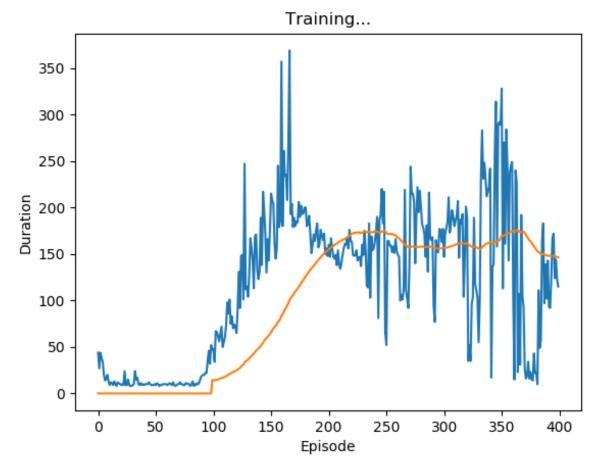
EECS 598 Deep Learning

Assignment 4

Shuyang HUANG 68621288

1. Deep Q-Network (DQN)

With all blocks filled, after trainning, we get the following result.



Notice that, the model we used is a combination of two full-connecte layers with dimension (4-64) and (64-2), plus a relu layer as the intermediate layer.

2. Policy Gradients

2.1

Notice,

$$abla_{ heta} p(au; heta) = p(au; heta) rac{
abla_{ heta} p(au; heta)}{p(au; heta)} = p(au; heta)
abla_{ heta} \log p(au; heta)$$

Also notice,

$$egin{aligned} p(au; heta) &= \prod_{t \geq 0} p\left(s_{t+1}|s_t, a_t
ight) \pi_{ heta}\left(a_t|s_t
ight) \ \log p(au; heta) &= \sum_{t \geq 0} \log p\left(s_{t+1}|s_t, a_t
ight) + \log \pi_{ heta}\left(a_t|s_t
ight) \
abla_{ heta} \log p(au; heta) &= \sum_{t \geq 0}
abla_{ heta} \log \pi_{ heta}\left(a_t|s_t
ight) \end{aligned}$$

Expand the agent's objective,

$$J(heta) = \mathbb{E}_{ au \sim p(au; heta)}[r(au)] = \int_{ au} r(au) p(au; heta) \mathrm{d} au$$

Take the gradient of θ on both side,

$$egin{aligned}
abla_{ heta} J(heta) &= \int_{ au} r(au)
abla_{ heta} p(au; heta) \mathrm{d} au \ &= \int_{ au} \left(r(au)
abla_{ heta} \log p(au; heta)
ight) p(au; heta) \mathrm{d} au \ &= \mathbb{E}_{ au \sim p(au; heta)} \left[r(au)
abla_{ heta} \log p(au; heta)
ight] \ &= \mathbb{E}_{ au \sim p(au; heta)} \left[r(au) \sum_{t \geq 0}
abla_{ heta} \log \pi_{ heta}(a_t | s_t)
ight] \end{aligned}$$

Consider a single episode τ^i is also $((a_1^i, s_1^i), \dots, (a_T^i, s_T^i))$, we have

$$egin{aligned}
abla_{ heta} J(heta) &pprox \sum_{t=1}^T r(au^i)
abla_{ heta} \log \pi_{ heta} \left(a_t^i | s_t^i
ight) \ &pprox rac{1}{N} \sum_{i=1}^N \sum_{t=1}^T
abla_{ heta} \log \pi_{ heta} \left(a_t^i | s_t^i
ight) r\left(au^i
ight) \end{aligned}$$

2.2

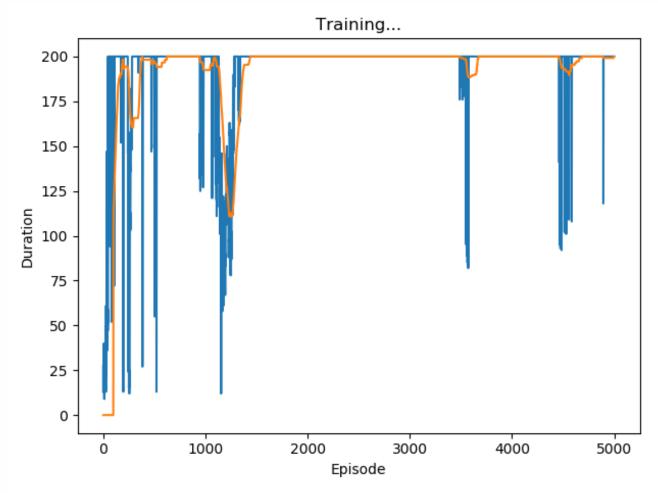
$$egin{aligned}
abla_{ heta} J(heta) &pprox rac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T}
abla_{ heta} \log \pi_{ heta} \left(a_{t}^{i} | s_{t}^{i}
ight) \sum_{t'=1}^{T} r_{t'}^{i} \ &= \sum_{t'=1}^{T}
abla_{ heta} \log \pi_{ heta} \left(a_{t} | s_{t}
ight) r(au) \ &= \sum_{t'=1}^{T} r_{t'} \sum_{t=1}^{T}
abla_{ heta} \log \pi_{ heta} \left(a_{t} | s_{t}
ight) \ &= \sum_{t'=1}^{T} r_{t'} \sum_{t=1}^{t'}
abla_{ heta} \log \pi_{ heta} \left(a_{t} | s_{t}
ight) \end{aligned}$$

Expand all, and reorganize them, we have,

$$egin{aligned}
abla_{ heta} J(heta) &pprox \sum_{t=1}^{T}
abla_{ heta} \log \pi_{ heta}\left(a_{t}|s_{t}
ight) \sum_{t'=t}^{T} r_{t'} \ &= rac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T}
abla_{ heta} \log \pi_{ heta}\left(a_{t}^{i}|s_{t}^{i}
ight) \sum_{t'=t}^{T} r_{t'}^{i} \end{aligned}$$

3. REINFORCE algorithm

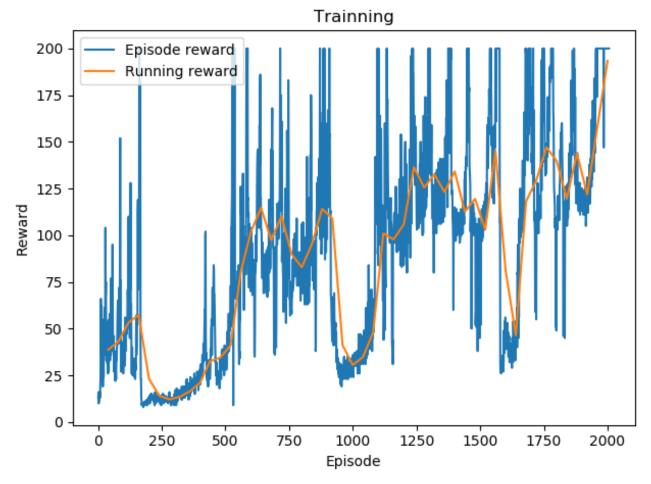
The reward curve is attached below.



As we can oberved, at last, nearly all episode will return a reward near 200, which is a very good performance.

4. Actor-Critic algorithm

The reward curve is attached below.



In above figure, the blue curve indicates the (total) reward for every episode, and the yellow curve indicates the running (average) reward.