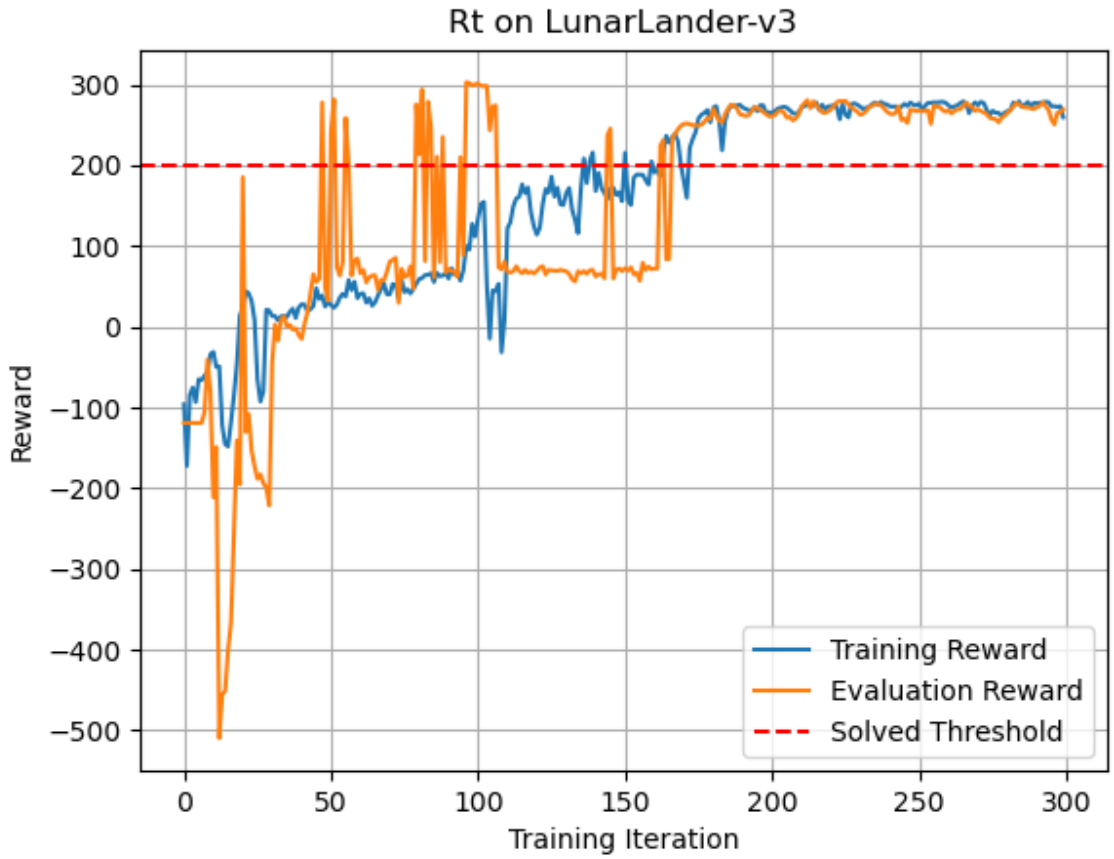
Name: V Raja Karthik **UIN: 635009384**

2.1.

As part of this question, I trained an agent using REINFORCE algorithm to play lunar lander game. The episodic undiscounted-cumulative reward plot is given below:



As seen in the graph, the training process is a bit unstable. Convergence was achieved after 175 iterations of training for the following set of hyper parameters:

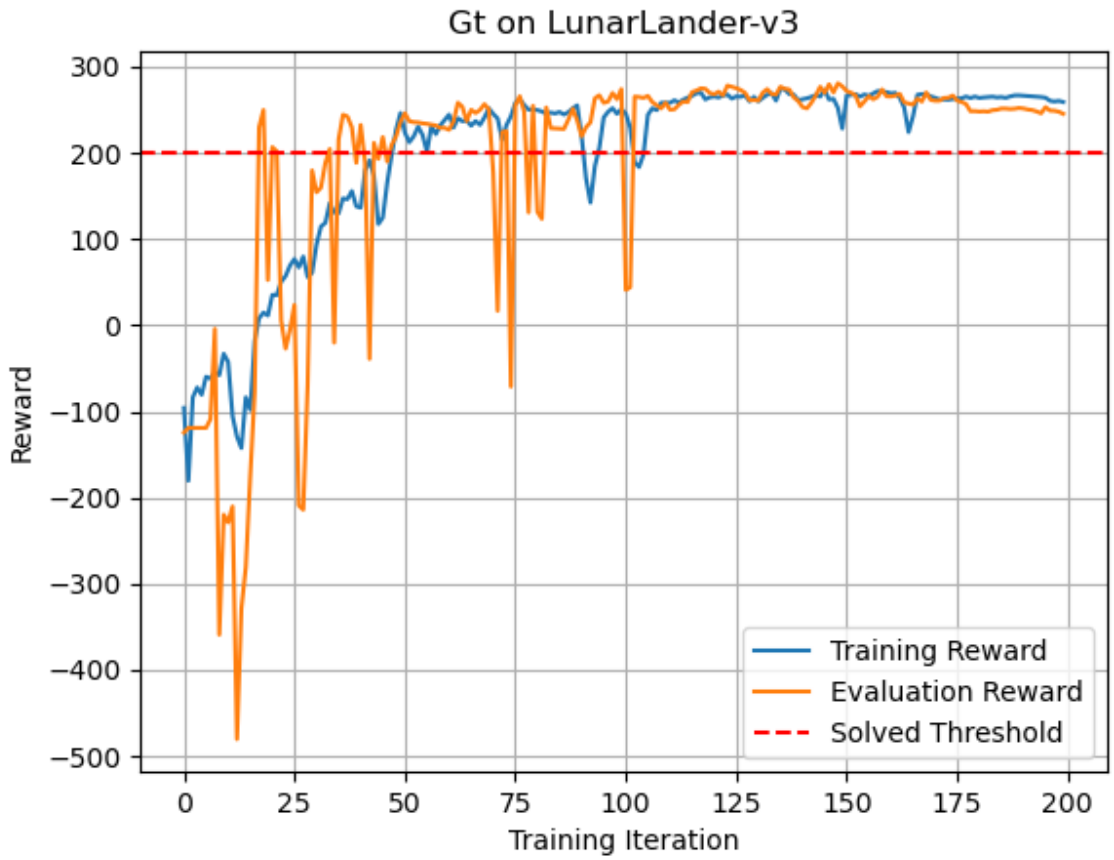
***N-iter**: 300.*	300 iterations
***batch_size:** 10000.*	10000 steps of play included for every iteration of training to perform gradient approximation
***learning_rate: 5**e-3.*	Learning rate of 0.005 was used to train the policy network

A variety of hyper parameters were tried out to achieve faster convergence and only the above mentioned parameters were tweaked. All the other hyper parameters were kept at their default values.
I tried training the agent with the following sets of hyper parameters:

- **(Learning rate: 1e-3, batch_size: 5000)** : Convergence was observed after ~750 epochs of training and the training was very unstable
- **(Learning rate: 1e-3, batch_size: 10000)** : Convergence was observed after ~700 epochs of training.
- **(Learning rate: 1e-4, batch_size: 10000)** : Convergence was observed after ~600 epochs of training and the training was much stable compared to when the batch size was 5k as more number of steps were used in this case for estimating gradient

As seen in the plot, The network converges to a reward of greater than 200 after around 175 epochs and this was the fastest convergence for this algorithm that I observed with the parameters mentioned.

2.2.
As part of this question, I trained an agent using Policy Gradient algorithm to play lunar lander game. The episodic undiscounted-cumulative reward plot is given below:



As seen in the graph convergence was achieved after 100 iterations of training for the following set of hyper parameters:

***N-iter**: 200.* 200 iterations

batch_size: *10000.*	10000 steps of play included for every iteration of training to perform gradient approximation
learning_rate: ***5****e-3.*	Learning rate of 0.005 was used to train the policy network

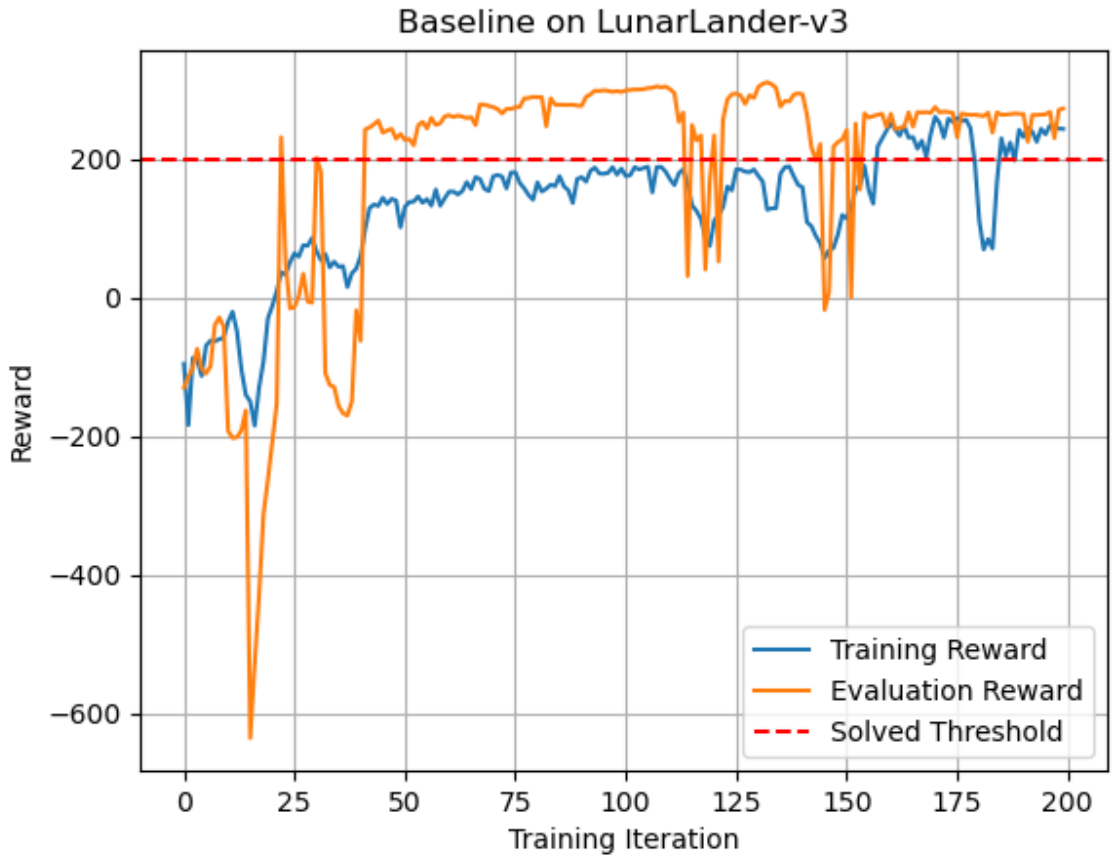
The training process was much more stable for this algorithm compared to the REINFORCE algorithm.

A variety of hyper parameters were tried out to achieve faster convergence and only the above mentioned parameters were tweaked. All the other hyper parameters were kept at their default values.
I tried training the agent with the following sets of hyper parameters:

- **(Learning rate: 1e-3, batch_size: 5000)** : Convergence was observed after ~650 epochs of training and the training was very unstable
- **(Learning rate: 1e-3, batch_size: 10000)** : Convergence was observed after ~500 epochs of training.
- **(Learning rate: 1e-4, batch_size: 10000)** : Convergence was observed after ~475 epochs of training and the training was much stable compared to when the batch size was 5k as more number of steps were used in this case for estimating gradient

As seen in the plot, The network converges to a reward of greater than 200 after around 100 epochs and this was the fastest convergence for this algorithm that I observed with the parameters mentioned.

2.3.



As part of this question, I trained an agent using Policy Gradient with baseline algorithm to play lunar lander game. The episodic undiscounted-cumulative reward plot is given above. The training process was very stable, I.e., the agent achieves higher rewards at a very early stage (~30 epochs) but it reached the threshold of 200 reward and stabilizes at that level after 150 epochs, which can be seen in the evaluation reward plot line in the figure

As seen in the graph, the training process is a bit unstable. Convergence was achieved after 150 iterations of training for the following set of hyper parameters:

***N-iter**: 200.*	*200* iterations
***batch_size:** 10000.*	10000 steps of play included for every iteration of training to perform gradient approximation
***learning_rate: 5**e-3.*	Learning rate of 0.005 was used to train the policy network

A variety of hyper parameters were tried out to achieve faster convergence and only the above mentioned parameters were tweaked. All the other hyper parameters were kept at their default values.
I tried training the agent with the following sets of hyper parameters:

- **(Learning rate: 1e-3, batch_size: 5000)** : Convergence was observed after ~300 epochs of training and the training was very unstable
- **(Learning rate: 1e-3, batch_size: 10000)** : Convergence was observed after ~250 epochs of training.
- **(Learning rate: 3e-4, batch_size: 10000)** : Convergence was observed after ~450 epochs of training and the training was much stable compared to when the batch size was 5k as more number of steps were used in this case for estimating gradient
- **(Learning rate: 5e-4, batch_size: 10000)** : Convergence was observed after ~250 epochs of training

The training logs can be seen in the following image:

```
Training Iteration 180 Training Reward: 110.66 Evaluation Reward: 262.35         Average Evaluation Reward: 261.55
Training Iteration 181 Training Reward: 68.91 Evaluation Reward: 261.34         Average Evaluation Reward: 261.01
Training Iteration 182 Training Reward: 83.95 Evaluation Reward: 266.15         Average Evaluation Reward: 260.86
Training Iteration 183 Training Reward: 70.42 Evaluation Reward: 237.43         Average Evaluation Reward: 257.88
Training Iteration 184 Training Reward: 166.36 Evaluation Reward: 266.60         Average Evaluation Reward: 257.95
Training Iteration 185 Training Reward: 229.36 Evaluation Reward: 263.30         Average Evaluation Reward: 261.21
Training Iteration 186 Training Reward: 203.96 Evaluation Reward: 263.43         Average Evaluation Reward: 261.05
Training Iteration 187 Training Reward: 223.09 Evaluation Reward: 264.12         Average Evaluation Reward: 261.09
Training Iteration 188 Training Reward: 197.57 Evaluation Reward: 264.94         Average Evaluation Reward: 261.29
Training Iteration 189 Training Reward: 241.48 Evaluation Reward: 264.57         Average Evaluation Reward: 261.42
Training Iteration 190 Training Reward: 231.16 Evaluation Reward: 264.22         Average Evaluation Reward: 261.61
Training Iteration 191 Training Reward: 244.30 Evaluation Reward: 224.33         Average Evaluation Reward: 257.91
Training Iteration 192 Training Reward: 238.91 Evaluation Reward: 262.86         Average Evaluation Reward: 257.58
Training Iteration 193 Training Reward: 224.24 Evaluation Reward: 262.85         Average Evaluation Reward: 260.12
Training Iteration 194 Training Reward: 242.93 Evaluation Reward: 263.40         Average Evaluation Reward: 259.80
Training Iteration 195 Training Reward: 235.09 Evaluation Reward: 263.82         Average Evaluation Reward: 259.85
Training Iteration 196 Training Reward: 247.94 Evaluation Reward: 267.34         Average Evaluation Reward: 260.25
Training Iteration 197 Training Reward: 245.04 Evaluation Reward: 229.58         Average Evaluation Reward: 256.79
Training Iteration 198 Training Reward: 243.27 Evaluation Reward: 270.31         Average Evaluation Reward: 257.33
Training Iteration 199 Training Reward: 242.78 Evaluation Reward: 272.23         Average Evaluation Reward: 258.09
/opt/anaconda3/envs/dl/lib/python3.9/site-packages/gymnasium/wrappers/rendering.py:283: UserWarning: WARN: Overwriting existi
ng videos at /Users/rajakarthikvobugari/Documents/RL_ECEN743/ECEN743-SP25-PG/videos/Baseline_LunarLander-v3_iter200 folder (t
ry specifying a different `video_folder` for the `RecordVideo` wrapper if this is not desired)
  logger.warn(
[Video] Iteration 200: Episode reward = 272.23
```

The cumulative undiscounted reward for the episode in the submitted video of performance is 272.23 which can be seen in the final line of the logs of the above image.
A video of performance has been included In the submitted file for this question along with the reward plot images for all algorithms

2.4

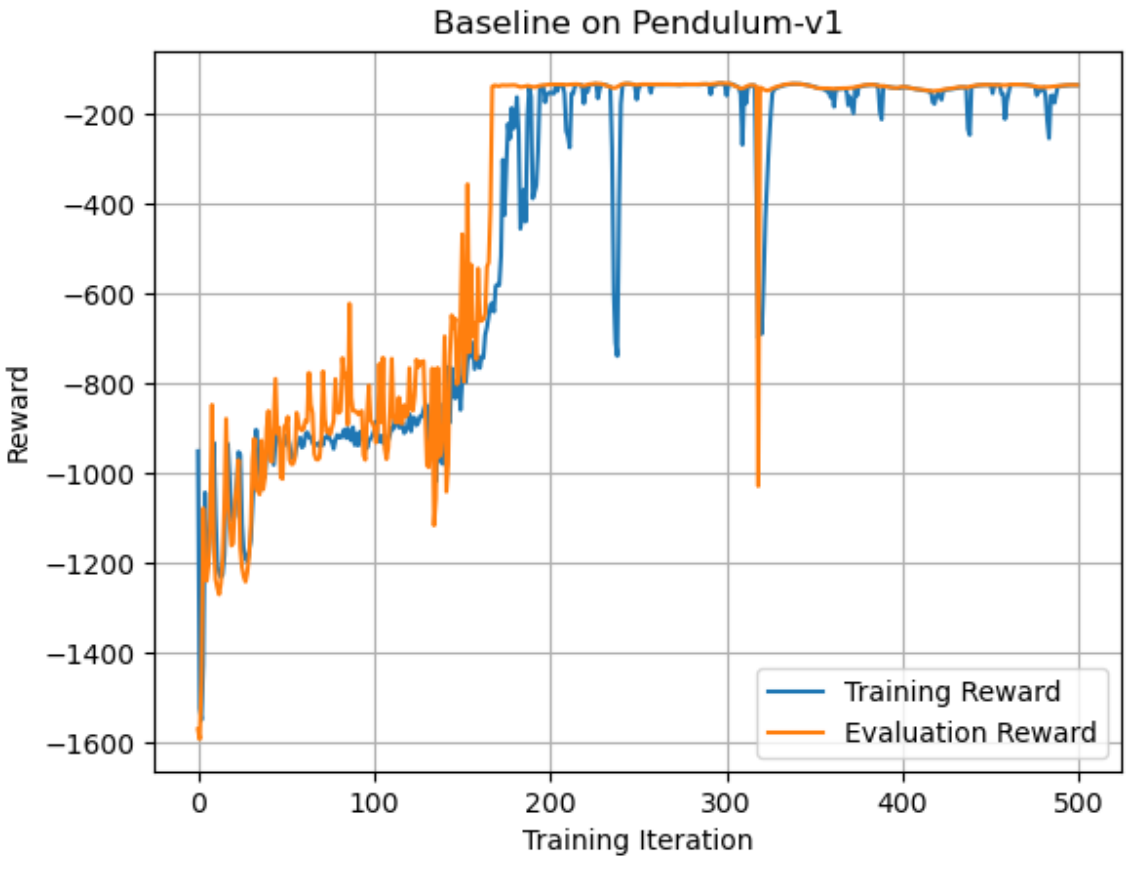
For this question, I have trained an agent to play the Pendulum game which belongs to the classic control set. The link to the environment is given below:

https://gymnasium.farama.org/environments/classic_control/pendulum/

The following are the hyper parameters used for training the agent:

Env: *"Pendulum-v1"*	*environment name*
N-iter*: 200.*	*200* iterations
batch_size: *10000.*	10000 steps of play included for every iteration of training to perform gradient approximation
learning_rate: 5*e-3.*	Learning rate of 0.005 was used to train the policy network
Discount*: 0.99.*	Discount factor of 0.99

The episodic undiscounted-cumulative reward plot is given below:



The highest reward achieved by the agent is ~ -133. The agent crosses the threshold after ~190 epochs of training. The training logs can be seen in the image below:

```
Training Iteration 473 Training Reward: -137.32 Evaluation Reward: -137.30         Average Evaluation Reward: -135.53
Training Iteration 474 Training Reward: -137.93 Evaluation Reward: -138.10         Average Evaluation Reward: -135.88
Training Iteration 475 Training Reward: -138.84 Evaluation Reward: -139.20         Average Evaluation Reward: -136.33
Training Iteration 476 Training Reward: -158.97 Evaluation Reward: -139.90         Average Evaluation Reward: -136.85
Training Iteration 477 Training Reward: -148.77 Evaluation Reward: -140.30         Average Evaluation Reward: -137.40
Training Iteration 478 Training Reward: -149.22 Evaluation Reward: -140.68         Average Evaluation Reward: -137.96
Training Iteration 479 Training Reward: -144.05 Evaluation Reward: -140.76         Average Evaluation Reward: -138.50
Training Iteration 480 Training Reward: -141.44 Evaluation Reward: -140.72         Average Evaluation Reward: -138.99
Training Iteration 481 Training Reward: -159.61 Evaluation Reward: -140.72         Average Evaluation Reward: -139.44
Training Iteration 482 Training Reward: -215.34 Evaluation Reward: -140.63         Average Evaluation Reward: -139.83
Training Iteration 483 Training Reward: -254.33 Evaluation Reward: -140.13         Average Evaluation Reward: -140.11
Training Iteration 484 Training Reward: -160.00 Evaluation Reward: -139.68         Average Evaluation Reward: -140.27
Training Iteration 485 Training Reward: -159.12 Evaluation Reward: -139.21         Average Evaluation Reward: -140.27
Training Iteration 486 Training Reward: -175.59 Evaluation Reward: -138.68         Average Evaluation Reward: -140.15
Training Iteration 487 Training Reward: -157.06 Evaluation Reward: -138.19         Average Evaluation Reward: -139.94
Training Iteration 488 Training Reward: -138.73 Evaluation Reward: -137.78         Average Evaluation Reward: -139.65
Training Iteration 489 Training Reward: -138.32 Evaluation Reward: -137.42         Average Evaluation Reward: -139.32
Training Iteration 490 Training Reward: -137.90 Evaluation Reward: -136.94         Average Evaluation Reward: -138.94
Training Iteration 491 Training Reward: -137.51 Evaluation Reward: -136.54         Average Evaluation Reward: -138.52
Training Iteration 492 Training Reward: -137.08 Evaluation Reward: -136.30         Average Evaluation Reward: -138.09
Training Iteration 493 Training Reward: -136.87 Evaluation Reward: -136.02         Average Evaluation Reward: -137.67
Training Iteration 494 Training Reward: -136.49 Evaluation Reward: -135.90         Average Evaluation Reward: -137.30
Training Iteration 495 Training Reward: -136.38 Evaluation Reward: -135.86         Average Evaluation Reward: -136.96
Training Iteration 496 Training Reward: -136.31 Evaluation Reward: -135.83         Average Evaluation Reward: -136.68
Training Iteration 497 Training Reward: -136.26 Evaluation Reward: -135.84         Average Evaluation Reward: -136.44
Training Iteration 498 Training Reward: -136.28 Evaluation Reward: -135.93         Average Evaluation Reward: -136.26
Training Iteration 499 Training Reward: -136.34 Evaluation Reward: -136.12         Average Evaluation Reward: -136.13
[Video] Iteration 500: Episode reward = -136.12
(dl) rajakarthikvobugari@Mac ECEN743-SP25-PG %
```

A video of performance has also been included in the submitted file. The reward for the episode in the submitted video is -136.12 which can be seen in the final line of the logs displayed above.