RL ASSIGNMENT Q3 Write Up :

CS21MDS14025

Sauradeep Debnath

Q3.a) part : Code for printing environment details :

I have trained Q3 in Google Colab. The notebook runs on colab --may or may not run on a Local machine depending on installed packages

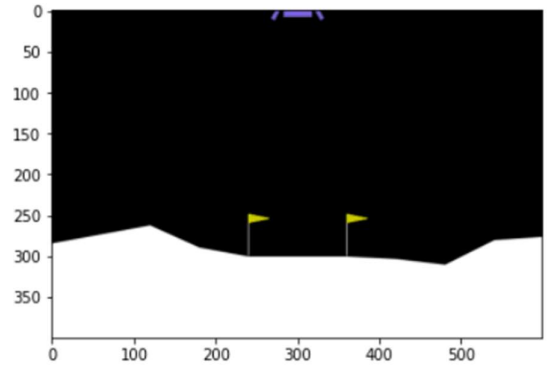
```
env = gym.make(environment)
print(environment)
env.reset()
#action_size = env.action_space.n
#state_size = env.observation_space.shape[0]
print('Action Space-->',env.action_space)
print('Observation Space-->',env.observation_space)
print('reward range-->',env.reward_range)
print('Meta data --> ',env.metadata)
print('Specifications -->',env.spec)
# set seed
prev_screen = env.render(mode='rgb_array')
plt.imshow(prev_screen)
try :
  for i in range(200):
      env.render()
      action = env.action_space.sample()
      obs, reward, done, info = env.step(action)
      if done:
          env.reset()
except :
  pass

env.close()
```

The output being :

FOR CARTPOLE -→
CartPole-v0
Action Space--> Discrete(2)
Observation Space--> Box([-4.8000002e+00 -3.4028235e+38 -4.1887903e-01 -3.4028235e+38], [4.8000002e+00 3.4028235e+38 4.1887903e-01 3.4028235e+38], (4,), float32)
reward range--> (-inf, inf)
Meta data --> {'render_modes': ['human', 'rgb_array', 'single_rgb_array'], 'render_fps': 50}
Specifications --> EnvSpec(id='CartPole-v0', entry_point='gym.envs.classic_control.cartpole:CartPoleEnv', reward_threshold=195.0, nondeterministic=False, max_episode_steps=200, order_enforce=True, autoreset=False, disable_env_checker=False, new_step_api=False, kwargs={}, namespace=None, name='CartPole', version=0)

FOR LUNAR LANDER→

```
LunarLander-v2
Action Space--> Discrete(4)
Observation Space--> Box([-1.5       -1.5       -5.        -5.        -3.1415927 -5.
 -0.        -0.       ], [1.5       1.5       5.        5.        3.1415927 5.
 1.        1.       ], (8,), float32)
reward range--> (-inf, inf)
Meta data -->  {'render_modes': ['human', 'rgb_array', 'single_rgb_array'], 'render_fps': 50}
Specifications --> EnvSpec(id='LunarLander-v2', entry_point='gym.envs.box2d.lunar_lander:LunarLander', rewar
```



Q 3.b) PG Implementation

Main Notebook : PG_LunarLander_Adv_No_R2g_Yes_File_CS21MDS14025.ipynb

USER Input for Env, Adv Normalization, Reward to Go, Number of iterations :

```
[ ]      get_inputs(variable, option_1, option_2)
       return variable
    message = 'For Cartpole Enter 1 , for Lunar Lander Enter 2 '
    environment = get_inputs( 'CartPole-v0', 'LunarLander-v2', message)
```

```
For Cartpole Enter 1 , for Lunar Lander Enter 2
2
You have selected LunarLander-v2
```

```
[ ] message = 'If you want to do Advantage Normalization Press 1 , Else Press 2 '
    adv_norm = get_inputs( 'adv_norm', 'None', message)

    message = 'If you want to do Reward to Go functionality Press 1 , Else Press 2 '
    r2go = get_inputs( 'r2go', 'None', message)
```

```
If you want to do Advantage Normalization Press 1 , Else Press 2
2
You have selected None
If you want to do Reward to Go functionality Press 1 , Else Press 2
1
You have selected r2go
```

```
[ ] num_iterations = input('Enter the number of episodes /Iterations in digits e.g. 5000--->')
    num_iterations = int(num_iterations.strip())
    print('num_iterations You Selected is -->  {}'.format(num_iterations))
```

```
Enter the number of episodes /Iterations in digits e.g. 5000--->2000
num_iterations You Selected is -->  2000
```

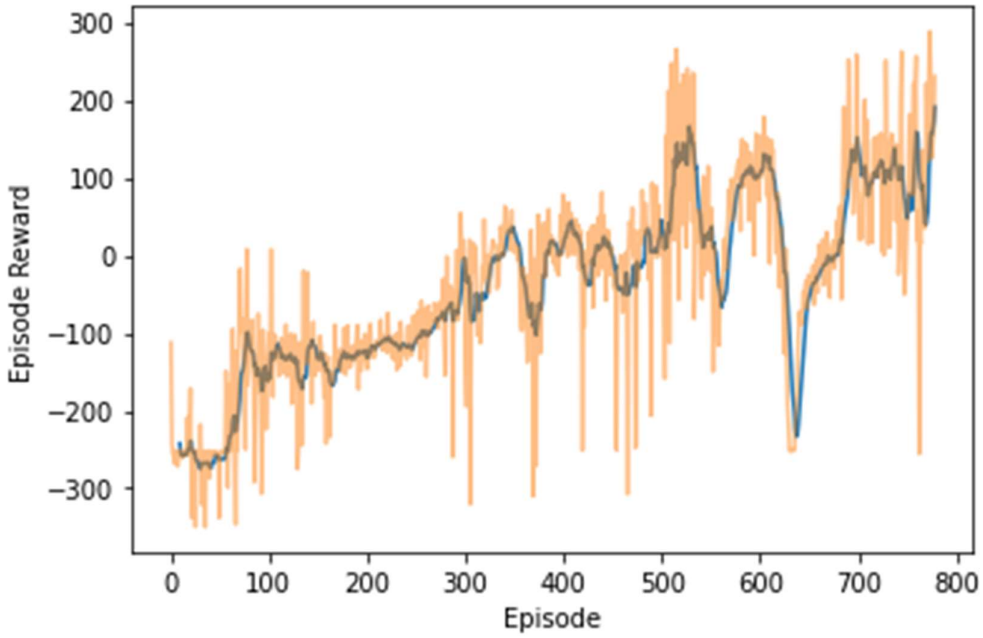
Convergence Plots:

NOTE : one of the termination criteria is getting 190 rewards

CASE 1 : For Adv Norm – No, Reward to Go – Yes : (NOTE currently only this version of Code is working currently→
So please choose these versions only)
(For **Lunar Lander**)
Notebook : PG_LunarLander_Adv_No_R2g_Yes_File_CS21MDS14025.ipynb

```
[5] message = 'If you want to do Advantage Normalization Press 1 , Else Press 2 '
    adv_norm = get_inputs( 'adv_norm', 'None', message)

    message = 'If you want to do Reward to Go functionality Press 1 , Else Press 2 '
    r2go = get_inputs( 'r2go', 'None', message)
```

```
If you want to do Advantage Normalization Press 1 , Else Press 2
2
You have selected None
If you want to do Reward to Go functionality Press 1 , Else Press 2
1
You have selected r2go
```



GAME 2 – CARTPOLE :
Notebook : PG_Cartpole_Adv_No_R2g_Yes_File_CS21MDS14025.ipynb
CASE 1 : For Adv Norm – No, Reward to Go – Yes : (In the User Inputs)

For the Same Hyperparameter combo – **Cartpole** converged much faster due to 1. the reward system being different
2. The graph is more monotonic, less zig zag as compared to Lunar Lander (GRAPH BELOW):

```
Episode: 0 Timestep: 21 Total reward: 21.0 Episode length: 21.0 Actor Loss: 0.0189 VF Loss: 0.0075
Episode: 10 Timestep: 180 Total reward: 15.9 Episode length: 15.9 Actor Loss: -0.0177 VF Loss: 0.0149
Episode: 20 Timestep: 349 Total reward: 16.9 Episode length: 16.9 Actor Loss: 0.0166 VF Loss: 0.0041
Episode: 30 Timestep: 502 Total reward: 15.3 Episode length: 15.3 Actor Loss: -0.0051 VF Loss: 0.0016
Episode: 40 Timestep: 657 Total reward: 15.5 Episode length: 15.5 Actor Loss: -0.0037 VF Loss: 0.0012
Episode: 50 Timestep: 962 Total reward: 30.5 Episode length: 30.5 Actor Loss: -0.0015 VF Loss: 0.0091
Episode: 60 Timestep: 1507 Total reward: 54.5 Episode length: 54.5 Actor Loss: 0.0138 VF Loss: 0.0279
Episode: 70 Timestep: 2043 Total reward: 53.6 Episode length: 53.6 Actor Loss: 0.0040 VF Loss: 0.0279
Episode: 80 Timestep: 3046 Total reward: 100.3 Episode length: 100.3 Actor Loss: 0.0147 VF Loss: 0.0369
Episode: 90 Timestep: 4684 Total reward: 163.8 Episode length: 163.8 Actor Loss: -0.0142 VF Loss: 0.0338
Episode: 100 Timestep: 6410 Total reward: 172.6 Episode length: 172.6 Actor Loss: -0.0111 VF Loss: 0.0606
Episode: 110 Timestep: 8038 Total reward: 162.8 Episode length: 162.8 Actor Loss: 0.0035 VF Loss: 0.0325
Episode: 120 Timestep: 9858 Total reward: 182.0 Episode length: 182.0 Actor Loss: -0.0058 VF Loss: 0.0359
Stopping at episode 128 with average rewards of 191.7 in last 10 episodes
```