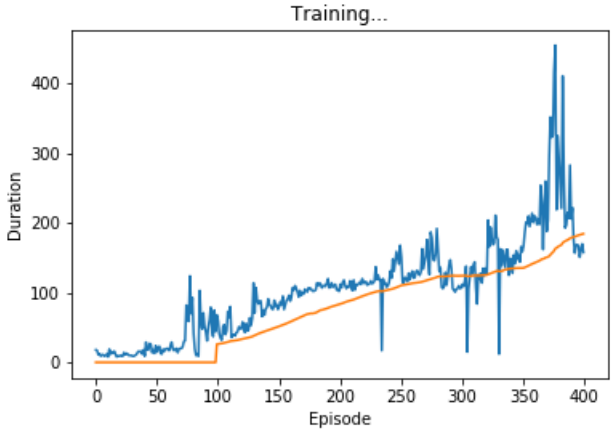
Author: sheng wang
study group: mengdi han, ruoxi yu

Homework 3

EECS 498
April 17, 2019

Problem 1

1. Learning Curve of episode duration.



1. $J(\theta) = \int_\tau r(\tau)\, p(\tau;\theta)\, d\tau$

$$\nabla_\theta J(\theta) = \nabla_\theta \int_\tau r(\tau)\, p(\tau;\theta)\, d\tau$$

$$= \int_\tau r(\tau)\, \nabla_\theta\, p(\tau;\theta)\, d\tau$$

$$= \int_\tau p(\tau;\theta)\, r(\tau)\, \nabla_\theta \log p(\tau;\theta)\, d\tau$$

$$= E\left[\, r(\tau)\, \nabla_\theta \log p(\tau;\theta)\,\right]$$

$$p(\tau;\theta) = \prod_{t\geq 0} p(s_{t+1}|s_t, a_t)\, \pi_\theta(a_t|s_t)$$

$$\nabla_\theta \log p(\tau;\theta) = \nabla_\theta \log p(s_{t+1}|s_t) + \nabla_\theta \log \pi_\theta(a_t|s_t)$$

$$= \nabla_\theta \log \pi_\theta(a_t|s_t)$$

According to Monte Carlo Sampling

$$\nabla_\theta J(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t^i|s_t^i)\, r(\tau^i)$$

2. $\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t^{(i)} | s_t^{(i)}) \, r(\tau^{(i)})$

$= E_\pi \left[\sum_{t=1}^{T} r(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t | s_t) \right]$

$= E_\pi \left[\sum_{t=1}^{T} \sum_{A_t} \pi(A_t | s_t, \theta) \, r(s_t, A_t) \frac{\nabla_\theta \pi_\theta(A_t | s_t, \theta)}{\pi_\theta(A_t | s_t, \theta)} \right]$

(↑ Replacing $a_t \sim \pi$ by A_t)

$= E_\pi \left[\sum_{t=1}^{T} \sum_{A_t} \nabla_\theta \pi_\theta(A_t | s_t) \, r(s_t, A_t) \right]$

$= E_\pi \left[\sum_{t=1}^{T} \sum_{A_t} \nabla_\theta \pi_\theta(A_t | s_t) \sum_{t'=1}^{T} r(s_{t'}, A_{t'}) \right]$

$= E_\pi \left[\sum_{t=1}^{T} \sum_{A_t} \nabla_\theta \pi_\theta(A_t | s_t) \left(\sum_{t'=t}^{T} r(s_{t'}, A_{t'}) + \sum_{t'=1}^{t-1} r(s_{t'}) \right) \right]$

(Because $r(s_t, A_t)$ is not ~~relacte~~ related to action when $t' < t$)

$E \left[\sum_{t=1}^{T} \sum_{A_t} R(s_t) \nabla_\theta \pi_\theta(A_t | s_t) \right]$

(Replacing $\sum_{t'=1}^{t-1} r(s_{t'})$ by $R(t)$)

$= E \left[\sum_{t=1}^{T} R(s_t) \nabla_\theta \sum_{A_t} \pi_\theta(A_t | s_t) \right] = 0$

(law of total probability)

$\therefore \nabla_\theta J(\theta) = E_\pi \left[\sum_{t=1}^{T} \sum_{A_t} \nabla_\theta \pi_\theta(A_t | s_t) \sum_{t'=t}^{T} r(s_{t'}, A_{t'}) \right]$

$\approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t^{(i)} | s_t^{(i)}) \sum_{t'=t}^{T} r_{t'}^{(i)}$.

5. According to 2.2, we can rewrite the equation as

$$\nabla_\theta J(\theta) = E_\pi \left[\sum_{t=1}^{T} \sum_{A_t} \nabla_\theta \pi_\theta(A_t|S_t) \left(\sum_{t'=t}^{T} r_{t'} - b(S_t) \right) \right]$$

$$= E_\pi \left[\sum_{t=1}^{T} \sum_{A_t} \nabla_\theta \pi_\theta(A_t|S_t) \sum_{t'=t}^{T} r_{t'} \right] -$$

$$E_\pi \left[\sum_{t=1}^{T} \sum_{A_t} \nabla_\theta \pi_\theta(A_t|S_t)\, b(S_t) \right]$$

the second term equals to..

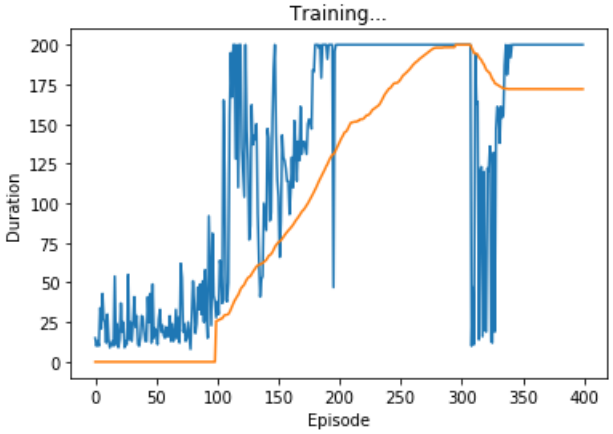
$$E_\pi \left[\sum_{t=1}^{T} b(S_t) \nabla_\theta \sum_{A_t} \pi_\theta(A_t|S_t) \right] = 0$$

~~$\therefore \nabla_\theta J(\theta) = E_\pi$~~

$\therefore$ Adding a baseline is an unbiased estimator of the gradient.

Problem 3

1. Learning curve of episode duration.



Problem 4

1. Learning curve of average reward.