A5_Deshmane_Aakash

April 27, 2023

1 AAKASH DESHMANE

UIN : 133008022

1.0.1 ECEN 743 ASSIGNMENT 5

1.0.2 POLICY GRADIENT IMPLEMENTATION OF LUNAR LANDER

[ ]:
```
!pip install gymnasium[box2d]
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Collecting gymnasium[box2d]
  Downloading gymnasium-0.28.1-py3-none-any.whl (925 kB)
                           925.5/925.5 kB
12.4 MB/s eta 0:00:00
Collecting farama-notifications>=0.0.1
  Downloading Farama_Notifications-0.0.4-py3-none-any.whl (2.5 kB)
Requirement already satisfied: typing-extensions>=4.3.0 in
/usr/local/lib/python3.10/dist-packages (from gymnasium[box2d]) (4.5.0)
Collecting jax-jumpy>=1.0.0
  Downloading jax_jumpy-1.0.0-py3-none-any.whl (20 kB)
Requirement already satisfied: numpy>=1.21.0 in /usr/local/lib/python3.10/dist-
packages (from gymnasium[box2d]) (1.22.4)
Requirement already satisfied: cloudpickle>=1.2.0 in
/usr/local/lib/python3.10/dist-packages (from gymnasium[box2d]) (2.2.1)
Collecting swig==4.*
  Downloading swig-4.1.1-py2.py3-none-manylinux_2_5_x86_64.manylinux1_x86_64.whl
(1.8 MB)
                            1.8/1.8 MB
40.8 MB/s eta 0:00:00
Collecting box2d-py==2.3.5
  Downloading box2d-py-2.3.5.tar.gz (374 kB)
                           374.4/374.4 kB
25.8 MB/s eta 0:00:00
  Preparing metadata (setup.py) … done
Collecting pygame==2.1.3
  Downloading
pygame-2.1.3-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (13.7
```

```
MB)
                                        13.7/13.7 MB
42.4 MB/s eta 0:00:00
Building wheels for collected packages: box2d-py
  error: subprocess-exited-with-error

  × python setup.py bdist_wheel did not run successfully.
    exit code: 1
   > See above for output.

  note: This error originates from a subprocess, and is likely not a
problem with pip.
  Building wheel for box2d-py (setup.py) … error
  ERROR: Failed building wheel for box2d-py

  Running setup.py clean for box2d-py
Failed to build box2d-py
Installing collected packages: swig, farama-notifications, box2d-py, pygame,
jax-jumpy, gymnasium
  Running setup.py install for box2d-py … done
  DEPRECATION: box2d-py was installed using the legacy 'setup.py install'
method, because a wheel could not be built for it. pip 23.1 will enforce this
behaviour change. A possible replacement is to fix the wheel build issue
reported above. Discussion can be found at
https://github.com/pypa/pip/issues/8368

  Attempting uninstall: pygame
    Found existing installation: pygame 2.3.0
    Uninstalling pygame-2.3.0:
      Successfully uninstalled pygame-2.3.0
Successfully installed box2d-py-2.3.5 farama-notifications-0.0.4
gymnasium-0.28.1 jax-jumpy-1.0.0 pygame-2.1.3 swig-4.1.1
```

[ ]:
```
"""
ECEN 743: Reinforcement Learning
Policy Gradient Assignment
Code tested using
        1. gymnasium 0.27.1
        2. box2d-py  2.3.5
        3. pytorch   2.0.0
        4. Python    3.9.12
1 & 2 can be installed using pip install gymnasium[box2d]

General Instructions
1. This code consists of TODO blocks, read them carefully and complete each of␣
 ↪the blocks
```

```
2. Type your code between the following lines
                        ###### TYPE YOUR CODE HERE ######
                        ##################################
3. The default hyperparameters should be able to solve LunarLander-v2 in the␣
 ↪continuous setting
4. It is not necessary to modify the rest of the code for this assignment, feel␣
 ↪free to do so if needed.

"""
import gymnasium as gym
import random
import torch
import torch.nn as nn
import torch.nn.functional as F
import torch.optim as optim
import argparse
import numpy as np
import math
from collections import deque
import matplotlib.pyplot as plt
```

VALUE NETWORK

[ ]:
```
class value_network(nn.Module):
        '''
        Value Network: Designed to take in state as input and give value as␣
 ↪output
        Used as a baseline in Policy Gradient (PG) algorithms
        '''
        def __init__(self,state_dim):
                '''
                        state_dim (int): state dimenssion
                '''
                super(value_network, self).__init__()
                self.l1 = nn.Linear(state_dim, 64)
                self.l2 = nn.Linear(64, 64)
                self.l3 = nn.Linear(64, 1)

        def forward(self,state):
                '''
                Input: State
                Output: Value of state
                '''
                v = F.tanh(self.l1(state))
                v = F.tanh(self.l2(v))
                return self.l3(v)
```

POLICY NETWORK

```
[ ]: class policy_network(nn.Module):
        '''
        Policy Network: Designed for continous action space, where given a
        state, the network outputs the mean and standard deviation of the action
        '''
        def __init__(self,state_dim,action_dim,log_std = 0.0):
                """
                        state_dim (int): state dimenssion
                        action_dim (int): action dimenssion
                        log_std (float): log of standard deviation (std)
                """
                super(policy_network, self).__init__()
                self.state_dim = state_dim
                self.action_dim = action_dim
                self.l1 = nn.Linear(state_dim,64)
                self.l2 = nn.Linear(64,64)
                self.mean = nn.Linear(64,action_dim)
                self.log_std = nn.Parameter(torch.ones(1, action_dim) * log_std)


        def forward(self,state):
                '''
                Input: State
                Output: Mean, log_std and std of action
                '''
                a = F.tanh(self.l1(state))
                a = F.tanh(self.l2(a))
                a_mean = self.mean(a)
                a_log_std = self.log_std.expand_as(a_mean)
                a_std = torch.exp(a_log_std)
                return a_mean, a_log_std, a_std

        def select_action(self, state):
                '''
                Input: State
                Output: Sample drawn from a normal disribution with mean and std
                '''
                a_mean, _, a_std = self.forward(state)
                action = torch.normal(a_mean, a_std)
                return action

        def get_log_prob(self, state, action):
                '''
                Input: State, Action
                Output: log probabilities
                '''
                mean, log_std, std = self.forward(state)
```

```
                var = std.pow(2)
                log_density = -(action - mean).pow(2) / (2 * var) - 0.5 * math.
↪log(2 * math.pi) - log_std
                return log_density.sum(1, keepdim=True)
```

POLICY GRADIENT AGENT

[ ]:
```
class PGAgent():
        '''
        An agent that performs different variants of the PG algorithm
        '''
        def __init__(self,
         state_dim,
         action_dim,
         discount=0.99,
         lr=1e-3,
         gpu_index=0,
         seed=0,
         env="LunarLander-v2"
         ):
                """
                        state_size (int): dimension of each state
                        action_size (int): dimension of each action
                        discount (float): discount factor
                        lr (float): learning rate
                        gpu_index (int): GPU used for training
                        seed (int): Seed of simulation
                        env (str): Name of environment
                """
                self.state_dim = state_dim
                self.action_dim = action_dim
                self.discount = discount
                self.lr = lr
                self.device = torch.device('cuda', index=gpu_index) if torch.
↪cuda.is_available() else torch.device('cpu')
                self.env_name = env
                self.seed = seed
                self.policy = policy_network(state_dim,action_dim)
                self.value = value_network(state_dim)
                self.optimizer_policy = torch.optim.Adam(self.policy.
↪parameters(), lr=self.lr)
                self.optimizer_value = torch.optim.Adam(self.value.
↪parameters(), lr=self.lr)

        def sample_traj(self,batch_size=2000,evaluate = False):
                '''
                Input:
```

```
                        batch_size: minimum batch size needed for update
                        evaluate: flag to be set during evaluation
                Output:
                        states, actions, rewards,not_dones, episodic reward␣
↪
                '''
                self.policy.to("cpu") #Move network to CPU for sampling
                env = gym.make(self.env_name,continuous=True)
                states = []
                actions = []
                rewards = []
                n_dones = []
                curr_reward_list = []
                while len(states) < batch_size:
                        state, _ = env.reset(seed=self.seed)
                        curr_reward = 0
                        for t in range(1000):
                                state_ten = torch.from_numpy(state).float().
↪unsqueeze(0)
                                with torch.no_grad():
                                        if evaluate:
                                                action = self.
↪policy(state_ten)[0][0].numpy() # Take mean action during evaluation
                                        else:
                                                action = self.policy.
↪select_action(state_ten)[0].numpy() # Sample from distribution during␣
↪training
                                action = action.astype(np.float64)
                                n_state,reward,terminated,truncated,_ = env.
↪step(action) # Execute action in the environment
                                done = terminated or truncated
                                states.append(state)
                                actions.append(action)
                                rewards.append(reward)
                                n_done = 0 if done else 1
                                n_dones.append(n_done)
                                state = n_state
                                curr_reward += reward
                                if done:
                                        break
                        curr_reward_list.append(curr_reward)
                if evaluate:
                        return np.mean(curr_reward_list)
                return states,actions,rewards,n_dones, np.mean(curr_reward_list)
```

```
	def update(self,states,actions,rewards,n_dones,update_type='Baseline'):
		'''
		TODO: Complete this block to update the policy using different variants of PG
		Inputs:
			states: list of states
			actions: list of actions
			rewards: list of rewards
			n_dones: list of not dones
			update_type: type of PG algorithm
		Output:
			None
		'''
		self.policy.to(self.device) #Move policy to GPU
		if update_type == "Baseline":
			self.value.to(self.device)		#Move value to GPU
		states_ten = torch.from_numpy(np.stack(states)).to(self.device)  #Convert to tensor and move to GPU
		action_ten = torch.from_numpy(np.stack(actions)).to(self.device)  #Convert to tensor and move to GPU
		rewards_ten = torch.from_numpy(np.stack(rewards)).to(self.device) #Convert to tensor and move to GPU
		n_dones_ten = torch.from_numpy(np.stack(n_dones)).to(self.device) #Convert to tensor and move to GPU


		if update_type == "Rt":

			rt = torch.zeros(rewards_ten.shape[0],1).to(self.device)
			rt_total = 0
			s = rewards_ten.shape[0]

			# CALCULATE REWARD
			for t in reversed(range(s)):
				rt_total = rewards_ten[t] + rt_total * self.discount *n_dones_ten[t]
				rt[t] = rt_total

			# CALCULATE LOG PROBABILITIES
			log_prob = self.policy.get_log_prob(states_ten, action_ten)
			l = log_prob * rt.detach()
			loss = -(l).mean()

			# UPDATE POLICY
```

```
                        self.optimizer_policy.zero_grad()
                        loss.backward()
                        self.optimizer_policy.step()

                if update_type == 'Gt':
                        '''
                        TODO: Peform PG using reward_to_go
                        1. Compute reward_to_go (gt) using rewards_ten and␣
↪n_dones_ten
                        2. gt should be of the same length as rewards_ten
                        3. Compute log probabilities using states_ten and␣
↪action_ten
                        4. Compute policy loss and update the policy
                        '''


                        ###### TYPE YOUR CODE HERE ######
                        ##################################

                        # STEP 1 : COMPUTE REWARD
                        g = 0

                        # STEP 2 : gt SHOULD BE OF THE SAME LENGTH AS␣
↪rewards_ten
                        gt = torch.zeros(rewards_ten.shape[0],1).to(self.device)

                        for i in reversed(range(rewards_ten.size(0))):
                                g = rewards_ten[i] + self.discount * g *␣
↪(n_dones_ten[i])
                                gt[i] = g

                        gt = (gt - gt.mean()) / gt.std() #Helps with learning␣
↪stablity

                        # STEP 3 : COMPUTE LOG PROBABILITIES USING states_ten␣
↪AND action_ten
                        log_prob = self.policy.get_log_prob(states_ten,␣
↪action_ten)
                        l = log_prob * gt.detach()
                        loss = -(l).mean()

                         # STEP 4 : COMPUTE POLICY LOSS AND UPDATE POLICY

                        self.optimizer_policy.zero_grad()
                        loss.backward()
                        self.optimizer_policy.step()
```

```
            if update_type == 'Gt_with_Baseline':
                    '''
                    TODO: Peform PG using reward_to_go and baseline
                    1. Compute values of states, this will be used as the␣
↪baseline
                    2. Compute reward_to_go (gt) using rewards_ten and␣
↪n_dones_ten
                    3. gt should be of the same length as rewards_ten
                    4. Compute advantages
                    5. Update the value network to predict gt for each␣
↪state (L2 norm)
                    6. Compute log probabilities using states_ten and␣
↪action_ten
                    7. Compute policy loss (using advantages) and update␣
↪the policy
                    '''
                    state_t = torch.FloatTensor(states).to(self.device)

                    # STEP 1 CALCULATE VALUES
                    with torch.no_grad():
                                    self.value.to(self.device)
                                    val = self.value(states_ten).to(self.
↪device)

                    # gt SHOULD HAVE THE SAME LENGTH AS rewards_ten
                    gt = torch.zeros(rewards_ten.shape[0],1).to(self.device)

                    g=0

                    # STEP 2 : COMPUTE REWARD-TO-GO (gt) and ADVANTAGES
                    returns = torch.zeros((rewards_ten.shape[0], 1)).
↪to(self.device)
                    advantages = torch.zeros((rewards_ten.shape[0], 1)).
↪to(self.device)
                    s = rewards_ten.size(0)
                    for i in reversed(range(s)):
                            g = rewards_ten[i] + self.discount * g *␣
↪n_dones_ten[i]
                            gt[i] = g

                    # STEP 4 : COMPUTE ADVANTAGES
                    advantages = gt - val


                    # Normalize advantages
```

```
                        advantages = (advantages - advantages.mean()) /␣
↪advantages.std()

                        # STEP 5 : UPDATE VALUE NETWORK TO PREDICT gt FOR EACH␣
↪STATE (L2 NORM)
                        loss = torch.nn.MSELoss()
                        value_loss = loss(self.value(states_ten), gt)
                        self.optimizer_value.zero_grad()
                        value_loss.backward()
                        self.optimizer_value.step()

                        # STEP 6 : COMPUTE LOG PROBABILITIES USING states_ten␣
↪and Compute log probabilities using states_ten and action_ten
                        log_probs = self.policy.get_log_prob(states_ten,␣
↪action_ten)

                        # STEP 7 : COMPUTE POLICY LOSS AND UPDATE POLICY
                        self.optimizer_policy.zero_grad()
                        l = log_probs * advantages.detach()
                        loss = -(l).mean()
                        loss.backward()
                        self.optimizer_policy.step()
            #␣
↪-------------------------------------------------------------------------
```

MAIN FUNCTION FOR THE CODE

[ ]:
```
def main_fn(algo):
        env_type = "LunarLander-v2"                                    #␣
↪Gymnasium environment name
        seed = 0                 ␣
↪                                                       # Sets Gym, PyTorch␣
↪and Numpy seeds
        n_iter = 200           ␣
↪                                                       # Maximum number of␣
↪training iterations
        discount = 0.99                                 # Discount factor
        batch_size = 5000      ␣
↪                                                       # Training samples␣
↪in each batch of training
        lr = 6e-3              ␣
↪                                                       # Learning rate
        gpu_index = 0                        ␣
↪                                                              # GPU index
        algo = algo                          ␣
↪                                                               # PG␣
↪algorithm type. Baseline_with_Gt/Gt/Rt
```

```
        # Making the environment
        env = gym.make(env_type,continuous=True)

        # Setting seeds
        torch.manual_seed(seed)
        np.random.seed(seed)
        random.seed(seed)

        state_dim = env.observation_space.shape[0]
        action_dim = env.action_space.shape[0]

        kwargs = {
                "state_dim":state_dim,
                "action_dim":action_dim,
                "discount":discount,
                "lr":lr,
                "gpu_index":gpu_index,
                "seed":seed,
                "env":env_type
        }
        learner = PGAgent(**kwargs) # Creating the PG learning agent
        av_rewards=[]
        moving_window = deque(maxlen=10)
        old_reward = 0
        old_eval_reward = 0
        old_train_reward = 0

        for e in range(n_iter):
                '''
                Steps of PG algorithm
                        1. Sample environment to gather data using a policy
                        2. Update the policy using the data
                        3. Evaluate the updated policy
                        4. Repeat 1-3
                '''
                states,actions,rewards,n_dones,train_reward = learner.
↪sample_traj(batch_size=batch_size)
                learner.update(states,actions,rewards,n_dones,algo)
                eval_reward= learner.sample_traj(evaluate=True)
                moving_window.append(eval_reward)
                print('Training Iteration {} Training Reward: {:.2f} Evaluation␣
↪Reward: {:.2f} \
                Average Evaluation Reward: {:.2f}'.
↪format(e,train_reward,eval_reward,np.mean(moving_window)))
```

```
                av_rewards.append(np.mean(moving_window))

                if eval_reward > old_eval_reward and train_reward >␣
↪old_train_reward and np.mean(moving_window)>old_reward and eval_reward > 210␣
↪and train_reward > 210 and np.mean(moving_window) > 210:
                        old_reward = np.mean(moving_window)
                        old_eval_reward = eval_reward
                        old_train_reward = train_reward
                        torch.save(learner.policy.state_dict(), (algo +␣
↪'_checkpoint.pth'))
                        print("Best result for training iteration {}".format(e))

        """
        TODO: Write code for
        1. Logging and plotting
        2. Rendering the trained agent
        """
        ###### TYPE YOUR CODE HERE ######
        ################################
        window_size = 20
        averages = []

        for i in range(len(av_rewards)-window_size + 1):
                        window = av_rewards[i:i+window_size]
                        average = sum(window) / window_size
                        averages.append(average)

        plt.plot(averages, color='g')
        plt.ylabel('Episodic Cumulative Reward')
        plt.xlabel('Episode #')
        plt.title('Curve for Episodic Cumulative Reward for algorithm = {}'.
↪format(algo))
        plt.show()
```

HYPERPARAMETERS TUNED TO: * ITERATIONS : 200 * DISCOUNT : 0.99 * BATCH SIZE : 5000 * LEARNING RATE : 6e-4

1) REINFORCE:

[ ]:
```
env_type = "LunarLander-v2"

# RUNNING CODE FOR 3 CONDITION :

# 1) REINFORCE ALGORITHM
print("1) REINFORCE ALGORITHM \n TRAINING FOR Rt: \n")
main_fn("Rt")
```

```
#_____________________________________________________________________________________________
#_____________________________________________________________________________________________
#_____________________________________________________________________________________________
```

```
1) REINFORCE ALGORITHM
 TRAINING FOR Rt:

Training Iteration 0 Training Reward: -80.02 Evaluation Reward: -232.22
Average Evaluation Reward: -232.22
Training Iteration 1 Training Reward: -235.48 Evaluation Reward: -88.72
Average Evaluation Reward: -160.47
Training Iteration 2 Training Reward: -143.70 Evaluation Reward: -102.33
Average Evaluation Reward: -141.09
Training Iteration 3 Training Reward: -71.00 Evaluation Reward: -393.46
Average Evaluation Reward: -204.18
Training Iteration 4 Training Reward: -94.70 Evaluation Reward: -151.58
Average Evaluation Reward: -193.66
Training Iteration 5 Training Reward: -128.36 Evaluation Reward: -88.98
Average Evaluation Reward: -176.21
Training Iteration 6 Training Reward: -80.86 Evaluation Reward: -46.37
Average Evaluation Reward: -157.66
Training Iteration 7 Training Reward: -52.87 Evaluation Reward: -36.45
Average Evaluation Reward: -142.51
Training Iteration 8 Training Reward: -43.91 Evaluation Reward: -64.98
Average Evaluation Reward: -133.90
Training Iteration 9 Training Reward: -81.69 Evaluation Reward: -269.85
Average Evaluation Reward: -147.49
Training Iteration 10 Training Reward: -94.08 Evaluation Reward: -224.97
Average Evaluation Reward: -146.77
Training Iteration 11 Training Reward: -85.30 Evaluation Reward: -190.05
Average Evaluation Reward: -156.90
Training Iteration 12 Training Reward: -60.17 Evaluation Reward: -158.32
Average Evaluation Reward: -162.50
Training Iteration 13 Training Reward: -15.68 Evaluation Reward: -136.03
Average Evaluation Reward: -136.76
Training Iteration 14 Training Reward: -0.08 Evaluation Reward: -123.18
Average Evaluation Reward: -133.92
Training Iteration 15 Training Reward: 14.35 Evaluation Reward: -134.15
Average Evaluation Reward: -138.44
Training Iteration 16 Training Reward: -0.30 Evaluation Reward: -121.52
Average Evaluation Reward: -145.95
Training Iteration 17 Training Reward: -15.97 Evaluation Reward: -164.99
Average Evaluation Reward: -158.81
Training Iteration 18 Training Reward: 4.91 Evaluation Reward: -154.70
Average Evaluation Reward: -167.78
Training Iteration 19 Training Reward: -1.98 Evaluation Reward: -202.33
Average Evaluation Reward: -161.02
```

```
Training Iteration 20 Training Reward: -45.22 Evaluation Reward: -154.63
Average Evaluation Reward: -153.99
Training Iteration 21 Training Reward: -43.27 Evaluation Reward: -159.74
Average Evaluation Reward: -150.96
Training Iteration 22 Training Reward: -64.52 Evaluation Reward: -199.49
Average Evaluation Reward: -155.08
Training Iteration 23 Training Reward: -80.04 Evaluation Reward: -137.02
Average Evaluation Reward: -155.18
Training Iteration 24 Training Reward: -54.08 Evaluation Reward: -144.99
Average Evaluation Reward: -157.36
Training Iteration 25 Training Reward: -49.66 Evaluation Reward: -48.26
Average Evaluation Reward: -148.77
Training Iteration 26 Training Reward: 4.57 Evaluation Reward: -44.67
Average Evaluation Reward: -141.08
Training Iteration 27 Training Reward: 11.79 Evaluation Reward: -60.95
Average Evaluation Reward: -130.68
Training Iteration 28 Training Reward: -21.37 Evaluation Reward: -96.22
Average Evaluation Reward: -124.83
Training Iteration 29 Training Reward: 25.44 Evaluation Reward: -209.37
Average Evaluation Reward: -125.53
Training Iteration 30 Training Reward: -28.38 Evaluation Reward: -148.79
Average Evaluation Reward: -124.95
Training Iteration 31 Training Reward: -93.89 Evaluation Reward: -157.34
Average Evaluation Reward: -124.71
Training Iteration 32 Training Reward: -39.25 Evaluation Reward: -160.96
Average Evaluation Reward: -120.86
Training Iteration 33 Training Reward: -73.10 Evaluation Reward: -161.22
Average Evaluation Reward: -123.28
Training Iteration 34 Training Reward: -90.95 Evaluation Reward: -165.27
Average Evaluation Reward: -125.30
Training Iteration 35 Training Reward: -4.38 Evaluation Reward: -164.36
Average Evaluation Reward: -136.92
Training Iteration 36 Training Reward: -66.20 Evaluation Reward: -181.01
Average Evaluation Reward: -150.55
Training Iteration 37 Training Reward: -38.24 Evaluation Reward: -194.75
Average Evaluation Reward: -163.93
Training Iteration 38 Training Reward: -70.63 Evaluation Reward: -221.73
Average Evaluation Reward: -176.48
Training Iteration 39 Training Reward: -18.02 Evaluation Reward: -250.06
Average Evaluation Reward: -180.55
Training Iteration 40 Training Reward: -11.30 Evaluation Reward: -239.05
Average Evaluation Reward: -189.58
Training Iteration 41 Training Reward: -0.98 Evaluation Reward: -241.27
Average Evaluation Reward: -197.97
Training Iteration 42 Training Reward: 17.11 Evaluation Reward: -227.81
Average Evaluation Reward: -204.65
Training Iteration 43 Training Reward: 13.45 Evaluation Reward: -219.08
Average Evaluation Reward: -210.44
```

```
Training Iteration 44 Training Reward: 21.80 Evaluation Reward: -170.83
Average Evaluation Reward: -211.00
Training Iteration 45 Training Reward: 24.79 Evaluation Reward: -240.88
Average Evaluation Reward: -218.65
Training Iteration 46 Training Reward: 26.15 Evaluation Reward: 15.65
Average Evaluation Reward: -198.98
Training Iteration 47 Training Reward: 26.24 Evaluation Reward: 20.03
Average Evaluation Reward: -177.50
Training Iteration 48 Training Reward: 41.03 Evaluation Reward: 262.40
Average Evaluation Reward: -129.09
Training Iteration 49 Training Reward: 41.05 Evaluation Reward: 271.05
Average Evaluation Reward: -76.98
Training Iteration 50 Training Reward: 53.04 Evaluation Reward: 249.24
Average Evaluation Reward: -28.15
Training Iteration 51 Training Reward: 57.61 Evaluation Reward: 288.34
Average Evaluation Reward: 24.81
Training Iteration 52 Training Reward: 66.42 Evaluation Reward: 296.66
Average Evaluation Reward: 77.26
Training Iteration 53 Training Reward: 68.99 Evaluation Reward: 300.54
Average Evaluation Reward: 129.22
Training Iteration 54 Training Reward: 70.14 Evaluation Reward: 299.00
Average Evaluation Reward: 176.20
Training Iteration 55 Training Reward: 81.85 Evaluation Reward: 280.98
Average Evaluation Reward: 228.39
Training Iteration 56 Training Reward: 108.71 Evaluation Reward: 279.46
Average Evaluation Reward: 254.77
Training Iteration 57 Training Reward: 127.53 Evaluation Reward: 217.31
Average Evaluation Reward: 274.50
Training Iteration 58 Training Reward: 117.23 Evaluation Reward: -265.29
Average Evaluation Reward: 221.73
Training Iteration 59 Training Reward: 30.01 Evaluation Reward: -220.06
Average Evaluation Reward: 172.62
Training Iteration 60 Training Reward: 86.86 Evaluation Reward: -175.05
Average Evaluation Reward: 130.19
Training Iteration 61 Training Reward: 83.24 Evaluation Reward: -203.66
Average Evaluation Reward: 80.99
Training Iteration 62 Training Reward: 78.52 Evaluation Reward: -205.27
Average Evaluation Reward: 30.80
Training Iteration 63 Training Reward: 49.78 Evaluation Reward: -220.95
Average Evaluation Reward: -21.35
Training Iteration 64 Training Reward: 117.14 Evaluation Reward: -223.91
Average Evaluation Reward: -73.64
Training Iteration 65 Training Reward: 55.12 Evaluation Reward: -177.87
Average Evaluation Reward: -119.53
Training Iteration 66 Training Reward: 135.24 Evaluation Reward: -225.84
Average Evaluation Reward: -170.06
Training Iteration 67 Training Reward: 120.30 Evaluation Reward: 220.68
Average Evaluation Reward: -169.72
```

```
Training Iteration 68 Training Reward: 158.47 Evaluation Reward: 207.75
Average Evaluation Reward: -122.42
Training Iteration 69 Training Reward: 133.58 Evaluation Reward: 263.47
Average Evaluation Reward: -74.07
Training Iteration 70 Training Reward: 158.63 Evaluation Reward: 252.74
Average Evaluation Reward: -31.29
Training Iteration 71 Training Reward: 130.82 Evaluation Reward: 250.27
Average Evaluation Reward: 14.11
Training Iteration 72 Training Reward: 119.39 Evaluation Reward: 253.07
Average Evaluation Reward: 59.94
Training Iteration 73 Training Reward: 123.99 Evaluation Reward: 241.93
Average Evaluation Reward: 106.23
Training Iteration 74 Training Reward: 154.87 Evaluation Reward: 219.95
Average Evaluation Reward: 150.62
Training Iteration 75 Training Reward: 148.77 Evaluation Reward: 254.51
Average Evaluation Reward: 193.85
Training Iteration 76 Training Reward: 154.21 Evaluation Reward: 219.97
Average Evaluation Reward: 238.43
Training Iteration 77 Training Reward: 152.53 Evaluation Reward: 212.58
Average Evaluation Reward: 237.62
Training Iteration 78 Training Reward: 156.32 Evaluation Reward: 219.09
Average Evaluation Reward: 238.76
Training Iteration 79 Training Reward: 155.62 Evaluation Reward: 224.75
Average Evaluation Reward: 234.89
Training Iteration 80 Training Reward: 139.93 Evaluation Reward: 261.07
Average Evaluation Reward: 235.72
Training Iteration 81 Training Reward: 153.87 Evaluation Reward: 229.30
Average Evaluation Reward: 233.62
Training Iteration 82 Training Reward: 156.82 Evaluation Reward: 233.49
Average Evaluation Reward: 231.66
Training Iteration 83 Training Reward: 161.21 Evaluation Reward: 241.81
Average Evaluation Reward: 231.65
Training Iteration 84 Training Reward: 166.15 Evaluation Reward: 261.81
Average Evaluation Reward: 235.84
Training Iteration 85 Training Reward: 153.05 Evaluation Reward: 261.79
Average Evaluation Reward: 236.57
Training Iteration 86 Training Reward: 163.38 Evaluation Reward: 222.32
Average Evaluation Reward: 236.80
Training Iteration 87 Training Reward: 142.23 Evaluation Reward: 257.24
Average Evaluation Reward: 241.27
Training Iteration 88 Training Reward: 95.27 Evaluation Reward: 212.48
Average Evaluation Reward: 240.61
Training Iteration 89 Training Reward: 138.04 Evaluation Reward: 232.31
Average Evaluation Reward: 241.36
Training Iteration 90 Training Reward: 162.01 Evaluation Reward: 231.20
Average Evaluation Reward: 238.38
Training Iteration 91 Training Reward: 162.26 Evaluation Reward: 177.57
Average Evaluation Reward: 233.20
```

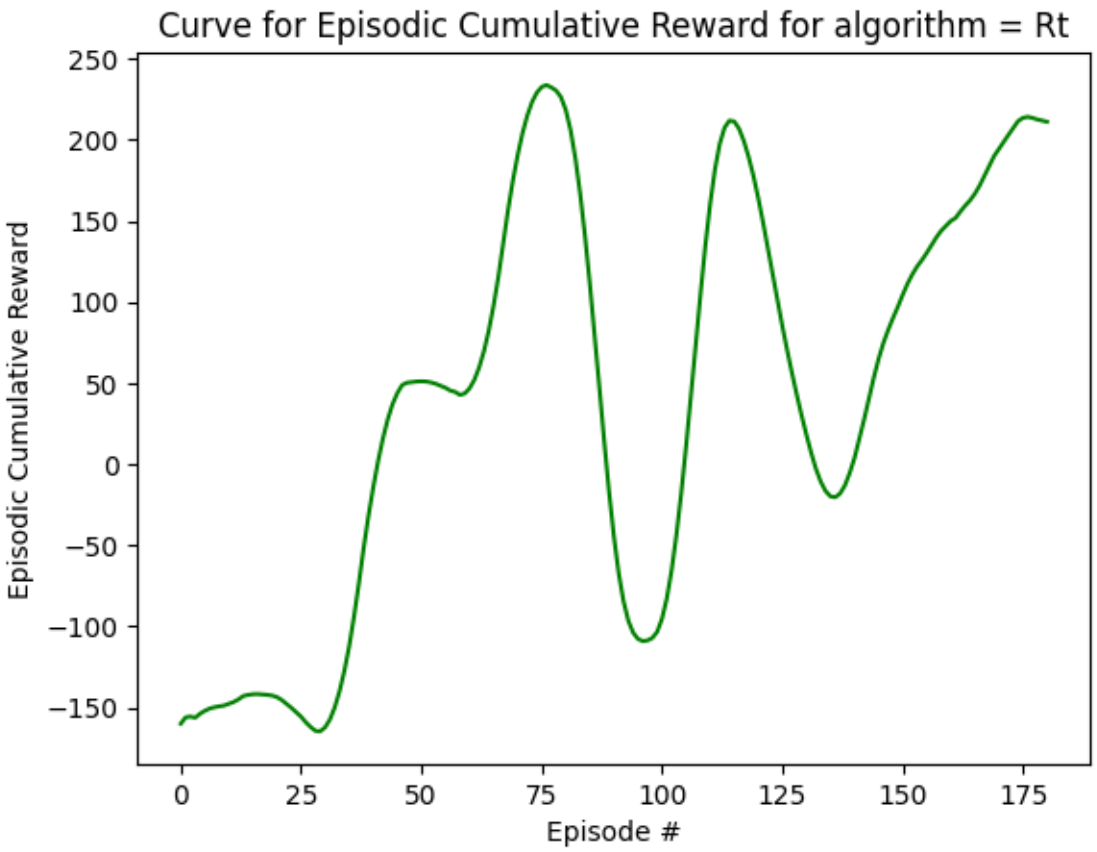
```
Training Iteration 92 Training Reward: 109.42 Evaluation Reward: 189.83
Average Evaluation Reward: 228.84
Training Iteration 93 Training Reward: 126.89 Evaluation Reward: 176.15
Average Evaluation Reward: 222.27
Training Iteration 94 Training Reward: 158.48 Evaluation Reward: 241.30
Average Evaluation Reward: 220.22
Training Iteration 95 Training Reward: 118.78 Evaluation Reward: 222.21
Average Evaluation Reward: 216.26
Training Iteration 96 Training Reward: 87.03 Evaluation Reward: 159.38
Average Evaluation Reward: 209.97
Training Iteration 97 Training Reward: 118.01 Evaluation Reward: 145.24
Average Evaluation Reward: 198.77
Training Iteration 98 Training Reward: 107.89 Evaluation Reward: -196.58
Average Evaluation Reward: 157.86
Training Iteration 99 Training Reward: 75.20 Evaluation Reward: -576.15
Average Evaluation Reward: 77.01
Training Iteration 100 Training Reward: -26.91 Evaluation Reward: -674.81
Average Evaluation Reward: -13.59
Training Iteration 101 Training Reward: -96.53 Evaluation Reward: -818.75
Average Evaluation Reward: -113.22
Training Iteration 102 Training Reward: -355.19 Evaluation Reward: -743.18
Average Evaluation Reward: -206.52
Training Iteration 103 Training Reward: -428.18 Evaluation Reward: -624.62
Average Evaluation Reward: -286.60
Training Iteration 104 Training Reward: -229.18 Evaluation Reward: -419.27
Average Evaluation Reward: -352.65
Training Iteration 105 Training Reward: -32.04 Evaluation Reward: -211.08
Average Evaluation Reward: -395.98
Training Iteration 106 Training Reward: -7.64 Evaluation Reward: -54.41
Average Evaluation Reward: -417.36
Training Iteration 107 Training Reward: 77.83 Evaluation Reward: 126.66
Average Evaluation Reward: -419.22
Training Iteration 108 Training Reward: 64.34 Evaluation Reward: 188.84
Average Evaluation Reward: -380.68
Training Iteration 109 Training Reward: 82.48 Evaluation Reward: 176.55
Average Evaluation Reward: -305.41
Training Iteration 110 Training Reward: 119.16 Evaluation Reward: 216.11
Average Evaluation Reward: -216.31
Training Iteration 111 Training Reward: 131.26 Evaluation Reward: 239.32
Average Evaluation Reward: -110.51
Training Iteration 112 Training Reward: 120.85 Evaluation Reward: 238.76
Average Evaluation Reward: -12.31
Training Iteration 113 Training Reward: 168.49 Evaluation Reward: 250.69
Average Evaluation Reward: 75.22
Training Iteration 114 Training Reward: 191.26 Evaluation Reward: 252.13
Average Evaluation Reward: 142.36
Training Iteration 115 Training Reward: 222.57 Evaluation Reward: 254.00
Average Evaluation Reward: 188.87
```

```
Training Iteration 116 Training Reward: 221.06 Evaluation Reward: 242.94
Average Evaluation Reward: 218.60
Best result for training iteration 116
Training Iteration 117 Training Reward: 177.67 Evaluation Reward: 237.05
Average Evaluation Reward: 229.64
Training Iteration 118 Training Reward: 201.88 Evaluation Reward: 225.30
Average Evaluation Reward: 233.29
Training Iteration 119 Training Reward: 183.80 Evaluation Reward: 232.42
Average Evaluation Reward: 238.87
Training Iteration 120 Training Reward: 222.66 Evaluation Reward: 199.19
Average Evaluation Reward: 237.18
Training Iteration 121 Training Reward: 229.04 Evaluation Reward: 201.67
Average Evaluation Reward: 233.41
Training Iteration 122 Training Reward: 229.66 Evaluation Reward: 208.80
Average Evaluation Reward: 230.42
Training Iteration 123 Training Reward: 204.37 Evaluation Reward: 209.53
Average Evaluation Reward: 226.30
Training Iteration 124 Training Reward: 219.77 Evaluation Reward: 207.45
Average Evaluation Reward: 221.84
Training Iteration 125 Training Reward: 221.12 Evaluation Reward: 214.23
Average Evaluation Reward: 217.86
Training Iteration 126 Training Reward: 232.80 Evaluation Reward: 218.47
Average Evaluation Reward: 215.41
Training Iteration 127 Training Reward: 236.72 Evaluation Reward: 217.64
Average Evaluation Reward: 213.47
Training Iteration 128 Training Reward: 236.85 Evaluation Reward: 223.70
Average Evaluation Reward: 213.31
Training Iteration 129 Training Reward: 229.30 Evaluation Reward: 216.41
Average Evaluation Reward: 211.71
Training Iteration 130 Training Reward: 214.52 Evaluation Reward: 179.25
Average Evaluation Reward: 209.72
Training Iteration 131 Training Reward: 190.39 Evaluation Reward: 159.04
Average Evaluation Reward: 205.45
Training Iteration 132 Training Reward: 170.94 Evaluation Reward: 9.96
Average Evaluation Reward: 185.57
Training Iteration 133 Training Reward: 71.10 Evaluation Reward: -26.56
Average Evaluation Reward: 161.96
Training Iteration 134 Training Reward: -1.85 Evaluation Reward: -79.45
Average Evaluation Reward: 133.27
Training Iteration 135 Training Reward: -39.37 Evaluation Reward: -116.84
Average Evaluation Reward: 100.16
Training Iteration 136 Training Reward: -85.87 Evaluation Reward: -116.69
Average Evaluation Reward: 66.65
Training Iteration 137 Training Reward: -84.29 Evaluation Reward: -116.27
Average Evaluation Reward: 33.26
Training Iteration 138 Training Reward: -75.28 Evaluation Reward: -121.35
Average Evaluation Reward: -1.25
Training Iteration 139 Training Reward: -88.99 Evaluation Reward: -120.79
```

```
Average Evaluation Reward: -34.97
Training Iteration 140 Training Reward: -89.42 Evaluation Reward: -100.86
Average Evaluation Reward: -62.98
Training Iteration 141 Training Reward: -76.07 Evaluation Reward: -78.11
Average Evaluation Reward: -86.69
Training Iteration 142 Training Reward: -53.44 Evaluation Reward: -58.26
Average Evaluation Reward: -93.52
Training Iteration 143 Training Reward: -36.98 Evaluation Reward: -38.98
Average Evaluation Reward: -94.76
Training Iteration 144 Training Reward: -22.05 Evaluation Reward: -21.85
Average Evaluation Reward: -89.00
Training Iteration 145 Training Reward: -10.95 Evaluation Reward: -11.26
Average Evaluation Reward: -78.44
Training Iteration 146 Training Reward: -0.94 Evaluation Reward: -2.12
Average Evaluation Reward: -66.99
Training Iteration 147 Training Reward: 9.07 Evaluation Reward: 8.40
Average Evaluation Reward: -54.52
Training Iteration 148 Training Reward: 12.15 Evaluation Reward: 13.68
Average Evaluation Reward: -41.01
Training Iteration 149 Training Reward: 27.32 Evaluation Reward: 17.06
Average Evaluation Reward: -27.23
Training Iteration 150 Training Reward: 13.13 Evaluation Reward: 24.54
Average Evaluation Reward: -14.69
Training Iteration 151 Training Reward: 55.63 Evaluation Reward: 132.13
Average Evaluation Reward: 6.33
Training Iteration 152 Training Reward: 51.83 Evaluation Reward: 152.35
Average Evaluation Reward: 27.39
Training Iteration 153 Training Reward: 130.26 Evaluation Reward: 166.15
Average Evaluation Reward: 47.91
Training Iteration 154 Training Reward: 160.35 Evaluation Reward: 179.51
Average Evaluation Reward: 68.04
Training Iteration 155 Training Reward: 195.35 Evaluation Reward: 204.96
Average Evaluation Reward: 89.67
Training Iteration 156 Training Reward: 207.38 Evaluation Reward: 219.96
Average Evaluation Reward: 111.87
Training Iteration 157 Training Reward: 210.55 Evaluation Reward: 221.35
Average Evaluation Reward: 133.17
Training Iteration 158 Training Reward: 224.84 Evaluation Reward: 207.89
Average Evaluation Reward: 152.59
Training Iteration 159 Training Reward: 221.24 Evaluation Reward: -7.00
Average Evaluation Reward: 150.18
Training Iteration 160 Training Reward: 142.27 Evaluation Reward: 237.27
Average Evaluation Reward: 171.46
Training Iteration 161 Training Reward: 195.69 Evaluation Reward: -37.13
Average Evaluation Reward: 154.53
Training Iteration 162 Training Reward: 224.75 Evaluation Reward: 199.44
Average Evaluation Reward: 159.24
Training Iteration 163 Training Reward: 229.26 Evaluation Reward: 217.09
```

```
Average Evaluation Reward: 164.33
Training Iteration 164 Training Reward: 134.46 Evaluation Reward: -36.50
Average Evaluation Reward: 142.73
Training Iteration 165 Training Reward: -30.45 Evaluation Reward: -29.30
Average Evaluation Reward: 119.31
Training Iteration 166 Training Reward: -21.52 Evaluation Reward: -21.27
Average Evaluation Reward: 95.18
Training Iteration 167 Training Reward: 25.97 Evaluation Reward: 187.62
Average Evaluation Reward: 91.81
Training Iteration 168 Training Reward: 82.16 Evaluation Reward: 243.70
Average Evaluation Reward: 95.39
Training Iteration 169 Training Reward: 239.29 Evaluation Reward: 247.11
Average Evaluation Reward: 120.80
Training Iteration 170 Training Reward: 239.02 Evaluation Reward: 246.36
Average Evaluation Reward: 121.71
Training Iteration 171 Training Reward: 226.56 Evaluation Reward: -37.12
Average Evaluation Reward: 121.71
Training Iteration 172 Training Reward: 216.68 Evaluation Reward: 234.64
Average Evaluation Reward: 125.23
Training Iteration 173 Training Reward: 209.63 Evaluation Reward: 236.12
Average Evaluation Reward: 127.14
Training Iteration 174 Training Reward: 240.67 Evaluation Reward: 253.68
Average Evaluation Reward: 156.15
Training Iteration 175 Training Reward: 241.34 Evaluation Reward: 248.47
Average Evaluation Reward: 183.93
Training Iteration 176 Training Reward: 194.04 Evaluation Reward: 233.75
Average Evaluation Reward: 209.43
Training Iteration 177 Training Reward: 202.95 Evaluation Reward: 234.86
Average Evaluation Reward: 214.16
Training Iteration 178 Training Reward: 174.02 Evaluation Reward: 232.34
Average Evaluation Reward: 213.02
Training Iteration 179 Training Reward: 220.52 Evaluation Reward: 227.54
Average Evaluation Reward: 211.06
Training Iteration 180 Training Reward: 206.69 Evaluation Reward: 222.85
Average Evaluation Reward: 208.71
Training Iteration 181 Training Reward: 212.57 Evaluation Reward: 216.30
Average Evaluation Reward: 234.05
Training Iteration 182 Training Reward: 217.76 Evaluation Reward: 212.18
Average Evaluation Reward: 231.81
Training Iteration 183 Training Reward: 220.70 Evaluation Reward: 202.25
Average Evaluation Reward: 228.42
Training Iteration 184 Training Reward: 197.84 Evaluation Reward: 212.46
Average Evaluation Reward: 224.30
Training Iteration 185 Training Reward: 206.39 Evaluation Reward: 210.76
Average Evaluation Reward: 220.53
Training Iteration 186 Training Reward: 207.56 Evaluation Reward: 211.64
Average Evaluation Reward: 218.32
Training Iteration 187 Training Reward: 201.56 Evaluation Reward: 209.75
```

```
Average Evaluation Reward: 215.81
Training Iteration 188 Training Reward: 184.91 Evaluation Reward: 191.13
Average Evaluation Reward: 211.69
Training Iteration 189 Training Reward: 166.42 Evaluation Reward: 210.36
Average Evaluation Reward: 209.97
Training Iteration 190 Training Reward: 99.93 Evaluation Reward: 212.01
Average Evaluation Reward: 208.88
Training Iteration 191 Training Reward: 151.32 Evaluation Reward: 216.14
Average Evaluation Reward: 208.87
Training Iteration 192 Training Reward: 90.48 Evaluation Reward: 219.46
Average Evaluation Reward: 209.60
Training Iteration 193 Training Reward: 112.59 Evaluation Reward: 221.78
Average Evaluation Reward: 211.55
Training Iteration 194 Training Reward: 62.13 Evaluation Reward: 51.36
Average Evaluation Reward: 195.44
Training Iteration 195 Training Reward: 70.05 Evaluation Reward: 222.48
Average Evaluation Reward: 196.61
Training Iteration 196 Training Reward: 43.02 Evaluation Reward: 202.76
Average Evaluation Reward: 195.72
Training Iteration 197 Training Reward: 62.48 Evaluation Reward: 205.07
Average Evaluation Reward: 195.26
Training Iteration 198 Training Reward: 158.86 Evaluation Reward: 217.71
Average Evaluation Reward: 197.91
Training Iteration 199 Training Reward: 185.59 Evaluation Reward: 213.76
Average Evaluation Reward: 198.25
```



2) POLICY GRADIENT WITH Gt

```
[ ]: print("2) POLICY GRADIENT \n TRAINING FOR Gt: \n")
     main_fn("Gt")

     #_____
     #_____
     #_____
```

```
2) POLICY GRADIENT
 TRAINING FOR Gt:

Training Iteration 0 Training Reward: -80.02 Evaluation Reward: -364.48
Average Evaluation Reward: -364.48
Training Iteration 1 Training Reward: -251.02 Evaluation Reward: -324.61
Average Evaluation Reward: -344.55
Training Iteration 2 Training Reward: -171.38 Evaluation Reward: -108.64
Average Evaluation Reward: -265.91
Training Iteration 3 Training Reward: -63.09 Evaluation Reward: -97.47
Average Evaluation Reward: -223.80
```

```
Training Iteration 4 Training Reward: -107.44 Evaluation Reward: -59.84
Average Evaluation Reward: -191.01
Training Iteration 5 Training Reward: -125.85 Evaluation Reward: -119.06
Average Evaluation Reward: -179.02
Training Iteration 6 Training Reward: -73.31 Evaluation Reward: -119.06
Average Evaluation Reward: -170.45
Training Iteration 7 Training Reward: -53.23 Evaluation Reward: -114.17
Average Evaluation Reward: -163.42
Training Iteration 8 Training Reward: -61.74 Evaluation Reward: -57.11
Average Evaluation Reward: -151.61
Training Iteration 9 Training Reward: -51.98 Evaluation Reward: -49.83
Average Evaluation Reward: -141.43
Training Iteration 10 Training Reward: -65.13 Evaluation Reward: -5.58
Average Evaluation Reward: -105.54
Training Iteration 11 Training Reward: -33.51 Evaluation Reward: 17.94
Average Evaluation Reward: -71.28
Training Iteration 12 Training Reward: -10.42 Evaluation Reward: -135.42
Average Evaluation Reward: -73.96
Training Iteration 13 Training Reward: 7.29 Evaluation Reward: -162.42
Average Evaluation Reward: -80.45
Training Iteration 14 Training Reward: 5.17 Evaluation Reward: -248.09
Average Evaluation Reward: -99.28
Training Iteration 15 Training Reward: -53.14 Evaluation Reward: -368.59
Average Evaluation Reward: -124.23
Training Iteration 16 Training Reward: -8.18 Evaluation Reward: -475.99
Average Evaluation Reward: -159.93
Training Iteration 17 Training Reward: -119.79 Evaluation Reward: -512.26
Average Evaluation Reward: -199.74
Training Iteration 18 Training Reward: -150.89 Evaluation Reward: -443.63
Average Evaluation Reward: -238.39
Training Iteration 19 Training Reward: -176.56 Evaluation Reward: -351.60
Average Evaluation Reward: -268.56
Training Iteration 20 Training Reward: -93.84 Evaluation Reward: -270.43
Average Evaluation Reward: -295.05
Training Iteration 21 Training Reward: -3.69 Evaluation Reward: -153.42
Average Evaluation Reward: -312.18
Training Iteration 22 Training Reward: 25.64 Evaluation Reward: -68.19
Average Evaluation Reward: -305.46
Training Iteration 23 Training Reward: 30.68 Evaluation Reward: 175.74
Average Evaluation Reward: -271.65
Training Iteration 24 Training Reward: 45.27 Evaluation Reward: -150.84
Average Evaluation Reward: -261.92
Training Iteration 25 Training Reward: 43.04 Evaluation Reward: -85.45
Average Evaluation Reward: -233.61
Training Iteration 26 Training Reward: 45.09 Evaluation Reward: -102.16
Average Evaluation Reward: -196.22
Training Iteration 27 Training Reward: 32.40 Evaluation Reward: 28.10
Average Evaluation Reward: -142.19
```

```
Training Iteration 28 Training Reward: 42.10 Evaluation Reward: 25.13
Average Evaluation Reward: -95.31
Training Iteration 29 Training Reward: 48.18 Evaluation Reward: 16.14
Average Evaluation Reward: -58.54
Training Iteration 30 Training Reward: 37.13 Evaluation Reward: -53.21
Average Evaluation Reward: -36.82
Training Iteration 31 Training Reward: 46.44 Evaluation Reward: -23.76
Average Evaluation Reward: -23.85
Training Iteration 32 Training Reward: 39.98 Evaluation Reward: 227.10
Average Evaluation Reward: 5.68
Training Iteration 33 Training Reward: 42.71 Evaluation Reward: -85.08
Average Evaluation Reward: -20.40
Training Iteration 34 Training Reward: 39.59 Evaluation Reward: 189.80
Average Evaluation Reward: 13.66
Training Iteration 35 Training Reward: 36.31 Evaluation Reward: 184.21
Average Evaluation Reward: 40.63
Training Iteration 36 Training Reward: 7.22 Evaluation Reward: 177.90
Average Evaluation Reward: 68.63
Training Iteration 37 Training Reward: 16.86 Evaluation Reward: 161.52
Average Evaluation Reward: 81.98
Training Iteration 38 Training Reward: -11.22 Evaluation Reward: 188.89
Average Evaluation Reward: 98.35
Training Iteration 39 Training Reward: 8.06 Evaluation Reward: 168.41
Average Evaluation Reward: 113.58
Training Iteration 40 Training Reward: 20.74 Evaluation Reward: 176.00
Average Evaluation Reward: 136.50
Training Iteration 41 Training Reward: 67.36 Evaluation Reward: 196.69
Average Evaluation Reward: 158.54
Training Iteration 42 Training Reward: 73.35 Evaluation Reward: -25.33
Average Evaluation Reward: 133.30
Training Iteration 43 Training Reward: 83.27 Evaluation Reward: -3.48
Average Evaluation Reward: 141.46
Training Iteration 44 Training Reward: 92.69 Evaluation Reward: 11.54
Average Evaluation Reward: 123.64
Training Iteration 45 Training Reward: 106.61 Evaluation Reward: 26.76
Average Evaluation Reward: 107.89
Training Iteration 46 Training Reward: 129.29 Evaluation Reward: 35.82
Average Evaluation Reward: 93.68
Training Iteration 47 Training Reward: 138.91 Evaluation Reward: 36.03
Average Evaluation Reward: 81.13
Training Iteration 48 Training Reward: 111.59 Evaluation Reward: 49.08
Average Evaluation Reward: 67.15
Training Iteration 49 Training Reward: 150.60 Evaluation Reward: 36.33
Average Evaluation Reward: 53.95
Training Iteration 50 Training Reward: 136.14 Evaluation Reward: 285.72
Average Evaluation Reward: 64.92
Training Iteration 51 Training Reward: 129.55 Evaluation Reward: 159.78
Average Evaluation Reward: 61.23
```

```
Training Iteration 52 Training Reward: 159.75 Evaluation Reward: 48.92
Average Evaluation Reward: 68.65
Training Iteration 53 Training Reward: 130.54 Evaluation Reward: 247.65
Average Evaluation Reward: 93.76
Training Iteration 54 Training Reward: 141.78 Evaluation Reward: 232.47
Average Evaluation Reward: 115.86
Training Iteration 55 Training Reward: 130.82 Evaluation Reward: 207.23
Average Evaluation Reward: 133.90
Training Iteration 56 Training Reward: 131.13 Evaluation Reward: 58.98
Average Evaluation Reward: 136.22
Training Iteration 57 Training Reward: 116.84 Evaluation Reward: 215.18
Average Evaluation Reward: 154.13
Training Iteration 58 Training Reward: 105.22 Evaluation Reward: 231.78
Average Evaluation Reward: 172.40
Training Iteration 59 Training Reward: 123.58 Evaluation Reward: 233.74
Average Evaluation Reward: 192.14
Training Iteration 60 Training Reward: 134.28 Evaluation Reward: 230.91
Average Evaluation Reward: 186.66
Training Iteration 61 Training Reward: 142.31 Evaluation Reward: 137.58
Average Evaluation Reward: 184.44
Training Iteration 62 Training Reward: 125.87 Evaluation Reward: 248.59
Average Evaluation Reward: 204.41
Training Iteration 63 Training Reward: 87.61 Evaluation Reward: 253.23
Average Evaluation Reward: 204.97
Training Iteration 64 Training Reward: 126.54 Evaluation Reward: 253.04
Average Evaluation Reward: 207.02
Training Iteration 65 Training Reward: 72.75 Evaluation Reward: 243.05
Average Evaluation Reward: 210.61
Training Iteration 66 Training Reward: 127.53 Evaluation Reward: 245.36
Average Evaluation Reward: 229.24
Training Iteration 67 Training Reward: 134.26 Evaluation Reward: 221.28
Average Evaluation Reward: 229.85
Training Iteration 68 Training Reward: 161.36 Evaluation Reward: 261.25
Average Evaluation Reward: 232.80
Training Iteration 69 Training Reward: 158.61 Evaluation Reward: 264.17
Average Evaluation Reward: 235.84
Training Iteration 70 Training Reward: 155.40 Evaluation Reward: 265.07
Average Evaluation Reward: 239.26
Training Iteration 71 Training Reward: 154.57 Evaluation Reward: 259.64
Average Evaluation Reward: 251.47
Training Iteration 72 Training Reward: 154.60 Evaluation Reward: 263.32
Average Evaluation Reward: 252.94
Training Iteration 73 Training Reward: 161.32 Evaluation Reward: 274.42
Average Evaluation Reward: 255.06
Training Iteration 74 Training Reward: 162.23 Evaluation Reward: 275.77
Average Evaluation Reward: 257.33
Training Iteration 75 Training Reward: 166.99 Evaluation Reward: 263.37
Average Evaluation Reward: 259.36
```

```
Training Iteration 76 Training Reward: 135.14 Evaluation Reward: 264.44
Average Evaluation Reward: 261.27
Training Iteration 77 Training Reward: 53.05 Evaluation Reward: 35.92
Average Evaluation Reward: 242.74
Training Iteration 78 Training Reward: 17.40 Evaluation Reward: 262.28
Average Evaluation Reward: 242.84
Training Iteration 79 Training Reward: 1.05 Evaluation Reward: 240.37
Average Evaluation Reward: 240.46
Training Iteration 80 Training Reward: 1.48 Evaluation Reward: 265.38
Average Evaluation Reward: 240.49
Training Iteration 81 Training Reward: 103.49 Evaluation Reward: 28.73
Average Evaluation Reward: 217.40
Training Iteration 82 Training Reward: 112.32 Evaluation Reward: 273.88
Average Evaluation Reward: 218.46
Training Iteration 83 Training Reward: 127.90 Evaluation Reward: 274.81
Average Evaluation Reward: 218.49
Training Iteration 84 Training Reward: 143.58 Evaluation Reward: 286.33
Average Evaluation Reward: 219.55
Training Iteration 85 Training Reward: 143.88 Evaluation Reward: 284.50
Average Evaluation Reward: 221.66
Training Iteration 86 Training Reward: 111.07 Evaluation Reward: 163.05
Average Evaluation Reward: 211.53
Training Iteration 87 Training Reward: 139.36 Evaluation Reward: 161.01
Average Evaluation Reward: 224.03
Training Iteration 88 Training Reward: 167.09 Evaluation Reward: 148.66
Average Evaluation Reward: 212.67
Training Iteration 89 Training Reward: 168.85 Evaluation Reward: 167.49
Average Evaluation Reward: 205.39
Training Iteration 90 Training Reward: 169.30 Evaluation Reward: 264.34
Average Evaluation Reward: 205.28
Training Iteration 91 Training Reward: 163.59 Evaluation Reward: 263.37
Average Evaluation Reward: 228.74
Training Iteration 92 Training Reward: 164.65 Evaluation Reward: 285.09
Average Evaluation Reward: 229.87
Training Iteration 93 Training Reward: 163.37 Evaluation Reward: 286.08
Average Evaluation Reward: 230.99
Training Iteration 94 Training Reward: 165.05 Evaluation Reward: 284.97
Average Evaluation Reward: 230.86
Training Iteration 95 Training Reward: 167.88 Evaluation Reward: 284.02
Average Evaluation Reward: 230.81
Training Iteration 96 Training Reward: 169.42 Evaluation Reward: 285.30
Average Evaluation Reward: 243.03
Training Iteration 97 Training Reward: 166.09 Evaluation Reward: 284.45
Average Evaluation Reward: 255.38
Training Iteration 98 Training Reward: 167.08 Evaluation Reward: 286.97
Average Evaluation Reward: 269.21
Training Iteration 99 Training Reward: 171.52 Evaluation Reward: 288.59
Average Evaluation Reward: 281.32
```

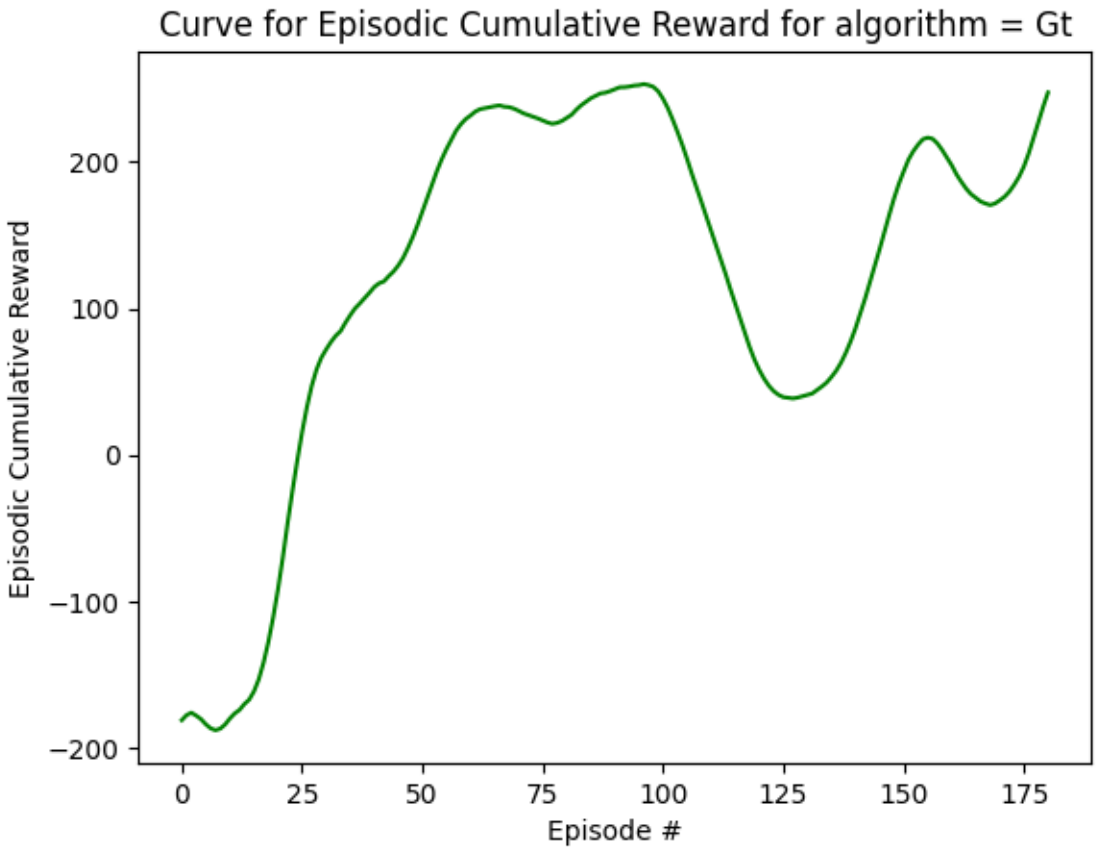
```
Training Iteration 100 Training Reward: 177.32 Evaluation Reward: 292.62
Average Evaluation Reward: 284.14
Training Iteration 101 Training Reward: 180.42 Evaluation Reward: 300.12
Average Evaluation Reward: 287.82
Training Iteration 102 Training Reward: 168.09 Evaluation Reward: 183.85
Average Evaluation Reward: 277.70
Training Iteration 103 Training Reward: 169.05 Evaluation Reward: 178.41
Average Evaluation Reward: 266.93
Training Iteration 104 Training Reward: 136.36 Evaluation Reward: 282.67
Average Evaluation Reward: 266.70
Training Iteration 105 Training Reward: 109.83 Evaluation Reward: 168.67
Average Evaluation Reward: 255.16
Training Iteration 106 Training Reward: 127.32 Evaluation Reward: 173.52
Average Evaluation Reward: 243.99
Training Iteration 107 Training Reward: 126.48 Evaluation Reward: 178.98
Average Evaluation Reward: 233.44
Training Iteration 108 Training Reward: 128.31 Evaluation Reward: 292.74
Average Evaluation Reward: 234.02
Training Iteration 109 Training Reward: 95.50 Evaluation Reward: 267.55
Average Evaluation Reward: 231.91
Training Iteration 110 Training Reward: 74.10 Evaluation Reward: 286.97
Average Evaluation Reward: 231.35
Training Iteration 111 Training Reward: 93.87 Evaluation Reward: 290.50
Average Evaluation Reward: 230.39
Training Iteration 112 Training Reward: 50.19 Evaluation Reward: 262.28
Average Evaluation Reward: 238.23
Training Iteration 113 Training Reward: 78.62 Evaluation Reward: 241.57
Average Evaluation Reward: 244.55
Training Iteration 114 Training Reward: 142.44 Evaluation Reward: 201.78
Average Evaluation Reward: 236.46
Training Iteration 115 Training Reward: 68.59 Evaluation Reward: 259.30
Average Evaluation Reward: 245.52
Training Iteration 116 Training Reward: 178.59 Evaluation Reward: 11.04
Average Evaluation Reward: 229.27
Training Iteration 117 Training Reward: 108.61 Evaluation Reward: 198.91
Average Evaluation Reward: 231.26
Training Iteration 118 Training Reward: 87.35 Evaluation Reward: 70.47
Average Evaluation Reward: 209.04
Training Iteration 119 Training Reward: 88.43 Evaluation Reward: -25.59
Average Evaluation Reward: 179.72
Training Iteration 120 Training Reward: 38.95 Evaluation Reward: 51.83
Average Evaluation Reward: 156.21
Training Iteration 121 Training Reward: 64.19 Evaluation Reward: 57.73
Average Evaluation Reward: 132.93
Training Iteration 122 Training Reward: 67.33 Evaluation Reward: 31.22
Average Evaluation Reward: 109.82
Training Iteration 123 Training Reward: 25.51 Evaluation Reward: 49.49
Average Evaluation Reward: 90.62
```

```
Training Iteration 124 Training Reward: 57.71 Evaluation Reward: 24.57
Average Evaluation Reward: 72.90
Training Iteration 125 Training Reward: 57.49 Evaluation Reward: 42.76
Average Evaluation Reward: 51.24
Training Iteration 126 Training Reward: 60.49 Evaluation Reward: 40.17
Average Evaluation Reward: 54.15
Training Iteration 127 Training Reward: 50.30 Evaluation Reward: 36.38
Average Evaluation Reward: 37.90
Training Iteration 128 Training Reward: 43.76 Evaluation Reward: 33.24
Average Evaluation Reward: 34.18
Training Iteration 129 Training Reward: 43.17 Evaluation Reward: 29.43
Average Evaluation Reward: 39.68
Training Iteration 130 Training Reward: 46.12 Evaluation Reward: 40.96
Average Evaluation Reward: 38.59
Training Iteration 131 Training Reward: 43.87 Evaluation Reward: 46.58
Average Evaluation Reward: 37.48
Training Iteration 132 Training Reward: 47.91 Evaluation Reward: 46.01
Average Evaluation Reward: 38.96
Training Iteration 133 Training Reward: 47.51 Evaluation Reward: 17.35
Average Evaluation Reward: 35.74
Training Iteration 134 Training Reward: 50.21 Evaluation Reward: 45.84
Average Evaluation Reward: 37.87
Training Iteration 135 Training Reward: 48.07 Evaluation Reward: 44.01
Average Evaluation Reward: 38.00
Training Iteration 136 Training Reward: 51.55 Evaluation Reward: 45.27
Average Evaluation Reward: 38.51
Training Iteration 137 Training Reward: 53.85 Evaluation Reward: 28.71
Average Evaluation Reward: 37.74
Training Iteration 138 Training Reward: 52.74 Evaluation Reward: 36.59
Average Evaluation Reward: 38.08
Training Iteration 139 Training Reward: 47.86 Evaluation Reward: 30.46
Average Evaluation Reward: 38.18
Training Iteration 140 Training Reward: 43.10 Evaluation Reward: 30.78
Average Evaluation Reward: 37.16
Training Iteration 141 Training Reward: 48.37 Evaluation Reward: 38.44
Average Evaluation Reward: 36.35
Training Iteration 142 Training Reward: 52.29 Evaluation Reward: 51.37
Average Evaluation Reward: 36.89
Training Iteration 143 Training Reward: 65.03 Evaluation Reward: 50.56
Average Evaluation Reward: 40.21
Training Iteration 144 Training Reward: 63.95 Evaluation Reward: 55.81
Average Evaluation Reward: 41.20
Training Iteration 145 Training Reward: 70.93 Evaluation Reward: 61.78
Average Evaluation Reward: 42.98
Training Iteration 146 Training Reward: 67.80 Evaluation Reward: 65.72
Average Evaluation Reward: 45.02
Training Iteration 147 Training Reward: 123.74 Evaluation Reward: 66.67
Average Evaluation Reward: 48.82
```

```
Training Iteration 148 Training Reward: 92.81 Evaluation Reward: 67.09
Average Evaluation Reward: 51.87
Training Iteration 149 Training Reward: 119.95 Evaluation Reward: 73.34
Average Evaluation Reward: 56.16
Training Iteration 150 Training Reward: 156.45 Evaluation Reward: 70.31
Average Evaluation Reward: 60.11
Training Iteration 151 Training Reward: 143.96 Evaluation Reward: 284.01
Average Evaluation Reward: 84.67
Training Iteration 152 Training Reward: 139.47 Evaluation Reward: 76.45
Average Evaluation Reward: 87.17
Training Iteration 153 Training Reward: 170.85 Evaluation Reward: 79.44
Average Evaluation Reward: 90.06
Training Iteration 154 Training Reward: 179.48 Evaluation Reward: 290.28
Average Evaluation Reward: 113.51
Training Iteration 155 Training Reward: 154.72 Evaluation Reward: 173.25
Average Evaluation Reward: 124.66
Training Iteration 156 Training Reward: 195.77 Evaluation Reward: 303.75
Average Evaluation Reward: 148.46
Training Iteration 157 Training Reward: 183.71 Evaluation Reward: 312.66
Average Evaluation Reward: 173.06
Training Iteration 158 Training Reward: 159.76 Evaluation Reward: 278.85
Average Evaluation Reward: 194.23
Training Iteration 159 Training Reward: 164.86 Evaluation Reward: 286.96
Average Evaluation Reward: 215.60
Training Iteration 160 Training Reward: 130.84 Evaluation Reward: 312.40
Average Evaluation Reward: 239.81
Training Iteration 161 Training Reward: 129.37 Evaluation Reward: 275.09
Average Evaluation Reward: 238.91
Training Iteration 162 Training Reward: 90.18 Evaluation Reward: 256.19
Average Evaluation Reward: 256.89
Training Iteration 163 Training Reward: 89.08 Evaluation Reward: 158.65
Average Evaluation Reward: 264.81
Training Iteration 164 Training Reward: 60.40 Evaluation Reward: 305.88
Average Evaluation Reward: 266.37
Training Iteration 165 Training Reward: 57.82 Evaluation Reward: 296.26
Average Evaluation Reward: 278.67
Training Iteration 166 Training Reward: 54.10 Evaluation Reward: 297.36
Average Evaluation Reward: 278.03
Training Iteration 167 Training Reward: 49.70 Evaluation Reward: 269.01
Average Evaluation Reward: 273.66
Training Iteration 168 Training Reward: 64.74 Evaluation Reward: 42.34
Average Evaluation Reward: 250.01
Training Iteration 169 Training Reward: 114.17 Evaluation Reward: 52.29
Average Evaluation Reward: 226.55
Training Iteration 170 Training Reward: 153.48 Evaluation Reward: 257.38
Average Evaluation Reward: 221.04
Training Iteration 171 Training Reward: 117.85 Evaluation Reward: 58.13
Average Evaluation Reward: 199.35
```

```
Training Iteration 172 Training Reward: 63.64 Evaluation Reward: 36.13
Average Evaluation Reward: 177.34
Training Iteration 173 Training Reward: 64.27 Evaluation Reward: 23.54
Average Evaluation Reward: 163.83
Training Iteration 174 Training Reward: 69.28 Evaluation Reward: 24.07
Average Evaluation Reward: 135.65
Training Iteration 175 Training Reward: 53.62 Evaluation Reward: 37.58
Average Evaluation Reward: 109.78
Training Iteration 176 Training Reward: 67.22 Evaluation Reward: 54.04
Average Evaluation Reward: 85.45
Training Iteration 177 Training Reward: 72.06 Evaluation Reward: 247.76
Average Evaluation Reward: 83.33
Training Iteration 178 Training Reward: 85.29 Evaluation Reward: 85.44
Average Evaluation Reward: 87.64
Training Iteration 179 Training Reward: 132.02 Evaluation Reward: 291.66
Average Evaluation Reward: 111.57
Training Iteration 180 Training Reward: 141.49 Evaluation Reward: 291.53
Average Evaluation Reward: 114.99
Training Iteration 181 Training Reward: 171.19 Evaluation Reward: 293.75
Average Evaluation Reward: 138.55
Training Iteration 182 Training Reward: 173.28 Evaluation Reward: 291.55
Average Evaluation Reward: 164.09
Training Iteration 183 Training Reward: 178.92 Evaluation Reward: 284.67
Average Evaluation Reward: 190.20
Training Iteration 184 Training Reward: 175.52 Evaluation Reward: 278.24
Average Evaluation Reward: 215.62
Training Iteration 185 Training Reward: 172.25 Evaluation Reward: 172.30
Average Evaluation Reward: 229.09
Training Iteration 186 Training Reward: 172.32 Evaluation Reward: 269.42
Average Evaluation Reward: 250.63
Training Iteration 187 Training Reward: 170.88 Evaluation Reward: 270.64
Average Evaluation Reward: 252.92
Training Iteration 188 Training Reward: 162.06 Evaluation Reward: 269.81
Average Evaluation Reward: 271.36
Training Iteration 189 Training Reward: 168.32 Evaluation Reward: 283.62
Average Evaluation Reward: 270.55
Training Iteration 190 Training Reward: 174.44 Evaluation Reward: 288.59
Average Evaluation Reward: 270.26
Training Iteration 191 Training Reward: 176.86 Evaluation Reward: 292.29
Average Evaluation Reward: 270.11
Training Iteration 192 Training Reward: 181.92 Evaluation Reward: 297.66
Average Evaluation Reward: 270.72
Training Iteration 193 Training Reward: 184.45 Evaluation Reward: 304.76
Average Evaluation Reward: 272.73
Training Iteration 194 Training Reward: 188.62 Evaluation Reward: 305.23
Average Evaluation Reward: 275.43
Training Iteration 195 Training Reward: 167.61 Evaluation Reward: 307.94
Average Evaluation Reward: 289.00
```

```
Training Iteration 196 Training Reward: 190.90 Evaluation Reward: 308.97
Average Evaluation Reward: 292.95
Training Iteration 197 Training Reward: 192.00 Evaluation Reward: 312.49
Average Evaluation Reward: 297.14
Training Iteration 198 Training Reward: 165.37 Evaluation Reward: 312.82
Average Evaluation Reward: 301.44
Training Iteration 199 Training Reward: 183.81 Evaluation Reward: 313.90
Average Evaluation Reward: 304.46
```



3) POLICY GRADIENT WITH Gt WITH BASELINE:

```
[ ]: print("1) POLICY GRADIENT WITH BASELINE \n TRAINING FOR Gt WITH BASELINE: \n")
     main_fn("Gt_with_Baseline")

     #-------------------------------------------------------------------------------------------
     #-------------------------------------------------------------------------------------------
     #-------------------------------------------------------------------------------------------
```

```
1) POLICY GRADIENT WITH BASELINE
 TRAINING FOR Gt WITH BASELINE:
```

```
Training Iteration 0 Training Reward: -80.02 Evaluation Reward: -364.16
Average Evaluation Reward: -364.16
Training Iteration 1 Training Reward: -250.53 Evaluation Reward: -255.62
Average Evaluation Reward: -309.89
Training Iteration 2 Training Reward: -161.20 Evaluation Reward: -47.97
Average Evaluation Reward: -222.58
Training Iteration 3 Training Reward: -74.79 Evaluation Reward: -46.67
Average Evaluation Reward: -178.60
Training Iteration 4 Training Reward: -61.00 Evaluation Reward: -98.40
Average Evaluation Reward: -162.56
Training Iteration 5 Training Reward: -105.56 Evaluation Reward: -105.51
Average Evaluation Reward: -153.05
Training Iteration 6 Training Reward: -65.21 Evaluation Reward: -117.78
Average Evaluation Reward: -148.02
Training Iteration 7 Training Reward: -47.50 Evaluation Reward: -56.65
Average Evaluation Reward: -136.59
Training Iteration 8 Training Reward: -44.17 Evaluation Reward: -59.12
Average Evaluation Reward: -127.99
Training Iteration 9 Training Reward: -26.44 Evaluation Reward: -12.33
Average Evaluation Reward: -116.42
Training Iteration 10 Training Reward: -6.59 Evaluation Reward: -127.86
Average Evaluation Reward: -92.79
Training Iteration 11 Training Reward: 7.02 Evaluation Reward: -157.30
Average Evaluation Reward: -82.96
Training Iteration 12 Training Reward: -2.47 Evaluation Reward: -200.55
Average Evaluation Reward: -98.22
Training Iteration 13 Training Reward: -24.33 Evaluation Reward: -239.10
Average Evaluation Reward: -117.46
Training Iteration 14 Training Reward: -71.28 Evaluation Reward: -271.18
Average Evaluation Reward: -134.74
Training Iteration 15 Training Reward: -119.40 Evaluation Reward: -294.50
Average Evaluation Reward: -153.64
Training Iteration 16 Training Reward: -162.75 Evaluation Reward: -281.48
Average Evaluation Reward: -170.01
Training Iteration 17 Training Reward: -208.15 Evaluation Reward: -262.99
Average Evaluation Reward: -190.64
Training Iteration 18 Training Reward: -258.46 Evaluation Reward: -138.12
Average Evaluation Reward: -198.54
Training Iteration 19 Training Reward: -101.26 Evaluation Reward: -77.74
Average Evaluation Reward: -205.08
Training Iteration 20 Training Reward: -138.46 Evaluation Reward: -60.25
Average Evaluation Reward: -198.32
Training Iteration 21 Training Reward: -72.87 Evaluation Reward: -69.76
Average Evaluation Reward: -189.57
Training Iteration 22 Training Reward: -25.53 Evaluation Reward: -252.12
Average Evaluation Reward: -194.72
Training Iteration 23 Training Reward: 21.22 Evaluation Reward: -1.69
Average Evaluation Reward: -170.98
```

```
Training Iteration 24 Training Reward: 24.56 Evaluation Reward: 19.81
Average Evaluation Reward: -141.88
Training Iteration 25 Training Reward: 29.58 Evaluation Reward: 29.19
Average Evaluation Reward: -109.51
Training Iteration 26 Training Reward: 30.52 Evaluation Reward: 256.73
Average Evaluation Reward: -55.69
Training Iteration 27 Training Reward: 32.69 Evaluation Reward: 40.07
Average Evaluation Reward: -25.39
Training Iteration 28 Training Reward: 41.81 Evaluation Reward: 49.13
Average Evaluation Reward: -6.66
Training Iteration 29 Training Reward: 48.47 Evaluation Reward: 197.45
Average Evaluation Reward: 20.86
Training Iteration 30 Training Reward: 44.72 Evaluation Reward: 282.88
Average Evaluation Reward: 55.17
Training Iteration 31 Training Reward: 56.92 Evaluation Reward: 105.71
Average Evaluation Reward: 72.72
Training Iteration 32 Training Reward: 60.32 Evaluation Reward: 109.63
Average Evaluation Reward: 108.89
Training Iteration 33 Training Reward: 31.47 Evaluation Reward: 22.37
Average Evaluation Reward: 111.30
Training Iteration 34 Training Reward: 44.81 Evaluation Reward: -18.89
Average Evaluation Reward: 107.43
Training Iteration 35 Training Reward: 16.24 Evaluation Reward: -25.25
Average Evaluation Reward: 101.98
Training Iteration 36 Training Reward: 2.20 Evaluation Reward: -14.34
Average Evaluation Reward: 74.88
Training Iteration 37 Training Reward: 14.75 Evaluation Reward: -17.95
Average Evaluation Reward: 69.07
Training Iteration 38 Training Reward: -7.63 Evaluation Reward: 8.40
Average Evaluation Reward: 65.00
Training Iteration 39 Training Reward: -1.13 Evaluation Reward: 20.54
Average Evaluation Reward: 47.31
Training Iteration 40 Training Reward: -5.41 Evaluation Reward: 19.89
Average Evaluation Reward: 21.01
Training Iteration 41 Training Reward: 2.73 Evaluation Reward: 180.00
Average Evaluation Reward: 28.44
Training Iteration 42 Training Reward: 26.67 Evaluation Reward: 244.91
Average Evaluation Reward: 41.97
Training Iteration 43 Training Reward: 37.74 Evaluation Reward: 43.52
Average Evaluation Reward: 44.08
Training Iteration 44 Training Reward: 43.58 Evaluation Reward: 27.51
Average Evaluation Reward: 48.72
Training Iteration 45 Training Reward: 36.66 Evaluation Reward: 25.36
Average Evaluation Reward: 53.78
Training Iteration 46 Training Reward: 51.95 Evaluation Reward: 18.07
Average Evaluation Reward: 57.02
Training Iteration 47 Training Reward: 52.36 Evaluation Reward: 33.58
Average Evaluation Reward: 62.18
```

```
Training Iteration 48 Training Reward: 73.30 Evaluation Reward: 42.34
Average Evaluation Reward: 65.57
Training Iteration 49 Training Reward: 79.07 Evaluation Reward: 45.60
Average Evaluation Reward: 68.08
Training Iteration 50 Training Reward: 87.78 Evaluation Reward: 57.90
Average Evaluation Reward: 71.88
Training Iteration 51 Training Reward: 91.07 Evaluation Reward: 199.57
Average Evaluation Reward: 73.84
Training Iteration 52 Training Reward: 142.37 Evaluation Reward: 198.54
Average Evaluation Reward: 69.20
Training Iteration 53 Training Reward: 147.60 Evaluation Reward: 293.25
Average Evaluation Reward: 94.17
Training Iteration 54 Training Reward: 116.87 Evaluation Reward: 295.03
Average Evaluation Reward: 120.93
Training Iteration 55 Training Reward: 156.17 Evaluation Reward: 198.19
Average Evaluation Reward: 138.21
Training Iteration 56 Training Reward: 136.43 Evaluation Reward: 67.51
Average Evaluation Reward: 143.15
Training Iteration 57 Training Reward: 154.77 Evaluation Reward: 52.58
Average Evaluation Reward: 145.05
Training Iteration 58 Training Reward: 170.54 Evaluation Reward: 38.49
Average Evaluation Reward: 144.67
Training Iteration 59 Training Reward: 166.78 Evaluation Reward: 35.99
Average Evaluation Reward: 143.71
Training Iteration 60 Training Reward: 132.30 Evaluation Reward: 31.84
Average Evaluation Reward: 141.10
Training Iteration 61 Training Reward: 128.90 Evaluation Reward: 30.25
Average Evaluation Reward: 124.17
Training Iteration 62 Training Reward: 127.57 Evaluation Reward: 18.11
Average Evaluation Reward: 106.13
Training Iteration 63 Training Reward: 114.17 Evaluation Reward: -5.33
Average Evaluation Reward: 76.27
Training Iteration 64 Training Reward: 97.51 Evaluation Reward: 193.55
Average Evaluation Reward: 66.12
Training Iteration 65 Training Reward: 135.11 Evaluation Reward: 201.29
Average Evaluation Reward: 66.43
Training Iteration 66 Training Reward: 72.66 Evaluation Reward: 194.36
Average Evaluation Reward: 79.11
Training Iteration 67 Training Reward: 76.01 Evaluation Reward: -32.14
Average Evaluation Reward: 70.64
Training Iteration 68 Training Reward: 67.97 Evaluation Reward: -39.73
Average Evaluation Reward: 62.82
Training Iteration 69 Training Reward: 69.01 Evaluation Reward: 202.10
Average Evaluation Reward: 79.43
Training Iteration 70 Training Reward: 73.52 Evaluation Reward: -17.23
Average Evaluation Reward: 74.52
Training Iteration 71 Training Reward: 76.88 Evaluation Reward: 206.75
Average Evaluation Reward: 92.17
```

```
Training Iteration 72 Training Reward: 69.27 Evaluation Reward: 194.63
Average Evaluation Reward: 109.83
Training Iteration 73 Training Reward: 96.04 Evaluation Reward: 176.63
Average Evaluation Reward: 128.02
Training Iteration 74 Training Reward: 88.30 Evaluation Reward: 189.41
Average Evaluation Reward: 127.61
Training Iteration 75 Training Reward: 87.68 Evaluation Reward: -12.29
Average Evaluation Reward: 106.25
Training Iteration 76 Training Reward: 80.96 Evaluation Reward: 180.19
Average Evaluation Reward: 104.83
Training Iteration 77 Training Reward: 105.22 Evaluation Reward: 16.10
Average Evaluation Reward: 109.66
Training Iteration 78 Training Reward: 91.91 Evaluation Reward: 25.16
Average Evaluation Reward: 116.15
Training Iteration 79 Training Reward: 141.98 Evaluation Reward: 237.65
Average Evaluation Reward: 119.70
Training Iteration 80 Training Reward: 165.40 Evaluation Reward: 251.30
Average Evaluation Reward: 146.55
Training Iteration 81 Training Reward: 171.72 Evaluation Reward: 267.71
Average Evaluation Reward: 152.65
Training Iteration 82 Training Reward: 174.40 Evaluation Reward: 251.79
Average Evaluation Reward: 158.36
Training Iteration 83 Training Reward: 169.17 Evaluation Reward: 182.82
Average Evaluation Reward: 158.98
Training Iteration 84 Training Reward: 172.18 Evaluation Reward: 168.02
Average Evaluation Reward: 156.84
Training Iteration 85 Training Reward: 176.29 Evaluation Reward: 168.10
Average Evaluation Reward: 174.88
Training Iteration 86 Training Reward: 132.96 Evaluation Reward: 181.08
Average Evaluation Reward: 174.97
Training Iteration 87 Training Reward: 88.96 Evaluation Reward: 279.76
Average Evaluation Reward: 201.34
Training Iteration 88 Training Reward: 53.66 Evaluation Reward: 174.97
Average Evaluation Reward: 216.32
Training Iteration 89 Training Reward: 44.89 Evaluation Reward: 176.28
Average Evaluation Reward: 210.18
Training Iteration 90 Training Reward: 34.71 Evaluation Reward: 178.80
Average Evaluation Reward: 202.93
Training Iteration 91 Training Reward: 28.99 Evaluation Reward: 175.20
Average Evaluation Reward: 193.68
Training Iteration 92 Training Reward: 76.56 Evaluation Reward: 238.36
Average Evaluation Reward: 192.34
Training Iteration 93 Training Reward: 112.57 Evaluation Reward: 169.96
Average Evaluation Reward: 191.05
Training Iteration 94 Training Reward: 135.51 Evaluation Reward: 159.63
Average Evaluation Reward: 190.21
Training Iteration 95 Training Reward: 184.08 Evaluation Reward: 168.80
Average Evaluation Reward: 190.28
```

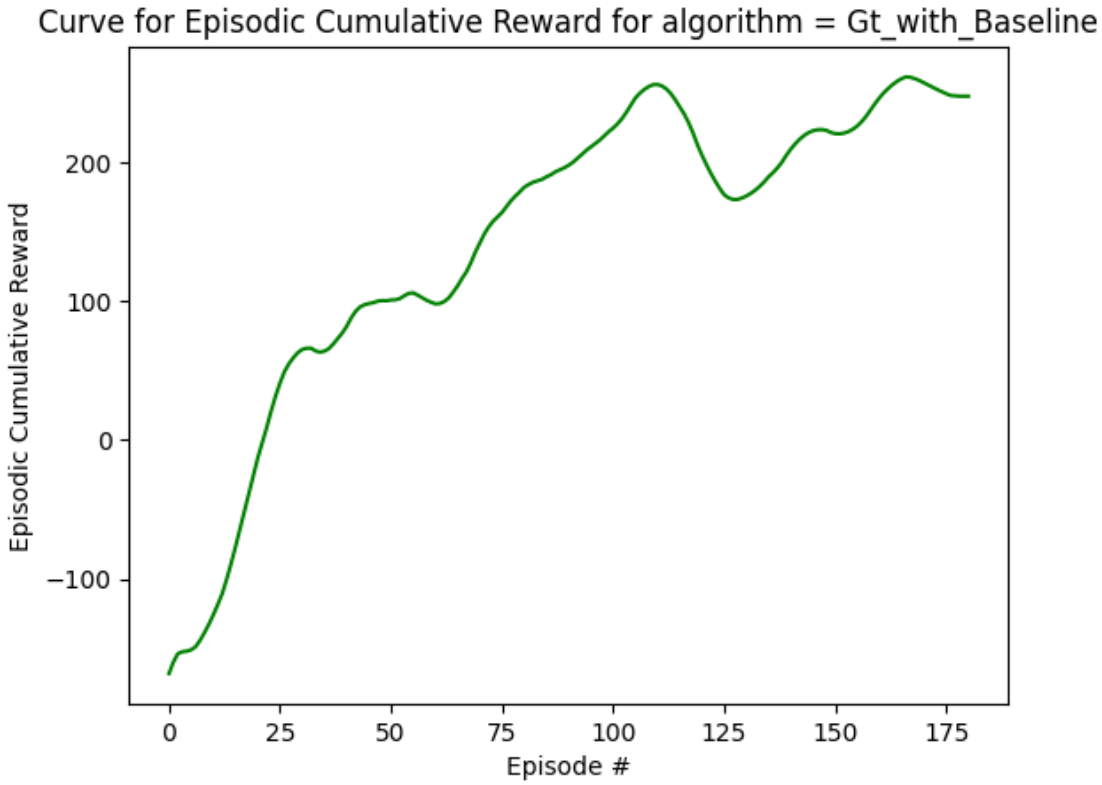
```
Training Iteration 96 Training Reward: 165.03 Evaluation Reward: 167.73
Average Evaluation Reward: 188.95
Training Iteration 97 Training Reward: 171.88 Evaluation Reward: 166.87
Average Evaluation Reward: 177.66
Training Iteration 98 Training Reward: 148.43 Evaluation Reward: 166.85
Average Evaluation Reward: 176.85
Training Iteration 99 Training Reward: 154.90 Evaluation Reward: 271.49
Average Evaluation Reward: 186.37
Training Iteration 100 Training Reward: 174.52 Evaluation Reward: 176.19
Average Evaluation Reward: 186.11
Training Iteration 101 Training Reward: 167.76 Evaluation Reward: 166.09
Average Evaluation Reward: 185.20
Training Iteration 102 Training Reward: 150.77 Evaluation Reward: 177.73
Average Evaluation Reward: 179.13
Training Iteration 103 Training Reward: 169.92 Evaluation Reward: 171.05
Average Evaluation Reward: 179.24
Training Iteration 104 Training Reward: 171.22 Evaluation Reward: 301.89
Average Evaluation Reward: 193.47
Training Iteration 105 Training Reward: 177.48 Evaluation Reward: 278.37
Average Evaluation Reward: 204.43
Training Iteration 106 Training Reward: 179.93 Evaluation Reward: 300.63
Average Evaluation Reward: 217.72
Training Iteration 107 Training Reward: 181.77 Evaluation Reward: 299.43
Average Evaluation Reward: 230.97
Training Iteration 108 Training Reward: 184.15 Evaluation Reward: 299.18
Average Evaluation Reward: 244.21
Training Iteration 109 Training Reward: 188.99 Evaluation Reward: 297.39
Average Evaluation Reward: 246.80
Training Iteration 110 Training Reward: 136.62 Evaluation Reward: 188.44
Average Evaluation Reward: 248.02
Training Iteration 111 Training Reward: 116.40 Evaluation Reward: 174.83
Average Evaluation Reward: 248.89
Training Iteration 112 Training Reward: 100.03 Evaluation Reward: 184.88
Average Evaluation Reward: 249.61
Training Iteration 113 Training Reward: 129.73 Evaluation Reward: 175.55
Average Evaluation Reward: 250.06
Training Iteration 114 Training Reward: 145.41 Evaluation Reward: 177.14
Average Evaluation Reward: 237.58
Training Iteration 115 Training Reward: 139.39 Evaluation Reward: 292.49
Average Evaluation Reward: 239.00
Training Iteration 116 Training Reward: 163.10 Evaluation Reward: 295.91
Average Evaluation Reward: 238.52
Training Iteration 117 Training Reward: 171.67 Evaluation Reward: 295.01
Average Evaluation Reward: 238.08
Training Iteration 118 Training Reward: 186.31 Evaluation Reward: 295.68
Average Evaluation Reward: 237.73
Training Iteration 119 Training Reward: 172.12 Evaluation Reward: 295.17
Average Evaluation Reward: 237.51
```

```
Training Iteration 120 Training Reward: 164.30 Evaluation Reward: 294.63
Average Evaluation Reward: 248.13
Training Iteration 121 Training Reward: 148.79 Evaluation Reward: 293.82
Average Evaluation Reward: 260.03
Training Iteration 122 Training Reward: 135.61 Evaluation Reward: 293.32
Average Evaluation Reward: 270.87
Training Iteration 123 Training Reward: 171.36 Evaluation Reward: 294.33
Average Evaluation Reward: 282.75
Training Iteration 124 Training Reward: 146.80 Evaluation Reward: 293.30
Average Evaluation Reward: 294.37
Training Iteration 125 Training Reward: 101.11 Evaluation Reward: 77.59
Average Evaluation Reward: 272.88
Training Iteration 126 Training Reward: 121.82 Evaluation Reward: 294.02
Average Evaluation Reward: 272.69
Training Iteration 127 Training Reward: 146.16 Evaluation Reward: 293.88
Average Evaluation Reward: 272.57
Training Iteration 128 Training Reward: 97.89 Evaluation Reward: 292.81
Average Evaluation Reward: 272.29
Training Iteration 129 Training Reward: 101.87 Evaluation Reward: 68.76
Average Evaluation Reward: 249.65
Training Iteration 130 Training Reward: 114.55 Evaluation Reward: 69.51
Average Evaluation Reward: 227.13
Training Iteration 131 Training Reward: 133.26 Evaluation Reward: 79.87
Average Evaluation Reward: 205.74
Training Iteration 132 Training Reward: 123.66 Evaluation Reward: 67.19
Average Evaluation Reward: 183.13
Training Iteration 133 Training Reward: 133.10 Evaluation Reward: 69.43
Average Evaluation Reward: 160.64
Training Iteration 134 Training Reward: 138.18 Evaluation Reward: 60.50
Average Evaluation Reward: 137.36
Training Iteration 135 Training Reward: 141.67 Evaluation Reward: 59.08
Average Evaluation Reward: 135.51
Training Iteration 136 Training Reward: 124.73 Evaluation Reward: 66.07
Average Evaluation Reward: 112.71
Training Iteration 137 Training Reward: 70.08 Evaluation Reward: 56.71
Average Evaluation Reward: 88.99
Training Iteration 138 Training Reward: 151.94 Evaluation Reward: 73.97
Average Evaluation Reward: 67.11
Training Iteration 139 Training Reward: 149.58 Evaluation Reward: 297.72
Average Evaluation Reward: 90.01
Training Iteration 140 Training Reward: 187.86 Evaluation Reward: 296.12
Average Evaluation Reward: 112.67
Training Iteration 141 Training Reward: 168.28 Evaluation Reward: 295.06
Average Evaluation Reward: 134.19
Training Iteration 142 Training Reward: 170.68 Evaluation Reward: 293.62
Average Evaluation Reward: 156.83
Training Iteration 143 Training Reward: 148.50 Evaluation Reward: 294.45
Average Evaluation Reward: 179.33
```

```
Training Iteration 144 Training Reward: 115.00 Evaluation Reward: 297.43
Average Evaluation Reward: 203.02
Training Iteration 145 Training Reward: 81.01 Evaluation Reward: 299.24
Average Evaluation Reward: 227.04
Training Iteration 146 Training Reward: 68.01 Evaluation Reward: 286.51
Average Evaluation Reward: 249.08
Training Iteration 147 Training Reward: 69.56 Evaluation Reward: 292.61
Average Evaluation Reward: 272.67
Training Iteration 148 Training Reward: 91.97 Evaluation Reward: 272.40
Average Evaluation Reward: 292.52
Training Iteration 149 Training Reward: 109.97 Evaluation Reward: 160.72
Average Evaluation Reward: 278.82
Training Iteration 150 Training Reward: 119.81 Evaluation Reward: 156.51
Average Evaluation Reward: 264.85
Training Iteration 151 Training Reward: 148.04 Evaluation Reward: 156.00
Average Evaluation Reward: 250.95
Training Iteration 152 Training Reward: 144.63 Evaluation Reward: 172.33
Average Evaluation Reward: 238.82
Training Iteration 153 Training Reward: 179.52 Evaluation Reward: 147.47
Average Evaluation Reward: 224.12
Training Iteration 154 Training Reward: 177.89 Evaluation Reward: 164.90
Average Evaluation Reward: 210.87
Training Iteration 155 Training Reward: 196.45 Evaluation Reward: 145.12
Average Evaluation Reward: 195.46
Training Iteration 156 Training Reward: 174.39 Evaluation Reward: 156.11
Average Evaluation Reward: 182.42
Training Iteration 157 Training Reward: 176.25 Evaluation Reward: 157.14
Average Evaluation Reward: 168.87
Training Iteration 158 Training Reward: 192.13 Evaluation Reward: 264.38
Average Evaluation Reward: 168.07
Training Iteration 159 Training Reward: 229.14 Evaluation Reward: 265.53
Average Evaluation Reward: 178.55
Training Iteration 160 Training Reward: 228.62 Evaluation Reward: 263.13
Average Evaluation Reward: 189.21
Training Iteration 161 Training Reward: 259.71 Evaluation Reward: 272.64
Average Evaluation Reward: 200.88
Training Iteration 162 Training Reward: 276.63 Evaluation Reward: 264.65
Average Evaluation Reward: 210.11
Best result for training iteration 162
Training Iteration 163 Training Reward: 275.01 Evaluation Reward: 254.68
Average Evaluation Reward: 220.83
Training Iteration 164 Training Reward: 268.98 Evaluation Reward: 260.50
Average Evaluation Reward: 230.39
Training Iteration 165 Training Reward: 266.60 Evaluation Reward: 254.46
Average Evaluation Reward: 241.32
Training Iteration 166 Training Reward: 265.43 Evaluation Reward: 259.13
Average Evaluation Reward: 251.62
Training Iteration 167 Training Reward: 267.08 Evaluation Reward: 263.29
```

```
Average Evaluation Reward: 262.24
Training Iteration 168 Training Reward: 248.03 Evaluation Reward: 261.56
Average Evaluation Reward: 261.96
Training Iteration 169 Training Reward: 218.03 Evaluation Reward: 264.25
Average Evaluation Reward: 261.83
Training Iteration 170 Training Reward: 198.03 Evaluation Reward: 266.13
Average Evaluation Reward: 262.13
Training Iteration 171 Training Reward: 221.34 Evaluation Reward: 262.58
Average Evaluation Reward: 261.12
Training Iteration 172 Training Reward: 273.93 Evaluation Reward: 252.28
Average Evaluation Reward: 259.89
Training Iteration 173 Training Reward: 267.06 Evaluation Reward: 247.87
Average Evaluation Reward: 259.20
Training Iteration 174 Training Reward: 267.45 Evaluation Reward: 250.22
Average Evaluation Reward: 258.18
Training Iteration 175 Training Reward: 268.06 Evaluation Reward: 258.70
Average Evaluation Reward: 258.60
Training Iteration 176 Training Reward: 272.88 Evaluation Reward: 268.50
Average Evaluation Reward: 259.54
Training Iteration 177 Training Reward: 270.41 Evaluation Reward: 261.95
Average Evaluation Reward: 259.40
Training Iteration 178 Training Reward: 253.43 Evaluation Reward: 273.48
Average Evaluation Reward: 260.60
Training Iteration 179 Training Reward: 182.64 Evaluation Reward: 275.66
Average Evaluation Reward: 261.74
Training Iteration 180 Training Reward: 156.54 Evaluation Reward: 275.23
Average Evaluation Reward: 262.65
Training Iteration 181 Training Reward: 166.21 Evaluation Reward: 274.06
Average Evaluation Reward: 263.79
Training Iteration 182 Training Reward: 211.93 Evaluation Reward: 268.00
Average Evaluation Reward: 265.37
Training Iteration 183 Training Reward: 248.42 Evaluation Reward: 265.20
Average Evaluation Reward: 267.10
Training Iteration 184 Training Reward: 246.49 Evaluation Reward: 263.99
Average Evaluation Reward: 268.48
Training Iteration 185 Training Reward: 246.62 Evaluation Reward: 227.66
Average Evaluation Reward: 265.37
Training Iteration 186 Training Reward: 242.68 Evaluation Reward: 62.93
Average Evaluation Reward: 244.81
Training Iteration 187 Training Reward: 210.15 Evaluation Reward: 228.67
Average Evaluation Reward: 241.49
Training Iteration 188 Training Reward: 188.18 Evaluation Reward: 221.49
Average Evaluation Reward: 236.29
Training Iteration 189 Training Reward: 187.33 Evaluation Reward: 215.69
Average Evaluation Reward: 230.29
Training Iteration 190 Training Reward: 162.12 Evaluation Reward: 261.40
Average Evaluation Reward: 228.91
Training Iteration 191 Training Reward: 187.48 Evaluation Reward: 263.61
```

```
Average Evaluation Reward: 227.86
Training Iteration 192 Training Reward: 155.54 Evaluation Reward: 272.09
Average Evaluation Reward: 228.27
Training Iteration 193 Training Reward: 198.70 Evaluation Reward: 272.84
Average Evaluation Reward: 229.04
Training Iteration 194 Training Reward: 225.62 Evaluation Reward: 261.99
Average Evaluation Reward: 228.84
Training Iteration 195 Training Reward: 237.28 Evaluation Reward: 266.52
Average Evaluation Reward: 232.72
Training Iteration 196 Training Reward: 229.88 Evaluation Reward: 269.19
Average Evaluation Reward: 253.35
Training Iteration 197 Training Reward: 239.40 Evaluation Reward: 252.15
Average Evaluation Reward: 255.70
Training Iteration 198 Training Reward: 149.92 Evaluation Reward: 253.57
Average Evaluation Reward: 258.91
Training Iteration 199 Training Reward: 147.06 Evaluation Reward: 252.95
Average Evaluation Reward: 262.63
```



THE CODE BELOW IS JUST TO EXPORT THE VIDEO AND DOES NOT TAKE PART IN THE ALGORITHM

[ ]:
```
 #For visualization
import gymnasium as gym
from gym.wrappers.monitoring import video_recorder
from IPython.display import HTML
from IPython import display
import glob
import cv2
```

VIDEO FUNCTION

[ ]:
```
def video_fn(agent, env_name, algo):
    env = gym.make(env_name, continuous = True, render_mode="rgb_array")
    fourcc = cv2.VideoWriter_fourcc(*'mp4v')
    video = cv2.VideoWriter(algo+'_video.mp4', fourcc, 30, (600, 400))
    agent.policy.load_state_dict(torch.load(algo+"_checkpoint.pth"))
    agent.policy.eval()
    state, _= env.reset()
    done = False
    while not done:
        frame = env.render()
        video.write(frame)
        state_ten = torch.from_numpy(state).float().unsqueeze(0)
        action = agent.policy.select_action(state_ten)[0].detach().numpy()
        action = action.astype(np.float64)
        n_state, reward, terminated, truncated, _ = env.step(action)
        done = terminated or truncated
        state = n_state
    env.close()
    video.release()
```

EXPORTING VIDEO

[ ]:
```
env_type = "LunarLander-v2"
env = gym.make(env_type, continuous=True)
state_dim = env.observation_space.shape[0]
action_dim = env.action_space.shape[0]
plotter_agent = PGAgent(state_dim,action_dim)
video_fn(plotter_agent, "LunarLander-v2", "Rt")
video_fn(plotter_agent, "LunarLander-v2", "Gt")
video_fn(plotter_agent, "LunarLander-v2", "Gt_with_Baseline")
```