1. (i) $L_{\pi_{\theta_1}}(\pi_{\theta_1}) = \eta(\pi_{\theta_1}) + \sum_{s\in S} d_\mu^{\pi_{\theta_1}}(s) \sum_{a\in A} \pi_{\theta_1}(a|s) A^{\pi_{\theta_1}}(s,a)$

$= \eta(\pi_{\theta_1}) + \sum_{s\in S} d_\mu^{\pi_{\theta_1}}(s) \sum_{a\in A} \pi_{\theta_1}(a|s)(Q^{\pi_{\theta_1}}(s,a) - V^{\pi_{\theta_1}}(s))$

$= \eta(\pi_{\theta_1}) + \sum_{s\in S} d_\mu^{\pi_{\theta_1}}(s) \cdot (V^{\pi_{\theta_1}}(s) - V^{\pi_{\theta_1}}(s)) = \eta(\pi_{\theta_1})$

(ii) First, $\nabla_\theta L_{\pi_{\theta_1}}(\pi_\theta)|_{\theta=\theta_1} = \sum_{s\in S} d_\mu^{\pi_{\theta_1}}(s) \sum_{a\in A} \nabla_\theta \pi_\theta(a|s) \cdot A^{\pi_{\theta_1}}(s,a)|_{\theta=\theta_1}$

And we know $\eta(\pi_\theta) = \eta(\pi_{\theta_1}) + \sum_{s\in S} d_\mu^{\pi_\theta}(s) \sum_{a\in A} \pi_\theta(a|s) \cdot A^{\pi_{\theta_1}}(s,a) \Rightarrow$

so $\nabla_\theta \eta(\pi_\theta)|_{\theta=\theta_1} = \sum_{s\in S} \Big(\nabla_\theta d_\mu^{\pi_\theta}(s) \underbrace{\sum_{a\in A} \pi_\theta(a|s) \cdot A^{\pi_{\theta_1}}(s,a)}_{\text{if } \theta=\theta_1,\ \text{this} = 0} + d_\mu^{\pi_{\theta_1}}(s) \cdot \sum_{a\in A} \nabla_\theta \pi_\theta(a|s) \cdot A^{\pi_{\theta_1}}(s,a) \Big)\Big|_{\theta=\theta_1} = \sum_{s\in S} d_\mu^{\pi_{\theta_1}}(s) \sum_{a\in A} \nabla_\theta \pi_\theta(a|s) \cdot A^{\pi_\theta}(s,a)|_{\theta=\theta_1} = \nabla_\theta L_{\theta_1}(\pi_\theta)|_{\theta=\theta_1}$

2.

(a) Because $D(\lambda) = \min_\theta L(\theta,\lambda)$, needs to make $\frac{\partial L(\theta,\lambda)}{\partial \theta} = 0$

$\Rightarrow \frac{\partial L(\theta,\lambda)}{\partial \theta} = -(\nabla_\theta L_{\theta_k}(\theta)|_{\theta=\theta_k}) + \lambda \cdot (\theta - \theta_k) H_{\theta_k} = 0$

$\Rightarrow \theta = \theta_k + \frac{\nabla_\theta L_{\theta_k}(\theta)|_{\theta=\theta_k}}{\lambda} \cdot H_{\theta_k}^{-1} \Rightarrow$ 把 θ 回去

$\Rightarrow D(\lambda) = -(\nabla_\theta L_{\theta_k}(\theta)|_{\theta=\theta_k})^T \left(\frac{\nabla_\theta L_{\theta_k}(\theta)|_{\theta=\theta_k}}{\lambda} \cdot H_{\theta_k}^{-1}\right) +$

$\frac{\lambda}{2} \cdot \left(\frac{\nabla_\theta L_{\theta_k}(\theta)|_{\theta=\theta_k}}{\lambda} \cdot H_{\theta_k}^{-1}\right)^T H_{\theta_k} \cdot \left(\frac{\nabla_\theta L_{\theta_k}(\theta)|_{\theta=\theta_k}}{\lambda} \cdot H_{\theta_k}^{-1}\right) - \lambda\delta$

$= -\frac{1}{\lambda} (\nabla_\theta L_{\theta_k}(\theta)|_{\theta=\theta_k})^T \cdot H_{\theta_k}^{-1} \cdot (\nabla_\theta L_{\theta_k}(\theta)|_{\theta=\theta_k}) +$

$\frac{1}{2\lambda} (\nabla_\theta L_{\theta_k}(\theta)|_{\theta=\theta_k})^T \cdot H_{\theta_k}^{-1} (\nabla_\theta L_{\theta_k}(\theta)|_{\theta=\theta_k}) - \lambda\delta$

$$= -\frac{1}{2\lambda}\left(\left(\nabla_\theta L_{\theta_k}(\theta)\big|_{\theta=\theta_k}\right)^T H_{\theta_k}^{-1}\left(\nabla_\theta L_{\theta_k}(\theta)\big|_{\theta=\theta_k}\right)\right) - \lambda\delta$$

$\Rightarrow$ To get λ^* $\Rightarrow$ needs to make $\frac{\partial D(\lambda)}{\partial \lambda} = 0$

$$\Rightarrow \frac{\partial D(\lambda)}{\partial \lambda} = \frac{1}{2\lambda^2}\left(\left(\nabla_\theta L_{\theta_k}(\theta)\big|_{\theta=\theta_k}\right)^T H_{\theta_k}^{-1}\left(\nabla_\theta L_{\theta_k}(\theta)\big|_{\theta=\theta_k}\right)\right) - \delta = 0$$

$$\Rightarrow \lambda^* = \sqrt{\frac{\left(\nabla_\theta L_{\theta_k}(\theta)\big|_{\theta=\theta_k}\right)^T H_{\theta_k}^{-1}\left(\nabla_\theta L_{\theta_k}(\theta)\big|_{\theta=\theta_k}\right)}{2\delta}} \quad (\because \lambda \geq 0)$$

(b) $L(\theta, \lambda^*) = -\left(\nabla_\theta L_{\theta_k}(\theta)\big|_{\theta=\theta_k}\right)^T \cdot (\theta - \theta_k) +$

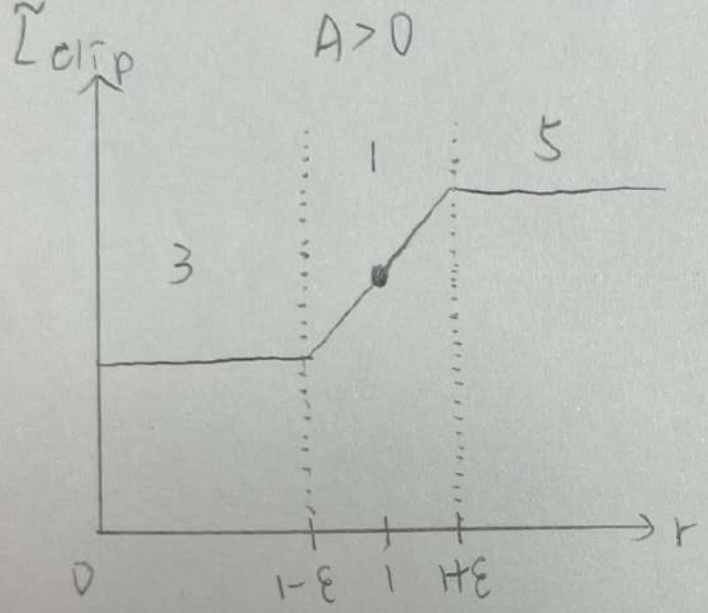
$\frac{\lambda^*}{2}\left((\theta-\theta_k)^T \cdot H_{\theta_k}(\theta-\theta_k) - \delta\right)$ $\Rightarrow$ To get θ^* $\Rightarrow$ needs to make $\frac{\partial L(\theta, \lambda^*)}{\partial \theta} = 0$ $\Rightarrow$ By part (a)

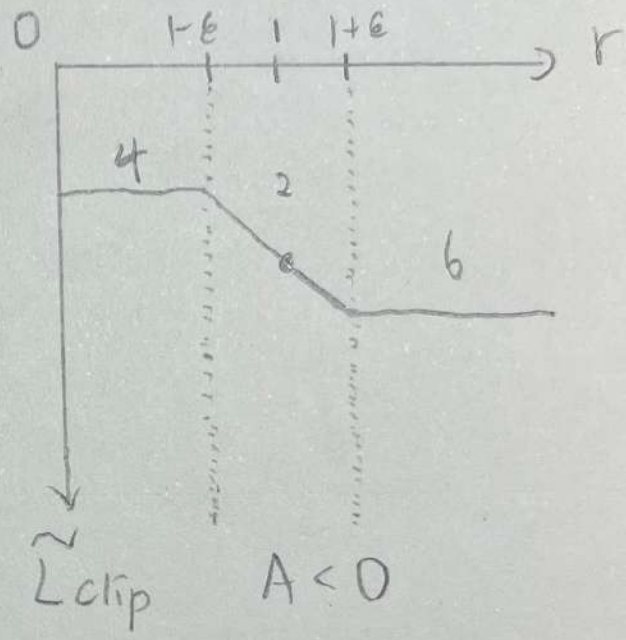
we know that $\theta^* = \theta_k + \frac{\nabla_\theta L_{\theta_k}(\theta)\big|_{\theta=\theta_k}}{\lambda^*} \cdot H_{\theta_k}^{-1}$,

where $\alpha = \frac{1}{\lambda^*} = \left(\frac{\left(\nabla_\theta L_{\theta_k}(\theta)\big|_{\theta=\theta_k}\right)^T \cdot H_{\theta_k}^{-1} \cdot \left(\nabla_\theta L_{\theta_k}(\theta)\big|_{\theta=\theta_k}\right)}{2\delta}\right)^{-\frac{1}{2}}$

3.

	$P_t(\theta) > 0$	A_t	Return Value	Objective is clipped	Sign of objective	Gradient
1	$\in [1-\varepsilon, 1+\varepsilon]$	+	$P_t(\theta) \cdot A_t$	no	+	✓
2	$\in [1-\varepsilon, 1+\varepsilon]$	−	$P_t(\theta) \cdot A_t$	no	−	✓
3	$P_t(\theta) < 1-\varepsilon$	+	$(1-\varepsilon) \cdot A_t$	yes	+	0
4	$P_t(\theta) < 1-\varepsilon$	−	$(1-\varepsilon) \cdot A_t$	yes	−	0
5	$P_t(\theta) > 1+\varepsilon$	+	$(1+\varepsilon) \cdot A_t$	yes	+	0
6	$P_t(\theta) > 1+\varepsilon$	−	$(1+\varepsilon) \cdot A_t$	yes	−	0





Above is the table and figure for $\tilde{L}^{clip}_{s,a}(\theta;\theta_k)$, the main difference between $L^{clip}_{s,a}(\theta;\theta_k)$ and $\tilde{L}^{clip}_{s,a}(\theta;\theta_k)$ is that once $P_t(\theta)$ not in $[1-\varepsilon, 1+\varepsilon]$, $\tilde{L}^{clip}_{s,a}(\theta;\theta_k)$ will clip but $L^{clip}_{s,a}(\theta;\theta_k)$ may not. So $\tilde{L}^{clip}_{s,a}(\theta;\theta_k)$ has more zero gradient regions than $L^{clip}_{s,a}(\theta;\theta_k)$.

Report

1. Pendulum

   My actor NN architecture has three layers, first two layers are with ReLU, last layer with tanh. My critic NN architecture has three layers too. First two layers with ReLU. I make a little change in layer2, I take layer1's output and action as layer2's input. I use kaiming normal distribution to initialize the models' weight. Below are the hyperparameters I focus on.

Actor learning rate	0.001
Critic learning rate	0.001
Hidden size	128
Batch size	512
Number of episodes	250
Gamma	0.995
Tau	0.002

It reaches well policy within 250 episodes, ewma reward is close to -200.

I test the model with 10, 20, 30 episodes and calculate the average results.

You can notice that average rewards are all above -250, these proved the model has a good policy.

```
Episode 1       Reward: -514.9689711827391
Episode 2       Reward: -244.6853753445574
Episode 3       Reward: -1.879456036511052
Episode 4       Reward: -122.44783343609598
Episode 5       Reward: -127.5885895341295
Episode 6       Reward: -241.26264620721363
Episode 7       Reward: -125.71313212245104
Episode 8       Reward: -233.62143527849065
Episode 9       Reward: -365.03945907030135
Episode 10      Reward: -368.96385641053
Average reward: -234.61707546230195
```

```
Episode 11      Reward: -246.57753633601138
Episode 12      Reward: -124.85543477577609
Episode 13      Reward: -1.8274410152748133
Episode 14      Reward: -121.70198720804875
Episode 15      Reward: -124.21634537438625
Episode 16      Reward: -358.98968485368067
Episode 17      Reward: -356.7115046091165
Episode 18      Reward: -119.19389690699332
Episode 19      Reward: -356.9794332026274
Episode 20      Reward: -125.67008534413401
Average reward: -214.1447052124535
```

```
Episode 21      Reward: -359.78789650432503
Episode 22      Reward: -122.29145802349485
Episode 23      Reward: -126.5710844948555
Episode 24      Reward: -126.42993826018478
Episode 25      Reward: -128.10645750828547
Episode 26      Reward: -246.55264322684053
Episode 27      Reward: -602.5350182252783
Episode 28      Reward: -358.8051679393101
Episode 29      Reward: -2.27547995035036
Episode 30      Reward: -126.35160603855529
Average reward: -216.08669514735163
```

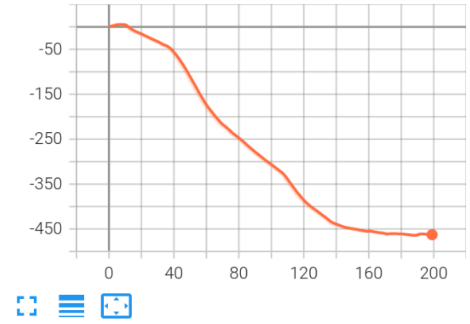
2. HalfCheetah

   My NN architecture is the same as above. I also use kaiming normal distribution to initialize the models' weight. Below are the hyperparameters I focus on.

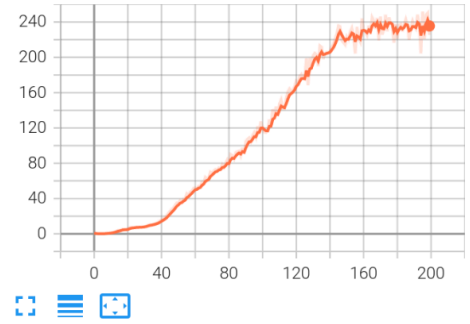
Actor learning rate	0.001
Critic learning rate	0.005
Hidden size	128
Batch size	512
Number of episodes	200
Gamma	0.99
Tau	0.005

Training 4

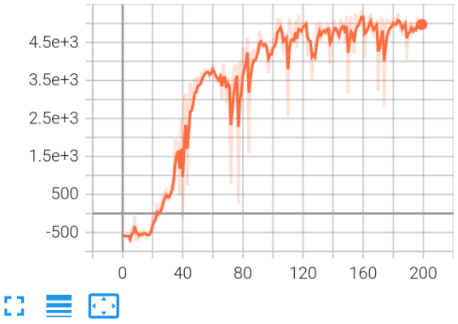
Training/actor loss
tag: Training/actor loss



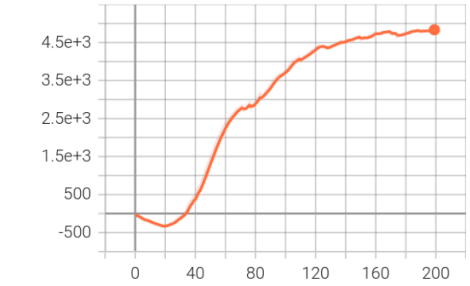
Training/critic loss
tag: Training/critic loss



Training/episode reward
tag: Training/episode reward



Training/ewma reward
tag: Training/ewma reward



It reaches well policy within 200 episodes, ewma reward is nearly 5000 and average reward is even above 5000 in multiple times. I test the model with 10, 20, 30 episodes and calculate the average results. You can notice that average rewards are all above 5000, these proved the model has a good policy.

```
Episode 1       Reward: 5366.468323872849
Episode 2       Reward: 5412.316717528854
Episode 3       Reward: 5289.504957303938
Episode 4       Reward: 5198.187455434627
Episode 5       Reward: 5194.849392253275
Episode 6       Reward: 5087.022904421965
Episode 7       Reward: 4973.915539689739
Episode 8       Reward: 4980.424453798801
Episode 9       Reward: 5220.459326313416
Episode 10      Reward: 5305.53987341521
Average reward: 5202.868894403267
```

```
Episode 11      Reward: 5240.458806874627
Episode 12      Reward: 5105.602001815002
Episode 13      Reward: 5248.3782609677655
Episode 14      Reward: 5186.908263817507
Episode 15      Reward: 3832.8675171059226
Episode 16      Reward: 4986.101035017243
Episode 17      Reward: 4845.117009720323
Episode 18      Reward: 5171.226032117566
Episode 19      Reward: 5098.727011303631
Episode 20      Reward: 5227.241481017506
Average reward: 5098.56581818949
```

```
Episode 21      Reward: 5227.109320837035
Episode 22      Reward: 5081.815225634772
Episode 23      Reward: 5010.45844825674
Episode 24      Reward: 5002.92245306253
Episode 25      Reward: 4933.96527131928
Episode 26      Reward: 5156.449180285608
Episode 27      Reward: 5278.347030191859
Episode 28      Reward: 5202.740012902692
Episode 29      Reward: 5039.273103236404
Episode 30      Reward: 4973.3494216868685
Average reward: 5095.924861040117
```