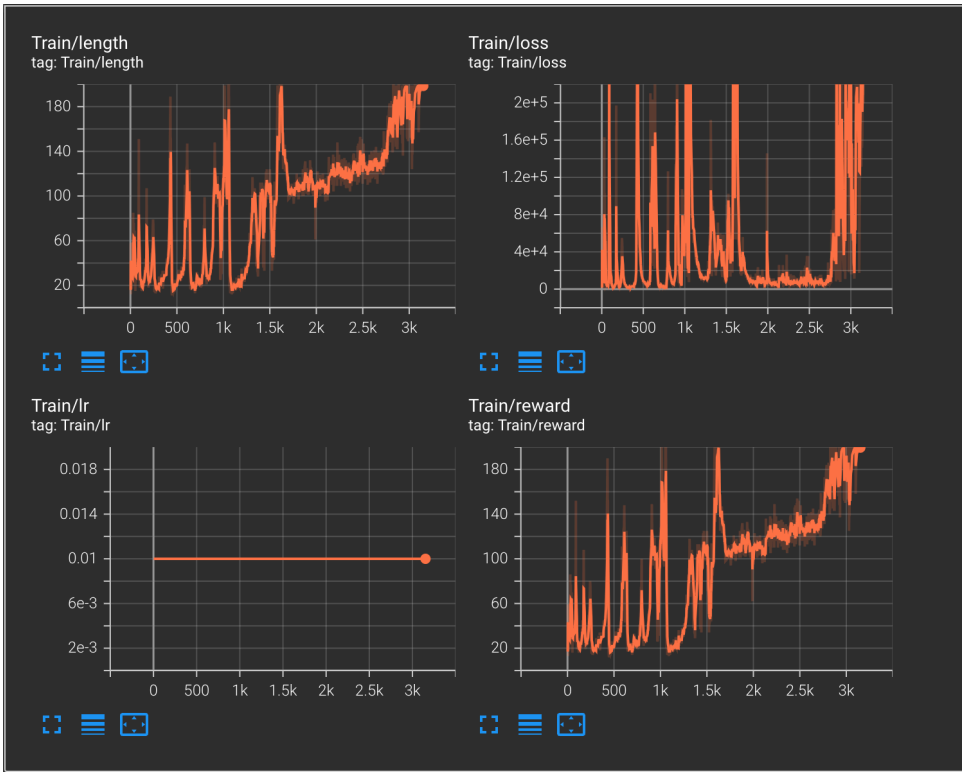
(a.) Vanilla REINFORCE

I simply implemented the algorithm with lr = 0.1, and it cost around 3000 episode to solve CartPole-v0.

The policy and value network are both in 2 linear layer and they share the input layer.

```
# Shared layer
self.shared_layer = nn.Linear(self.observation_dim, self.hidden_size)
# Action layer for policy
self.action_layer = nn.Linear(self.hidden_size, self.action_dim)
# Value layer
self.value_layer = nn.Linear(self.hidden_size, 1)
```

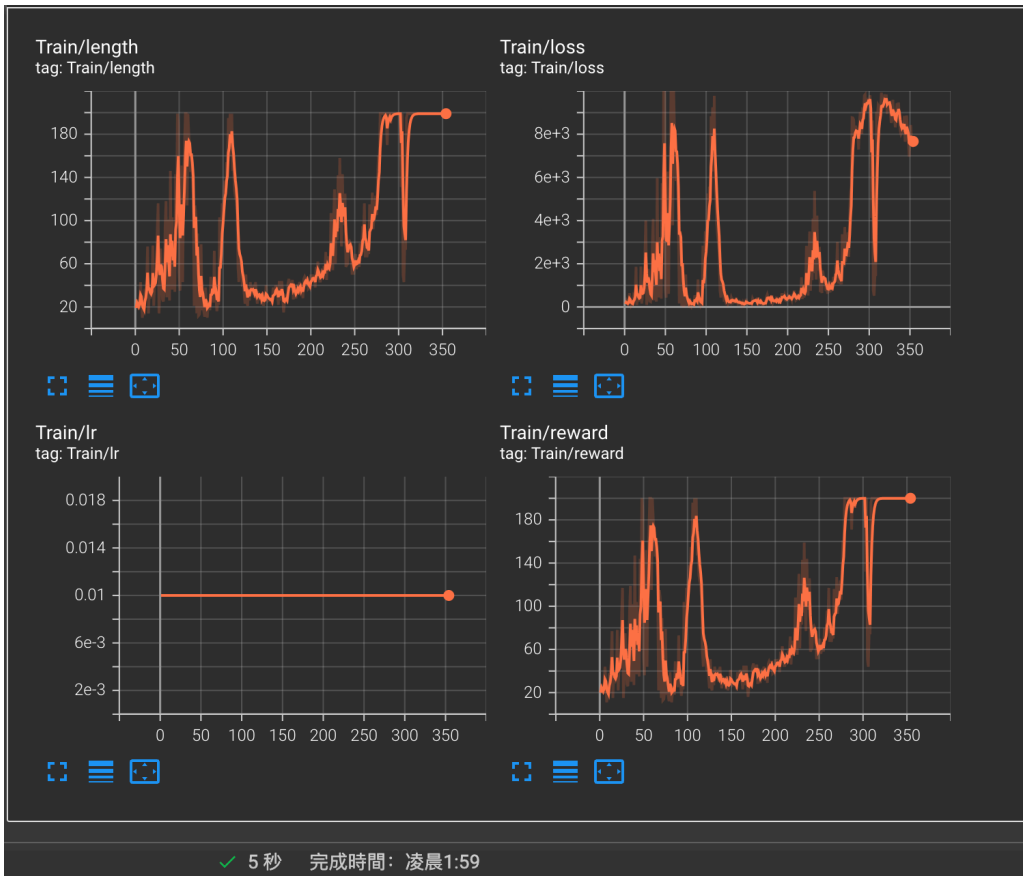
Tensor board:



(b.) Baseline

I choose the learned V(s) as the baseline, since to calculate true V(s) may be too time-demanding. The value is basically from the value stored in SavedAction. To clarify the performance with or without baseline, I didn't change other parameters and network structure and it turned out that the baseline model outperformed the original one, which only run about 350 episodes to pass the threshold.

Tensor board:



Unfortunately, I haven't figure out how to run LunarLander-v2 locally(macOS); furthermore, it took too long to train the model on Google colab, so I didn't have enough time to tune the model for LunarLander-v2. (The performance seems not so ideal on LunarLander-v2)

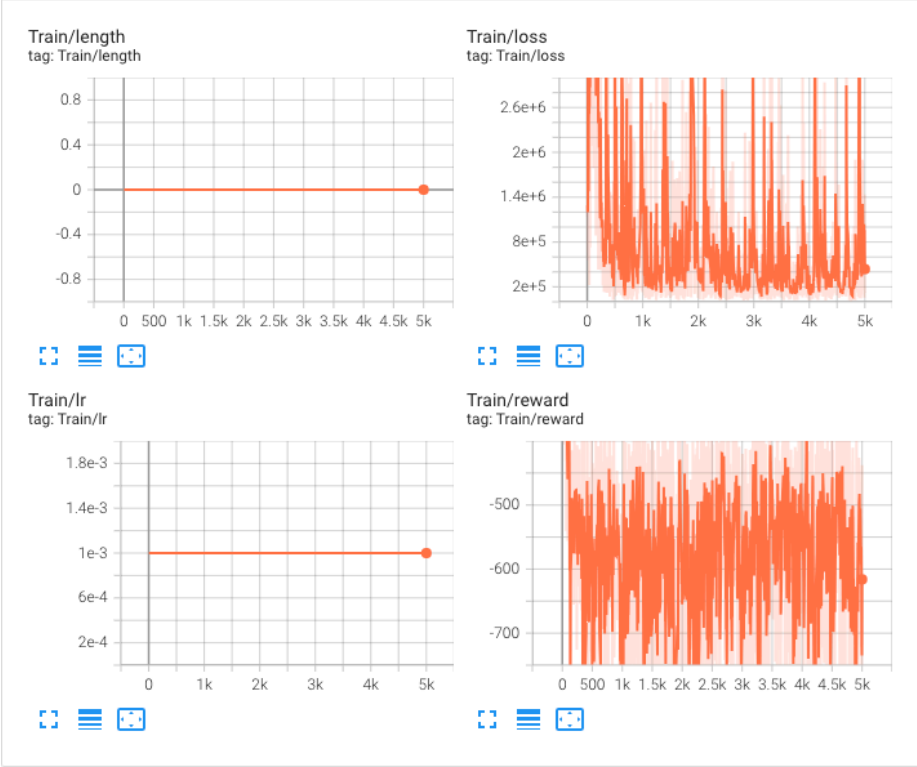
(c.) GAE

NN architecture:

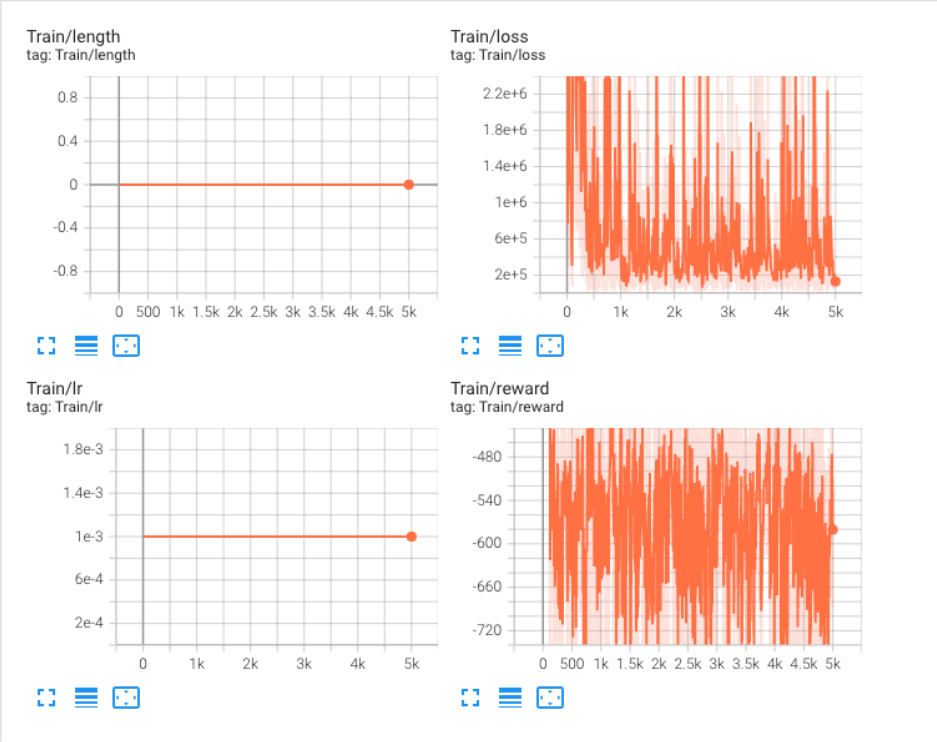
```
# Shared layer
self.shared_layer = nn.Linear(self.observation_dim, self.hidden_size)
self.drop_1 = nn.Dropout(0.1)
self.linear_layer = nn.Linear(self.hidden_size, self.hidden_size)
# Action layer for policy
self.action_layer = nn.Linear(self.hidden_size, self.action_dim)
# Value layer
self.value_layer = nn.Linear(self.hidden_size, 1)
```

I have tried lr = 0.001, lambda_ = 0.9, 0.99, and 0.8. The following charts are their tensor boards respectively. Even though I have tried multiple combination of parameters, I still can't pass the threshold. (episode limit = 5000)

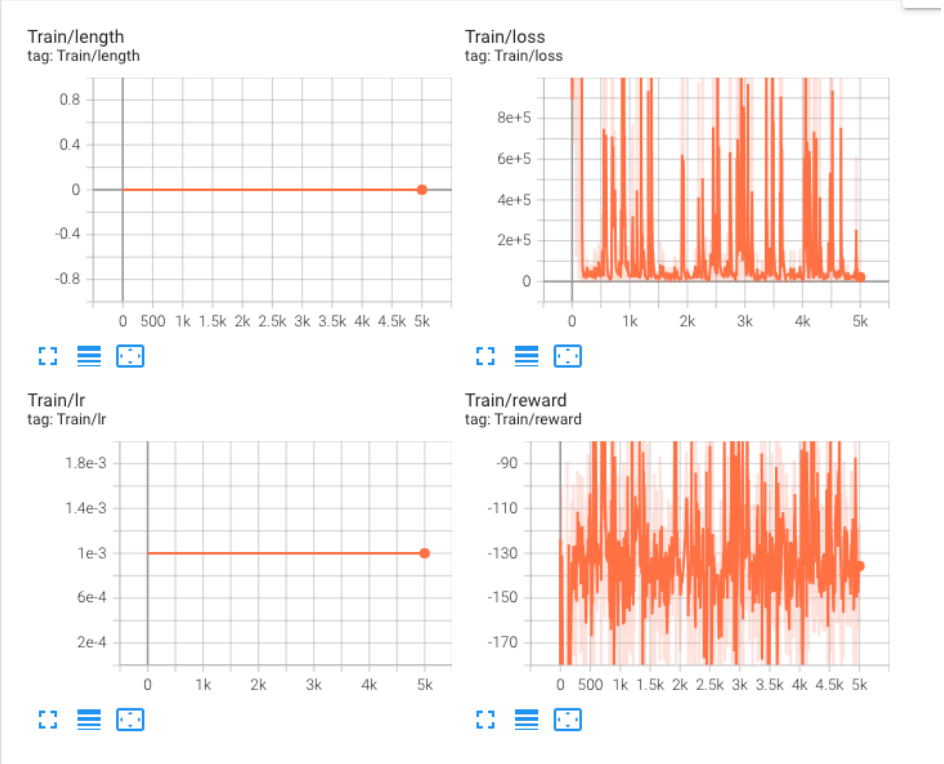
Tensor board of 0.9:



Tensor board of 0.99:



Tensor board of 0.8:



GAE with lambda = 0.8 has the maximum rewards.