Amina Tabassum

NUID:002190127

Exercise 09

Name: Amina Tabassum
NUID: 002190127
Exercise #09

Question 01:

Since, the equation is:

$$\eta(s) = h(s) + \sum_{\bar{s}} \eta(\bar{s}) \sum_{a} \pi(a|\bar{s}) p(s|\bar{s}, a)$$

where

$h(s)$ = probability that episode begins

$\eta(s)$ = number of time steps spent

So, generalisation of this recursion equation that governs expected time in each state:

$$\eta(s) = h(s) + \gamma \sum_{\bar{s}} \eta(\bar{s}) \sum_{a} \pi(a|\bar{s}) p(s|\bar{s}, a)$$

$$\eta(s) = h(s) + \gamma \sum_{\bar{s},a} \pi(a|\bar{s}) p(s|\bar{s}, a) + \gamma \left[\sum_{\bar{s},a} \pi(a|\bar{s}) p(s|\bar{s}, a) \times \sum_{x,a'} \pi(a'|x) p(\bar{s}|x, a) + \cdots \right]$$

So, this generalisation just changes solution for $\eta(s)$ but the equation for on-policy distribution is still

$$\mu(s) = \frac{\eta(s)}{\sum_{s'} \eta(s')}$$

The generalisation for proof of policy gradient theorem is:

$$v_\pi(s) = \sum_a \pi(a|s) q_\pi(s,a) \quad \text{for all } s \in S$$

Now, taking gradient w.r.t θ and unfolding Bellman's equation gives:

$$\nabla_\theta v_\pi(s) = \nabla_\theta \left[\sum_a \pi(a|s) q_\pi(s,a) \right]$$

using product rule:

$$\because \frac{d}{dx}(f(x)g(x)) = f(x)g'(x) + f'(x)g(x)$$

$$= \sum_a \left[\nabla_\theta (\pi(a|s) q_\pi(s,a)) \right]$$

$$= \sum_a \left[\nabla_\theta \pi(a|s) q_\pi(s,a) + \pi(a|s) \nabla_\theta q_\pi(s,a) \right]$$

$$= \sum_a \left[\nabla_\theta \pi(a|s) q_\pi(s,a) + \pi(a|s) \nabla_\theta \sum_{s',r} p(s',r|s,a)(r + \gamma v_\pi(s')) \right]$$

$$= \sum_a \left[\nabla_\theta \pi(a|s) q_\pi(s,a) + \pi(a|s) \sum_{s'} p(s'|s,a) \nabla_\theta v_\pi(s') \right]$$

$$= \sum_a \left[\nabla_\theta \pi(a|s) q_\pi(s,a) + \pi(a|s) \sum_{s'} p(s'|s,a) \nabla_\theta \left(\sum_{a'} \pi(a'|s') q_\pi(s',a') \right) \right]$$

$$= \sum_a \left[\nabla_\theta \pi(a|s) q_\pi(s,a) + \pi(a|s) \sum_{s'} p(s'|s,a) \sum_{a'} \left(\nabla_\theta \pi(a'|s') q_\pi(s',a') + \pi(a'|s') \nabla_\theta q_\pi(s',a') \right) \right]$$

$$= \sum_a \left[\nabla_\theta \pi(a|s) q_\pi(s,a) + \pi(a|s) \sum_{s'} p(s'|s,a) \sum_{a'} \left(\nabla_\theta \pi(a'|s') q_\pi(s',a') + \pi(a'|s') \sum_{s''} p(s''|s',a') \nabla_\theta v_\pi(s'') \right) \right]$$

$$= \sum_{x \in S} \sum_{k=0}^{\infty} Pr(s \to x, k, \pi) \sum_a \nabla_\theta \pi(a|x) q_\pi(x,a)$$

To consider discounting, we need to view it as form of termination. So, policy gradient theorem becomes:

$$\nabla_\theta J(\theta) = E_\pi \left[\gamma_t \sum_a q_\pi(s_t, a) \nabla_\theta \pi(a|s_t, \theta) \right]$$

we can derive it as:

$$\nabla_\theta J(\theta) = \nabla v_\pi(s_0)$$

$$= \sum_s \left(\sum_{k=0}^{\infty} P_\gamma(s_0 \to s, k, \pi) \right) \sum_a \nabla \pi(a|s) q_\pi(s, a)$$

from page 1:

$$= \sum_s \gamma \mu(s) \sum_a \pi(a|s) q_\pi(s, a)$$

$$\propto \sum_s \gamma \mu(s) \sum_a \nabla_\theta \pi(a|s) q_\pi(s, a)$$

Hence,

$$\nabla_\theta J(\theta) \propto \sum_s \gamma \mu(s) \sum_a q_\pi(s, a) \nabla_\theta \pi(a|s, \theta)$$

$$\boxed{\nabla_\theta J(\theta) = \gamma_t E_\pi \left[\sum_a q_\pi(s_t, a) \nabla_\theta \pi(a|s_t, \theta) \right]}$$

The factor γ_t then follows through when we apply SGD. proof is not given in book.

Question NO:02

$$\nabla \log \pi(a|s,\theta) = x(s,a) - \sum_b \pi(b|s,\theta)\, x(s,b)$$

Proof:

Since, softmax policy can be written as:

$$\pi(a|s,\theta) = \frac{\exp(h(s,a,\theta))}{\sum_b \exp(h(s,a,\theta))}$$

and

$$h(s,a,\theta) = \theta^T x(s,a)$$

So,

$$\nabla \log \pi(a|s,\theta) = \log\left[\frac{\exp(h(s,a,\theta)}{\sum_b \exp(h(s,a,\theta)}\right]$$

$$= \log e^{h(s,a,\theta)} - \log \sum_b e^{h(s,a,\theta)}$$

$$= h(s,a,\theta) - \log \sum_b e^{\theta^T x(s,a)}$$

$$\log \pi(a|s,\theta) = \theta^T x(s,a) - \log \sum_b e$$

$$\nabla \log(\pi(a|s,\theta)) = x(s,a) - \sum_b x(s,b)\,\pi(b|s,\theta)$$

Hence, proved

Question #03

Gaussian policy can be written as:

$$\pi(a|s,\theta) = \frac{1}{\sigma(s,\theta)\sqrt{2\pi}} e^{\left(-\frac{(a-\mu(s,\theta))^2}{2\sigma(s,\theta)^2}\right)}$$

where

$$\mu(s,\theta_\mu) = \theta_\mu^T x_\mu(s)$$

$$\sigma(s,\theta_\sigma) = e^{\theta_\sigma^T x_\sigma(s)}$$

$$\log \pi(a|s,\theta) = \log\left\{\frac{1 \times e^{\left(-\frac{(a-\theta_\mu^T x_\mu(s))^2}{2\times(e^{\theta_\sigma^T x_\sigma(s)})^2}\right)}}{(e^{\theta_\sigma^T x_\mu(s)})\sqrt{2\pi}}\right\}$$

$$= -\log\sqrt{2\pi} - \log\sigma - \frac{(a-\mu)^2}{2\sigma^2}$$

So,

$$\nabla_{\theta_\sigma} \log \pi(a|s,\theta) = -\frac{\nabla_{\theta_\sigma}\sigma}{\sigma} + \frac{(a-\mu)^2}{\sigma^2}\nabla_{\theta_\sigma}\sigma$$

$$= \left(\frac{(a-\mu)^2}{\sigma^2} - 1\right) x_\sigma(s)$$

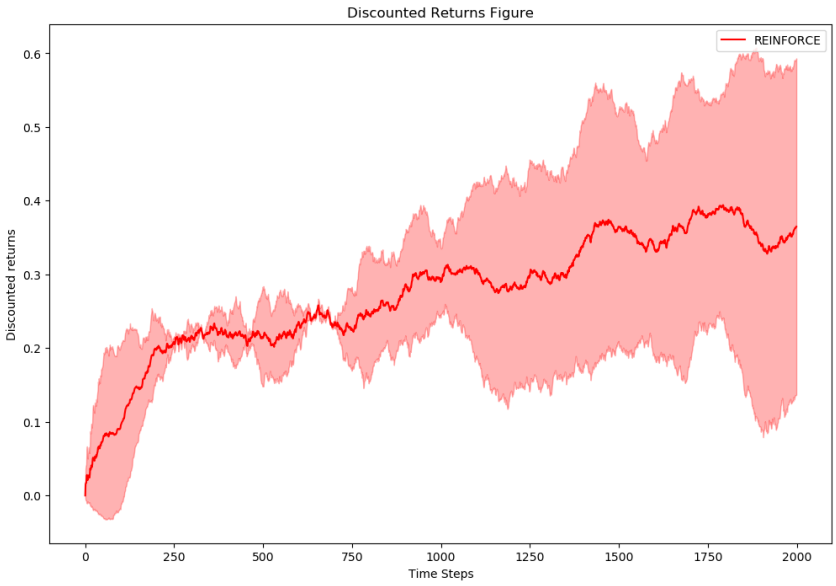
because

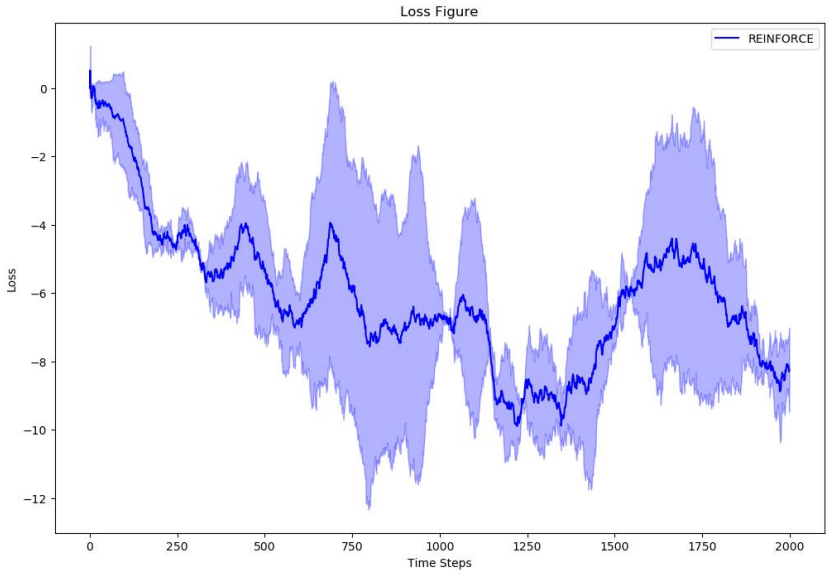
$$\nabla_{\theta_\sigma}\sigma = x_\sigma(s)\sigma$$

Hence, proved!

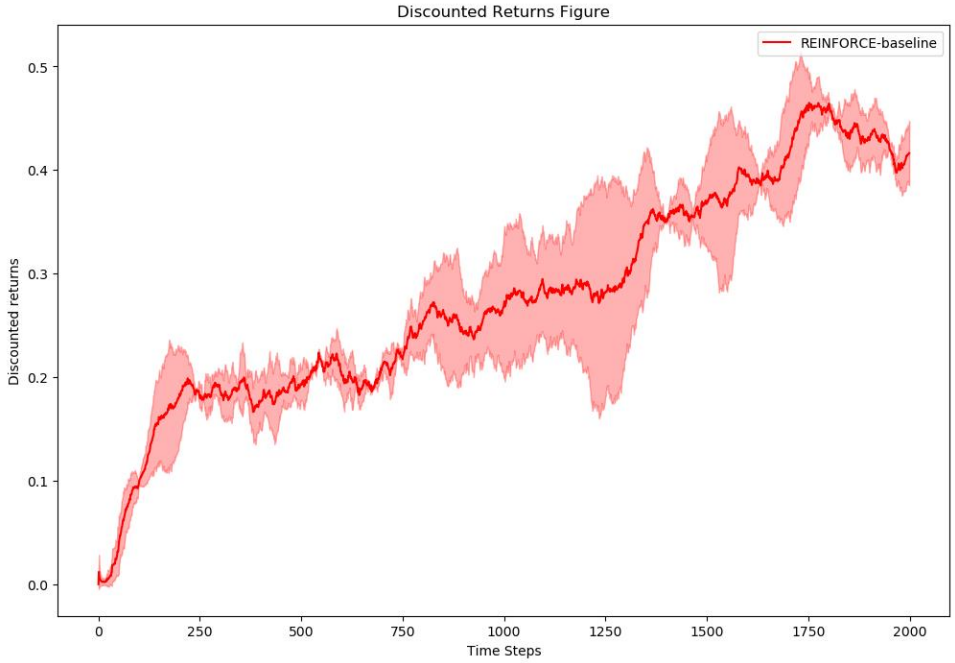
Question 04:

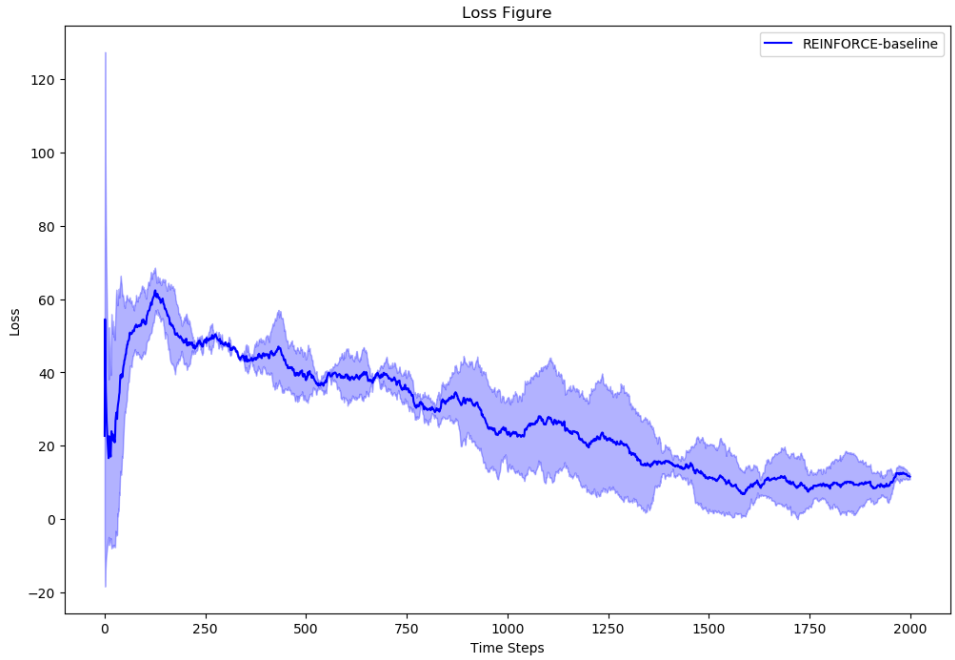
a)

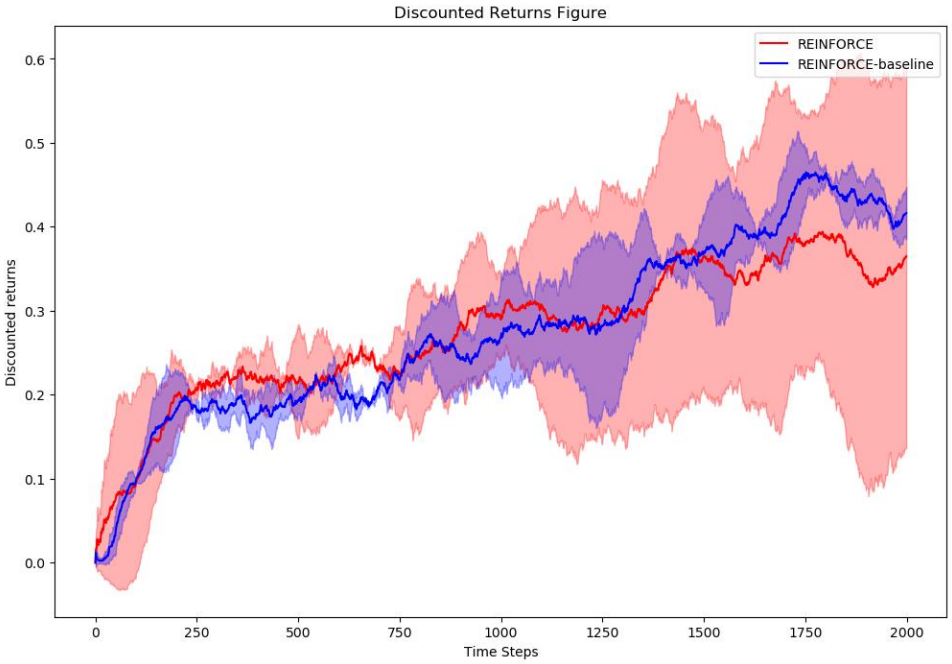


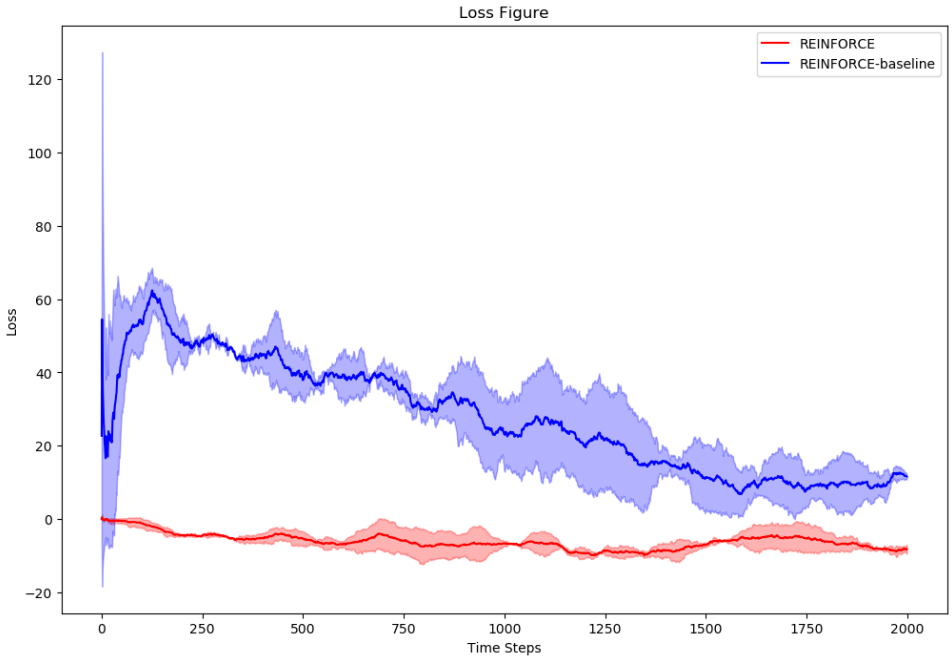


b)


Discounted Returns Figure
REINFORCE-baseline
Discounted returns
Time Steps
0.0
0.1
0.2
0.3
0.4
0.5
0
250
500
750
1000
1250
1500
1750
2000


Loss Figure
REINFORCE-baseline
Loss
Time Steps
−20
0
20
40
60
80
100
120
0
250
500
750
1000
1250
1500
1750
2000


Discounted Returns Figure
REINFORCE
REINFORCE-baseline
Discounted returns
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
250
500
750
1000
1250
1500
1750
2000
Time Steps


Loss Figure
REINFORCE
REINFORCE-baseline
Loss
120
100
80
60
40
20
0
−20
0
250
500
750
1000
1250
1500
1750
2000
Time Steps

c)

One important thing to notice is that variance is lower for REINFROCE with baseline as compared to simple REINFROCE algorithm. The variance is higher for simple REINFROCE algorithm because of the reward scale. Ideally, we think of policy gradient methods as it increases the probability of taking good actions and reduces the probability of bad actions. But mostly this is not true. For example, consider the return of good episode is 40 and return for bad episode is 15 so, simple REINFORCE will increase probability of both actions which is not the case we want. So, when a baseline function is added to expectation, it does not change the expected value but reduces variance in the sense that it selects actions better than average and increases their probability. At the same time, it reduces the probability of actions with return lower than average and hence reduces their probability. In this way, it does not change expected return but reduces variance.

In other words, we can say that REINFORCE with baseline is more stable than simple REIFNORCE algorithm. Because here we subtract the average expected return from action-values. Contrast to vanilla policy gradient, where Q-values continuously increase and hence it leads to situation when minor incremental update to one of the actions causes vast changes in policy. Hence, you can see from the above plots that baseline REINFORCE algorithm performs better and result is consistent than standard REINFORCE algorithm. We can observe loss function as well that for vanilla PG algorithm, the loss function decreases from zero and it goes negative. Whereas REINFORCE with baseline, loss drastically reduces from 60 to 20 initially, and then agent learns to select better actions in one direction and hence this algorithm is more consistent and stable.