1, $\pi_\theta(a|s)=0.1$, $\pi_\theta(b|s)=0.5$, $\pi_\theta(c|s)=0.4$

(a) $\hat{\nabla} V = \begin{bmatrix} \hat{\nabla} V_A \\ \hat{\nabla} V_B \\ \hat{\nabla} V_C \end{bmatrix}$

$$E[\hat{\nabla} V] = \begin{bmatrix} 100(1-0.1)\cdot 0.1 + 98(-0.1)0.5 + 95(-0.1)0.4 \\ 100(-0.5)0.1 + 98(0.5)0.5 + 95(-0.5)0.4 \\ 100(-0.4)0.1 + 98(-0.4)0.5 + 95(0.6)0.4 \end{bmatrix} = \begin{bmatrix} 0.3 \\ 0.5 \\ -0.8 \end{bmatrix}$$

$$Var[\hat{\nabla} V] = E[\hat{\nabla} V (\hat{\nabla} V)^T] - E[\hat{\nabla} V] E[\hat{\nabla} V]^T$$

$$= \begin{bmatrix} 993.3666 & -509.75 & -384.28 \\ -509.75 & 2614.166 & -1843 \\ -384.28 & -1843 & 2474.755 \end{bmatrix}$$

(b) $V^{\pi_\theta}(s) = 100\times 0.1 + 98\times 0.5 + 95\times 0.4 = 97$

$r'(s,a)=3$, $r'(s,b)=1$, $r'(s,c)$

$$\tilde{\nabla} V = \begin{bmatrix} \tilde{\nabla} V_A \\ \tilde{\nabla} V_B \\ \tilde{\nabla} V_C \end{bmatrix} \quad \tilde{\nabla} V_A = \begin{cases} 2.7, & p=0.1 \\ -0.1, & p=0.5 \\ 0.2, & p=0.4 \end{cases} \quad \tilde{\nabla} V_B = \begin{cases} -1.5, & p=0.1 \\ 0.5, & p=0.5 \\ 1, & p=0.4 \end{cases} \quad \tilde{\nabla} V_C = \begin{cases} -1.2, & p=0.1 \\ -0.4, & p=0.5 \\ -1.2, & p=0.4 \end{cases}$$

$$E[\tilde{\nabla} V] = \begin{bmatrix} 0.3 \\ 0.5 \\ -0.8 \end{bmatrix}, \quad Var[\tilde{\nabla} V] = E[\tilde{\nabla} V (\tilde{\nabla} V)^T] - E[\tilde{\nabla} V] E[\hat{\nabla} V]^T = \begin{bmatrix} 0.7333 & -0.5 & -0.16 \\ -0.5 & 0.5555 & 0 \\ -0.16 & 0 & 0.1777 \end{bmatrix}$$

(c)

$f(B(s)) = Var(\nabla V_{Ba}) + Var(\nabla V_{Bb}) + Var(\nabla V_{Bc})$

minimum trace of $Var[\nabla V_B] \Rightarrow \min f(B(s))$

$$f(B(s)) = 0.1\,(100-B(s))^2 (0.9^2+0.5^2+0.4^2) + 0.5\,(98-B(s))^2 (0.1^2+0.5^2+0.4^2) + 0.4\,(95-B(s))^2 (0.1^2+0.5^2+0.6^2)$$

$$f'(B(s)) = 0 \Rightarrow 0.244\,(100-B(s)) + 0.42\,(98-B(s)) + 0.496\,(95-B(s)) = 0 \Rightarrow 0.58\,B(s) = 56.34$$

$$\Rightarrow B(s) = 97.137931\ldots$$

2,

$$E_{\tau\sim p_M^{\pi_\theta}}\left[\sum_{t=0}^{\infty}\gamma^t f(s_t,a_t)\right] = \underset{s_0\sim M}{E}\left[\sum_{a_0}\pi_\theta(a_0|s_0)\sum_{t=0}^{\infty}\gamma^t\sum_{s_{t+1}}P(s_{t+1}|s_t,a_t)\cdot f(s_t,a_t)\right]\cdots①$$

$$\frac{1}{1-\gamma}E_{s\sim d_M^{\pi_\theta}}E_{a\sim\pi_\theta(\cdot|s)}[f(s,a)] = \frac{1}{1-\gamma}E_{s\sim d_M^{\pi_\theta}}\left[\sum_a \pi_\theta(a|s)f(s,a)\right]$$

$$d_{s_0}^{\pi_\theta}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P(s_t=s|s_0,\pi)$$

$$= \underset{s_0\sim M}{E}\left[\sum_{a_0}\pi_\theta(a_0|s_0)\sum_{t=0}^{\infty}\gamma^t\sum_{s_{t+1}}P(s_{t+1}|s_t,a_t)\cdot f(s_t,a_t)\right]\cdots②$$

① equals to ②. The equivalence of eq.(1) is true.

3,

Property 1:

$$V(S) = R_T P_T + (R_S + R_T) P_S P_T + (R_S + R_S + R_T) P_S^2 P_T + \cdots$$

$$= \frac{P_T}{1-P_S} R_T + R_S P_S P_T + 2R_S P_S^2 P_T + 3R_S P_S^3 P_T + \cdots$$

$$= \frac{P_T}{1-P_S} R_T + R_S \frac{P_S}{1-P_S} = R_T + \frac{P_S}{P_T} R_S \quad \#$$

Property 2:

$$E_\tau[\hat{V}_{MC}(S;\tau)] = \sum_{i=0}^{\infty} P_T P_S^i \left(\frac{R_S + 2R_S + \cdots + iR_S + (i+1)R_T}{i+1} \right)$$

$$= \sum_{i=0}^{\infty} P_T P_S^i \frac{1}{i+1} \sum_{j=1}^{i} j R_S + \sum_{i=0}^{\infty} P_T P_S^i R_T$$

$$= \sum_{i=0}^{\infty} P_T P_S^i \frac{i}{2} R_S + R_T = \frac{P_T R_S}{2} \cdot \frac{P_S}{(1-P_S)^2} + R_T = \frac{P_S R_S}{2P_T} + R_T$$

Homework 1 Report

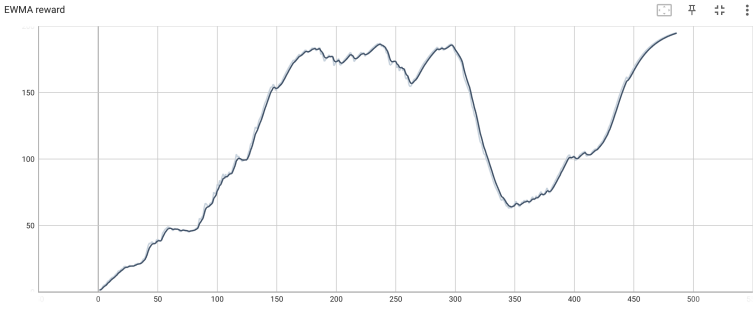
Reinforcement Learning

Name: Kai-Jie Lin
Student ID: 110652019
March 31, 2023

1 Experiment

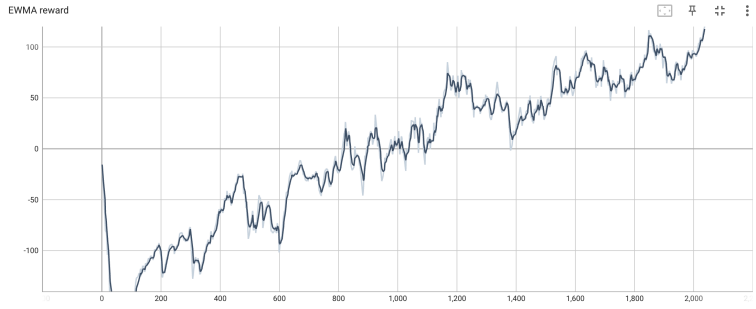
1.1 Vanilla REINFORCE

Training result of the Vanilla REINFORCE.
Hyperparameters: Learning rate = 0.001, decay rate of learning rate = 0.9, discounted factor = 0.999.
Environment: Cartpole-v0.
Result(EWMA reward): Converge in near 500 steps.



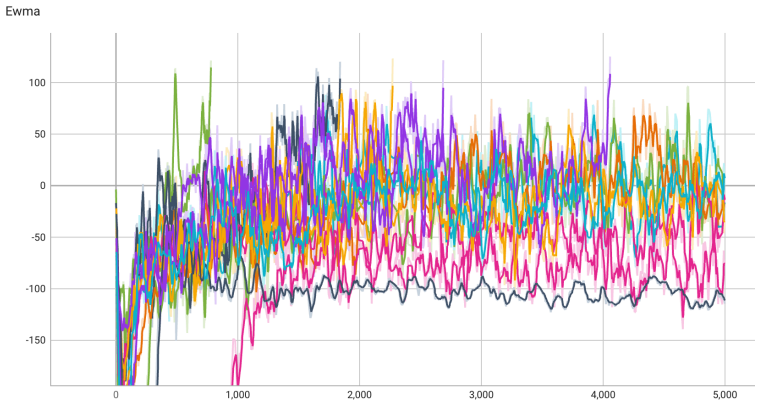
1.2 REINFORCE with baseline

Training result of REINFORCE with baseline. Hyperparameters: Learning rate = 0.002, decay rate of learning rate = 0.95, discounted factor = 0.999.
Environment: LunarLander-v2.
Result(EWMA reward): Converge in 2037 steps.



1.3 REINFORCE with GAE

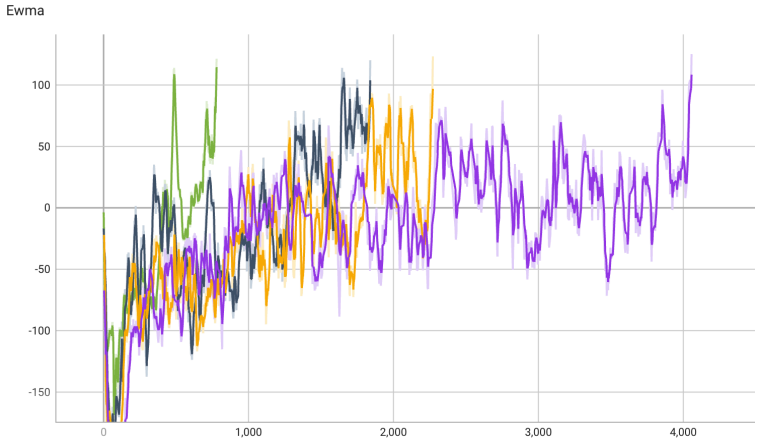
Training result of REINFORCE with GAE. I have used bayesian optimization to tuning the hyper-parameters to reach better result. Bayesian optimization have tried 15 parameters combinations(learning rate, learning rate decay, lambda for GAE). The episodes would stop in 5000 steps once it have not converge to 120. Otherwise the running time will be too long. I have spent on this tuning process for 17 hours. All trajectory:



I selected some of them which have converged to 120 EWMA reward and list below also for comparing the difference of different λ (picture at the bottom):

Learning Rate	**LR Decay**	λ	**Episodes spent**
0.0065	0.8104	0.9700	780
0.0080	0.8561	0.7752	1840
0.0047	0.9436	0.6358	2272
0.0039	0.9508	0.5361	4057

Table 1: Hyperparameters and Results for REINFORCE with GAE



2 Discussion

2.1 Neural Network

Three model used same architecture. Two shared linear layer, one action layer and one value layer. The activation function is ReLU. The vannila version have hidden size 128. Baseline and GAE version have hidden size 256.

2.2 Implementation of GAE

My implementation of GAE is nothing special. It is like the formula below:

$$\hat{A}_t^{GAE(\gamma,\lambda)} = \sum_{k=0}^{\infty} (\gamma\lambda)^k (\delta_{t+k})$$

$\hat{A}_t^{GAE(\gamma,\lambda)}$ is the GAE estimate of the advantage at time step t. γ is the discount factor. λ is the GAE parameter. δ_{t+k} is the temporal difference error at time step $t+k$, defined as $\delta_{t+k} = R_{t+k} + \gamma V(S_{t+k+1}) - V(S_{t+k})$, where R_{t+k} is the reward received at time step $t+k$, S_{t+k} is the state observed at time step $t+k$, and $V(s)$ is the value function estimated for state s.

2.3 About λ in GAE

From the section 1.3, we can see that when λ is large, the converge time is short, λ is smaller, the converge time is longer. In my oppinion, since the reward distribution of environment LunarLander is farsighted, that is we need to sample whole trajectory to get important reward, thus the model sould consider more about the end of the trajectory. λ controls how far we see the whole trajectory. When the λ is large, the model may not be myopic.