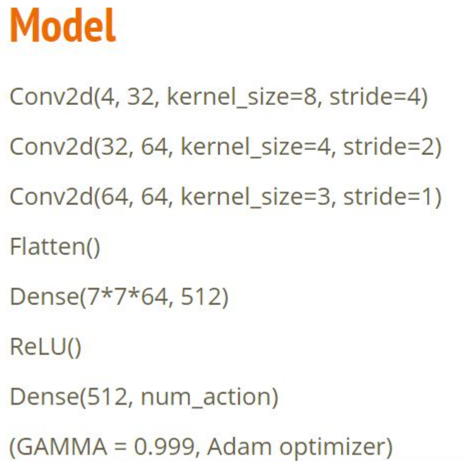
Machine Learning and Having it Deep and Structured
HW4-3 Actor-Critic

組員：陳泓均、陳欽安、詹書愷、丁昱升

Original Model

- Without parameter-sharing
- After each episode, update actor, critic once each
- Failed
    - Reward <= 2



Model

Conv2d(4, 32, kernel_size=8, stride=4)

Conv2d(32, 64, kernel_size=4, stride=2)

Conv2d(64, 64, kernel_size=3, stride=1)

Flatten()

Dense(7*7*64, 512)

ReLU()

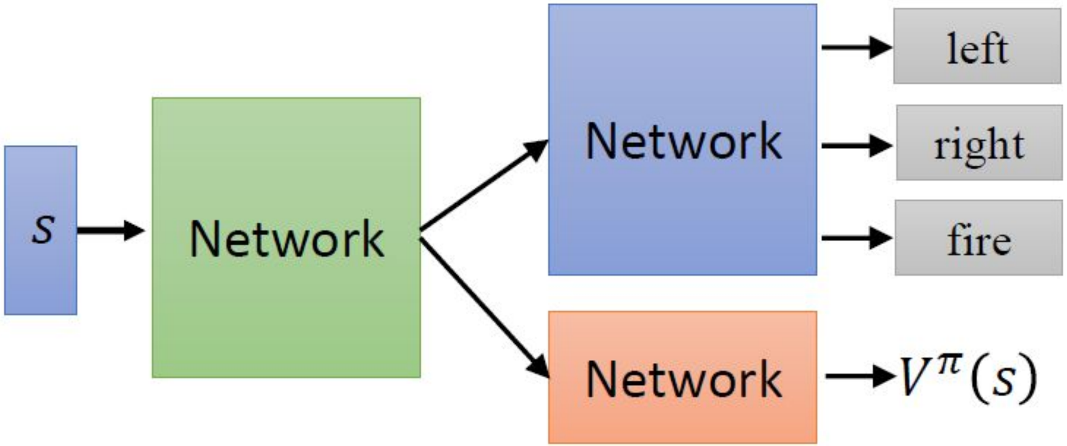
Dense(512, num_action)

(GAMMA = 0.999, Adam optimizer)

Original Model

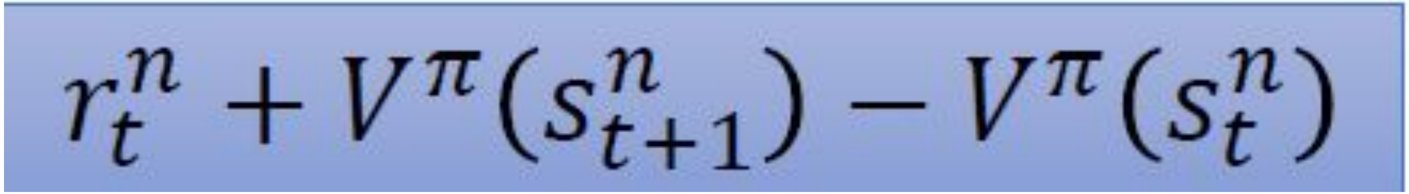
- Parameter-sharing
- Fail

$$r_t^n + V^\pi(s_{t+1}^n) - V^\pi(s_t^n)$$



Add target critic

- tried both parameter-sharing and not sharing
-



$$r_t^n + V^\pi(s_{t+1}^n) - V^\pi(s_t^n)$$

target(fixed) not fixed

-> Still failed

Accumulated Reward

- Start to train
- But very very slow…

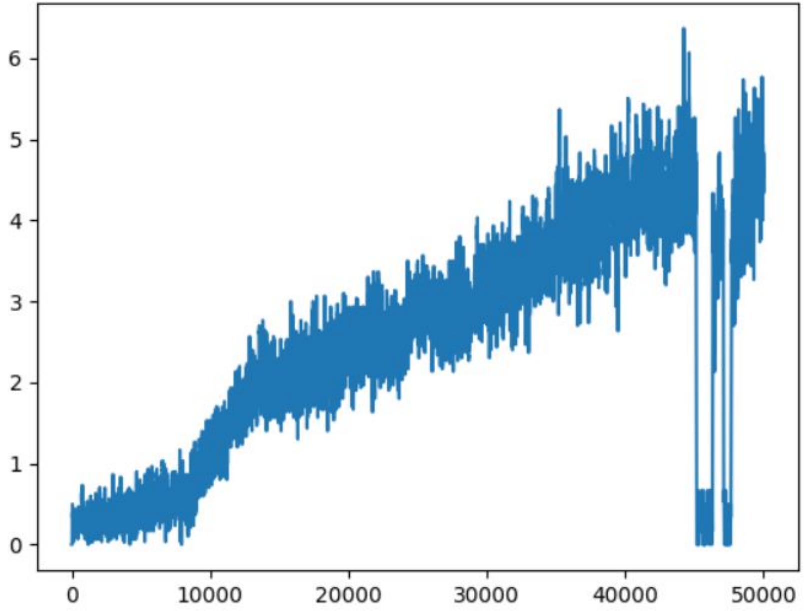
$$\nabla \bar{R}_\theta \approx \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n}\left(\boxed{\underline{\sum_{t'=t}^{T_n}\gamma^{t'-t} r_{t'}^{n}}} - \underline{b}\right)\nabla log p_\theta(a_t^n|s_t^n)$$

Entropy Regularization

- Use output entropy as regularization for actor
    - Larger entropy is preferred
        - exploration

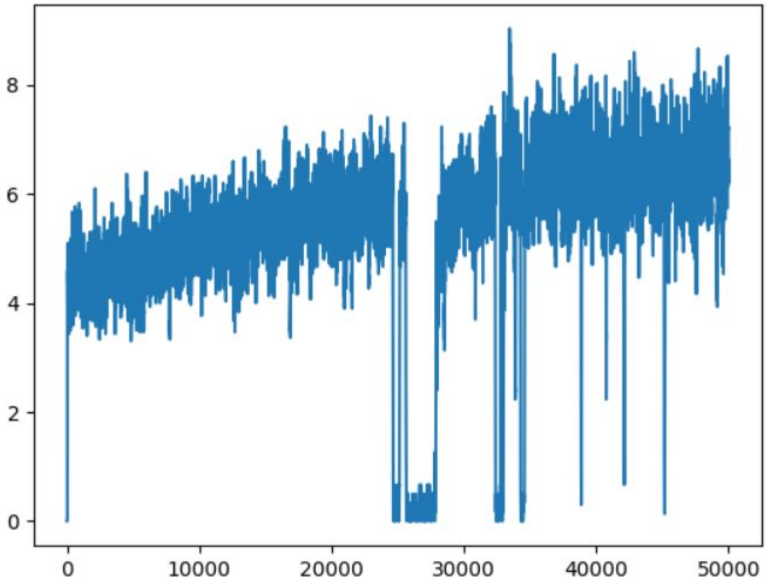
With Entropy Regularization-50000 episode

- Score : 6.04
  - Still learning



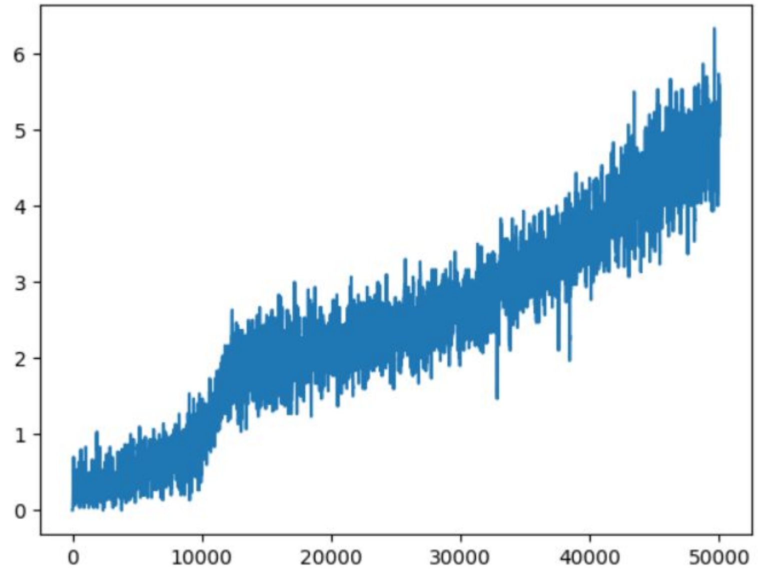
With Entropy Regularization-100000 episode

- Score : 13.11
  - Still learning



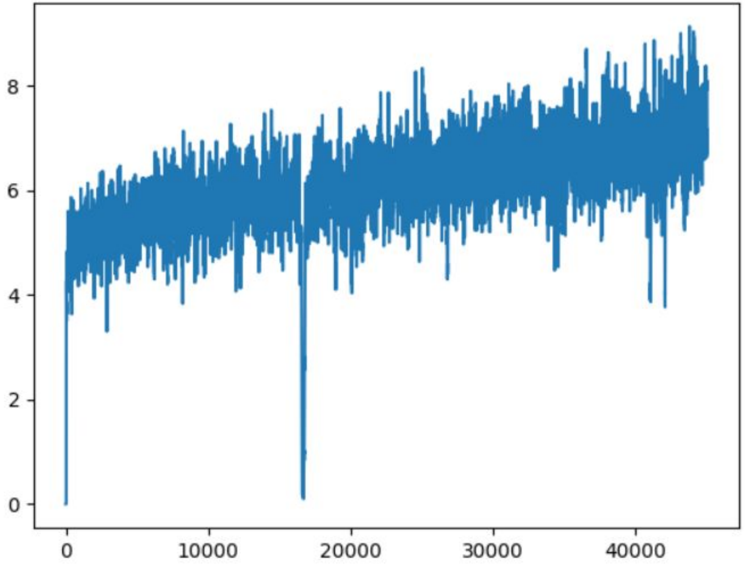
Without Entropy Regularization-50000 episode

- Score : 9.22
  - Still learning
  - Better



Without Entropy Regularization- 95299 episode

- Score : 14.18

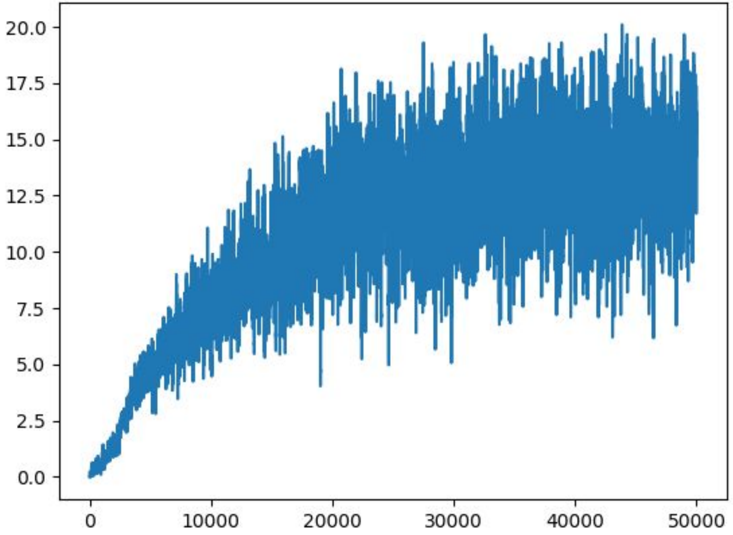


Entropy Regularization

- Use output entropy as regularization for actor
    - Not better?
        - Always exploration

Compared with Q-Learning

- Converge in 50000 episodes



20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
10000
20000
30000
40000
50000