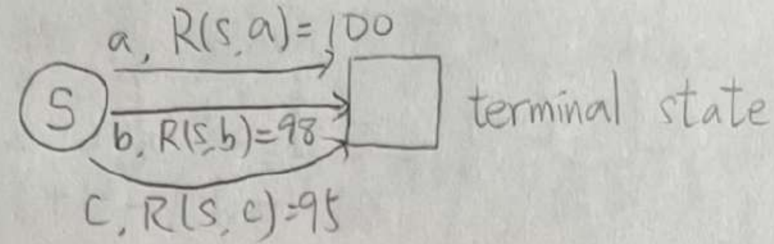
1.



$$\pi_\theta(a|s)=\frac{1}{10}=0.1,\ \pi_\theta(b|s)=\frac{5}{10}=0.5,\ \pi(c|s)=0.4$$

(a) We know there are only 3 trajectories, and

$Q^{\pi_\theta}(s,a)=100$, $Q^{\pi_\theta}(s,b)=98$, $Q^{\pi_\theta}(s,c)=95$,

so we need to compute $\nabla_\theta \log\pi_\theta(\cdot|s)$ first.

$$\frac{\partial \log\pi_\theta(s,a)}{\partial\theta_a}=\frac{\partial\left(\log e^{\theta_a}-\log(e^{\theta_a}+e^{\theta_b}+e^{\theta_c}\right)}{\partial\theta_a}$$

$$=1-\frac{e^{\theta_a}}{e^{\theta_a}+e^{\theta_b}+e^{\theta_c}}=1-\pi_\theta(a|s),$$

$$\frac{\partial \log\pi_\theta(s,a)}{\partial\theta_b}=-\pi_\theta(b|s),\ \text{so}\ \frac{\partial\log\pi_\theta(s,a)}{\partial\theta_c}=-\pi_\theta(c|s)$$

$$\Rightarrow \nabla_\theta\log\pi_\theta(a|s)=\begin{bmatrix}0.9\\-0.5\\-0.4\end{bmatrix},\quad \text{similarly}\ \nabla_\theta\log\pi_\theta(b|s)=\begin{bmatrix}-0.1\\0.5\\-0.4\end{bmatrix},$$

$$\nabla_\theta\log\pi_\theta(c|s)=\begin{bmatrix}-0.1\\-0.5\\0.6\end{bmatrix},\quad \text{so}\ E[\hat\nabla V]=\underbrace{0.1\times100\times\begin{bmatrix}0.9\\-0.5\\-0.4\end{bmatrix}}_{\text{traj. } s\to a\to \text{terminal}}+$$

$$\underbrace{0.5\times98\times\begin{bmatrix}-0.1\\0.5\\-0.4\end{bmatrix}}_{\text{traj. } s\to b\to\text{terminal}}+\underbrace{0.4\times95\times\begin{bmatrix}-0.1\\-0.5\\0.6\end{bmatrix}}_{\text{traj. } s\to c\to\text{terminal}}=\begin{bmatrix}0.3\\0.5\\-0.8\end{bmatrix}$$

so the covariance matrix will be $\Longrightarrow$

$$0.1 \times \begin{bmatrix} 89.7 \\ -50.5 \\ -39.2 \end{bmatrix} \begin{bmatrix} 89.7 & -50.5 & -39.2 \end{bmatrix} + 0.5 \times \begin{bmatrix} -10.1 \\ 48.5 \\ -38.4 \end{bmatrix} \begin{bmatrix} -10.1 & 48.5 & -38.4 \end{bmatrix} +$$

$$0.4 \times \begin{bmatrix} -9.8 \\ -48 \\ 57.8 \end{bmatrix} \begin{bmatrix} -9.8 & -48 & 57.8 \end{bmatrix} = \begin{bmatrix} 804.609 & -452.985 & -351.624 \\ -452.985 & 255.025 & 197.96 \\ -351.624 & 197.96 & 153.664 \end{bmatrix} +$$

$$\begin{bmatrix} 51.005 & -244.925 & 193.92 \\ -244.925 & 1176.125 & -931.2 \\ 193.92 & -931.2 & 737.28 \end{bmatrix} + \begin{bmatrix} 38.416 & 188.16 & -226.576 \\ 188.16 & 921.6 & -1109.76 \\ -226.576 & -1109.76 & 1336.336 \end{bmatrix}$$

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

(b)

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

$\Rightarrow R'(s,a) = 3, \; R'(s,b) = 1, \; R'(s,c) = -2$

$$\Rightarrow E[\tilde{\nabla} V] = 0.1 \times 3 \times \begin{bmatrix} 0.9 \\ -0.5 \\ -0.4 \end{bmatrix} + 0.5 \times 1 \times \begin{bmatrix} -0.1 \\ 0.5 \\ -0.4 \end{bmatrix} + 0.4 \times (-2) \times \begin{bmatrix} -0.1 \\ -0.5 \\ 0.6 \end{bmatrix}$$

$$= \begin{bmatrix} 0.3 \\ 0.5 \\ -0.8 \end{bmatrix}$$

So the covariance matrix will be $\Rightarrow$

$$0.1 \times \begin{bmatrix} 2.4 \\ -2 \\ -0.4 \end{bmatrix} \begin{bmatrix} 2.4 & -2 & -0.4 \end{bmatrix} + 0.5 \times \begin{bmatrix} -0.4 \\ 0 \\ 0.4 \end{bmatrix} \begin{bmatrix} -0.4 & 0 & 0.4 \end{bmatrix} +$$

$$0.4 \times \begin{bmatrix} -0.1 \\ 0.5 \\ -0.4 \end{bmatrix} \begin{bmatrix} -0.1 & 0.5 & -0.4 \end{bmatrix} = \begin{bmatrix} 0.576 & -0.48 & -0.096 \\ -0.48 & 0.4 & 0.08 \\ -0.096 & 0.08 & 0.016 \end{bmatrix} +$$

$$\begin{bmatrix} 0.08 & 0 & -0.08 \\ 0 & 0 & 0 \\ -0.08 & 0 & 0.08 \end{bmatrix} + \begin{bmatrix} 0.004 & -0.02 & 0.016 \\ -0.02 & 0.1 & -0.08 \\ 0.016 & -0.08 & 0.064 \end{bmatrix}$$

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

(C)

We need to minimize the trace of covariance matrix, let M be the covariance matrix and $f(M)$ be the trace, so $f(M) =$

$$= 0.1 \times \left\{ \left[(100 - B(s)) \times 0.9 - 0.3 \right]^2 + \left[(100 - B(s)) \times (-0.5) - 0.5 \right]^2 + \left[(100 - B(s)) \times (-0.4) + 0.8 \right]^2 \right\}$$

$$+ 0.5 \times \left\{ \left[(98 - B(s)) \times (-0.1) - 0.3 \right]^2 + \left[(98 - B(s)) \times 0.5 - 0.5 \right]^2 + \left[(98 - B(s)) \times (-0.4) + 0.8 \right]^2 \right\}$$

$$+ 0.4 \times \left\{ \left[(95 - B(s)) \times (-0.1) - 0.3 \right]^2 + \left[(95 - B(s)) \times (-0.5) - 0.5 \right]^2 + \left[(95 - B(s)) \times 0.6 + 0.8 \right]^2 \right\}$$

let $\frac{df(M)}{dB(s)} = 0 \Rightarrow 0.2 \times (121.66 - 1.22B(s)) + 1 \times (40.62 - 0.42B(s))$

$+ 0.8 \times (59.66 - 0.62B(s)) = 0 \Rightarrow 1.16 B(s) = 112.68$

$\Rightarrow B(s) = 97.137931$, which is optimal

2.

(a)

$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 = \sqrt{\sum_{s,a} \left(\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} \right)^2} \geq \sqrt{\sum_s \left(\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s, a^*(s))} \right)^2},$$

and by Cauchy-Schwarz inequality we have

$$\sqrt{\sum_s \left(\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s, a^*(s))} \right)^2 \cdot \sum_s 1^2} \geq \sqrt{\left(\sum_s \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s, a^*(s))} \right)^2},$$

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

$$= \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \cdot \sum_s \left| d_\mu^{\pi_\theta}(s) \pi_\theta(a^*(s)|s) A^{\pi_\theta}(s, a^*(s)) \right|$$

(by PG under softmax policies)

$$= \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \cdot \sum_s d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a^*(s)|s) \cdot \left| A^{\pi_\theta}(s, a^*(s)) \right|$$

(because $d_\mu^{\pi_\theta}(s)$, $\pi_\theta(a^*(s)|s) \geq 0$)

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

(b) By (a), We know

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

$$\geq \frac{1}{1-\gamma}\cdot\frac{1}{\sqrt{S}}\cdot\underbrace{\left\|\frac{d_\mu^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1}}_{\Downarrow\ \left(\max_s \frac{d_\mu^{\pi^*}(s)}{d_\mu^{\pi_\theta}(s)}\right)}\cdot\sum_s d_\mu^{\pi^*}(s)\cdot\pi_\theta(a^*(s)|s)\cdot\underbrace{A^{\pi_\theta}(s,a^*(s))}_{\text{may be } <0}$$

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

$$= \frac{1}{\sqrt{S}}\cdot\left\|\frac{d_\mu^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1}\cdot\min_s \pi_\theta(a^*(s)|s)\cdot\frac{1}{1-\gamma}\cdot\sum_s d_\mu^{\pi^*}(s)\cdot\sum_a \pi^*(a|s)\cdot A^{\pi_\theta}(s,a)$$

$$= \frac{1}{\sqrt{S}}\cdot\left\|\frac{d_\mu^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1}\cdot\min_s \pi_\theta(a^*(s)|s)\cdot\frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi^*}}\mathbb{E}_{a\sim\pi^*(a|s)}\left[A^{\pi_\theta}(s,a)\right]$$

$$= \frac{1}{\sqrt{S}}\cdot\left\|\frac{d_\mu^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1}\cdot\min_s \pi_\theta(a^*(s)|s)\cdot\left[V^{\pi^*}(\mu)-V^{\pi_\theta}(\mu)\right]$$

(by Performance difference lemma)

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

3.

Property 1: Consider possible scenarios:

$S \to T$, $S \to S \to T$, $S \to S \to S \to T$, ...

and $P_T = 1 - P_S$

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

$$= \sum_{n=0}^{\infty} P_S^n P_T (nR_S + R_T) = \sum_{n=0}^{\infty} P_S^n P_T \cdot nR_S + \sum_{n=0}^{\infty} P_S^n P_T \cdot R_T$$

$$\boxed{\sum_{i=0}^{\infty} x^i = \frac{1}{1-x} \xRightarrow{\text{左右微分}} \sum_{i=1}^{\infty} i \cdot x^{i-1} = \frac{1}{(1-x)^2} \Rightarrow \sum_{i=0}^{\infty} i \cdot x^i = \frac{x}{(1-x)^2}}$$

$$= \sum_{n=0}^{\infty} nP_S^n \cdot P_T \cdot R_S + \frac{P_T}{1-P_S} \cdot R_T$$

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

$$= \frac{P_S}{P_T} \cdot R_S + R_T$$

Property 2: Again, consider possible scenarios

$$\underbrace{S \to S \to \cdots \to S}_{k \text{ times}} \to T \Rightarrow \text{Every-visit MC estimate}$$

$$= \frac{[(k-1)R_S + R_T] + [(k-2)R_S + R_T] + \cdots + R_T}{k} \quad \text{with } p = P_S^{k-1} P_T.$$

$$\text{so } E_T[\hat{V}_{MC}(S;T)] = \sum_{k=0}^{\infty} P_T P_S^k \left(\frac{R_S + 2R_S + 3R_S + \cdots + kR_S + (k+1)R_T}{k+1} \right)$$

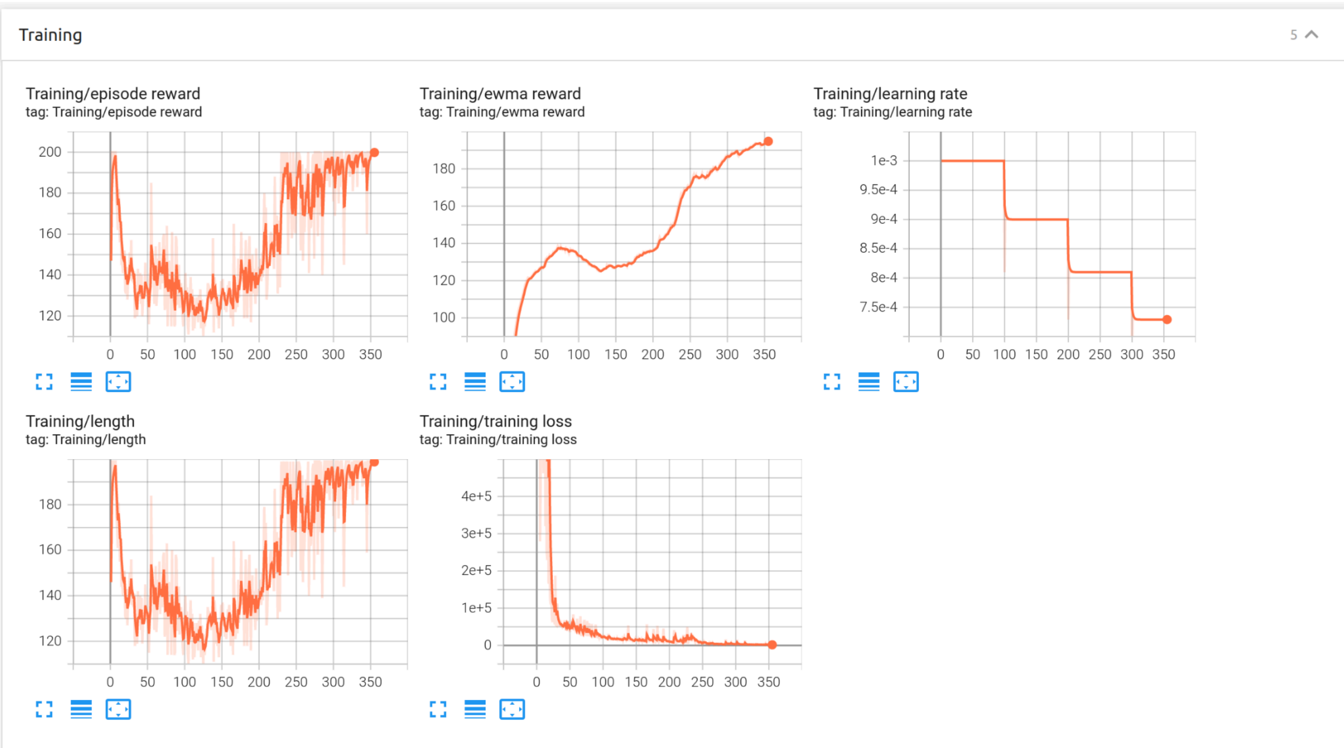
$$= \sum_{k=0}^{\infty} P_T P_S^k \left[\left(\frac{\frac{k(k+1)}{2} R_S}{k+1} \right) + R_T \right] = \underbrace{\sum_{k=0}^{\infty} P_S^k P_T \left(\frac{kR_S}{2} + R_T \right)}_{\text{just proved in property 1}} = \boxed{\frac{P_S}{2P_T} R_S + R_T}$$

Report

1. Vanilla REINFORCE

   I only change two hyperparameters which are learning rate and discounted factor. The NN architecture has two shared layers with ReLU, and one layer for action with Softmax, one layer for value, and hidden size is unchanged(128). I use normal distribution to initialize the model's weight.

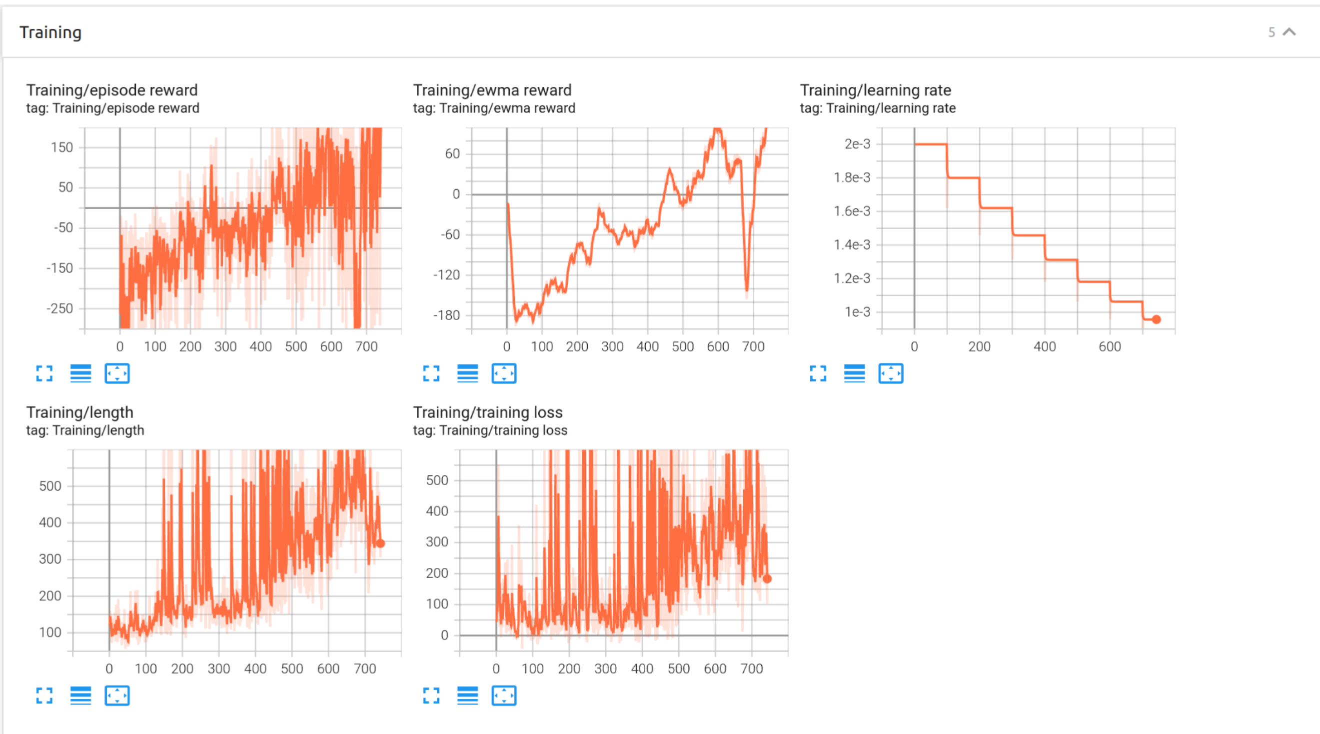
Learning rate	**Discounted factor**	**Episodes**
0.001	0.9	355

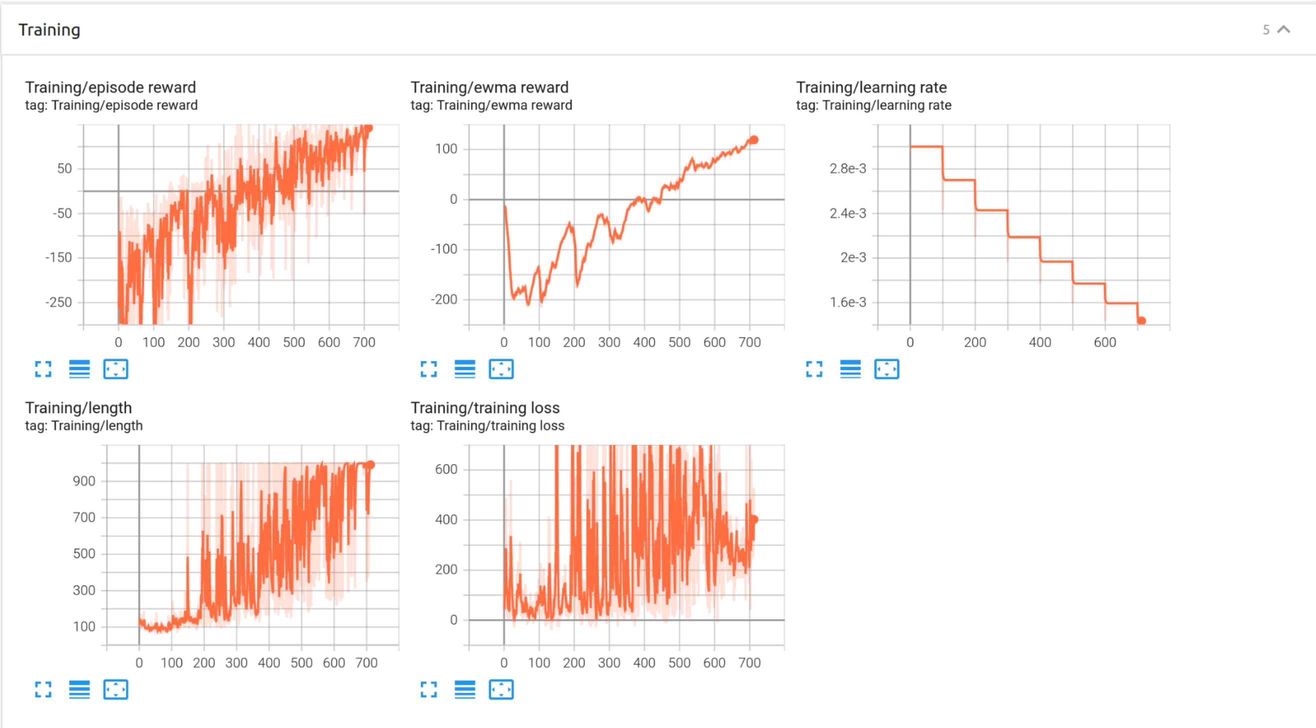


2. REINFORCE with baseline

I only change two hyperparameters which are learning rate and discounted factor. The NN architecture is the same as vanilla REINFORCE, but I am using kaiming normal distribution to initialize the model's weight.I choose the value function to be my baseline because it's a popular choice. I train two models and compare their results.

Learning rate	Discounted factor	Episodes
0.002	0.97	742
0.003	0.97	712





I test them with 10 episodes and calculate their average results. You can notice that although lr =0.002 converges slower, it has better results.

```
Episode 1       Reward: 235.68028808122457
Episode 2       Reward: 36.100586712950246
Episode 3       Reward: 271.19657414505207
Episode 4       Reward: 278.3323964838671
Episode 5       Reward: -4.507550244671137
Episode 6       Reward: 23.593921227674528
Episode 7       Reward: 231.6715025522946
Episode 8       Reward: 255.3399065390551
Episode 9       Reward: 235.19430925770087
Episode 10      Reward: 56.76879801004554
Average reward: 161.93707327651933
```

lr = 0.002

```
Episode 1       Reward: 114.56449420639106
Episode 2       Reward: 97.75347495217362
Episode 3       Reward: 138.37363443361517
Episode 4       Reward: 84.90679310819628
Episode 5       Reward: 114.93849251666288
Episode 6       Reward: 87.43212349229341
Episode 7       Reward: 16.861820809045966
Episode 8       Reward: 221.80700847724586
Episode 9       Reward: 158.15169231595078
Episode 10      Reward: 161.6758777941402
Average reward: 119.64654121057154
```

lr = 0.003

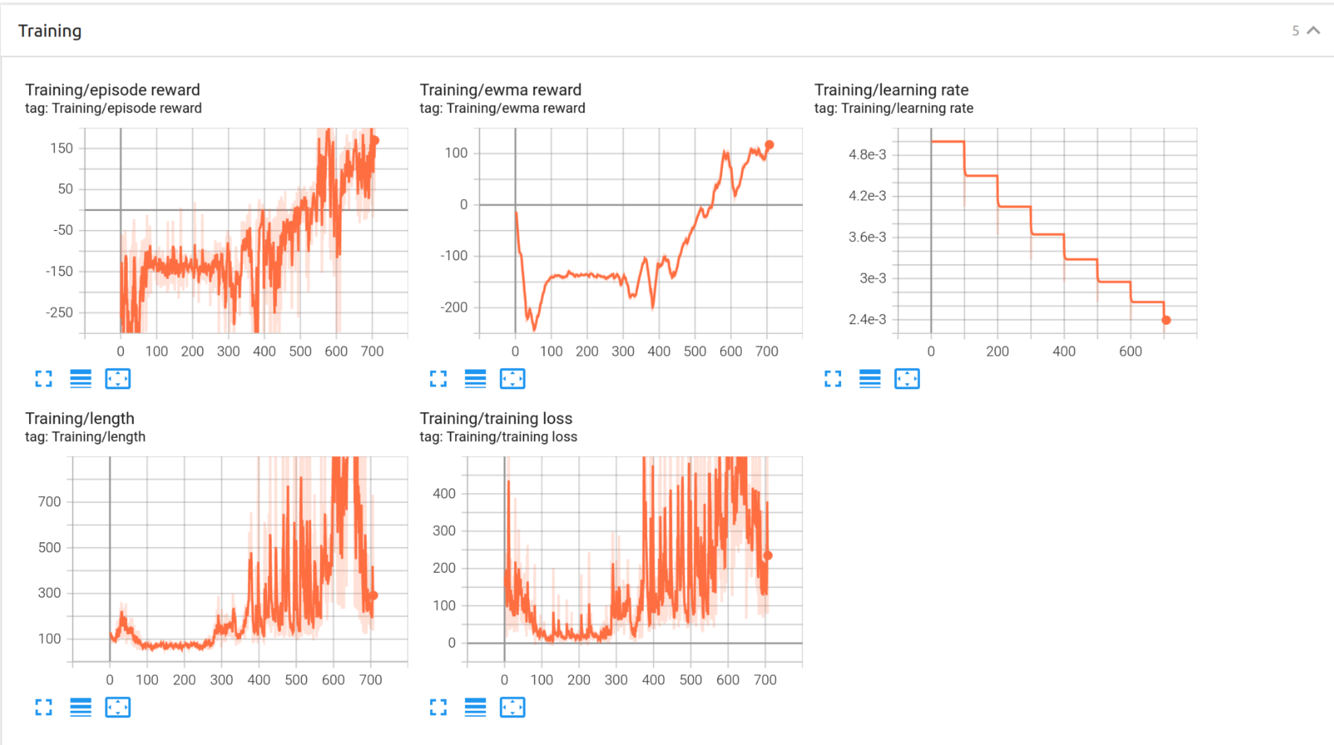
3. REINFORCE with GAE

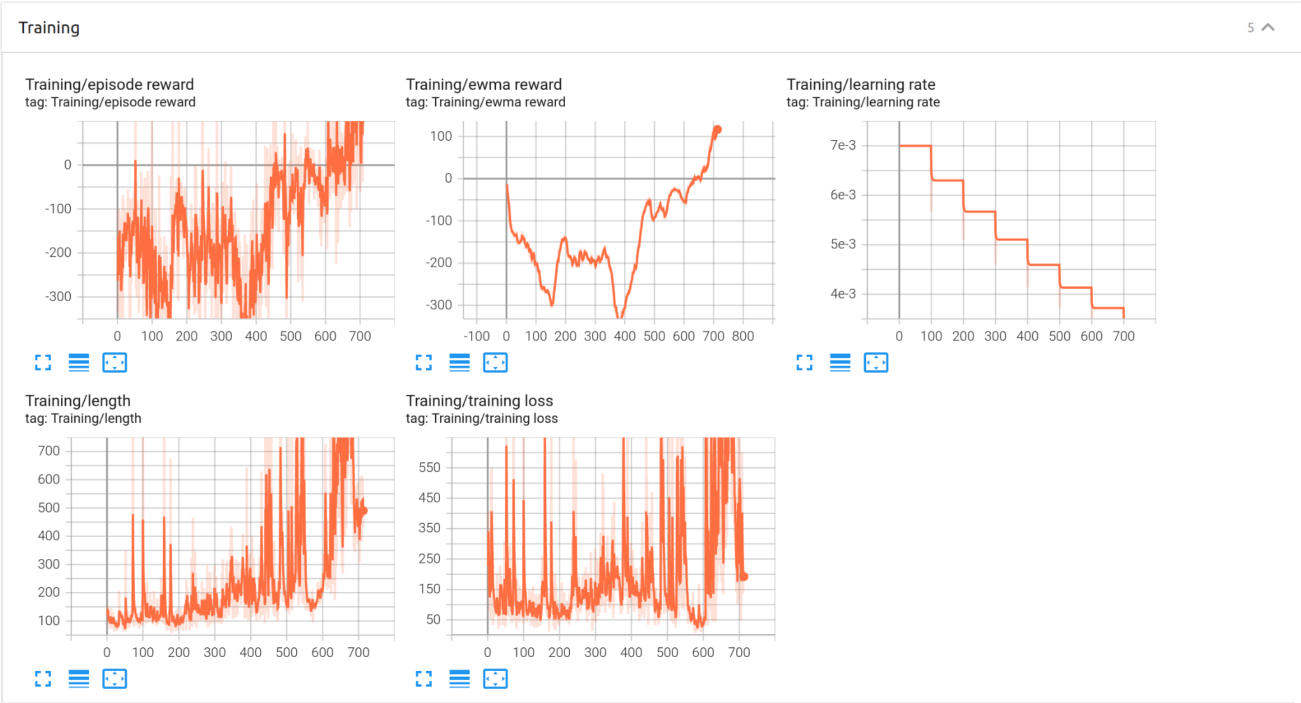
I only change three hyperparameters which are learning rate, discounted factor and lambda. The NN architecture is the same as REINFORCE with baseline. My implementation of GAE is the formula in lecture.

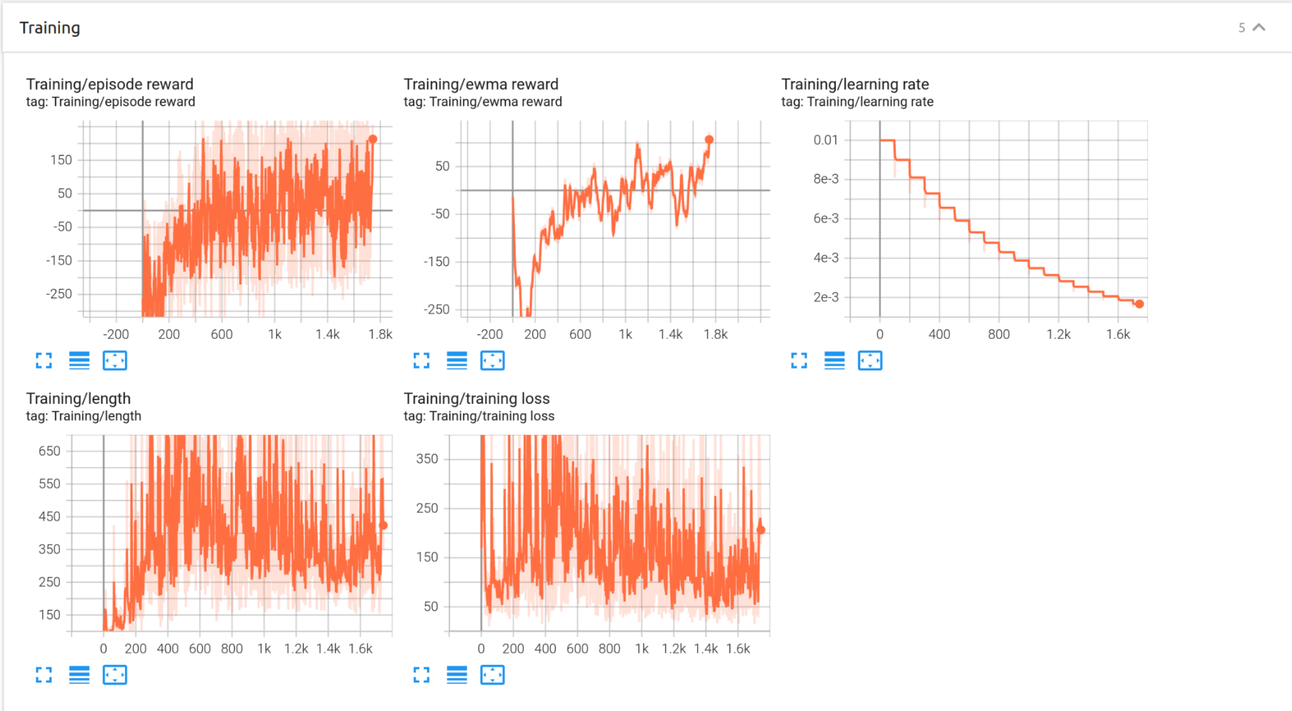
$$\hat{A}_t^{GAE(\gamma,\lambda)} = (1-\lambda)\left(\hat{A}_t^{(1)} + \lambda \hat{A}_t^{(2)} + \lambda^2 \hat{A}_t^{(3)} + \cdots\right) = \sum_{\ell=0}^{\infty} (\gamma\lambda)^{\ell} \delta_{t+\ell}$$

But I calculate it in reverse to reduce computation, that is ADV_t = TD_ERROR + gamma * lambda * ADV_t+1.

Learning rate	**Discounted factor**	**Lambda**	**Episodes**
0.005	0.97	0.99	707
0.007	0.99	0.95	713
0.01	0.8	0.87	1743







We can notice that if lambda is large, then it converges faster, and I test them with 10 episodes and calculate their average results. We can notice that if lambda is large, its result may be better, but lambda = 0.95 worse than lambda = 0.87. It might be because the discounted factor is too large for that model.

```
Episode 1        Reward: 258.2458056714496
Episode 2        Reward: 42.10706413233544
Episode 3        Reward: 266.2285872284804
Episode 4        Reward: 45.563059766118556
Episode 5        Reward: 277.97293998287273
Episode 6        Reward: 234.07958010285085
Episode 7        Reward: 280.433020993243
Episode 8        Reward: 19.35088533284241
Episode 9        Reward: 240.90078146682748
Episode 10       Reward: 261.4940281204607
Average reward: 192.63757527974812
```

lr = 0.005, lambda = 0.99

```
Episode 1        Reward: -86.09689831131416
Episode 2        Reward: 290.202604370014
Episode 3        Reward: 213.73233702799885
Episode 4        Reward: 218.95291736699613
Episode 5        Reward: 178.3679924598219
Episode 6        Reward: -22.358784808036404
Episode 7        Reward: 122.69390606303143
Episode 8        Reward: 129.73512175082732
Episode 9        Reward: 249.53945018267916
Episode 10       Reward: 167.81749293163364
Average reward: 146.2586139033652
```

lr = 0.007, lambda = 0.95

```
Episode 1        Reward: 116.64886289924667
Episode 2        Reward: 265.3332916474075
Episode 3        Reward: 245.88833821460798
Episode 4        Reward: 29.440604726675843
Episode 5        Reward: 245.56190519481493
Episode 6        Reward: -82.06732046759338
Episode 7        Reward: 170.5358646378007
Episode 8        Reward: 236.3366292447455
Episode 9        Reward: -9.120516542632998
Episode 10       Reward: 271.9263564835892
Average reward: 149.0484016038662
```

lr = 0.01, lambda = 0.87