EECS 598 Deep Learning

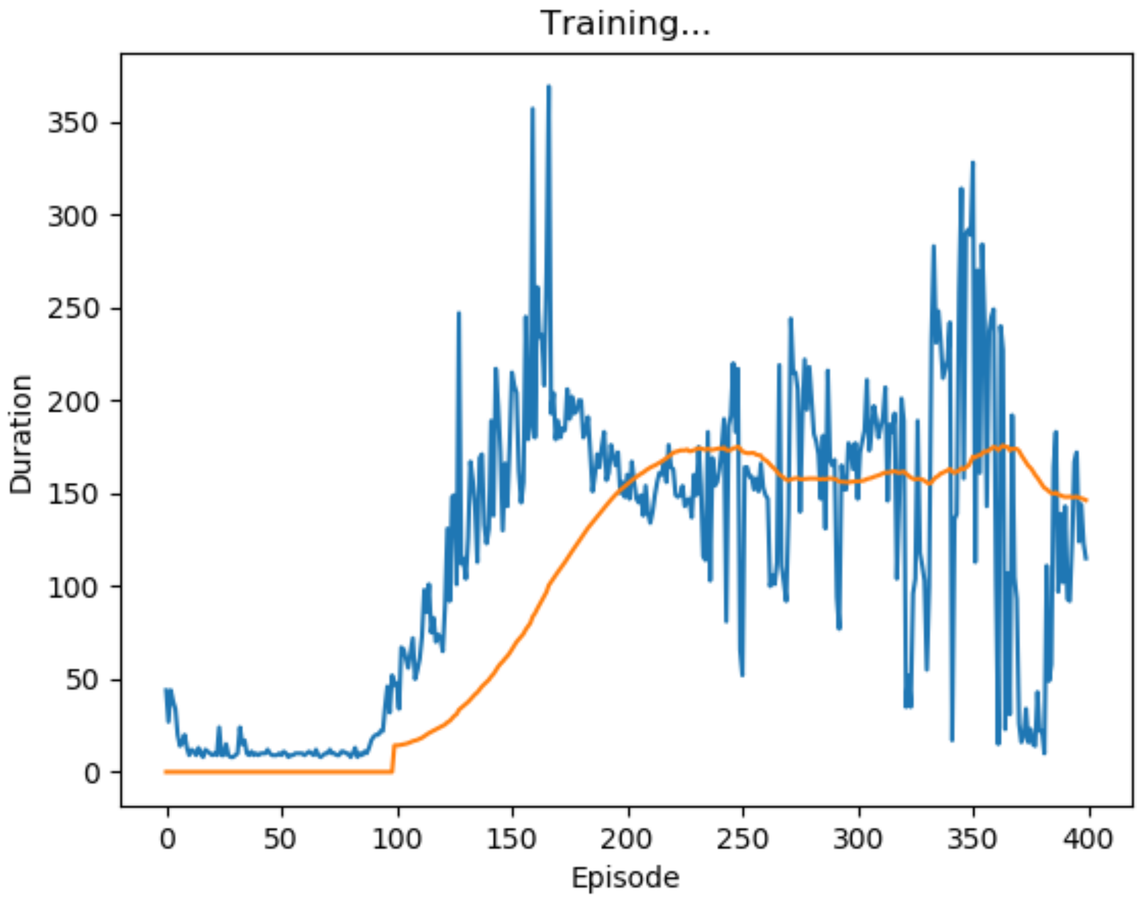
Assignment 4

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,

$$\nabla_\theta p(\tau;\theta) = p(\tau;\theta)\frac{\nabla_\theta p(\tau;\theta)}{p(\tau;\theta)} = p(\tau;\theta)\nabla_\theta \log p(\tau;\theta)$$

Also notice,

$$\begin{aligned}
p(\tau;\theta) &= \prod_{t\geq 0} p\left(s_{t+1} \middle| s_t, a_t\right) \pi_\theta\left(a_t \middle| s_t\right) \\
\log p(\tau;\theta) &= \sum_{t\geq 0} \log p\left(s_{t+1} \middle| s_t, a_t\right) + \log \pi_\theta\left(a_t \middle| s_t\right) \\
\nabla_\theta \log p(\tau;\theta) &= \sum_{t\geq 0} \nabla_\theta \log \pi_\theta\left(a_t \middle| s_t\right)
\end{aligned}$$

Expand the agent's objective,

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

Take the gradient of θ on both side,

$$\begin{aligned}
\nabla_\theta J(\theta) &= \int_{\tau} r(\tau) \nabla_\theta p(\tau;\theta) \mathrm{d}\tau \\
&= \int_{\tau} \left(r(\tau) \nabla_\theta \log p(\tau;\theta) \right) p(\tau;\theta) \mathrm{d}\tau \\
&= \mathbb{E}_{\tau \sim p(\tau;\theta)} \left[r(\tau) \nabla_\theta \log p(\tau;\theta) \right] \\
&= \mathbb{E}_{\tau \sim p(\tau;\theta)} [r(\tau) \sum_{t \geq 0} \nabla_\theta \log \pi_\theta (a_t | s_t)]
\end{aligned}$$

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

$$\begin{aligned}
\nabla_\theta J(\theta) &\approx \sum_{t=1}^{T} r(\tau^i) \nabla_\theta \log \pi_\theta \left(a_t^i | s_t^i \right) \\
&\approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta \left(a_t^i | s_t^i \right) r\left(\tau^i \right)
\end{aligned}$$

2.2

$$\begin{aligned}
\nabla_\theta J(\theta) &\approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta \left(a_t^i | s_t^i \right) \sum_{t'=1}^{T} r_{t'}^i \\
&= \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta \left(a_t | s_t \right) r(\tau) \\
&= \sum_{t'=1}^{T} r_{t'} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta \left(a_t | s_t \right) \\
&= \sum_{t'=1}^{T} r_{t'} \sum_{t=1}^{t'} \nabla_\theta \log \pi_\theta \left(a_t | s_t \right)
\end{aligned}$$

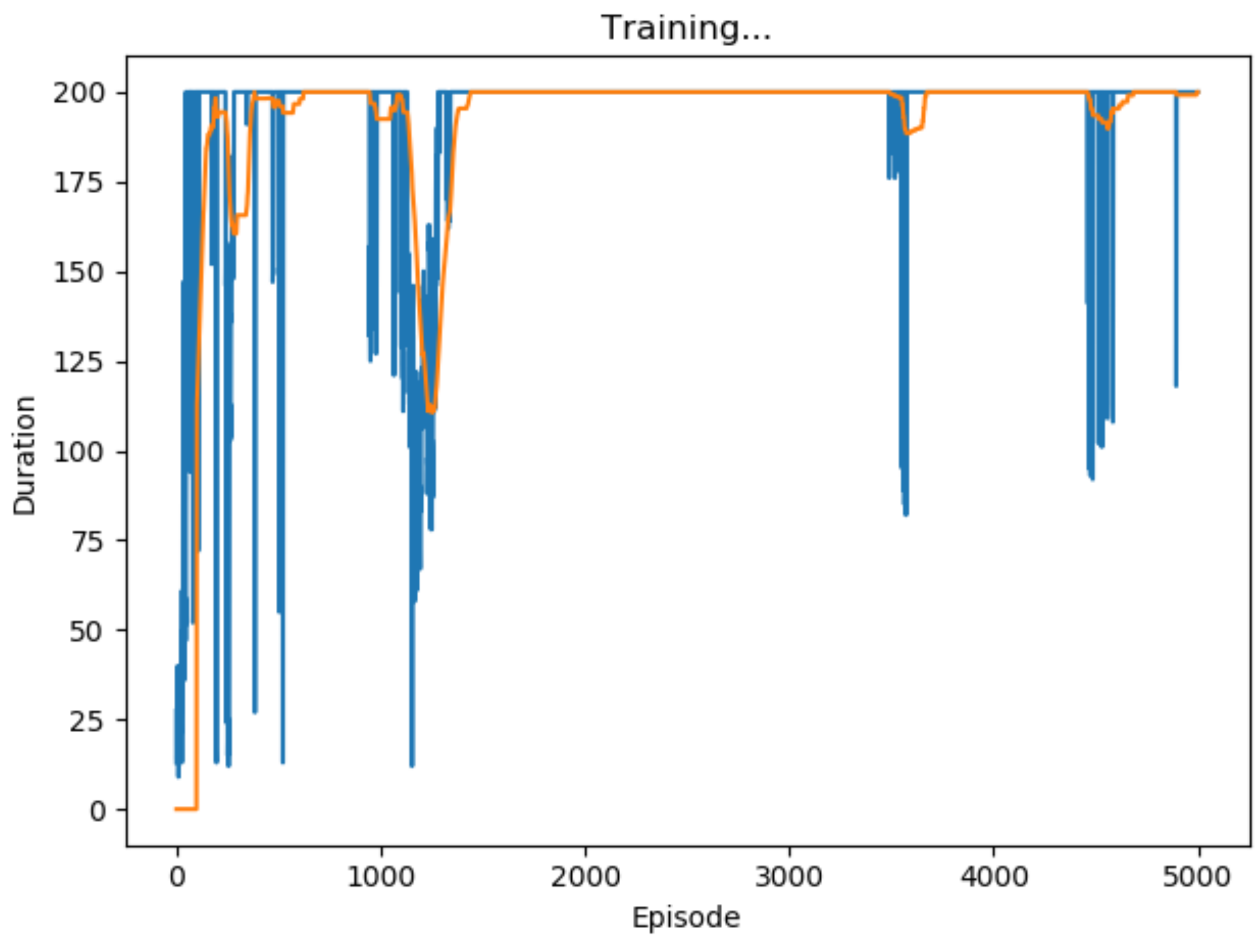
Expand all, and reorganize them, we have,

$$\begin{aligned}
\nabla_\theta J(\theta) &\approx \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta \left(a_t | s_t \right) \sum_{t'=t}^{T} r_{t'} \\
&= \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta \left(a_t^i | s_t^i \right) \sum_{t'=t}^{T} r_{t'}^i
\end{aligned}$$

Q.E.D.

3. REINFORCE algorithm

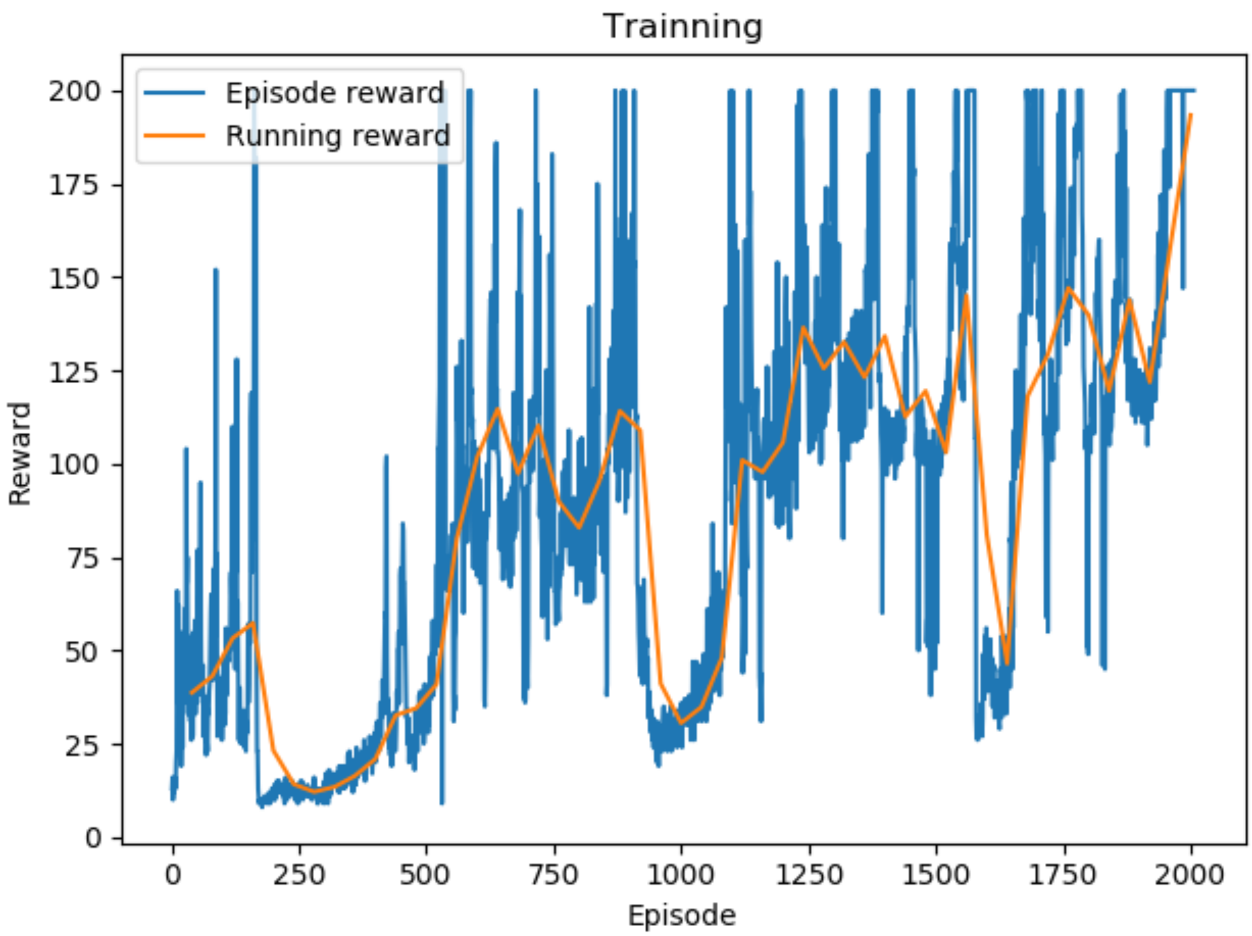
The reward curve is attached below.



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.