Problem1

Property1 $L_{\pi_{\theta_1}}(\pi_{\theta_1}) = \eta(\pi_{\theta_1})$

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

(by definition of $L_{\pi_{\theta_1}}(\pi_\theta)$ in (1))

$= \eta(\pi_{\theta_1}) + \left(\sum_{s\in S} d_\mu^{\pi_{\theta_1}}(s)\right) \cdot 0$

$= \eta(\pi_{\theta_1})$

Claim: ___ $= 0$

___ $= \sum_{a\in A} \pi_{\theta_1}(a|s)\left(Q^{\pi_1}(s,a) - V^{\pi_1}(s,a)\right)$ $A^{\pi_{\theta_1}}(s,a)$ by defition of A^{π_θ}

$= \sum_{a\in A} \pi_{\theta_1}(a|s) Q^{\pi_1}(s,a) - \sum_{a\in A} \pi_{\theta_1}(a|s) V^{\pi_1}(s,a)$ $= 1$

$= V^{\pi_1}(s) - V^{\pi_1}(s)$

(by bellman equation)

$= 0$ □

Property2 $\nabla_\theta L_{\pi_{\theta_1}}(\pi_\theta)|_{\theta=\theta_1} = \nabla_\theta \eta(\pi_\theta)|_{\theta=\theta_1}$

pf By Performance Lemma and (1), we have

$\eta(\pi_\theta) = \eta(\pi_{\theta_1}) + \sum_s d_\mu^{\pi_\theta}(s) \sum_a \pi_\theta(a|s) A^{\pi_{\theta_1}}(s,a)$ RHS1

$L_{\pi_{\theta_1}}(\pi_\theta) = \eta(\pi_{\theta_1}) + \sum_s d_\mu^{\pi_{\theta_1}}(s) \sum_a \pi_\theta(a|s) A^{\pi_{\theta_1}}(s,a)$ RHS2

$\Rightarrow$ If $\frac{\partial RHS1}{\partial\theta}\Big|_{\theta=\theta_1} = \frac{\partial RHS2}{\partial\theta}\Big|_{\theta=\theta_1}$, then Property2 holds.

$\frac{\partial RHS1}{\partial\theta}\Big|_{\theta=\theta_1} = \left(\sum_s \nabla_\theta d_\mu^{\pi_\theta}(s) \sum_a \pi_\theta(a|s) A^{\pi_{\theta_1}}(s,a)\right)\Big|_{\theta=\theta_1}$ $= 0$

$+ \left(\sum_s d_\mu^{\pi_\theta}(s) \sum_a \nabla_\theta \left[\pi_\theta(a|s) A^{\pi_{\theta_1}}(s,a)\right]\right)\Big|_{\theta=\theta_1}$

(by Chain rule) $\because \sum_a \pi_{\theta_1}(a|s) A^{\pi_{\theta_1}}(s,a) = 0$

$$\frac{\partial RHS2}{\partial\theta}\Big|_{\theta=\theta_1}=\Big(\sum_s \nabla d_\mu^{\pi_{\theta_1}}(s)\underbrace{\sum \pi_\theta(a|s)A^{\pi_{\theta_1}}(s,a)}_{=0}\Big)\Big|_{\theta=\theta_1}$$

$$+\Big(\sum_s d_\mu^{\pi_{\theta_1}}(s)\sum_a \nabla_\theta\big[\pi_\theta(a|s)A^{\pi_{\theta_1}}(s,a)\big]\Big)\Big|_{\theta=\theta_1}$$

$$\Rightarrow \frac{\partial RHS1}{\partial\theta}\Big|_{\theta=\theta_1}=\frac{\partial RHS2}{\partial\theta}\Big|_{\theta=\theta_2}$$

$\Rightarrow$ Property2 holds

Problem2

(a) We use two Lemma to solve.

Lemma1

Let $f:\mathbb{R}^n\to\mathbb{R}$ given by $f(x)=X^TAX$, where A: symmetric and $X=(X_1\cdots X_n)^T$. Then $\frac{\partial f}{\partial X}=2AX$.

pf $y=f(x)=X^TAX=\sum_{i=1}^n\sum_{j=1}^n a_{ij}X_iX_j$

$$=\sum_{i=1}^n a_{ip}X_iX_p+\sum_{j=1}^n a_{pj}X_pX_j+\sum_{\substack{i=1\\ i,j\neq p}}^n\sum_{j=1}^n a_{ij}X_iX_j$$

$$\frac{\partial y}{\partial X_p}=\sum_{i=1}^n a_{ip}X_i+\sum_{j=1}^n a_{pj}X_j$$

$$=2\sum_{i=1}^n a_{pi}X_i \quad (\because A\text{: symmetric})$$

$$\Rightarrow \frac{\partial f}{\partial X}=\begin{pmatrix}2\sum_{i=1}^n a_{1i}X_i\\ \vdots \\ 2\sum_{i=1}^n a_{ni}X_i\end{pmatrix}=2\begin{pmatrix}a_{11}&a_{12}&\cdots&a_{1n}\\ \vdots&\vdots&\ddots&\vdots\\ a_{n1}&a_{n2}&\cdots&a_{nn}\end{pmatrix}\begin{pmatrix}X_1\\ \vdots\\ X_n\end{pmatrix}$$

$$=2AX \quad \square$$

Lemma2

$$\boxed{\frac{\partial A^T x}{\partial x} = A, \text{ where } A: \text{metrix}}$$

pf similar to above proof. □

Let $\frac{\partial \mathcal{L}(\theta,\lambda)}{\partial \theta} = 0.$

$$\Rightarrow -(\nabla_\theta L_{\theta_k}|_{\theta=\theta_k}) + \frac{\lambda}{\cancel{2}}(\cancel{2}H_{\theta_k}(\theta-\theta_k)) = 0$$

by Lemma1, Lemma2,

$$\Rightarrow (\theta-\theta_k) = \frac{1}{\lambda} H_{\theta_k}^{-1}(\nabla_\theta L_{\theta_k}|_{\theta=\theta_k}) \text{ 代回 } (4)$$

$$\Rightarrow \min_\theta \mathcal{L}(\theta,\lambda) = -(\nabla_\theta L_{\theta_k}(\theta)|_{\theta=\theta_k})^T \frac{1}{\lambda} H_{\theta_k}^{-1}(\nabla_\theta L_{\theta_k}|_{\theta=\theta_k})$$
$$+\lambda\left(\frac{1}{2}\left(\frac{1}{\lambda} H_{\theta_k}^{-1}(\nabla_\theta L_{\theta_k}|_{\theta=\theta_k})\right)^T H_{\theta_k}(\theta-\theta_k) - \delta\right)$$

$$= -(\nabla_\theta L_{\theta_k}(\theta)|_{\theta=\theta_k})^T \frac{1}{\lambda} H_{\theta_k}^{-1}(\nabla_\theta L_{\theta_k}|_{\theta=\theta_k})$$
$$+\cancel{\lambda}\left(\frac{1}{2}(\nabla_\theta L_{\theta_k}(\theta)|_{\theta=\theta_k})^T (\cancel{H_{\theta_k}^{-1}})^T \frac{1}{\cancel{\lambda}} \cancel{H}_{\theta_k}(\theta-\theta_k)\right.$$
$$-\lambda\delta \quad \because H: \text{symmetric}$$

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

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

$$= \min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta,\lambda) \quad \square$$

(∵ 它是 LP transformation, strong duality holds

② Let $\frac{dD(\omega)}{d\lambda}=0$

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

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

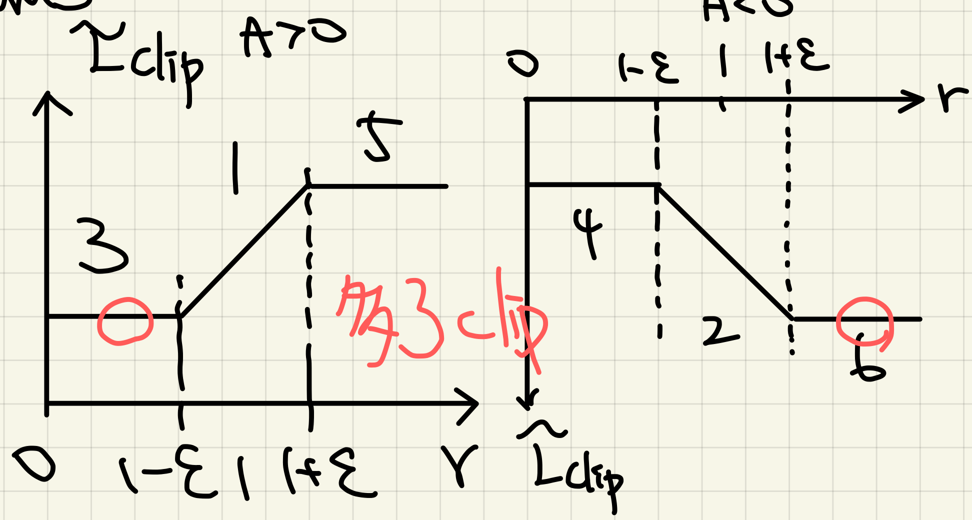
$$\Rightarrow \lambda^*=\sqrt{\frac{(\nabla_\theta L_{\theta_k}(\theta)|_{\theta=\theta_k})^T H_{\theta_k}^{-1}(\nabla_\theta L_{\theta_k}(\theta)|_{\theta=\theta_k})}{2\delta}} \quad \square$$

(b) ① Let $\frac{\partial \mathcal{L}(\theta,\lambda)}{\partial\theta}=0$, we have

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

② $$\alpha=\frac{1}{\lambda^*}=\sqrt{\frac{2\delta}{(\nabla_\theta L_{\theta_k}(\theta)|_{\theta=\theta_k})^T H_{\theta_k}^{-1}(\nabla_\theta L_{\theta_k}(\theta)|_{\theta=\theta_k})}} \quad \square$$

Problem3



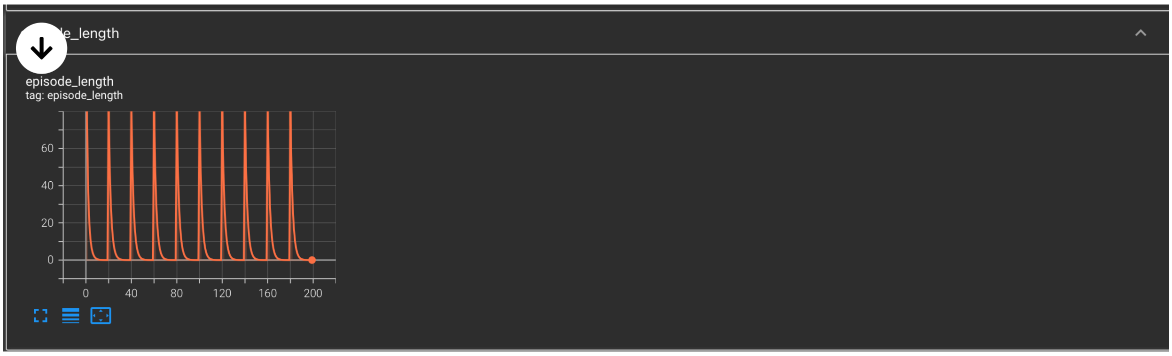
num	$p_t(\theta)>0$	A_t	Return Value of min	Objective is Clipped	Sign of Objective	Gradient
1	$p_t(\theta)\in[1-\epsilon,1+\epsilon]$	+	$p_t(\theta)A_t$	no	+	✅
2	$p_t(\theta)\in[1-\epsilon,1+\epsilon]$	-	$p_t(\theta)A_t$	no	-	✅
3	$p_t(\theta)<1-\epsilon$	+	$(1-\epsilon)p_t(\theta)A_t$	yes	+	0
4	$p_t(\theta)<1-\epsilon$	-	$(1-\epsilon)p_t(\theta)A_t$	yes	-	0
5	$p_t(\theta)>1+\epsilon$	+	$(1+\epsilon)p_t(\theta)A_t$	yes	+	0
5	$p_t(\theta)>1+\epsilon$	-	$(1+\epsilon)p_t(\theta)A_t$	yes	-	0

tags: `2024 年 下學期讀書計畫` `Reinforcement Learning`

RL Homework 3: DDPG, TRPO, and PPO

a

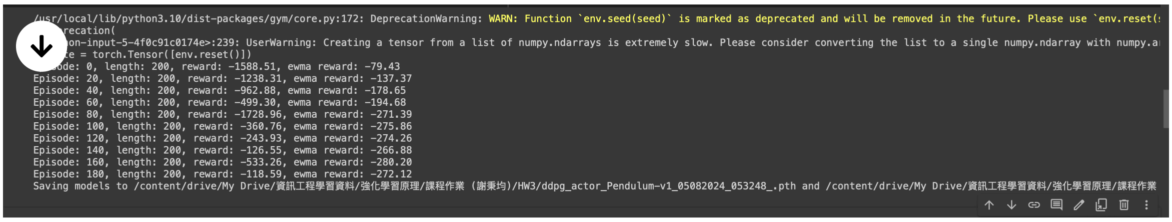
ep_len



ep_reward

reward

train



```
/usr/local/lib/python3.10/dist-packages/gym/core.py:172: DeprecationWarning: WARN: Function `env.seed(seed)` is marked as deprecated and will be removed in the future. Please use `env.reset(s
  recation(
  on-input-5-4f0c91c0174e>:239: UserWarning: Creating a tensor from a list of numpy.ndarrays is extremely slow. Please consider converting the list to a single numpy.ndarray with numpy.a
  te = torch.Tensor([env.reset()])
  ode: 0, length: 200, reward: -1588.51, ewma reward: -79.43
Episode: 20, length: 200, reward: -1238.31, ewma reward: -137.37
Episode: 40, length: 200, reward: -962.88, ewma reward: -178.65
Episode: 60, length: 200, reward: -499.30, ewma reward: -194.68
Episode: 80, length: 200, reward: -1728.96, ewma reward: -271.39
Episode: 100, length: 200, reward: -360.76, ewma reward: -275.86
Episode: 120, length: 200, reward: -243.93, ewma reward: -274.26
Episode: 140, length: 200, reward: -126.55, ewma reward: -266.88
Episode: 160, length: 200, reward: -533.26, ewma reward: -280.20
Episode: 180, length: 200, reward: -118.59, ewma reward: -272.12
Saving models to /content/drive/My Drive/資訊工程學習資料/強化學習原理/課程作業 (謝秉均)/HW3/ddpg_actor_Pendulum-v1_05082024_053248_.pth and /content/drive/My Drive/資訊工程學習資料/強化學習原理/課程作業
```

hyperparameters

```
num_episodes = 200
gamma = 0.995
tau = 0.002
hidden_size = 128
noise_scale = 0.3
replay_size = 100000
batch_size = 128
updates_per_step = 1
print_freq = 20
ewma_reward = 0
rewards = []
ewma_reward_history = []
total_numsteps = 0
updates = 0
```

learning rates

```
def __init__(self, num_inputs, action_space, gamma=0.995, tau=0.0005, hidden_size=128, lr_a=1e-4, lr_c=1e-3):
```

NN architecture

```
class Actor(nn.Module):
    def __init__(self, hidden_size, num_inputs, action_space):
        super(Actor, self).__init__()
        self.action_space = action_space
        num_outputs = action_space.shape[0]

        ########## YOUR CODE HERE (5~10 lines) ##########
        # Construct your own actor network
        self.fc1 = nn.Linear(num_inputs, hidden_size)
        self.fc2 = nn.Linear(hidden_size, hidden_size)
        self.fc3 = nn.Linear(hidden_size, num_outputs)
        self.relu = nn.ReLU()
        self.tanh = nn.Tanh()
```

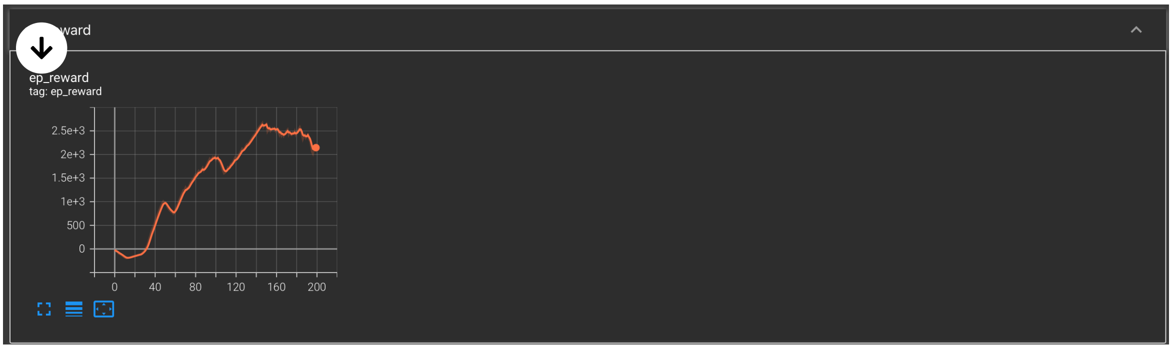
```python
class Critic(nn.Module):
    def __init__(self, hidden_size, num_inputs, action_space):
        super(Critic, self).__init__()
        self.action_space = action_space
        num_outputs = action_space.shape[0]

        ########## YOUR CODE HERE (5~10 lines) ##########
        # Construct your own critic network
        self.fc1 = nn.Linear(num_inputs + num_outputs, hidden_size)
        self.fc2 = nn.Linear(hidden_size, hidden_size)
        self.fc3 = nn.Linear(hidden_size, 1)
        self.relu = nn.ReLU()
```

b

ep_len

ep_reward



reward

train

```
[16] Episode: 157, length: 1000, reward: 2083.81, ewma reward: 2534.96
     Episode: 158, length: 1000, reward: 3020.91, ewma reward: 2559.26
     Episode: 159, length: 1000, reward: 1305.35, ewma reward: 2496.56
     Episode: 160, length: 1000, reward: 3318.22, ewma reward: 2537.64
     Episode: 161, length: 1000, reward: 2742.44, ewma reward: 2547.88
     Episode: 162, length: 1000, reward: 877.94, ewma reward: 2464.39
     Episode: 163, length: 1000, reward: 1312.17, ewma reward: 2406.78
     Episode: 164, length: 1000, reward: 3624.49, ewma reward: 2467.66
     Episode: 165, length: 1000, reward: 967.19, ewma reward: 2392.64
     Episode: 166, length: 1000, reward: 3241.97, ewma reward: 2435.10
     Episode: 167, length: 1000, reward: 1462.13, ewma reward: 2386.46
     Episode: 168, length: 1000, reward: 3333.04, ewma reward: 2433.78
     Episode: 169, length: 1000, reward: 3142.67, ewma reward: 2469.23
     Episode: 170, length: 1000, reward: 3363.18, ewma reward: 2513.93
     Episode: 171, length: 1000, reward: 2942.58, ewma reward: 2535.36
     Episode: 172, length: 1000, reward: 236.81, ewma reward: 2420.43
     Episode: 173, length: 1000, reward: 3423.65, ewma reward: 2470.59
     Episode: 174, length: 1000, reward: 1410.40, ewma reward: 2417.58
     Episode: 175, length: 1000, reward: 2252.22, ewma reward: 2409.32
     Episode: 176, length: 1000, reward: 3493.72, ewma reward: 2463.54
     Episode: 177, length: 1000, reward: 1959.56, ewma reward: 2438.34
     Episode: 178, length: 1000, reward: 3597.95, ewma reward: 2496.32
     Episode: 179, length: 1000, reward: 856.66, ewma reward: 2414.33
     Episode: 180, length: 1000, reward: 3307.08, ewma reward: 2458.97
     Episode: 181, length: 1000, reward: 3396.18, ewma reward: 2505.83
     Episode: 182, length: 1000, reward: 3294.55, ewma reward: 2545.27
     Episode: 183, length: 1000, reward: 3342.33, ewma reward: 2585.12
     Episode: 184, length: 1000, reward: 1332.18, ewma reward: 2522.47
     Episode: 185, length: 1000, reward: 338.65, ewma reward: 2413.28
     Episode: 186, length: 1000, reward: 457.40, ewma reward: 2315.49
     Episode: 187, length: 1000, reward: 3412.56, ewma reward: 2370.34
     Episode: 188, length: 1000, reward: 3340.93, ewma reward: 2418.87
     Episode: 189, length: 1000, reward: 1001.08, ewma reward: 2347.98
     Episode: 190, length: 1000, reward: 3126.40, ewma reward: 2386.90
     Episode: 191, length: 1000, reward: 3617.68, ewma reward: 2448.44
     Episode: 192, length: 1000, reward: -137.54, ewma reward: 2319.14
     Episode: 193, length: 1000, reward: 1711.94, ewma reward: 2288.78
     Episode: 194, length: 1000, reward: 52.80, ewma reward: 2176.98
     Episode: 195, length: 1000, reward: -246.57, ewma reward: 2055.81
     Episode: 196, length: 1000, reward: 1137.88, ewma reward: 2009.91
     Episode: 197, length: 1000, reward: 3773.08, ewma reward: 2098.07
     Episode: 198, length: 1000, reward: 3467.57, ewma reward: 2166.54
     Episode: 199, length: 1000, reward: 2095.95, ewma reward: 2163.01
     Saving models to /content/drive/My Drive/資訊工程學習資料/強化學習原理/課程作業 (謝秉均)/HW3/ddpg_cheetah_actor_HalfCheetah-v2_05082024_142324_.pth and /content/drive/My Drive/資訊工程學習資料/強化學習
```

hyperparameters

```
num_episodes = 200
gamma = 0.995
tau = 0.002
hidden_size = 128
noise_scale = 0.3
replay_size = 100000
batch_size = 128
updates_per_step = 1
print_freq = 20
ewma_reward = 0
rewards = []
ewma_reward_history = []
total_numsteps = 0
updates = 0
```

learning rates

```
def __init__(self, num_inputs, action_space, gamma=0.995, tau=0.0005, hidden_size=128, lr_a=1e-4, lr_c=1e-3):
```

NN architecture

```python
ass Actor(nn.Module):
    def __init__(self, hidden_size, num_inputs, action_space):
        super(Actor, self).__init__()
        self.action_space = action_space
        num_outputs = action_space.shape[0]

        ########## YOUR CODE HERE (5~10 lines) ##########
        # Construct your own actor network
        self.fc1 = nn.Linear(num_inputs, hidden_size)
        self.fc2 = nn.Linear(hidden_size, hidden_size)
        self.fc3 = nn.Linear(hidden_size, num_outputs)
        self.relu = nn.ReLU()
        self.tanh = nn.Tanh()
```

```python
ass Critic(nn.Module):
    def __init__(self, hidden_size, num_inputs, action_space):
        super(Critic, self).__init__()
        self.action_space = action_space
        num_outputs = action_space.shape[0]

        ########## YOUR CODE HERE (5~10 lines) ##########
        # Construct your own critic network
        self.fc1 = nn.Linear(num_inputs + num_outputs, hidden_size)
        self.fc2 = nn.Linear(hidden_size, hidden_size)
        self.fc3 = nn.Linear(hidden_size, 1)
        self.relu = nn.ReLU()
```