LunarLanderDQN2025AssignmentTargetQN

March 10, 2025

0.1 Lunar lander with DQN-style neural function approximator using PyTorch

0.1.1 Christian Igel, 2025

If you have suggestions for improvement, let me know.

I took inspiration from https://github.com/udacity/deep-learning/blob/master/reinforcement/Q-learning-cart.ipynb.

Imports:

[3]:
```
import gymnasium as gym

from tqdm.notebook import tqdm  # Progress bar

import torch
import torch.nn as nn
import torch.nn.functional as F

import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
```

Create the game environment (you need the `gym` package):

[4]:
```
env_visual = gym.make('LunarLander-v3', render_mode="human")
action_size = 4
state_size = 8
```

Let's just test the environment first:

[5]:
```
test_episodes = 0
for _ in range(test_episodes):
    R = 0
    state, _ = env_visual.reset()  # Environment starts in a random state, cart␣
 ↪and pole are moving
    print("initial state:", state)
    while True:  # Environment sets "truncated" to true after 500 steps
        env_visual.render()
        state, reward, terminated, truncated, _ = env_visual.step(env_visual.
 ↪action_space.sample()) #  Take a random action
```

```
        R += reward  # Accumulate reward
        if terminated or truncated:
            print("return: ", R)
            env_visual.reset()
            break
```

[6]:
```
#env.close()  # Closes the visualization window
```

Define Q network architecture:

[7]:
```
class QNetwork(nn.Module):
    def __init__(self, state_size=8, action_size=4, hidden_size=10, bias=True):
        super(QNetwork, self).__init__()
        self.fc1 = nn.Linear(state_size, hidden_size, bias)
        self.fc2 = nn.Linear(hidden_size, hidden_size, bias)
        self.output_layer = nn.Linear(hidden_size + state_size, action_size,␣
 ↪bias)

    def forward(self, x_input):
        x = F.tanh(self.fc1(x_input))
        x = F.tanh(self.fc2(x))
        x = torch.cat((x_input, x), dim=1)
        x = self.output_layer(x)
        return x
```

Data structure for storing experiences:

[8]:
```
from collections import deque
class Memory():
    def __init__(self, max_size = 1000):
        self.buffer = deque(maxlen=max_size)

    def add(self, experience):
        self.buffer.append(experience)

    def sample(self, batch_size):
        idx = np.random.choice(np.arange(len(self.buffer)),
                               size=batch_size,
                               replace=False)
        return [self.buffer[ii] for ii in idx]
```

Define basic constants:

[9]:
```
train_episodes = 400          # Max number of episodes to learn from
gamma = 0.99                  # Future reward discount
learning_rate = 0.001         # Q-network learning rate
tau = .01                     # learning rate for target network

# Exploration parameters
```

```
explore_start = 1.0            # Exploration probability at start
explore_stop = 0.0001          # Minimum exploration probability
decay_rate = 0.05              # Exponential decay rate for exploration prob

# Network parameters
hidden_size = 64               # Number of units in each Q-network hidden layer

# Memory parameters
memory_size = 10000            # Memory capacity
batch_size = 128               # Experience mini-batch size
pretrain_length = batch_size   # Number experiences to pretrain the memory

log_path = "/tmp/deep_Q_network"
```

Instantiate network:

[10]:
```
mainQN = QNetwork(hidden_size=hidden_size)

# Create the target network
targetQN = QNetwork(hidden_size=hidden_size)

# Copy parameters from mainQN to targetQN so they start identical
targetQN.load_state_dict(mainQN.state_dict())

print(mainQN)
```

```
QNetwork(
  (fc1): Linear(in_features=8, out_features=64, bias=True)
  (fc2): Linear(in_features=64, out_features=64, bias=True)
  (output_layer): Linear(in_features=72, out_features=4, bias=True)
)
```

Initialize the experience memory:

[11]:
```
# Initialize the simulation
env = gym.make('LunarLander-v3')
state = env.reset()[0]

memory = Memory(max_size=memory_size)

# Make a bunch of random actions and store the experiences
for _ in tqdm(range(pretrain_length)):
    # Make a random action
    action = env.action_space.sample()
    next_state, reward, terminated, truncated, _ = env.step(action)

    if terminated or truncated:
        # The simulation fails, so no next state
        next_state = np.zeros(state.shape)
```

```
        # Add experience to memory
        memory.add((state, action, reward, next_state))

        # Start new episode
        env.reset()
        # Take one random step to get the pole and cart moving
        state, reward, terminated, truncated, _ = env.step(env.action_space.
 ↪sample())
    else:
        # Add experience to memory
        memory.add((state, action, reward, next_state))
        state = next_state
```

```
  0%|          | 0/128 [00:00<?, ?it/s]
```

Now train with experiences:

[12]:
```
total_reward_list = []  # Returns for the individual episodes

optimizer = torch.optim.AdamW(mainQN.parameters(), lr=learning_rate) # AdamW␣
 ↪uses weight decay by default
loss_fn = torch.nn.MSELoss()

for ep in range(train_episodes):
    total_reward = 0  # Return / accumulated rewards
    state = env.reset()[0]  # Reset and get initial state
    while True:
        # Explore or exploit
        explore_p = explore_stop + (explore_start - explore_stop)*np.
 ↪exp(-decay_rate*ep)
        if explore_p > np.random.rand():
            # Pick a random action
            action = env.action_space.sample()
        else:
            # Get action from Q-network
            state_tensor = torch.from_numpy(np.resize(state, (1, state_size)).
 ↪astype(np.float32))
            Qs = mainQN(state_tensor)
            action = torch.argmax(Qs).item()

        # Take action, get new state and reward
        next_state, reward, terminated, truncated, _ = env.step(action)

        total_reward += reward  # Return / accumulated rewards

        if terminated or truncated:
            # Episode ends because of failure, so no next state
            next_state = np.zeros(state.shape)
```

```
            print('Episode: {}'.format(ep), 'Total reward: {}'.
↪format(total_reward),
                  'Training loss: {:.4f}'.format(loss), 'Explore P: {:.4f}'.
↪format(explore_p))
            total_reward_list.append((ep, total_reward))

            # Add experience to memory
            memory.add((state, action, reward, next_state))
            break; # End of episode
        else:
            # Add experience to memory
            memory.add((state, action, reward, next_state))
            state = next_state

        # Sample mini-batch from memory
        batch = memory.sample(batch_size)
        next_states_np = np.array([each[3] for each in batch], dtype=np.float32)
        next_states = torch.as_tensor(next_states_np)  # as_tensor does not␣
↪copy the data
        rewards     = torch.as_tensor(np.array([each[2] for each in batch],␣
↪dtype=np.float32))
        states      = torch.as_tensor(np.array([each[0] for each in batch],␣
↪dtype=np.float32))
        actions     = torch.as_tensor(np.array([each[1] for each in batch]))

        # Compute Q values for all actions in the new state
        target_Qs = mainQN(next_states)

        # Set target_Qs to 0 for states where episode ended because of failure
        episode_ends = (next_states_np == np.zeros(states[0].shape)).all(axis=1)
        target_Qs[episode_ends] = torch.zeros(action_size)

        # Compute targets
        max_elements = torch.max(target_Qs, dim=1)[0].detach()
        y = rewards + gamma * max_elements

        # Network learning starts here
        optimizer.zero_grad()

        # Compute the Q values of the actions taken
        main_Qs = mainQN(states)  # Q values for all action in each state
        Q = torch.gather(main_Qs, 1, actions.unsqueeze(-1)).squeeze()  # Only␣
↪the Q values for the actions taken

        # Gradient-based update
```

```
        loss = loss_fn(Q, y)
        loss.backward()
        optimizer.step()

        with torch.no_grad():
            for target_param, main_param in zip(targetQN.parameters(), mainQN.
↪parameters()):
                target_param.data.copy_(tau * main_param.data + (1.0 - tau) *␣
↪target_param.data)
```

```
Episode: 0 Total reward: -99.3205000879208 Training loss: 88.0973 Explore P:
1.0000
Episode: 1 Total reward: -174.97402225880785 Training loss: 73.8131 Explore P:
0.9512
Episode: 2 Total reward: -191.13501635517017 Training loss: 21.7698 Explore P:
0.9048
Episode: 3 Total reward: -178.01023777367948 Training loss: 127.4457 Explore P:
0.8607
Episode: 4 Total reward: -225.6772079352046 Training loss: 4.5616 Explore P:
0.8187
Episode: 5 Total reward: -266.15643705513963 Training loss: 168.2151 Explore P:
0.7788
Episode: 6 Total reward: -208.56963371240914 Training loss: 21.6141 Explore P:
0.7408
Episode: 7 Total reward: -70.3224054125693 Training loss: 68.9120 Explore P:
0.7047
Episode: 8 Total reward: -161.5716629171195 Training loss: 33.2243 Explore P:
0.6704
Episode: 9 Total reward: -243.7142076101546 Training loss: 80.3784 Explore P:
0.6377
Episode: 10 Total reward: -67.4640453950085 Training loss: 15.7446 Explore P:
0.6066
Episode: 11 Total reward: -221.72851875242992 Training loss: 77.9527 Explore P:
0.5770
Episode: 12 Total reward: -265.1078480635706 Training loss: 63.3333 Explore P:
0.5489
Episode: 13 Total reward: -403.65428876484134 Training loss: 91.2110 Explore P:
0.5221
Episode: 14 Total reward: -199.55494491383678 Training loss: 29.9264 Explore P:
0.4966
Episode: 15 Total reward: -181.0886093903519 Training loss: 175.0772 Explore P:
0.4724
Episode: 16 Total reward: -138.7362185218033 Training loss: 110.2616 Explore P:
0.4494
Episode: 17 Total reward: -179.4988533924493 Training loss: 109.8987 Explore P:
0.4275
Episode: 18 Total reward: -223.15320900528314 Training loss: 43.5677 Explore P:
0.4066
```

```
Episode: 19 Total reward: -245.34981645962895 Training loss: 30.7327 Explore P:
0.3868
Episode: 20 Total reward: -406.47716848265947 Training loss: 135.3825 Explore P:
0.3679
Episode: 21 Total reward: -424.485553974266 Training loss: 38.0095 Explore P:
0.3500
Episode: 22 Total reward: -222.78730808696304 Training loss: 60.3646 Explore P:
0.3329
Episode: 23 Total reward: -301.03854058577843 Training loss: 2.8499 Explore P:
0.3167
Episode: 24 Total reward: -332.2014725700492 Training loss: 56.5470 Explore P:
0.3013
Episode: 25 Total reward: -127.55980305339911 Training loss: 6.2772 Explore P:
0.2866
Episode: 26 Total reward: -18.281569410231885 Training loss: 9.8822 Explore P:
0.2726
Episode: 27 Total reward: -313.6084867327977 Training loss: 177.6756 Explore P:
0.2593
Episode: 28 Total reward: -199.4950041069768 Training loss: 14.6729 Explore P:
0.2467
Episode: 29 Total reward: -154.00718029810653 Training loss: 3.5560 Explore P:
0.2346
Episode: 30 Total reward: -113.1365368488953 Training loss: 10.1284 Explore P:
0.2232
Episode: 31 Total reward: -104.2545703798711 Training loss: 10.2227 Explore P:
0.2123
Episode: 32 Total reward: 66.44487224371315 Training loss: 2.8354 Explore P:
0.2020
Episode: 33 Total reward: -29.32413427194996 Training loss: 10.2297 Explore P:
0.1921
Episode: 34 Total reward: -82.45687192884245 Training loss: 3.7988 Explore P:
0.1828
Episode: 35 Total reward: -87.63121702080454 Training loss: 24.8240 Explore P:
0.1739
Episode: 36 Total reward: 181.3589829931115 Training loss: 4.2851 Explore P:
0.1654
Episode: 37 Total reward: -52.29801903490669 Training loss: 7.8421 Explore P:
0.1573
Episode: 38 Total reward: 130.7038817274038 Training loss: 3.9514 Explore P:
0.1497
Episode: 39 Total reward: 159.6978301617588 Training loss: 3.0151 Explore P:
0.1424
Episode: 40 Total reward: -52.186145082643165 Training loss: 1.8133 Explore P:
0.1354
Episode: 41 Total reward: 29.900746076402672 Training loss: 2.1694 Explore P:
0.1288
Episode: 42 Total reward: -27.268286108344867 Training loss: 1.5494 Explore P:
0.1225
```

```
Episode: 43 Total reward: -602.2349499694593 Training loss: 4.6306 Explore P:
0.1166
Episode: 44 Total reward: -56.230893535221014 Training loss: 2.0848 Explore P:
0.1109
Episode: 45 Total reward: -194.16019384890438 Training loss: 1.3630 Explore P:
0.1055
Episode: 46 Total reward: 94.81696843807458 Training loss: 2.0675 Explore P:
0.1003
Episode: 47 Total reward: -8.402194417497702 Training loss: 3.4430 Explore P:
0.0955
Episode: 48 Total reward: -69.61291927927128 Training loss: 6.5633 Explore P:
0.0908
Episode: 49 Total reward: 0.05951035590910103 Training loss: 2.1521 Explore P:
0.0864
Episode: 50 Total reward: 11.562176001302209 Training loss: 2.1256 Explore P:
0.0822
Episode: 51 Total reward: -55.94745215294644 Training loss: 2.5156 Explore P:
0.0782
Episode: 52 Total reward: -67.77990116639035 Training loss: 51.9763 Explore P:
0.0744
Episode: 53 Total reward: -54.696166849919464 Training loss: 1.6026 Explore P:
0.0707
Episode: 54 Total reward: -1.698992856540945 Training loss: 1.7345 Explore P:
0.0673
Episode: 55 Total reward: 133.0535497982109 Training loss: 2.1156 Explore P:
0.0640
Episode: 56 Total reward: -92.07061532306727 Training loss: 1.3483 Explore P:
0.0609
Episode: 57 Total reward: 99.78392772020182 Training loss: 4.8307 Explore P:
0.0579
Episode: 58 Total reward: -664.6746173215191 Training loss: 1.7339 Explore P:
0.0551
Episode: 59 Total reward: -208.21574826016567 Training loss: 4.6588 Explore P:
0.0524
Episode: 60 Total reward: -200.13947214798202 Training loss: 3.5679 Explore P:
0.0499
Episode: 61 Total reward: -143.48618444101402 Training loss: 3.3446 Explore P:
0.0475
Episode: 62 Total reward: -202.6157896504074 Training loss: 26.2821 Explore P:
0.0451
Episode: 63 Total reward: -76.73281887279019 Training loss: 1.8294 Explore P:
0.0429
Episode: 64 Total reward: 4.891795201081686 Training loss: 1.3979 Explore P:
0.0409
Episode: 65 Total reward: 47.63769568735007 Training loss: 53.2458 Explore P:
0.0389
Episode: 66 Total reward: 99.30690053819286 Training loss: 4.2525 Explore P:
0.0370
```

```
Episode: 67 Total reward: 170.75905126034556 Training loss: 1.5458 Explore P:
0.0352
Episode: 68 Total reward: -285.02023367732727 Training loss: 2.0020 Explore P:
0.0335
Episode: 69 Total reward: -27.059433109800427 Training loss: 1.3402 Explore P:
0.0318
Episode: 70 Total reward: -114.09100667606171 Training loss: 1.3902 Explore P:
0.0303
Episode: 71 Total reward: 112.20700612752631 Training loss: 0.7770 Explore P:
0.0288
Episode: 72 Total reward: 188.16954642671828 Training loss: 1.2896 Explore P:
0.0274
Episode: 73 Total reward: -39.527897947358795 Training loss: 3.3595 Explore P:
0.0261
Episode: 74 Total reward: -107.42919811152629 Training loss: 4.4846 Explore P:
0.0248
Episode: 75 Total reward: -95.82313817643667 Training loss: 1.3219 Explore P:
0.0236
Episode: 76 Total reward: 93.80583535067993 Training loss: 6.4797 Explore P:
0.0225
Episode: 77 Total reward: -67.60765633995841 Training loss: 1.3378 Explore P:
0.0214
Episode: 78 Total reward: -278.68090682764273 Training loss: 1.1103 Explore P:
0.0203
Episode: 79 Total reward: -276.6416311761851 Training loss: 1.6123 Explore P:
0.0194
Episode: 80 Total reward: -167.11673818136185 Training loss: 2.1955 Explore P:
0.0184
Episode: 81 Total reward: -31.831422885914595 Training loss: 1.3455 Explore P:
0.0175
Episode: 82 Total reward: 163.84659104597114 Training loss: 1.8581 Explore P:
0.0167
Episode: 83 Total reward: 206.71030529540874 Training loss: 3.1397 Explore P:
0.0159
Episode: 84 Total reward: 82.09740662313395 Training loss: 2.3832 Explore P:
0.0151
Episode: 85 Total reward: 171.89705442045687 Training loss: 6.5808 Explore P:
0.0144
Episode: 86 Total reward: 177.11855396793035 Training loss: 74.1993 Explore P:
0.0137
Episode: 87 Total reward: 233.32588918554353 Training loss: 4.4357 Explore P:
0.0130
Episode: 88 Total reward: 177.50435025543175 Training loss: 1.4219 Explore P:
0.0124
Episode: 89 Total reward: -88.48404580939037 Training loss: 1.5051 Explore P:
0.0118
Episode: 90 Total reward: -192.38022009145288 Training loss: 8.0373 Explore P:
0.0112
```

```
Episode: 91 Total reward: 12.379681608088305 Training loss: 0.8093 Explore P:
0.0107
Episode: 92 Total reward: -203.64275810410618 Training loss: 1.5400 Explore P:
0.0102
Episode: 93 Total reward: 272.67397057521475 Training loss: 1.6843 Explore P:
0.0097
Episode: 94 Total reward: 108.48855583457883 Training loss: 3.1423 Explore P:
0.0092
Episode: 95 Total reward: -741.5707324871958 Training loss: 1.9751 Explore P:
0.0088
Episode: 96 Total reward: 269.9898851244594 Training loss: 10.0387 Explore P:
0.0083
Episode: 97 Total reward: 226.05240112066383 Training loss: 1.0313 Explore P:
0.0079
Episode: 98 Total reward: -54.8637032061568 Training loss: 1.6919 Explore P:
0.0075
Episode: 99 Total reward: -64.56315900792188 Training loss: 1.6324 Explore P:
0.0072
Episode: 100 Total reward: 112.46333595622889 Training loss: 1.3880 Explore P:
0.0068
Episode: 101 Total reward: 244.65721125430264 Training loss: 1.1033 Explore P:
0.0065
Episode: 102 Total reward: 160.6996086160575 Training loss: 4.9663 Explore P:
0.0062
Episode: 103 Total reward: -26.286221660267998 Training loss: 12.4195 Explore P:
0.0059
Episode: 104 Total reward: 232.4834752217259 Training loss: 1.4710 Explore P:
0.0056
Episode: 105 Total reward: -54.842813212511444 Training loss: 7.4157 Explore P:
0.0053
Episode: 106 Total reward: -431.9321269350536 Training loss: 5.2000 Explore P:
0.0051
Episode: 107 Total reward: -483.9090755054134 Training loss: 77.1109 Explore P:
0.0048
Episode: 108 Total reward: -88.16999912385661 Training loss: 2.7994 Explore P:
0.0046
Episode: 109 Total reward: -370.66981379089077 Training loss: 6.2837 Explore P:
0.0044
Episode: 110 Total reward: 181.22332417769806 Training loss: 1.9045 Explore P:
0.0042
Episode: 111 Total reward: -431.1200252935068 Training loss: 7.0552 Explore P:
0.0040
Episode: 112 Total reward: -121.33920994774896 Training loss: 24.7386 Explore P:
0.0038
Episode: 113 Total reward: 75.31322439572295 Training loss: 4.8282 Explore P:
0.0036
Episode: 114 Total reward: 12.88878573873414 Training loss: 2.1527 Explore P:
0.0034
```

```
Episode: 115 Total reward: 137.40268565940795 Training loss: 5.0030 Explore P:
0.0033
Episode: 116 Total reward: 120.64426776468572 Training loss: 3.0331 Explore P:
0.0031
Episode: 117 Total reward: 241.47167144207464 Training loss: 3.4919 Explore P:
0.0030
Episode: 118 Total reward: 145.48800114236622 Training loss: 4.3347 Explore P:
0.0028
Episode: 119 Total reward: -215.9691910543108 Training loss: 141.9843 Explore P:
0.0027
Episode: 120 Total reward: -107.44235946111591 Training loss: 3.3497 Explore P:
0.0026
Episode: 121 Total reward: -204.5046955146639 Training loss: 3.1848 Explore P:
0.0025
Episode: 122 Total reward: -429.07384811080766 Training loss: 7.3606 Explore P:
0.0023
Episode: 123 Total reward: -89.05392153064486 Training loss: 11.7691 Explore P:
0.0022
Episode: 124 Total reward: -121.85259680772528 Training loss: 3.3330 Explore P:
0.0021
Episode: 125 Total reward: -219.36349031765388 Training loss: 1.5840 Explore P:
0.0020
Episode: 126 Total reward: -302.88395094826353 Training loss: 3.3537 Explore P:
0.0019
Episode: 127 Total reward: 23.7938726556871 Training loss: 7.3625 Explore P:
0.0018
Episode: 128 Total reward: 19.27195700831947 Training loss: 2.6710 Explore P:
0.0018
Episode: 129 Total reward: 200.0597228075644 Training loss: 2.9930 Explore P:
0.0017
Episode: 130 Total reward: 205.66366923694102 Training loss: 2.7651 Explore P:
0.0016
Episode: 131 Total reward: -195.7830513247801 Training loss: 1.8318 Explore P:
0.0015
Episode: 132 Total reward: 185.87284231553843 Training loss: 20.8895 Explore P:
0.0015
Episode: 133 Total reward: 157.65723909171405 Training loss: 1.6288 Explore P:
0.0014
Episode: 134 Total reward: 54.84679300320624 Training loss: 1.4475 Explore P:
0.0013
Episode: 135 Total reward: -118.28780657010728 Training loss: 1.2961 Explore P:
0.0013
Episode: 136 Total reward: 158.68601661944814 Training loss: 4.7535 Explore P:
0.0012
Episode: 137 Total reward: 146.56838588149364 Training loss: 1.5946 Explore P:
0.0012
Episode: 138 Total reward: 7.613179635792793 Training loss: 1.2309 Explore P:
0.0011
```

```
Episode: 139 Total reward: 138.5658876829045 Training loss: 1.0102 Explore P:
0.0011
Episode: 140 Total reward: 5.819623645406111 Training loss: 1.0446 Explore P:
0.0010
Episode: 141 Total reward: 163.63750613877875 Training loss: 1.0342 Explore P:
0.0010
Episode: 142 Total reward: 223.23236529787965 Training loss: 1.2581 Explore P:
0.0009
Episode: 143 Total reward: 16.105185407539565 Training loss: 7.8521 Explore P:
0.0009
Episode: 144 Total reward: 233.18869855609896 Training loss: 1.2109 Explore P:
0.0008
Episode: 145 Total reward: 149.61588124370894 Training loss: 1.1638 Explore P:
0.0008
Episode: 146 Total reward: 221.75869798372906 Training loss: 1.4399 Explore P:
0.0008
Episode: 147 Total reward: -411.99134058706665 Training loss: 1.2371 Explore P:
0.0007
Episode: 148 Total reward: -97.01749799522888 Training loss: 21.5071 Explore P:
0.0007
Episode: 149 Total reward: -76.63441841722609 Training loss: 4.4617 Explore P:
0.0007
Episode: 150 Total reward: -73.47619545204594 Training loss: 6.5750 Explore P:
0.0007
Episode: 151 Total reward: -197.32406919458754 Training loss: 1.4401 Explore P:
0.0006
Episode: 152 Total reward: 231.43182968823726 Training loss: 1.6933 Explore P:
0.0006
Episode: 153 Total reward: 222.75519275629904 Training loss: 1.6045 Explore P:
0.0006
Episode: 154 Total reward: 210.24068139561447 Training loss: 3.8488 Explore P:
0.0006
Episode: 155 Total reward: -62.212222957692276 Training loss: 1.5850 Explore P:
0.0005
Episode: 156 Total reward: 210.02142056798925 Training loss: 1.8676 Explore P:
0.0005
Episode: 157 Total reward: 196.48345302979342 Training loss: 0.9607 Explore P:
0.0005
Episode: 158 Total reward: 204.20223652383612 Training loss: 1.1472 Explore P:
0.0005
Episode: 159 Total reward: 116.76102612258582 Training loss: 1.2740 Explore P:
0.0005
Episode: 160 Total reward: 246.9796620910533 Training loss: 2.5608 Explore P:
0.0004
Episode: 161 Total reward: -164.8236162071709 Training loss: 14.4820 Explore P:
0.0004
Episode: 162 Total reward: -121.00869244958689 Training loss: 1.4976 Explore P:
0.0004
```

```
Episode: 163 Total reward: 161.8994990962083 Training loss: 13.0123 Explore P:
0.0004
Episode: 164 Total reward: -47.12036755141432 Training loss: 1.1857 Explore P:
0.0004
Episode: 165 Total reward: 282.18361331468407 Training loss: 1.9352 Explore P:
0.0004
Episode: 166 Total reward: 205.30872968228948 Training loss: 10.7570 Explore P:
0.0003
Episode: 167 Total reward: 245.34018417025746 Training loss: 4.0711 Explore P:
0.0003
Episode: 168 Total reward: -255.97025353265653 Training loss: 3.1504 Explore P:
0.0003
Episode: 169 Total reward: 258.82687230547424 Training loss: 7.1575 Explore P:
0.0003
Episode: 170 Total reward: 213.05737677065014 Training loss: 1.8050 Explore P:
0.0003
Episode: 171 Total reward: 205.5254459895649 Training loss: 3.5386 Explore P:
0.0003
Episode: 172 Total reward: 174.24936684273953 Training loss: 2.0152 Explore P:
0.0003
Episode: 173 Total reward: -75.05422284872667 Training loss: 4.9090 Explore P:
0.0003
Episode: 174 Total reward: 189.39983949016533 Training loss: 9.0757 Explore P:
0.0003
Episode: 175 Total reward: 263.17551504263986 Training loss: 4.0356 Explore P:
0.0003
Episode: 176 Total reward: -60.32246134383756 Training loss: 3.6174 Explore P:
0.0003
Episode: 177 Total reward: 196.64951109690446 Training loss: 2.4705 Explore P:
0.0002
Episode: 178 Total reward: 164.17257185734258 Training loss: 5.2858 Explore P:
0.0002
Episode: 179 Total reward: 168.19431793285918 Training loss: 2.5379 Explore P:
0.0002
Episode: 180 Total reward: -345.8741749019421 Training loss: 6.7882 Explore P:
0.0002
Episode: 181 Total reward: 185.82873060175626 Training loss: 1.9628 Explore P:
0.0002
Episode: 182 Total reward: 167.47217443743196 Training loss: 3.6819 Explore P:
0.0002
Episode: 183 Total reward: -79.18702638410093 Training loss: 3.5407 Explore P:
0.0002
Episode: 184 Total reward: 235.50323263173095 Training loss: 1.9849 Explore P:
0.0002
Episode: 185 Total reward: -176.92038225645288 Training loss: 3.9100 Explore P:
0.0002
Episode: 186 Total reward: -120.65616556878614 Training loss: 2.6765 Explore P:
0.0002
```

```
Episode: 187 Total reward: 188.1069994607206 Training loss: 2.6595 Explore P:
0.0002
Episode: 188 Total reward: -116.50951526999417 Training loss: 5.9877 Explore P:
0.0002
Episode: 189 Total reward: 257.1942439269071 Training loss: 5.9695 Explore P:
0.0002
Episode: 190 Total reward: 187.45340211064695 Training loss: 2.4114 Explore P:
0.0002
Episode: 191 Total reward: 211.00739571131078 Training loss: 1.5688 Explore P:
0.0002
Episode: 192 Total reward: 221.76189659987534 Training loss: 3.3348 Explore P:
0.0002
Episode: 193 Total reward: 206.64496849884776 Training loss: 9.1871 Explore P:
0.0002
Episode: 194 Total reward: 221.85797120969607 Training loss: 1.2723 Explore P:
0.0002
Episode: 195 Total reward: 276.94513559812725 Training loss: 1.3300 Explore P:
0.0002
Episode: 196 Total reward: 177.09707538777246 Training loss: 1.1113 Explore P:
0.0002
Episode: 197 Total reward: 240.9396457240758 Training loss: 4.0852 Explore P:
0.0002
Episode: 198 Total reward: 173.90901705221324 Training loss: 2.1049 Explore P:
0.0002
Episode: 199 Total reward: 190.00204613315134 Training loss: 8.5564 Explore P:
0.0001
Episode: 200 Total reward: 274.0444620870257 Training loss: 1.5760 Explore P:
0.0001
Episode: 201 Total reward: 238.63199021482748 Training loss: 2.0376 Explore P:
0.0001
Episode: 202 Total reward: 253.4453798488251 Training loss: 3.5689 Explore P:
0.0001
Episode: 203 Total reward: 246.55371392640012 Training loss: 1.4801 Explore P:
0.0001
Episode: 204 Total reward: 207.9406988957774 Training loss: 1.9302 Explore P:
0.0001
Episode: 205 Total reward: -192.5441828405518 Training loss: 5.9883 Explore P:
0.0001
Episode: 206 Total reward: 254.06051268131193 Training loss: 1.7535 Explore P:
0.0001
Episode: 207 Total reward: 243.17873344312153 Training loss: 4.0580 Explore P:
0.0001
Episode: 208 Total reward: 277.9798971088567 Training loss: 1.1754 Explore P:
0.0001
Episode: 209 Total reward: 237.99386545031476 Training loss: 1.3212 Explore P:
0.0001
Episode: 210 Total reward: 221.05617356520096 Training loss: 9.0670 Explore P:
0.0001
```

```
Episode: 211 Total reward: 251.1225451292891 Training loss: 2.2464 Explore P:
0.0001
Episode: 212 Total reward: 218.96687061802675 Training loss: 1.3737 Explore P:
0.0001
Episode: 213 Total reward: 269.84142041327306 Training loss: 6.6544 Explore P:
0.0001
Episode: 214 Total reward: 275.8714126177596 Training loss: 1.1443 Explore P:
0.0001
Episode: 215 Total reward: 231.3548635191621 Training loss: 1.2933 Explore P:
0.0001
Episode: 216 Total reward: 249.63642346428279 Training loss: 1.1196 Explore P:
0.0001
Episode: 217 Total reward: 245.325650431378 Training loss: 1.6487 Explore P:
0.0001
Episode: 218 Total reward: 276.4656474070481 Training loss: 1.4843 Explore P:
0.0001
Episode: 219 Total reward: 254.62516416741525 Training loss: 1.9077 Explore P:
0.0001
Episode: 220 Total reward: 269.6244517736296 Training loss: 0.6520 Explore P:
0.0001
Episode: 221 Total reward: 258.9000242458684 Training loss: 1.3239 Explore P:
0.0001
Episode: 222 Total reward: 251.60726939021703 Training loss: 0.9626 Explore P:
0.0001
Episode: 223 Total reward: 254.2364523904847 Training loss: 10.2774 Explore P:
0.0001
Episode: 224 Total reward: 271.35751384743185 Training loss: 7.1167 Explore P:
0.0001
Episode: 225 Total reward: 72.00878595462302 Training loss: 5.1862 Explore P:
0.0001
Episode: 226 Total reward: 258.4530931508294 Training loss: 1.1945 Explore P:
0.0001
Episode: 227 Total reward: 267.55801373341535 Training loss: 0.7207 Explore P:
0.0001
Episode: 228 Total reward: 237.70010402106627 Training loss: 9.9630 Explore P:
0.0001
Episode: 229 Total reward: 271.0849485450037 Training loss: 9.4462 Explore P:
0.0001
Episode: 230 Total reward: -11.736906111612498 Training loss: 8.0166 Explore P:
0.0001
Episode: 231 Total reward: -69.17382220458917 Training loss: 11.5228 Explore P:
0.0001
Episode: 232 Total reward: 277.7771981703849 Training loss: 2.4255 Explore P:
0.0001
Episode: 233 Total reward: 241.6966057731724 Training loss: 1.1564 Explore P:
0.0001
Episode: 234 Total reward: 231.16660385040512 Training loss: 2.2492 Explore P:
0.0001
```

```
Episode: 235 Total reward: 239.45884820906696 Training loss: 52.1738 Explore P:
0.0001
Episode: 236 Total reward: 229.35370948017967 Training loss: 15.3353 Explore P:
0.0001
Episode: 237 Total reward: 275.9096819678492 Training loss: 1.3062 Explore P:
0.0001
Episode: 238 Total reward: -34.96363449456132 Training loss: 20.0375 Explore P:
0.0001
Episode: 239 Total reward: 12.977851690058202 Training loss: 0.8049 Explore P:
0.0001
Episode: 240 Total reward: -0.6868531485155387 Training loss: 12.0924 Explore P:
0.0001
Episode: 241 Total reward: -94.86196760630018 Training loss: 15.0152 Explore P:
0.0001
Episode: 242 Total reward: 215.25624811872132 Training loss: 1.3048 Explore P:
0.0001
Episode: 243 Total reward: 209.4481304913113 Training loss: 15.8735 Explore P:
0.0001
Episode: 244 Total reward: 178.34539957644873 Training loss: 3.1427 Explore P:
0.0001
Episode: 245 Total reward: 244.96518534038154 Training loss: 5.8856 Explore P:
0.0001
Episode: 246 Total reward: -89.51696272096837 Training loss: 93.1966 Explore P:
0.0001
Episode: 247 Total reward: 247.82252269755304 Training loss: 2.5840 Explore P:
0.0001
Episode: 248 Total reward: 233.95543196086066 Training loss: 3.1383 Explore P:
0.0001
Episode: 249 Total reward: -395.46330216112693 Training loss: 2.7266 Explore P:
0.0001
Episode: 250 Total reward: 255.27858548807097 Training loss: 7.4672 Explore P:
0.0001
Episode: 251 Total reward: -23.44873513901888 Training loss: 22.0879 Explore P:
0.0001
Episode: 252 Total reward: -569.9820773702794 Training loss: 4.0458 Explore P:
0.0001
Episode: 253 Total reward: 215.66423926832783 Training loss: 4.3860 Explore P:
0.0001
Episode: 254 Total reward: -257.82527346771656 Training loss: 7.2625 Explore P:
0.0001
Episode: 255 Total reward: 36.523225275574305 Training loss: 19.6815 Explore P:
0.0001
Episode: 256 Total reward: -14.983538625305329 Training loss: 5.2571 Explore P:
0.0001
Episode: 257 Total reward: -200.90251348475002 Training loss: 34.6068 Explore P:
0.0001
Episode: 258 Total reward: 145.39902834911308 Training loss: 15.8608 Explore P:
0.0001
```

```
Episode: 259 Total reward: 195.8828186884452 Training loss: 17.2173 Explore P:
0.0001
Episode: 260 Total reward: 24.05234419232471 Training loss: 3.4388 Explore P:
0.0001
Episode: 261 Total reward: 256.9350629919278 Training loss: 2.5316 Explore P:
0.0001
Episode: 262 Total reward: -450.9463842089181 Training loss: 2.1945 Explore P:
0.0001
Episode: 263 Total reward: 267.15600015361684 Training loss: 5.9098 Explore P:
0.0001
Episode: 264 Total reward: 249.54618184062593 Training loss: 11.3953 Explore P:
0.0001
Episode: 265 Total reward: 233.51552563932225 Training loss: 14.4534 Explore P:
0.0001
Episode: 266 Total reward: 226.08691279532928 Training loss: 22.8484 Explore P:
0.0001
Episode: 267 Total reward: 229.23921424011937 Training loss: 3.4318 Explore P:
0.0001
Episode: 268 Total reward: 252.00037898950237 Training loss: 6.8074 Explore P:
0.0001
Episode: 269 Total reward: 220.64551800530418 Training loss: 2.0934 Explore P:
0.0001
Episode: 270 Total reward: 272.33986840685 Training loss: 8.5890 Explore P:
0.0001
Episode: 271 Total reward: 282.75398750074714 Training loss: 5.3516 Explore P:
0.0001
Episode: 272 Total reward: -412.4627459959237 Training loss: 114.7426 Explore P:
0.0001
Episode: 273 Total reward: 241.95349495516658 Training loss: 16.0005 Explore P:
0.0001
Episode: 274 Total reward: 265.49148743690637 Training loss: 2.7493 Explore P:
0.0001
Episode: 275 Total reward: 261.92682958367055 Training loss: 6.9450 Explore P:
0.0001
Episode: 276 Total reward: 260.15850085131615 Training loss: 4.8292 Explore P:
0.0001
Episode: 277 Total reward: 250.7816731833316 Training loss: 11.6070 Explore P:
0.0001
Episode: 278 Total reward: 217.77594336146046 Training loss: 8.1114 Explore P:
0.0001
Episode: 279 Total reward: -152.40803322555493 Training loss: 86.6231 Explore P:
0.0001
Episode: 280 Total reward: 220.80144108304466 Training loss: 2.7710 Explore P:
0.0001
Episode: 281 Total reward: 260.4485958966537 Training loss: 7.0382 Explore P:
0.0001
Episode: 282 Total reward: 229.27010098020543 Training loss: 2.4211 Explore P:
0.0001
```

```
Episode: 283 Total reward: 246.9829731349422 Training loss: 5.5663 Explore P:
0.0001
Episode: 284 Total reward: 244.21564743521452 Training loss: 2.4724 Explore P:
0.0001
Episode: 285 Total reward: 147.449583300308 Training loss: 21.0961 Explore P:
0.0001
Episode: 286 Total reward: 248.46337839881255 Training loss: 12.7227 Explore P:
0.0001
Episode: 287 Total reward: 231.71436278834386 Training loss: 1.4302 Explore P:
0.0001
Episode: 288 Total reward: 274.33016887073813 Training loss: 0.8810 Explore P:
0.0001
Episode: 289 Total reward: 269.32330522234764 Training loss: 1.1267 Explore P:
0.0001
Episode: 290 Total reward: 199.07323232986636 Training loss: 3.5271 Explore P:
0.0001
Episode: 291 Total reward: 254.33867415975482 Training loss: 2.1258 Explore P:
0.0001
Episode: 292 Total reward: 253.17698337741362 Training loss: 12.2865 Explore P:
0.0001
Episode: 293 Total reward: 229.6827838246529 Training loss: 11.8773 Explore P:
0.0001
Episode: 294 Total reward: 254.93210860234038 Training loss: 11.9193 Explore P:
0.0001
Episode: 295 Total reward: 234.32554587482738 Training loss: 0.5915 Explore P:
0.0001
Episode: 296 Total reward: 296.904817568648 Training loss: 14.7847 Explore P:
0.0001
Episode: 297 Total reward: 253.67089233231096 Training loss: 1.9609 Explore P:
0.0001
Episode: 298 Total reward: 263.1975467761156 Training loss: 0.6871 Explore P:
0.0001
Episode: 299 Total reward: 259.10467288269035 Training loss: 7.4594 Explore P:
0.0001
Episode: 300 Total reward: 262.0150676708927 Training loss: 7.3230 Explore P:
0.0001
Episode: 301 Total reward: 266.0414654900704 Training loss: 0.8709 Explore P:
0.0001
Episode: 302 Total reward: 227.47606602368435 Training loss: 0.7496 Explore P:
0.0001
Episode: 303 Total reward: 279.36967049454955 Training loss: 1.2695 Explore P:
0.0001
Episode: 304 Total reward: 233.51072869904255 Training loss: 1.6641 Explore P:
0.0001
Episode: 305 Total reward: 275.40739787898667 Training loss: 1.7036 Explore P:
0.0001
Episode: 306 Total reward: 274.3312131772245 Training loss: 25.8481 Explore P:
0.0001
```

```
Episode: 307 Total reward: 233.5501398793249 Training loss: 0.8159 Explore P:
0.0001
Episode: 308 Total reward: 267.5468619054561 Training loss: 7.8934 Explore P:
0.0001
Episode: 309 Total reward: 264.91258956904426 Training loss: 12.0504 Explore P:
0.0001
Episode: 310 Total reward: 252.48802379582241 Training loss: 0.6327 Explore P:
0.0001
Episode: 311 Total reward: 253.01925745549175 Training loss: 7.8930 Explore P:
0.0001
Episode: 312 Total reward: 262.8660894027262 Training loss: 0.8833 Explore P:
0.0001
Episode: 313 Total reward: 255.1864137587241 Training loss: 0.4481 Explore P:
0.0001
Episode: 314 Total reward: 269.2313510909232 Training loss: 0.5430 Explore P:
0.0001
Episode: 315 Total reward: 255.05781036221518 Training loss: 4.7640 Explore P:
0.0001
Episode: 316 Total reward: 246.53213017023887 Training loss: 7.1784 Explore P:
0.0001
Episode: 317 Total reward: 241.51664875578416 Training loss: 1.3758 Explore P:
0.0001
Episode: 318 Total reward: 193.68703586039248 Training loss: 0.7801 Explore P:
0.0001
Episode: 319 Total reward: 224.95988278055688 Training loss: 0.6442 Explore P:
0.0001
Episode: 320 Total reward: 126.0784418100647 Training loss: 2.2479 Explore P:
0.0001
Episode: 321 Total reward: -108.60931663191437 Training loss: 1.3427 Explore P:
0.0001
Episode: 322 Total reward: -4.588097449299767 Training loss: 7.3586 Explore P:
0.0001
Episode: 323 Total reward: 197.9165470921635 Training loss: 3.6309 Explore P:
0.0001
Episode: 324 Total reward: 272.99159472669953 Training loss: 0.6857 Explore P:
0.0001
Episode: 325 Total reward: 196.77343921287508 Training loss: 8.0300 Explore P:
0.0001
Episode: 326 Total reward: -3.7028840743166427 Training loss: 1.7359 Explore P:
0.0001
Episode: 327 Total reward: 206.04358749621665 Training loss: 1.8277 Explore P:
0.0001
Episode: 328 Total reward: 248.34869163838624 Training loss: 5.2861 Explore P:
0.0001
Episode: 329 Total reward: 151.3033671548344 Training loss: 4.6834 Explore P:
0.0001
Episode: 330 Total reward: -26.627492911992334 Training loss: 0.8595 Explore P:
0.0001
```

```
Episode: 331 Total reward: 248.88349724577805 Training loss: 1.4682 Explore P:
0.0001
Episode: 332 Total reward: -16.858819732335846 Training loss: 1.5018 Explore P:
0.0001
Episode: 333 Total reward: 264.8150538919199 Training loss: 7.2505 Explore P:
0.0001
Episode: 334 Total reward: 266.81103258924986 Training loss: 1.0719 Explore P:
0.0001
Episode: 335 Total reward: 259.87141001847937 Training loss: 1.9001 Explore P:
0.0001
Episode: 336 Total reward: 252.44816969566577 Training loss: 1.4307 Explore P:
0.0001
Episode: 337 Total reward: 263.820286136258 Training loss: 1.7552 Explore P:
0.0001
Episode: 338 Total reward: 282.41432280999703 Training loss: 1.1994 Explore P:
0.0001
Episode: 339 Total reward: 230.62534785733425 Training loss: 0.8416 Explore P:
0.0001
Episode: 340 Total reward: 262.61505715219744 Training loss: 1.1503 Explore P:
0.0001
Episode: 341 Total reward: -93.79160625609413 Training loss: 13.5436 Explore P:
0.0001
Episode: 342 Total reward: 39.473354461930796 Training loss: 12.7334 Explore P:
0.0001
Episode: 343 Total reward: 288.193754553021 Training loss: 2.6583 Explore P:
0.0001
Episode: 344 Total reward: -278.6108899899184 Training loss: 2.0790 Explore P:
0.0001
Episode: 345 Total reward: 288.0052612756879 Training loss: 2.1431 Explore P:
0.0001
Episode: 346 Total reward: -0.0409133582990755 Training loss: 4.4240 Explore P:
0.0001
Episode: 347 Total reward: 173.17194669237878 Training loss: 1.4826 Explore P:
0.0001
Episode: 348 Total reward: 236.59728721822165 Training loss: 2.6669 Explore P:
0.0001
Episode: 349 Total reward: 243.66776357221087 Training loss: 2.1357 Explore P:
0.0001
Episode: 350 Total reward: 95.87947271640797 Training loss: 10.7268 Explore P:
0.0001
Episode: 351 Total reward: 229.473626214104 Training loss: 4.4834 Explore P:
0.0001
Episode: 352 Total reward: 261.9310039656809 Training loss: 5.6154 Explore P:
0.0001
Episode: 353 Total reward: 266.8971513345224 Training loss: 1.4621 Explore P:
0.0001
Episode: 354 Total reward: 239.761353210154 Training loss: 1.2936 Explore P:
0.0001
```

```
Episode: 355 Total reward: 261.23814372379445 Training loss: 4.8797 Explore P:
0.0001
Episode: 356 Total reward: 234.8897728071561 Training loss: 1.7139 Explore P:
0.0001
Episode: 357 Total reward: 227.1756473239734 Training loss: 1.1237 Explore P:
0.0001
Episode: 358 Total reward: 269.2071521641926 Training loss: 1.8087 Explore P:
0.0001
Episode: 359 Total reward: 183.0310257411043 Training loss: 2.1013 Explore P:
0.0001
Episode: 360 Total reward: 231.3392441697526 Training loss: 138.6706 Explore P:
0.0001
Episode: 361 Total reward: 266.2040527488431 Training loss: 1.8545 Explore P:
0.0001
Episode: 362 Total reward: 263.96719063535477 Training loss: 2.8067 Explore P:
0.0001
Episode: 363 Total reward: 263.2754568296598 Training loss: 3.1508 Explore P:
0.0001
Episode: 364 Total reward: 252.91911342641356 Training loss: 1.7644 Explore P:
0.0001
Episode: 365 Total reward: 204.1076150294901 Training loss: 10.2854 Explore P:
0.0001
Episode: 366 Total reward: -63.28796113234537 Training loss: 2.4092 Explore P:
0.0001
Episode: 367 Total reward: 244.1187421213331 Training loss: 5.4250 Explore P:
0.0001
Episode: 368 Total reward: -72.1550919200429 Training loss: 10.2924 Explore P:
0.0001
Episode: 369 Total reward: -35.86856638032536 Training loss: 2.5871 Explore P:
0.0001
Episode: 370 Total reward: -61.409840853017315 Training loss: 2.1836 Explore P:
0.0001
Episode: 371 Total reward: 268.01320953574157 Training loss: 24.4353 Explore P:
0.0001
Episode: 372 Total reward: 253.7945476288443 Training loss: 7.3869 Explore P:
0.0001
Episode: 373 Total reward: 257.1640416760673 Training loss: 7.5469 Explore P:
0.0001
Episode: 374 Total reward: 276.87202684386295 Training loss: 1.7318 Explore P:
0.0001
Episode: 375 Total reward: 256.0048501900179 Training loss: 3.5284 Explore P:
0.0001
Episode: 376 Total reward: 275.61807680133154 Training loss: 31.5344 Explore P:
0.0001
Episode: 377 Total reward: 248.30028469894768 Training loss: 2.9925 Explore P:
0.0001
Episode: 378 Total reward: -22.041430372012414 Training loss: 1.8270 Explore P:
0.0001
```

```
Episode: 379 Total reward: 206.42291798375084 Training loss: 6.3292 Explore P:
0.0001
Episode: 380 Total reward: 259.50065070354333 Training loss: 109.4701 Explore P:
0.0001
Episode: 381 Total reward: 286.08405624767965 Training loss: 1.3179 Explore P:
0.0001
Episode: 382 Total reward: 284.5752111524892 Training loss: 1.8853 Explore P:
0.0001
Episode: 383 Total reward: 291.8822049467291 Training loss: 5.4500 Explore P:
0.0001
Episode: 384 Total reward: 195.65470282811626 Training loss: 3.3716 Explore P:
0.0001
Episode: 385 Total reward: 206.3283038720566 Training loss: 6.3473 Explore P:
0.0001
Episode: 386 Total reward: 263.3842235534667 Training loss: 2.4152 Explore P:
0.0001
Episode: 387 Total reward: 6.1151244233702755 Training loss: 11.6051 Explore P:
0.0001
Episode: 388 Total reward: 278.4022596215427 Training loss: 7.2675 Explore P:
0.0001
Episode: 389 Total reward: 238.18163479146918 Training loss: 3.0193 Explore P:
0.0001
Episode: 390 Total reward: 192.32840502829248 Training loss: 10.6258 Explore P:
0.0001
Episode: 391 Total reward: 230.29230258859394 Training loss: 6.1398 Explore P:
0.0001
Episode: 392 Total reward: -150.7856146830696 Training loss: 31.8688 Explore P:
0.0001
Episode: 393 Total reward: 250.27534616522416 Training loss: 8.2621 Explore P:
0.0001
Episode: 394 Total reward: 260.52514552242997 Training loss: 155.2674 Explore P:
0.0001
Episode: 395 Total reward: 193.0895500745516 Training loss: 44.0661 Explore P:
0.0001
Episode: 396 Total reward: 289.3816909767333 Training loss: 1.6421 Explore P:
0.0001
Episode: 397 Total reward: 228.0353752371612 Training loss: 46.6632 Explore P:
0.0001
Episode: 398 Total reward: 217.4726903678871 Training loss: 21.5769 Explore P:
0.0001
Episode: 399 Total reward: -28.086123583428204 Training loss: 20.1241 Explore P:
0.0001
```

Save policy network:

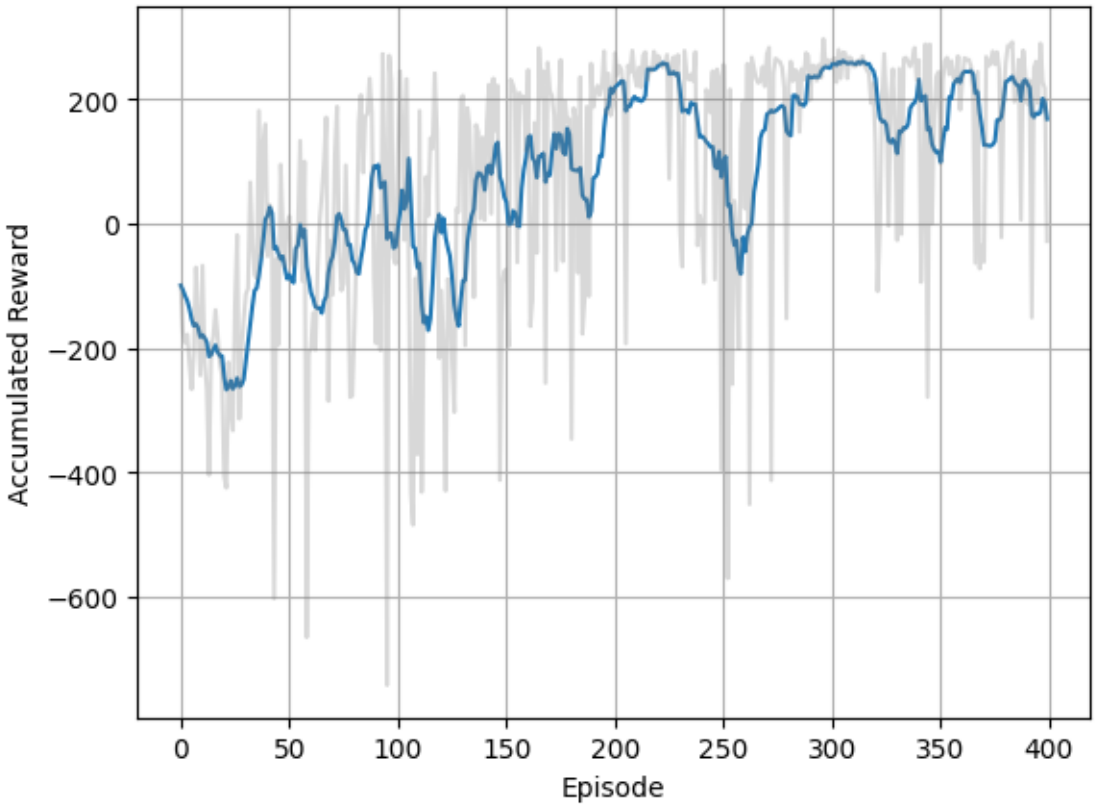
[13]:
```
torch.save(mainQN, log_path)
```

Plot learning process:

[14]:
```
# Moving average for smoothing plot
def running_mean(x, N):
    cumsum = np.cumsum(np.insert(x, 0, x[0]*np.ones(N)))
    return (cumsum[N:] - cumsum[:-N]) / N

eps, rews = np.array(total_reward_list).T
smoothed_rews = running_mean(rews, 10)

plt.plot(eps, smoothed_rews)
plt.grid()
plt.plot(eps, rews, color='grey', alpha=0.3)
plt.xlabel('Episode')
plt.ylabel('Accumulated Reward')
plt.savefig('deepQ.pdf')
```



Evaluate stored policy:

[15]:
```
testQN = torch.load(log_path, weights_only=False)

test_episodes = 5
```

```
for ep in range(test_episodes):
    state = env_visual.reset()[0]
    print("initial state:", state)
    R = 0
    while True:
        # Get action from Q-network
        # Hm, the following line could perhaps be more elegant ...
        state_tensor = torch.from_numpy(np.resize(state, (1, state_size)).
 ↪astype(np.float32))
        Qs = testQN(state_tensor)
        action = torch.argmax(Qs).item()

        # Take action, get new state and reward
        next_state, reward, terminated, truncated, _ = env_visual.step(action)
        R += reward

        if terminated or truncated:
            print("reward:", R)
            break
        else:
            state = next_state
```

```
initial state: [ 1.3621331e-03  1.4192797e+00  1.3795212e-01  3.7153926e-01
 -1.5715744e-03 -3.1248260e-02  0.0000000e+00  0.0000000e+00]
reward: 192.39073100689671
initial state: [ 0.00471525  1.415199    0.47759208  0.19016033 -0.00545705
-0.10818183
  0.          0.        ]
reward: 230.9202474194195
initial state: [ 0.00724916  1.4106017   0.7342621  -0.01416572 -0.00839332
-0.16632129
  0.          0.        ]
reward: 227.1218682752121
initial state: [-5.5055617e-04  1.4104651e+00 -5.5771094e-02 -2.0226300e-02
  6.4463663e-04  1.2632990e-02  0.0000000e+00  0.0000000e+00]
reward: 107.97261482022796
initial state: [-0.00430498  1.4099395  -0.436064   -0.04358765  0.0049952
0.09877495
  0.          0.        ]
reward: 273.38600920969657
```

[ ]: