IERG 5350 Assignment 4: Advanced Algorithms for Continuous Control in RL

Welcome to assignment 4 of our RL course!

2020-2021 Term 1, IERG 5350: Reinforcement Learning. Department of Information Engineering, The Chinese University of Hong Kong. Course Instructor: Professor ZHOU Bolei. Assignment author: PENG Zhenghao, SUN Hao, ZHAN Xiaohang.

Student Name	Student ID
Wanli Wang	1155160517

In this assignment, we will implement a system of RL that allows us to train and evaluate RL agents formally and efficiently.

In this notebook, you will go through the following components of the whole system:

- Preparation: Colab, and Environment
- Section 1: Training with algorithm PPO
- Section 2: Training with algorithm DDPG
- Section 3: Training with algorithm TD3
- Section 4: Transfer your PPO/ DDPG/ TD3 to another task: Four-Solution-Maze

The author of this assignment is SUN, Hao (sh018 AT ie.cuhk.edu.hk).

Colab

Introduction to Google Colab:

From now on, our assignment as well as the final project will be based on the Google Colab, where you can apply for free GPU resources to accelerate the learning of your RL models.

Here are some resources as intro to the Colab.

- YouTube Video: https://www.youtube.com/watch?v=inN8seMm7UI (https://www.youtube.com/watch?v=inN8seMm7UI)
- Colab Intro: https://colab.research.google.com/notebooks/intro.ipynb (https://colab.research.google.com/notebooks/intro.ipynb) (you may need to login with your google account)

Gym Continuous Control Tasks

Introduction to the Gym Continuous Control Envirionments

In the last assignment, you have already used the gym[atari] benchmarks, where the action space is discrete so that normal approach is value-based methods e.g., DQN.

In this assignment, we will try to implement three prevailing RL algorithms for continuous control tasks, namely the PPO(https://arxiv.org/abs/1707.06347 (https://arxiv.org/abs/1707.06347)), DDPG(https://arxiv.org/abs/1509.02971 (https://arxiv.org/abs/1509.02971)) and TD3(https://arxiv.org/abs/1802.09477 (https://arxiv.org/abs/1802.09477)).

We will now begin with a gym environment for continuous control,

The Pendulum-v0

In [13]:

```
import gym
ENV_NAME = "Pendulum-v0"
env = gym.make(ENV_NAME)
state = env.reset()
print('the state space is like', state)
print('the max and min action is: ',env.action_space.high,env.action_space.low)

'''so that you may need to use action value re-size if you want to use the tanh activation functions'''
```

```
the state space is like [-0.25874714 -0.96594509 -0.12174252]
the max and min action is:  [2.] [-2.]
```

Out[13]:

```
'so that you may need to use action value re-size if you want to use the tanh acti
vation functions'
```

PPO

The Proximal Policy Optimization Algorithms is the most prevailing on-policy learning method. Although its sample efficiency is not as high as the off-policy methods, the PPO is relatively easy to implement and the learning is much more stable than off-policy methods. Whenever you have a task you want to try whether RL works, you may try to run a PPO agent at first. It is worth mentioning even the most challenging game, the StarCraftII agent AlphaStar is trained based on PPO (with lots of improvements, ofcourse).

TODOs for You

The ppo has the benfitsof trust region policy optimization (TRPO) but is much simpler to implement, and with some implementation engeneering, the sample complexity of TRPO is further improved.

The key idea of PPO optimization is *Not Optimize the Policy Too Much in a Certain Step*, which follows the key insight of the method of TRPO.

In TRPO, the optimization objective of policy is to learn a policy such that

$$\max_{\theta} \hat{\mathbb{E}}_t [\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \hat{A}_t]$$

subject to

$$\hat{\mathbb{E}}_t [KL[\pi_{\theta_{old}}(\cdot|s_t), \pi_\theta(\cdot|s_t)]] \leq \delta$$

where $\hat{A}$ denotes the advantage function, rather than optimize the objective function of

$$L^{PG}(\theta) = \hat{\mathbb{E}}_t [\log \pi_\theta(a_t|s_t) \hat{A}_t]$$

in the normal policy gradint methods.

The PPO proposed two alternative approaches to solve the constrained optimization above, namely the Clipped Surrogated Objective and the Adaptive KL penalty Coefficient. The former one is more generally used in practice as it's more convenient to implement, more efficient and owns stable performance.

The Clipped Surrogated Objective approach replace the surrogate objective

$$L^{CPI}(\theta) = \hat{\mathbb{E}}_t [\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \hat{A}_t] = \hat{\mathbb{E}}_t [r_t(\theta) \hat{A}_t]$$

of TRPO (CPI: Conservative Policy Iteration) by

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t [\min(r_t(\theta) \hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon) \hat{A}_t)]$$

You can check that $L^{CLIP}(\theta) = L^{CPI}(\theta)$ around the old policy parameter θ_{old}, i.e., when r = 1.

TODOs here:

In this section, your task is to finish the code of a PPO algorithm and evaluate its performance in the Pendulum-v0 environment.

Specifically, you need to

- Q1. finish building up the ActorCritic "__init__" function, i.e., build up the neural network.
- Q2. finish the foward function, in this part, there are two functions need to finish: the _forward_actor function and the _forward_critic function

- Q3. finish the select_action function, which is called during interacting with the environment, so that you may need to return an action as well as the (log-)probability of getting that action for future optimization
- Q4. finish the optimization steps for your PPO agent, that means you need to build up the surrogate loss through your saved tuples in previous episodes and optimize it with current network parameters.

In [14]:

```
# You need not to rivese this unless you want to try other hyper-parameter settings
# in which case you may revise the default values of class args()
from IPython import display
import torch
import torch.nn as nn
import torch.optim as opt
from torch import Tensor
from torch.autograd import Variable
from collections import namedtuple
from itertools import count
import torch.nn.functional as F
import matplotlib.pyplot as plt
from os.path import join as joindir
from os import makedirs as mkdir
import pandas as pd
import numpy as np
import argparse
import datetime
import math
import random

Transition = namedtuple('Transition', ('state', 'value', 'action', 'logproba', 'mask', 'next_sta
te', 'reward'))
env = gym.make(ENV_NAME)
env.reset()

EPS = 1e-10 # you may need this tiny value somewhere, and think about why?
RESULT_DIR = 'Result_PPO'
mkdir(RESULT_DIR, exist_ok=True)
mkdir(ENV_NAME.split('-')[0]+'/CheckPoints', exist_ok=True)
mkdir(ENV_NAME.split('-')[0]+'/Rwds', exist_ok=True)
rwds = []
rwds_history = []

class args(object):
    hid_num = 256
    drop_prob = 0.1
    env_name = ENV_NAME
    seed = 1234
    num_episode = 1000
    batch_size = 5120
    max_step_per_round = 2000
    gamma = 0.995
    lamda = 0.97
    log_num_episode = 1
    num_epoch = 10
    minibatch_size = 256
    clip = 0.2
    loss_coeff_value = 0.5
    loss_coeff_entropy = 0.01
    lr = 3e-4
    num_parallel_run = 1
    # tricks
    schedule_adam = 'linear'
    schedule_clip = 'linear'
    layer_norm = True
    state_norm = False
    advantage_norm = True
    lossvalue_norm = True
```

In [15]:

```
# You need not to rivese this, these classes are used for normalization
class RunningStat(object):
    def __init__(self, shape):
        self._n = 0
        self._M = np.zeros(shape)
        self._S = np.zeros(shape)

    def push(self, x):
        x = np.asarray(x)
        assert x.shape == self._M.shape
        self._n += 1
        if self._n == 1:
            self._M[...] = x
        else:
            oldM = self._M.copy()
            self._M[...] = oldM + (x - oldM) / self._n
            self._S[...] = self._S + (x - oldM) * (x - self._M)

    @property
    def n(self):
        return self._n

    @property
    def mean(self):
        return self._M

    @property
    def var(self):
        return self._S / (self._n - 1) if self._n > 1 else np.square(self._M)

    @property
    def std(self):
        return np.sqrt(self.var)

    @property
    def shape(self):
        return self._M.shape


class ZFilter:
    """
    y = (x-mean)/std
    using running estimates of mean,std
    """

    def __init__(self, shape, demean=True, destd=True, clip=10.0):
        self.demean = demean
        self.destd = destd
        self.clip = clip

        self.rs = RunningStat(shape)

    def __call__(self, x, update=True):
        if update: self.rs.push(x)
        if self.demean:
            x = x - self.rs.mean
        if self.destd:
            x = x / (self.rs.std + 1e-8)
        if self.clip:
```

```
            x = np.clip(x, -self.clip, self.clip)
        return x

    def output_shape(self, input_space):
        return input_space.shape
```

In [16]:

```
# Here, you need to finish the first 5 tasks.
class ActorCritic(nn.Module):
    def __init__(self, num_inputs, num_outputs, layer_norm=True):
        super(ActorCritic, self).__init__()
        '''
        Q1:
        Initialize your networks
        '''
        self.actor_fc1 = nn.Linear(num_inputs, 64)
        self.actor_fc2 = nn.Linear(64, 64)
        self.actor_fc3 = nn.Linear(64, num_outputs)
        self.actor_logstd = nn.Parameter(torch.zeros(1, num_outputs))

        self.critic_fc1 = nn.Linear(num_inputs, 64)
        self.critic_fc2 = nn.Linear(64, 64)
        self.critic_fc3 = nn.Linear(64, 1)

        if layer_norm:
            self.layer_norm(self.actor_fc1, std=1.0)
            self.layer_norm(self.actor_fc2, std=1.0)
            self.layer_norm(self.actor_fc3, std=0.01)

            self.layer_norm(self.critic_fc1, std=1.0)
            self.layer_norm(self.critic_fc2, std=1.0)
            self.layer_norm(self.critic_fc3, std=1.0)

    @staticmethod
    def layer_norm(layer, std=1.0, bias_const=0.0):
        torch.nn.init.orthogonal_(layer.weight, std)
        torch.nn.init.constant_(layer.bias, bias_const)

    def forward(self, states):
        """
        Q2.1:
        run policy network (actor) as well as value network (critic)
        :param states: a tensor represents states
        :return: 3 Tensor2
        your _forward_actor() function should return both the mean value of action and the log-s
tandard deviation of the action
        """

        action_mean, action_logstd = self._forward_actor(states)
        critic_value = self._forward_critic(states)
        return action_mean, action_logstd, critic_value

    def _forward_actor(self, states):
        '''
        Q2.2:
        build something like
        x = activation (actor_fc(state))
        the logstd output has already been provided
        '''
        x = torch.tanh(self.actor_fc1(states))
        x = torch.tanh(self.actor_fc2(x))
        action_mean = self.actor_fc3(x)
        action_logstd = self.actor_logstd.expand_as(action_mean)
        return action_mean, action_logstd

    def _forward_critic(self, states):
```

```
        '''
        Q2.3:
        build something like
        x = activation (critic_fc(state))'''

        x = torch.tanh(self.critic_fc1(states))
        x = torch.tanh(self.critic_fc2(x))
        critic_value = self.critic_fc3(x)
        return critic_value

    def select_action(self, action_mean, action_logstd, return_logproba=True):
        """
        Q3.1:
        given mean and std, sample an action from normal(mean, std)
        also returns probability of the given chosen
        """
        action_std = torch.exp(action_logstd)
        action = torch.normal(action_mean, action_std)
        if return_logproba:
            logproba = self._normal_logproba(action, action_mean, action_logstd, action_std)
        return action, logproba

    @staticmethod
    def _normal_logproba(x, mean, logstd, std=None):
        '''
        Q3.2:
        given a mean and logstd of a gaussian,
        calculate the log-probability of a given x'''

        if std is None:
            std = torch.exp(logstd)

        std_sq = std.pow(2)
        logproba = - 0.5 * math.log(2 * math.pi) - logstd - (x - mean).pow(2) / (2 * std_sq)
        return logproba.sum(1)

    def get_logproba(self, states, actions):
        """
        return probability of chosen the given actions under corresponding states of current net
work
        :param states: Tensor
        :param actions: Tensor
        """
        action_mean, action_logstd = self._forward_actor(states)
        action_mean = action_mean.cpu()
        action_logstd = action_logstd.cpu()
        logproba = self._normal_logproba(actions, action_mean, action_logstd)
        return logproba


class Memory(object):
    def __init__(self):
        self.memory = []

    def push(self, *args):
        self.memory.append(Transition(*args))

    def sample(self):
        return Transition(*zip(*self.memory))

    def __len__(self):
```

```
        return len(self.memory)

env = gym.make(ENV_NAME)
num_inputs = env.observation_space.shape[0]
num_actions = env.action_space.shape[0]
network = ActorCritic(num_inputs, num_actions, layer_norm=args.layer_norm)
network.train()
def ppo(args):
    env = gym.make(args.env_name)
    num_inputs = env.observation_space.shape[0]
    num_actions = env.action_space.shape[0]

    env.seed(args.seed)
    torch.manual_seed(args.seed)

    #network = ActorCritic(num_inputs, num_actions, layer_norm=args.layer_norm)
    optimizer = opt.Adam(network.parameters(), lr=args.lr)

    running_state = ZFilter((num_inputs,), clip=5.0)

    # record average 1-round cumulative reward in every episode
    reward_record = []
    global_steps = 0

    lr_now = args.lr
    clip_now = args.clip

    for i_episode in range(args.num_episode):
        # step1: perform current policy to collect trajectories
        # this is an on-policy method!
        memory = Memory()
        num_steps = 0
        reward_list = []
        len_list = []
        while num_steps < args.batch_size:
            state = env.reset()
            if args.state_norm:
                state = running_state(state)
            reward_sum = 0
            for t in range(args.max_step_per_round):
                action_mean, action_logstd, value = network(Tensor(state).unsqueeze(0))
                action, logproba = network.select_action(action_mean, action_logstd)
                action = action.cpu().data.numpy()[0]
                logproba = logproba.cpu().data.numpy()[0]
                next_state, reward, done, _ = env.step(action)

                reward_sum += reward
                if args.state_norm:
                    next_state = running_state(next_state)
                mask = 0 if done else 1

                memory.push(state, value, action, logproba, mask, next_state, reward)

                if done:
                    break

                state = next_state

            num_steps += (t + 1)
            global_steps += (t + 1)
            reward_list.append(reward_sum)
```

```
            len_list.append(t + 1)
        reward_record.append({
            'episode': i_episode,
            'steps': global_steps,
            'meanepreward': np.mean(reward_list),
            'meaneplen': np.mean(len_list)})
        rwds.extend(reward_list)
        batch = memory.sample()
        batch_size = len(memory)

        # step2: extract variables from trajectories
        rewards = Tensor(batch.reward)
        values = Tensor(batch.value)
        masks = Tensor(batch.mask)
        actions = Tensor(batch.action)
        states = Tensor(batch.state)
        oldlogproba = Tensor(batch.logproba)

        returns = Tensor(batch_size)
        deltas = Tensor(batch_size)
        advantages = Tensor(batch_size)

        prev_return = 0
        prev_value = 0
        prev_advantage = 0
        for i in reversed(range(batch_size)):
            returns[i] = rewards[i] + args.gamma * prev_return * masks[i]
            deltas[i] = rewards[i] + args.gamma * prev_value * masks[i] - values[i]
            # ref: https://arxiv.org/pdf/1506.02438.pdf (generalization advantage estimate)
            advantages[i] = deltas[i] + args.gamma * args.lamda * prev_advantage * masks[i]

            prev_return = returns[i]
            prev_value = values[i]
            prev_advantage = advantages[i]
        if args.advantage_norm:
            advantages = (advantages - advantages.mean()) / (advantages.std() + EPS)

        for i_epoch in range(int(args.num_epoch * batch_size / args.minibatch_size)):
            # sample from current batch
            minibatch_ind = np.random.choice(batch_size, args.minibatch_size, replace=False)
            minibatch_states = states[minibatch_ind]
            minibatch_actions = actions[minibatch_ind]
            minibatch_oldlogproba = oldlogproba[minibatch_ind]
            minibatch_newlogproba = network.get_logproba(minibatch_states, minibatch_actions)
            minibatch_advantages = advantages[minibatch_ind]
            minibatch_returns = returns[minibatch_ind]
            minibatch_newvalues = network._forward_critic(minibatch_states).flatten()



            '''
            Q4:

            HERE:
            now you have the advantages, and log-probabilities (both pi_new and pi_old)
            you need to do optimization according to the CLIP loss

            '''
            ratio =  torch.exp(minibatch_newlogproba - minibatch_oldlogproba)
            surr1 = ratio * minibatch_advantages
            surr2 = ratio.clamp(1 - clip_now, 1 + clip_now) * minibatch_advantages
```

```
            loss_surr = - torch.mean(torch.min(surr1, surr2))

            if args.lossvalue_norm:
                minibatch_return_6std = 6 * minibatch_returns.std()
                loss_value = torch.mean((minibatch_newvalues - minibatch_returns).pow(2)) / mini
batch_return_6std
            else:
                loss_value = torch.mean((minibatch_newvalues - minibatch_returns).pow(2))

            loss_entropy = torch.mean(torch.exp(minibatch_newlogproba) * minibatch_newlogproba)

            total_loss = loss_surr + args.loss_coeff_value * loss_value + args.loss_coeff_entrop
y * loss_entropy
            optimizer.zero_grad()
            total_loss.backward()
            optimizer.step()

        if args.schedule_clip == 'linear':
            ep_ratio = 1 - (i_episode / args.num_episode)
            clip_now = args.clip * ep_ratio

        if args.schedule_adam == 'linear':
            ep_ratio = 1 - (i_episode / args.num_episode)
            lr_now = args.lr * ep_ratio
            for g in optimizer.param_groups:
                g['lr'] = lr_now

        if i_episode % args.log_num_episode == 0:
            print('Finished episode: {} Reward: {:.4f} total_loss = {:.4f} = {:.4f} + {} *
{:.4f} + {} * {:.4f}' \
                .format(i_episode, reward_record[-1]['meanepreward'], total_loss.data, loss_surr
.data, args.loss_coeff_value,
                        loss_value.data, args.loss_coeff_entropy, loss_entropy.data))
            print('------------------')

    return reward_record

def test(args):
    record_dfs = []
    for i in range(args.num_parallel_run):
        args.seed += 1
        reward_record = pd.DataFrame(ppo(args))
        reward_record['#parallel_run'] = i
        record_dfs.append(reward_record)
    record_dfs = pd.concat(record_dfs, axis=0)
    record_dfs.to_csv(joindir(RESULT_DIR, 'ppo-record-{}.csv'.format(args.env_name)))

if __name__ == '__main__':
    for envname in [ENV_NAME]:
        args.env_name = envname
        test(args)

#torch.save(network.state_dict(),args.env_name.split('-')[0]+'/CheckPoints/checkpoint_new_{0}hid
den_{1}drop_prob_{2}repeat'.format(args.hid_num,config.drop_prob,repeat))
#np.savetxt(args.env_name.split('-')[0]+'/Rwds/rwds_new_{0}hidden_{1}drop_prob_{2}repeat'.format
(args.hid_num,config.drop_prob,repeat),rwds)
```

```
Finished episode: 0 Reward: -1241.0188 total_loss = 93.8050 = 0.0296 + 0.5 * 187.5
574 + 0.01 * -0.3292
-----------------
Finished episode: 1 Reward: -1224.5668 total_loss = 76.6035 = -0.1211 + 0.5 * 153.
4558 + 0.01 * -0.3230
-----------------
Finished episode: 2 Reward: -1254.0797 total_loss = 90.7226 = 0.0793 + 0.5 * 181.2
933 + 0.01 * -0.3356
-----------------
Finished episode: 3 Reward: -1233.3350 total_loss = 76.7953 = -0.0861 + 0.5 * 153.
7695 + 0.01 * -0.3289
-----------------
Finished episode: 4 Reward: -1240.8401 total_loss = 91.7659 = 0.1497 + 0.5 * 183.2
390 + 0.01 * -0.3306
-----------------
Finished episode: 5 Reward: -1217.0088 total_loss = 76.1486 = 0.0149 + 0.5 * 152.2
741 + 0.01 * -0.3287
-----------------
Finished episode: 6 Reward: -1163.9988 total_loss = 79.3931 = 0.1055 + 0.5 * 158.5
819 + 0.01 * -0.3340
-----------------
Finished episode: 7 Reward: -1125.8915 total_loss = 66.0185 = -0.0210 + 0.5 * 132.
0856 + 0.01 * -0.3299
-----------------
Finished episode: 8 Reward: -1119.4932 total_loss = 66.3297 = 0.0390 + 0.5 * 132.5
877 + 0.01 * -0.3192
-----------------
Finished episode: 9 Reward: -1190.5774 total_loss = 75.1554 = 0.0800 + 0.5 * 150.1
576 + 0.01 * -0.3321
-----------------
Finished episode: 10 Reward: -1194.5046 total_loss = 71.4228 = 0.0379 + 0.5 * 142.
7763 + 0.01 * -0.3276
-----------------
Finished episode: 11 Reward: -1078.3222 total_loss = 66.4062 = 0.0183 + 0.5 * 132.
7825 + 0.01 * -0.3345
-----------------
Finished episode: 12 Reward: -1195.3488 total_loss = 62.1268 = -0.1172 + 0.5 * 12
4.4946 + 0.01 * -0.3271
-----------------
Finished episode: 13 Reward: -1162.1914 total_loss = 62.5841 = 0.0048 + 0.5 * 125.
1652 + 0.01 * -0.3306
-----------------
Finished episode: 14 Reward: -1129.7535 total_loss = 62.2287 = 0.0558 + 0.5 * 124.
3524 + 0.01 * -0.3315
-----------------
Finished episode: 15 Reward: -1148.5334 total_loss = 65.6847 = 0.1166 + 0.5 * 131.
1428 + 0.01 * -0.3290
-----------------
Finished episode: 16 Reward: -1119.6907 total_loss = 59.2672 = -0.0039 + 0.5 * 11
8.5490 + 0.01 * -0.3356
-----------------
Finished episode: 17 Reward: -1195.2228 total_loss = 66.2078 = 0.0676 + 0.5 * 132.
2871 + 0.01 * -0.3286
-----------------
Finished episode: 18 Reward: -1090.5022 total_loss = 55.3381 = 0.0076 + 0.5 * 110.
6677 + 0.01 * -0.3332
-----------------
Finished episode: 19 Reward: -1156.0991 total_loss = 52.9419 = -0.0924 + 0.5 * 10
6.0753 + 0.01 * -0.3345
-----------------
Finished episode: 20 Reward: -1092.4346 total_loss = 53.9828 = 0.0021 + 0.5 * 107.
```

```
9680 + 0.01 * -0.3305
----------------
Finished episode: 21 Reward: -1054.3504 total_loss = 51.7754 = 0.0319 + 0.5 * 103.
4937 + 0.01 * -0.3336
----------------
Finished episode: 22 Reward: -1120.1745 total_loss = 50.4124 = -0.0709 + 0.5 * 10
0.9731 + 0.01 * -0.3340
----------------
Finished episode: 23 Reward: -1140.7857 total_loss = 51.2343 = -0.0524 + 0.5 * 10
2.5800 + 0.01 * -0.3277
----------------
Finished episode: 24 Reward: -1080.5187 total_loss = 49.7447 = -0.0182 + 0.5 * 99.
5324 + 0.01 * -0.3311
----------------
Finished episode: 25 Reward: -1063.3737 total_loss = 49.1654 = -0.0533 + 0.5 * 98.
4441 + 0.01 * -0.3355
----------------
Finished episode: 26 Reward: -1067.5558 total_loss = 42.6288 = -0.0965 + 0.5 * 85.
4572 + 0.01 * -0.3330
----------------
Finished episode: 27 Reward: -1106.4913 total_loss = 44.9831 = -0.0495 + 0.5 * 90.
0717 + 0.01 * -0.3338
----------------
Finished episode: 28 Reward: -1063.5406 total_loss = 47.4453 = 0.0205 + 0.5 * 94.8
564 + 0.01 * -0.3373
----------------
Finished episode: 29 Reward: -1054.0328 total_loss = 44.8415 = 0.0763 + 0.5 * 89.5
372 + 0.01 * -0.3347
----------------
Finished episode: 30 Reward: -1041.3125 total_loss = 38.9531 = -0.0152 + 0.5 * 77.
9434 + 0.01 * -0.3371
----------------
Finished episode: 31 Reward: -1020.6483 total_loss = 34.5583 = -0.1061 + 0.5 * 69.
3354 + 0.01 * -0.3314
----------------
Finished episode: 32 Reward: -1009.4946 total_loss = 41.0179 = 0.0830 + 0.5 * 81.8
766 + 0.01 * -0.3348
----------------
Finished episode: 33 Reward: -1042.1476 total_loss = 38.5662 = -0.0159 + 0.5 * 77.
1707 + 0.01 * -0.3290
----------------
Finished episode: 34 Reward: -1022.5686 total_loss = 39.0521 = -0.0411 + 0.5 * 78.
1931 + 0.01 * -0.3363
----------------
Finished episode: 35 Reward: -1031.7049 total_loss = 44.0403 = 0.0922 + 0.5 * 87.9
031 + 0.01 * -0.3419
----------------
Finished episode: 36 Reward: -1061.6280 total_loss = 34.1204 = -0.0928 + 0.5 * 68.
4331 + 0.01 * -0.3320
----------------
Finished episode: 37 Reward: -1068.7876 total_loss = 35.9801 = -0.0352 + 0.5 * 72.
0372 + 0.01 * -0.3337
----------------
Finished episode: 38 Reward: -1066.0515 total_loss = 38.5121 = -0.0218 + 0.5 * 77.
0745 + 0.01 * -0.3340
----------------
Finished episode: 39 Reward: -1064.5640 total_loss = 37.7240 = 0.0011 + 0.5 * 75.4
526 + 0.01 * -0.3297
----------------
Finished episode: 40 Reward: -1070.3039 total_loss = 38.9909 = 0.0354 + 0.5 * 77.9
177 + 0.01 * -0.3328
```

```
----------------
Finished episode: 41 Reward: -1073.4590 total_loss = 40.1523 = 0.0643 + 0.5 * 80.1
826 + 0.01 * -0.3307
----------------
Finished episode: 42 Reward: -1072.6515 total_loss = 32.6418 = -0.1148 + 0.5 * 65.
5198 + 0.01 * -0.3287
----------------
Finished episode: 43 Reward: -1020.1784 total_loss = 34.7483 = 0.0176 + 0.5 * 69.4
683 + 0.01 * -0.3409
----------------
Finished episode: 44 Reward: -1021.0788 total_loss = 31.4444 = 0.0066 + 0.5 * 62.8
822 + 0.01 * -0.3257
----------------
Finished episode: 45 Reward: -1005.7094 total_loss = 26.6496 = -0.0772 + 0.5 * 53.
4602 + 0.01 * -0.3332
----------------
Finished episode: 46 Reward: -995.5171 total_loss = 28.2212 = -0.0325 + 0.5 * 56.5
142 + 0.01 * -0.3408
----------------
Finished episode: 47 Reward: -1027.9951 total_loss = 28.3155 = -0.0021 + 0.5 * 56.
6419 + 0.01 * -0.3350
----------------
Finished episode: 48 Reward: -1024.6861 total_loss = 29.2050 = 0.0403 + 0.5 * 58.3
360 + 0.01 * -0.3318
----------------
Finished episode: 49 Reward: -1008.6553 total_loss = 29.1847 = 0.0476 + 0.5 * 58.2
808 + 0.01 * -0.3331
----------------
Finished episode: 50 Reward: -984.9269 total_loss = 26.1072 = -0.0539 + 0.5 * 52.3
290 + 0.01 * -0.3369
----------------
Finished episode: 51 Reward: -1039.6798 total_loss = 29.2008 = -0.0184 + 0.5 * 58.
4451 + 0.01 * -0.3338
----------------
Finished episode: 52 Reward: -999.8316 total_loss = 24.1719 = -0.1030 + 0.5 * 48.5
563 + 0.01 * -0.3259
----------------
Finished episode: 53 Reward: -1025.2405 total_loss = 27.6728 = 0.0137 + 0.5 * 55.3
247 + 0.01 * -0.3300
----------------
Finished episode: 54 Reward: -1000.7535 total_loss = 25.1649 = -0.0855 + 0.5 * 50.
5074 + 0.01 * -0.3238
----------------
Finished episode: 55 Reward: -1035.6597 total_loss = 25.4210 = -0.0149 + 0.5 * 50.
8783 + 0.01 * -0.3310
----------------
Finished episode: 56 Reward: -1009.8373 total_loss = 27.2632 = 0.0586 + 0.5 * 54.4
159 + 0.01 * -0.3338
----------------
Finished episode: 57 Reward: -1051.1362 total_loss = 24.5948 = -0.0676 + 0.5 * 49.
3312 + 0.01 * -0.3242
----------------
Finished episode: 58 Reward: -998.9812 total_loss = 25.0258 = 0.0723 + 0.5 * 49.91
37 + 0.01 * -0.3330
----------------
Finished episode: 59 Reward: -951.3628 total_loss = 20.9732 = 0.0103 + 0.5 * 41.93
24 + 0.01 * -0.3290
----------------
Finished episode: 60 Reward: -1033.1991 total_loss = 24.8993 = -0.0185 + 0.5 * 49.
8423 + 0.01 * -0.3366
----------------
```

```
Finished episode: 61 Reward: -1031.6657 total_loss = 25.7171 = 0.0480 + 0.5 * 51.3
450 + 0.01 * -0.3410
-----------------
Finished episode: 62 Reward: -1047.2166 total_loss = 24.5665 = 0.0150 + 0.5 * 49.1
096 + 0.01 * -0.3298
-----------------
Finished episode: 63 Reward: -1043.4940 total_loss = 23.5214 = -0.0139 + 0.5 * 47.
0772 + 0.01 * -0.3351
-----------------
Finished episode: 64 Reward: -1060.2899 total_loss = 25.1323 = -0.0698 + 0.5 * 50.
4110 + 0.01 * -0.3337
-----------------
Finished episode: 65 Reward: -1049.6542 total_loss = 23.8360 = 0.0134 + 0.5 * 47.6
520 + 0.01 * -0.3417
-----------------
Finished episode: 66 Reward: -1160.0595 total_loss = 31.2076 = 0.1046 + 0.5 * 62.2
126 + 0.01 * -0.3283
-----------------
Finished episode: 67 Reward: -1110.1447 total_loss = 25.5873 = -0.0775 + 0.5 * 51.
3363 + 0.01 * -0.3307
-----------------
Finished episode: 68 Reward: -1049.4694 total_loss = 23.4393 = 0.0444 + 0.5 * 46.7
964 + 0.01 * -0.3285
-----------------
Finished episode: 69 Reward: -1035.3532 total_loss = 20.6979 = -0.0526 + 0.5 * 41.
5076 + 0.01 * -0.3312
-----------------
Finished episode: 70 Reward: -1038.2068 total_loss = 19.9933 = -0.1136 + 0.5 * 40.
2206 + 0.01 * -0.3347
-----------------
Finished episode: 71 Reward: -1091.7781 total_loss = 23.4858 = 0.0473 + 0.5 * 46.8
836 + 0.01 * -0.3303
-----------------
Finished episode: 72 Reward: -1104.8371 total_loss = 24.8251 = 0.0235 + 0.5 * 49.6
100 + 0.01 * -0.3315
-----------------
Finished episode: 73 Reward: -1047.6178 total_loss = 20.4546 = -0.0250 + 0.5 * 40.
9658 + 0.01 * -0.3278
-----------------
Finished episode: 74 Reward: -1049.2467 total_loss = 20.3534 = 0.0785 + 0.5 * 40.5
566 + 0.01 * -0.3398
-----------------
Finished episode: 75 Reward: -1012.1580 total_loss = 18.5902 = -0.0758 + 0.5 * 37.
3386 + 0.01 * -0.3298
-----------------
Finished episode: 76 Reward: -1055.0061 total_loss = 19.8461 = -0.0515 + 0.5 * 39.
8018 + 0.01 * -0.3333
-----------------
Finished episode: 77 Reward: -1037.5261 total_loss = 20.1575 = -0.0526 + 0.5 * 40.
4270 + 0.01 * -0.3337
-----------------
Finished episode: 78 Reward: -1043.2890 total_loss = 18.5787 = -0.0251 + 0.5 * 37.
2143 + 0.01 * -0.3310
-----------------
Finished episode: 79 Reward: -1008.2385 total_loss = 17.9174 = 0.0214 + 0.5 * 35.7
988 + 0.01 * -0.3355
-----------------
Finished episode: 80 Reward: -1023.2896 total_loss = 18.2133 = -0.1122 + 0.5 * 36.
6576 + 0.01 * -0.3299
-----------------
Finished episode: 81 Reward: -1024.3549 total_loss = 16.6343 = -0.0619 + 0.5 * 33.
```

```
3991 + 0.01 * -0.3367
----------------
Finished episode: 82 Reward: -1019.5792 total_loss = 16.0986 = -0.0224 + 0.5 * 32.
2486 + 0.01 * -0.3257
----------------
Finished episode: 83 Reward: -1017.0847 total_loss = 17.5234 = -0.0756 + 0.5 * 35.
2046 + 0.01 * -0.3327
----------------
Finished episode: 84 Reward: -1028.8255 total_loss = 18.0828 = 0.0300 + 0.5 * 36.1
123 + 0.01 * -0.3312
----------------
Finished episode: 85 Reward: -990.3927 total_loss = 17.3793 = 0.0637 + 0.5 * 34.63
79 + 0.01 * -0.3399
----------------
Finished episode: 86 Reward: -1009.8676 total_loss = 16.6415 = -0.0415 + 0.5 * 33.
3728 + 0.01 * -0.3348
----------------
Finished episode: 87 Reward: -1041.3907 total_loss = 15.9974 = -0.0567 + 0.5 * 32.
1148 + 0.01 * -0.3378
----------------
Finished episode: 88 Reward: -1029.3677 total_loss = 15.4095 = -0.1244 + 0.5 * 31.
0743 + 0.01 * -0.3293
----------------
Finished episode: 89 Reward: -1017.7444 total_loss = 17.0144 = 0.0744 + 0.5 * 33.8
867 + 0.01 * -0.3335
----------------
Finished episode: 90 Reward: -1062.2605 total_loss = 17.9954 = -0.0394 + 0.5 * 36.
0764 + 0.01 * -0.3391
----------------
Finished episode: 91 Reward: -1062.9034 total_loss = 17.0589 = -0.0672 + 0.5 * 34.
2592 + 0.01 * -0.3404
----------------
Finished episode: 92 Reward: -1037.6513 total_loss = 17.4373 = 0.0232 + 0.5 * 34.8
349 + 0.01 * -0.3346
----------------
Finished episode: 93 Reward: -1067.4300 total_loss = 17.9819 = 0.0533 + 0.5 * 35.8
638 + 0.01 * -0.3297
----------------
Finished episode: 94 Reward: -1053.3505 total_loss = 16.2428 = 0.0197 + 0.5 * 32.4
529 + 0.01 * -0.3321
----------------
Finished episode: 95 Reward: -1062.4083 total_loss = 17.9702 = -0.0153 + 0.5 * 35.
9776 + 0.01 * -0.3304
----------------
Finished episode: 96 Reward: -1075.9387 total_loss = 16.7266 = -0.0004 + 0.5 * 33.
4606 + 0.01 * -0.3358
----------------
Finished episode: 97 Reward: -1021.4597 total_loss = 15.0018 = -0.0294 + 0.5 * 30.
0690 + 0.01 * -0.3289
----------------
Finished episode: 98 Reward: -1071.9034 total_loss = 15.3508 = -0.0302 + 0.5 * 30.
7685 + 0.01 * -0.3284
----------------
Finished episode: 99 Reward: -1070.5343 total_loss = 17.2542 = 0.0603 + 0.5 * 34.3
945 + 0.01 * -0.3263
----------------
Finished episode: 100 Reward: -1058.4479 total_loss = 15.7098 = 0.1087 + 0.5 * 31.
2088 + 0.01 * -0.3329
----------------
Finished episode: 101 Reward: -1069.4320 total_loss = 13.8890 = 0.0104 + 0.5 * 27.
7638 + 0.01 * -0.3328
```

```
----------
Finished episode: 102 Reward: -1038.9786 total_loss = 16.5500 = 0.0546 + 0.5 * 32.
9973 + 0.01 * -0.3314
----------
Finished episode: 103 Reward: -1048.2427 total_loss = 15.3833 = -0.0174 + 0.5 * 3
0.8081 + 0.01 * -0.3246
----------
Finished episode: 104 Reward: -1045.1793 total_loss = 15.3596 = 0.0136 + 0.5 * 30.
6985 + 0.01 * -0.3285
----------
Finished episode: 105 Reward: -1102.4238 total_loss = 16.3967 = 0.0253 + 0.5 * 32.
7495 + 0.01 * -0.3263
----------
Finished episode: 106 Reward: -1091.8135 total_loss = 16.4118 = 0.0315 + 0.5 * 32.
7673 + 0.01 * -0.3353
----------
Finished episode: 107 Reward: -1121.8376 total_loss = 17.0631 = 0.0238 + 0.5 * 34.
0850 + 0.01 * -0.3272
----------
Finished episode: 108 Reward: -1105.9608 total_loss = 16.4854 = -0.0064 + 0.5 * 3
2.9902 + 0.01 * -0.3304
----------
Finished episode: 109 Reward: -1125.2070 total_loss = 16.5051 = -0.0252 + 0.5 * 3
3.0673 + 0.01 * -0.3305
----------
Finished episode: 110 Reward: -1124.4593 total_loss = 14.8365 = 0.0630 + 0.5 * 29.
5536 + 0.01 * -0.3279
----------
Finished episode: 111 Reward: -1088.7090 total_loss = 15.4733 = 0.0553 + 0.5 * 30.
8426 + 0.01 * -0.3317
----------
Finished episode: 112 Reward: -1113.6978 total_loss = 16.6096 = -0.0001 + 0.5 * 3
3.2261 + 0.01 * -0.3366
----------
Finished episode: 113 Reward: -1086.5504 total_loss = 15.8306 = 0.0008 + 0.5 * 31.
6661 + 0.01 * -0.3258
----------
Finished episode: 114 Reward: -1079.0380 total_loss = 14.6884 = 0.0499 + 0.5 * 29.
2836 + 0.01 * -0.3273
----------
Finished episode: 115 Reward: -1132.6216 total_loss = 16.1753 = 0.0498 + 0.5 * 32.
2577 + 0.01 * -0.3300
----------
Finished episode: 116 Reward: -1146.3347 total_loss = 15.1852 = -0.0925 + 0.5 * 3
0.5619 + 0.01 * -0.3252
----------
Finished episode: 117 Reward: -1092.4022 total_loss = 14.3682 = -0.0607 + 0.5 * 2
8.8643 + 0.01 * -0.3263
----------
Finished episode: 118 Reward: -1137.0020 total_loss = 15.7594 = 0.0069 + 0.5 * 31.
5116 + 0.01 * -0.3246
----------
Finished episode: 119 Reward: -1089.5133 total_loss = 14.6219 = 0.1386 + 0.5 * 28.
9730 + 0.01 * -0.3248
----------
Finished episode: 120 Reward: -1112.0446 total_loss = 14.8595 = -0.0420 + 0.5 * 2
9.8094 + 0.01 * -0.3205
----------
Finished episode: 121 Reward: -1129.9822 total_loss = 15.0376 = 0.0293 + 0.5 * 30.
0231 + 0.01 * -0.3260
----------
```

```
Finished episode: 122 Reward: -1145.2556 total_loss = 15.7610 = -0.0069 + 0.5 * 3
1.5423 + 0.01 * -0.3222
-----------------
Finished episode: 123 Reward: -1140.3742 total_loss = 15.3547 = 0.0881 + 0.5 * 30.
5395 + 0.01 * -0.3199
-----------------
Finished episode: 124 Reward: -1167.0252 total_loss = 16.2115 = -0.0152 + 0.5 * 3
2.4599 + 0.01 * -0.3277
-----------------
Finished episode: 125 Reward: -1186.2198 total_loss = 14.4892 = 0.0561 + 0.5 * 28.
8724 + 0.01 * -0.3185
-----------------
Finished episode: 126 Reward: -1162.0746 total_loss = 16.5783 = 0.0809 + 0.5 * 33.
0015 + 0.01 * -0.3259
-----------------
Finished episode: 127 Reward: -1169.9384 total_loss = 17.3637 = 0.0153 + 0.5 * 34.
7034 + 0.01 * -0.3271
-----------------
Finished episode: 128 Reward: -1189.3594 total_loss = 17.3066 = 0.0937 + 0.5 * 34.
4323 + 0.01 * -0.3243
-----------------
Finished episode: 129 Reward: -1195.9593 total_loss = 16.6447 = 0.0012 + 0.5 * 33.
2935 + 0.01 * -0.3235
-----------------
Finished episode: 130 Reward: -1167.4358 total_loss = 15.1247 = -0.0230 + 0.5 * 3
0.3019 + 0.01 * -0.3283
-----------------
Finished episode: 131 Reward: -1162.2131 total_loss = 14.9481 = -0.0238 + 0.5 * 2
9.9501 + 0.01 * -0.3214
-----------------
Finished episode: 132 Reward: -1163.6473 total_loss = 15.5716 = 0.0168 + 0.5 * 31.
1161 + 0.01 * -0.3262
-----------------
Finished episode: 133 Reward: -1175.5049 total_loss = 15.0623 = 0.0493 + 0.5 * 30.
0325 + 0.01 * -0.3229
-----------------
Finished episode: 134 Reward: -1157.3070 total_loss = 15.3771 = 0.0535 + 0.5 * 30.
6537 + 0.01 * -0.3232
-----------------
Finished episode: 135 Reward: -1183.7634 total_loss = 16.2839 = -0.0324 + 0.5 * 3
2.6390 + 0.01 * -0.3175
-----------------
Finished episode: 136 Reward: -1194.9264 total_loss = 15.0240 = -0.0294 + 0.5 * 3
0.1132 + 0.01 * -0.3199
-----------------
Finished episode: 137 Reward: -1195.4477 total_loss = 15.1177 = 0.0332 + 0.5 * 30.
1753 + 0.01 * -0.3211
-----------------
Finished episode: 138 Reward: -1183.9406 total_loss = 14.7316 = 0.0925 + 0.5 * 29.
2846 + 0.01 * -0.3202
-----------------
Finished episode: 139 Reward: -1174.3701 total_loss = 16.1326 = 0.0089 + 0.5 * 32.
2538 + 0.01 * -0.3202
-----------------
Finished episode: 140 Reward: -1147.4754 total_loss = 15.7741 = 0.0626 + 0.5 * 31.
4295 + 0.01 * -0.3211
-----------------
Finished episode: 141 Reward: -1153.2143 total_loss = 15.0316 = 0.0190 + 0.5 * 30.
0317 + 0.01 * -0.3232
-----------------
Finished episode: 142 Reward: -1129.9464 total_loss = 15.4042 = -0.0040 + 0.5 * 3
```

```
0.8228 + 0.01 * -0.3127
----------
Finished episode: 143 Reward: -1147.3170 total_loss = 16.0924 = 0.0278 + 0.5 * 32.
1355 + 0.01 * -0.3157
----------
Finished episode: 144 Reward: -1180.5007 total_loss = 15.2974 = 0.0202 + 0.5 * 30.
5608 + 0.01 * -0.3165
----------
Finished episode: 145 Reward: -1206.1599 total_loss = 15.3193 = 0.0104 + 0.5 * 30.
6241 + 0.01 * -0.3133
----------
Finished episode: 146 Reward: -1190.4634 total_loss = 16.2415 = -0.0091 + 0.5 * 3
2.5074 + 0.01 * -0.3179
----------
Finished episode: 147 Reward: -1195.2286 total_loss = 17.2574 = 0.0186 + 0.5 * 34.
4839 + 0.01 * -0.3132
----------
Finished episode: 148 Reward: -1185.8111 total_loss = 16.5242 = 0.0120 + 0.5 * 33.
0305 + 0.01 * -0.3117
----------
Finished episode: 149 Reward: -1215.1350 total_loss = 17.1838 = 0.0066 + 0.5 * 34.
3604 + 0.01 * -0.3066
----------
Finished episode: 150 Reward: -1265.0649 total_loss = 18.9786 = 0.0528 + 0.5 * 37.
8576 + 0.01 * -0.3000
----------
Finished episode: 151 Reward: -1250.3358 total_loss = 17.3870 = 0.0953 + 0.5 * 34.
5896 + 0.01 * -0.3082
----------
Finished episode: 152 Reward: -1273.4069 total_loss = 18.6491 = 0.1192 + 0.5 * 37.
0659 + 0.01 * -0.3061
----------
Finished episode: 153 Reward: -1297.9920 total_loss = 18.1635 = -0.0333 + 0.5 * 3
6.3997 + 0.01 * -0.3021
----------
Finished episode: 154 Reward: -1268.8669 total_loss = 19.2827 = -0.1324 + 0.5 * 3
8.8363 + 0.01 * -0.2962
----------
Finished episode: 155 Reward: -1263.2009 total_loss = 16.4561 = 0.0542 + 0.5 * 32.
8096 + 0.01 * -0.2911
----------
Finished episode: 156 Reward: -1212.3171 total_loss = 19.4364 = -0.0443 + 0.5 * 3
8.9673 + 0.01 * -0.2913
----------
Finished episode: 157 Reward: -1246.1448 total_loss = 19.0845 = -0.0496 + 0.5 * 3
8.2742 + 0.01 * -0.2992
----------
Finished episode: 158 Reward: -1255.5120 total_loss = 17.1615 = 0.0120 + 0.5 * 34.
3048 + 0.01 * -0.2952
----------
Finished episode: 159 Reward: -1269.7028 total_loss = 18.1440 = -0.1162 + 0.5 * 3
6.5263 + 0.01 * -0.2901
----------
Finished episode: 160 Reward: -1268.4910 total_loss = 18.0003 = 0.0452 + 0.5 * 35.
9159 + 0.01 * -0.2906
----------
Finished episode: 161 Reward: -1258.6529 total_loss = 18.3549 = 0.0873 + 0.5 * 36.
5411 + 0.01 * -0.2887
----------
Finished episode: 162 Reward: -1275.1076 total_loss = 16.8546 = -0.0213 + 0.5 * 3
3.7578 + 0.01 * -0.3005
```

```
----------------
Finished episode: 163 Reward: -1210.5021 total_loss = 18.5061 = -0.0539 + 0.5 * 3
7.1258 + 0.01 * -0.2907
----------------
Finished episode: 164 Reward: -1259.2408 total_loss = 16.2182 = 0.0416 + 0.5 * 32.
3592 + 0.01 * -0.3022
----------------
Finished episode: 165 Reward: -1271.8147 total_loss = 17.8640 = -0.0659 + 0.5 * 3
5.8656 + 0.01 * -0.2893
----------------
Finished episode: 166 Reward: -1233.6792 total_loss = 18.4363 = -0.0517 + 0.5 * 3
6.9818 + 0.01 * -0.2929
----------------
Finished episode: 167 Reward: -1263.2088 total_loss = 15.3625 = -0.0173 + 0.5 * 3
0.7655 + 0.01 * -0.2938
----------------
Finished episode: 168 Reward: -1250.4483 total_loss = 18.0078 = 0.0756 + 0.5 * 35.
8700 + 0.01 * -0.2789
----------------
Finished episode: 169 Reward: -1267.3642 total_loss = 16.7821 = 0.0894 + 0.5 * 33.
3911 + 0.01 * -0.2847
----------------
Finished episode: 170 Reward: -1248.9346 total_loss = 17.8132 = 0.0015 + 0.5 * 35.
6292 + 0.01 * -0.2892
----------------
Finished episode: 171 Reward: -1232.7472 total_loss = 18.9462 = -0.0270 + 0.5 * 3
7.9522 + 0.01 * -0.2899
----------------
Finished episode: 172 Reward: -1220.8978 total_loss = 17.0639 = 0.0049 + 0.5 * 34.
1236 + 0.01 * -0.2799
----------------
Finished episode: 173 Reward: -1236.1622 total_loss = 16.6787 = 0.0170 + 0.5 * 33.
3289 + 0.01 * -0.2814
----------------
Finished episode: 174 Reward: -1264.9893 total_loss = 17.9508 = -0.0116 + 0.5 * 3
5.9305 + 0.01 * -0.2813
----------------
Finished episode: 175 Reward: -1254.0960 total_loss = 16.1223 = 0.1007 + 0.5 * 32.
0488 + 0.01 * -0.2813
----------------
Finished episode: 176 Reward: -1256.6456 total_loss = 15.9687 = 0.1093 + 0.5 * 31.
7244 + 0.01 * -0.2880
----------------
Finished episode: 177 Reward: -1230.0281 total_loss = 18.8388 = -0.0536 + 0.5 * 3
7.7906 + 0.01 * -0.2898
----------------
Finished episode: 178 Reward: -1237.1827 total_loss = 17.8255 = -0.0506 + 0.5 * 3
5.7580 + 0.01 * -0.2853
----------------
Finished episode: 179 Reward: -1245.0216 total_loss = 18.2948 = -0.0423 + 0.5 * 3
6.6800 + 0.01 * -0.2856
----------------
Finished episode: 180 Reward: -1279.0829 total_loss = 17.7612 = 0.0366 + 0.5 * 35.
4548 + 0.01 * -0.2790
----------------
Finished episode: 181 Reward: -1280.6508 total_loss = 15.7489 = 0.0985 + 0.5 * 31.
3065 + 0.01 * -0.2798
----------------
Finished episode: 182 Reward: -1275.6864 total_loss = 18.4541 = 0.0329 + 0.5 * 36.
8479 + 0.01 * -0.2804
----------------
```

```
Finished episode: 183 Reward: -1237.5119 total_loss = 17.8342 = 0.0285 + 0.5 * 35.
6171 + 0.01 * -0.2832
-----------------
Finished episode: 184 Reward: -1250.1791 total_loss = 19.0051 = -0.0185 + 0.5 * 3
8.0527 + 0.01 * -0.2738
-----------------
Finished episode: 185 Reward: -1286.5354 total_loss = 17.9622 = -0.0652 + 0.5 * 3
6.0604 + 0.01 * -0.2779
-----------------
Finished episode: 186 Reward: -1286.2053 total_loss = 17.3223 = -0.0026 + 0.5 * 3
4.6554 + 0.01 * -0.2766
-----------------
Finished episode: 187 Reward: -1308.4868 total_loss = 17.3986 = -0.0328 + 0.5 * 3
4.8684 + 0.01 * -0.2791
-----------------
Finished episode: 188 Reward: -1284.3610 total_loss = 17.9988 = -0.0249 + 0.5 * 3
6.0530 + 0.01 * -0.2793
-----------------
Finished episode: 189 Reward: -1262.9897 total_loss = 17.9584 = 0.0417 + 0.5 * 35.
8390 + 0.01 * -0.2809
-----------------
Finished episode: 190 Reward: -1279.4031 total_loss = 16.0512 = 0.0094 + 0.5 * 32.
0889 + 0.01 * -0.2725
-----------------
Finished episode: 191 Reward: -1227.0379 total_loss = 17.4240 = 0.0419 + 0.5 * 34.
7699 + 0.01 * -0.2802
-----------------
Finished episode: 192 Reward: -1272.9237 total_loss = 17.6492 = -0.0184 + 0.5 * 3
5.3408 + 0.01 * -0.2744
-----------------
Finished episode: 193 Reward: -1285.4061 total_loss = 17.3964 = -0.0947 + 0.5 * 3
4.9875 + 0.01 * -0.2692
-----------------
Finished episode: 194 Reward: -1213.6820 total_loss = 18.7305 = -0.0278 + 0.5 * 3
7.5219 + 0.01 * -0.2606
-----------------
Finished episode: 195 Reward: -1260.8944 total_loss = 16.6684 = -0.0334 + 0.5 * 3
3.4090 + 0.01 * -0.2656
-----------------
Finished episode: 196 Reward: -1225.1126 total_loss = 18.6452 = 0.0577 + 0.5 * 37.
1801 + 0.01 * -0.2535
-----------------
Finished episode: 197 Reward: -1261.2165 total_loss = 16.6875 = -0.0481 + 0.5 * 3
3.4766 + 0.01 * -0.2661
-----------------
Finished episode: 198 Reward: -1234.2513 total_loss = 17.9768 = -0.0110 + 0.5 * 3
5.9806 + 0.01 * -0.2511
-----------------
Finished episode: 199 Reward: -1236.8160 total_loss = 17.3616 = 0.0681 + 0.5 * 34.
5923 + 0.01 * -0.2646
-----------------
Finished episode: 200 Reward: -1243.8425 total_loss = 17.5702 = 0.0527 + 0.5 * 35.
0400 + 0.01 * -0.2526
-----------------
Finished episode: 201 Reward: -1250.9194 total_loss = 15.3455 = 0.0818 + 0.5 * 30.
5326 + 0.01 * -0.2615
-----------------
Finished episode: 202 Reward: -1245.1063 total_loss = 18.9375 = -0.0673 + 0.5 * 3
8.0147 + 0.01 * -0.2589
-----------------
Finished episode: 203 Reward: -1252.5492 total_loss = 16.5717 = -0.0196 + 0.5 * 3
```

```
3.1876 + 0.01 * -0.2471
----------------
Finished episode: 204 Reward: -1254.7680 total_loss = 17.4260 = -0.0349 + 0.5 * 3
4.9268 + 0.01 * -0.2476
----------------
Finished episode: 205 Reward: -1240.3071 total_loss = 15.8178 = 0.0072 + 0.5 * 31.
6262 + 0.01 * -0.2540
----------------
Finished episode: 206 Reward: -1237.1510 total_loss = 17.1812 = 0.0314 + 0.5 * 34.
3047 + 0.01 * -0.2569
----------------
Finished episode: 207 Reward: -1243.8579 total_loss = 17.0943 = 0.0891 + 0.5 * 34.
0154 + 0.01 * -0.2453
----------------
Finished episode: 208 Reward: -1209.3452 total_loss = 19.0764 = 0.0564 + 0.5 * 38.
0448 + 0.01 * -0.2412
----------------
Finished episode: 209 Reward: -1244.5229 total_loss = 16.7197 = -0.0258 + 0.5 * 3
3.4960 + 0.01 * -0.2460
----------------
Finished episode: 210 Reward: -1293.3207 total_loss = 17.1116 = 0.0511 + 0.5 * 34.
1259 + 0.01 * -0.2468
----------------
Finished episode: 211 Reward: -1290.6093 total_loss = 18.2734 = 0.0444 + 0.5 * 36.
4626 + 0.01 * -0.2386
----------------
Finished episode: 212 Reward: -1270.0945 total_loss = 18.1585 = -0.0531 + 0.5 * 3
6.4281 + 0.01 * -0.2377
----------------
Finished episode: 213 Reward: -1253.8444 total_loss = 18.3219 = 0.0830 + 0.5 * 36.
4822 + 0.01 * -0.2241
----------------
Finished episode: 214 Reward: -1258.1920 total_loss = 17.7303 = 0.0099 + 0.5 * 35.
4456 + 0.01 * -0.2391
----------------
Finished episode: 215 Reward: -1255.1358 total_loss = 17.8614 = 0.0531 + 0.5 * 35.
6213 + 0.01 * -0.2325
----------------
Finished episode: 216 Reward: -1266.5881 total_loss = 17.8976 = 0.0037 + 0.5 * 35.
7924 + 0.01 * -0.2309
----------------
Finished episode: 217 Reward: -1270.7256 total_loss = 18.0101 = 0.0185 + 0.5 * 35.
9882 + 0.01 * -0.2454
----------------
Finished episode: 218 Reward: -1214.8628 total_loss = 19.4030 = -0.0094 + 0.5 * 3
8.8294 + 0.01 * -0.2315
----------------
Finished episode: 219 Reward: -1253.8874 total_loss = 18.7366 = 0.0673 + 0.5 * 37.
3430 + 0.01 * -0.2314
----------------
Finished episode: 220 Reward: -1268.0437 total_loss = 17.5134 = 0.0089 + 0.5 * 35.
0139 + 0.01 * -0.2421
----------------
Finished episode: 221 Reward: -1248.9761 total_loss = 18.5795 = -0.0447 + 0.5 * 3
7.2531 + 0.01 * -0.2349
----------------
Finished episode: 222 Reward: -1277.3125 total_loss = 16.7753 = -0.0040 + 0.5 * 3
3.5636 + 0.01 * -0.2495
----------------
Finished episode: 223 Reward: -1285.7698 total_loss = 18.2631 = -0.0334 + 0.5 * 3
6.5977 + 0.01 * -0.2372
```

```
----------------
Finished episode: 224 Reward: -1240.4446 total_loss = 18.9454 = -0.0678 + 0.5 * 3
8.0312 + 0.01 * -0.2359
----------------
Finished episode: 225 Reward: -1235.2058 total_loss = 19.4160 = -0.0358 + 0.5 * 3
8.9084 + 0.01 * -0.2392
----------------
Finished episode: 226 Reward: -1280.1232 total_loss = 17.1241 = -0.0709 + 0.5 * 3
4.3946 + 0.01 * -0.2286
----------------
Finished episode: 227 Reward: -1258.6567 total_loss = 18.4342 = -0.0541 + 0.5 * 3
6.9812 + 0.01 * -0.2382
----------------
Finished episode: 228 Reward: -1288.9082 total_loss = 19.6301 = 0.0010 + 0.5 * 39.
2629 + 0.01 * -0.2376
----------------
Finished episode: 229 Reward: -1298.4547 total_loss = 18.4157 = -0.0539 + 0.5 * 3
6.9438 + 0.01 * -0.2306
----------------
Finished episode: 230 Reward: -1283.5902 total_loss = 19.2152 = 0.0088 + 0.5 * 38.
4171 + 0.01 * -0.2167
----------------
Finished episode: 231 Reward: -1281.4824 total_loss = 18.2008 = 0.0779 + 0.5 * 36.
2503 + 0.01 * -0.2233
----------------
Finished episode: 232 Reward: -1276.5450 total_loss = 20.0267 = -0.0113 + 0.5 * 4
0.0806 + 0.01 * -0.2273
----------------
Finished episode: 233 Reward: -1302.1014 total_loss = 20.0454 = 0.0576 + 0.5 * 39.
9799 + 0.01 * -0.2169
----------------
Finished episode: 234 Reward: -1250.4597 total_loss = 18.4737 = 0.0674 + 0.5 * 36.
8170 + 0.01 * -0.2177
----------------
Finished episode: 235 Reward: -1271.9175 total_loss = 15.0430 = 0.0432 + 0.5 * 30.
0040 + 0.01 * -0.2213
----------------
Finished episode: 236 Reward: -1286.4099 total_loss = 17.8705 = -0.0320 + 0.5 * 3
5.8093 + 0.01 * -0.2163
----------------
Finished episode: 237 Reward: -1266.1976 total_loss = 17.6532 = -0.0048 + 0.5 * 3
5.3201 + 0.01 * -0.2072
----------------
Finished episode: 238 Reward: -1235.2192 total_loss = 20.1134 = -0.1308 + 0.5 * 4
0.4928 + 0.01 * -0.2139
----------------
Finished episode: 239 Reward: -1263.9343 total_loss = 18.1504 = -0.0430 + 0.5 * 3
6.3912 + 0.01 * -0.2222
----------------
Finished episode: 240 Reward: -1255.3564 total_loss = 17.3343 = -0.0937 + 0.5 * 3
4.8604 + 0.01 * -0.2202
----------------
Finished episode: 241 Reward: -1235.2043 total_loss = 18.5312 = 0.0339 + 0.5 * 36.
9986 + 0.01 * -0.2021
----------------
Finished episode: 242 Reward: -1275.2438 total_loss = 17.3426 = -0.0814 + 0.5 * 3
4.8523 + 0.01 * -0.2159
----------------
Finished episode: 243 Reward: -1220.8203 total_loss = 17.7648 = 0.0382 + 0.5 * 35.
4574 + 0.01 * -0.2158
----------------
```

```
Finished episode: 244 Reward: -1251.0078 total_loss = 18.2160 = 0.0794 + 0.5 * 36.
2773 + 0.01 * -0.2030
-----------------
Finished episode: 245 Reward: -1238.1702 total_loss = 16.1160 = 0.0733 + 0.5 * 32.
0894 + 0.01 * -0.1999
-----------------
Finished episode: 246 Reward: -1209.5419 total_loss = 17.1418 = -0.0684 + 0.5 * 3
4.4248 + 0.01 * -0.2170
-----------------
Finished episode: 247 Reward: -1217.4200 total_loss = 17.6636 = 0.0387 + 0.5 * 35.
2538 + 0.01 * -0.2044
-----------------
Finished episode: 248 Reward: -1232.7680 total_loss = 18.4296 = -0.1124 + 0.5 * 3
7.0881 + 0.01 * -0.2055
-----------------
Finished episode: 249 Reward: -1206.5842 total_loss = 19.3180 = -0.0923 + 0.5 * 3
8.8248 + 0.01 * -0.2071
-----------------
Finished episode: 250 Reward: -1226.6851 total_loss = 17.2604 = 0.0094 + 0.5 * 34.
5060 + 0.01 * -0.1968
-----------------
Finished episode: 251 Reward: -1232.7014 total_loss = 16.7098 = 0.0859 + 0.5 * 33.
2515 + 0.01 * -0.1873
-----------------
Finished episode: 252 Reward: -1235.1519 total_loss = 17.8247 = -0.0500 + 0.5 * 3
5.7531 + 0.01 * -0.1815
-----------------
Finished episode: 253 Reward: -1224.0226 total_loss = 15.2098 = 0.1029 + 0.5 * 30.
2178 + 0.01 * -0.2061
-----------------
Finished episode: 254 Reward: -1238.6822 total_loss = 17.2545 = 0.1384 + 0.5 * 34.
2362 + 0.01 * -0.1987
-----------------
Finished episode: 255 Reward: -1258.1823 total_loss = 17.5027 = 0.0188 + 0.5 * 34.
9717 + 0.01 * -0.1934
-----------------
Finished episode: 256 Reward: -1236.6419 total_loss = 17.5460 = 0.1385 + 0.5 * 34.
8190 + 0.01 * -0.1980
-----------------
Finished episode: 257 Reward: -1207.6372 total_loss = 15.9962 = 0.0400 + 0.5 * 31.
9162 + 0.01 * -0.1841
-----------------
Finished episode: 258 Reward: -1211.1892 total_loss = 16.2725 = -0.0342 + 0.5 * 3
2.6168 + 0.01 * -0.1778
-----------------
Finished episode: 259 Reward: -1225.9464 total_loss = 16.9943 = 0.0204 + 0.5 * 33.
9515 + 0.01 * -0.1785
-----------------
Finished episode: 260 Reward: -1234.2005 total_loss = 18.2295 = -0.0136 + 0.5 * 3
6.4900 + 0.01 * -0.1940
-----------------
Finished episode: 261 Reward: -1221.5007 total_loss = 16.9228 = -0.1079 + 0.5 * 3
4.0651 + 0.01 * -0.1800
-----------------
Finished episode: 262 Reward: -1229.7012 total_loss = 16.2678 = 0.0536 + 0.5 * 32.
4320 + 0.01 * -0.1851
-----------------
Finished episode: 263 Reward: -1240.6690 total_loss = 16.6978 = 0.0480 + 0.5 * 33.
3032 + 0.01 * -0.1808
-----------------
Finished episode: 264 Reward: -1249.0076 total_loss = 17.7103 = -0.0273 + 0.5 * 3
```

```
5.4787 + 0.01 * -0.1775
----------------
Finished episode: 265 Reward: -1220.0243 total_loss = 17.4769 = 0.0679 + 0.5 * 34.
8212 + 0.01 * -0.1611
----------------
Finished episode: 266 Reward: -1232.6717 total_loss = 16.2571 = 0.0441 + 0.5 * 32.
4294 + 0.01 * -0.1688
----------------
Finished episode: 267 Reward: -1252.8847 total_loss = 18.2685 = -0.0629 + 0.5 * 3
6.6662 + 0.01 * -0.1706
----------------
Finished episode: 268 Reward: -1240.2770 total_loss = 16.5051 = 0.0105 + 0.5 * 32.
9923 + 0.01 * -0.1570
----------------
Finished episode: 269 Reward: -1260.2556 total_loss = 16.6177 = -0.0713 + 0.5 * 3
3.3812 + 0.01 * -0.1581
----------------
Finished episode: 270 Reward: -1241.9969 total_loss = 17.4027 = 0.0232 + 0.5 * 34.
7623 + 0.01 * -0.1593
----------------
Finished episode: 271 Reward: -1223.3511 total_loss = 18.1610 = -0.1370 + 0.5 * 3
6.5989 + 0.01 * -0.1437
----------------
Finished episode: 272 Reward: -1238.4974 total_loss = 14.5961 = 0.2210 + 0.5 * 28.
7531 + 0.01 * -0.1517
----------------
Finished episode: 273 Reward: -1260.9425 total_loss = 13.9895 = 0.0198 + 0.5 * 27.
9425 + 0.01 * -0.1617
----------------
Finished episode: 274 Reward: -1230.9083 total_loss = 16.7731 = -0.1456 + 0.5 * 3
3.8406 + 0.01 * -0.1590
----------------
Finished episode: 275 Reward: -1252.2983 total_loss = 16.7795 = 0.0803 + 0.5 * 33.
4016 + 0.01 * -0.1524
----------------
Finished episode: 276 Reward: -1261.6155 total_loss = 14.3545 = 0.0271 + 0.5 * 28.
6580 + 0.01 * -0.1631
----------------
Finished episode: 277 Reward: -1252.2462 total_loss = 17.0198 = 0.0234 + 0.5 * 33.
9961 + 0.01 * -0.1658
----------------
Finished episode: 278 Reward: -1250.1533 total_loss = 17.8837 = 0.0302 + 0.5 * 35.
7099 + 0.01 * -0.1406
----------------
Finished episode: 279 Reward: -1231.0868 total_loss = 18.4524 = 0.0224 + 0.5 * 36.
8632 + 0.01 * -0.1596
----------------
Finished episode: 280 Reward: -1229.8953 total_loss = 16.7645 = 0.0684 + 0.5 * 33.
3954 + 0.01 * -0.1584
----------------
Finished episode: 281 Reward: -1260.4775 total_loss = 18.4162 = 0.0264 + 0.5 * 36.
7827 + 0.01 * -0.1505
----------------
Finished episode: 282 Reward: -1268.5422 total_loss = 17.0220 = 0.1371 + 0.5 * 33.
7728 + 0.01 * -0.1546
----------------
Finished episode: 283 Reward: -1244.3047 total_loss = 16.9595 = 0.0013 + 0.5 * 33.
9194 + 0.01 * -0.1563
----------------
Finished episode: 284 Reward: -1279.0657 total_loss = 16.6213 = -0.0495 + 0.5 * 3
3.3448 + 0.01 * -0.1499
```

```
----------
Finished episode: 285 Reward: -1279.4307 total_loss = 17.0382 = -0.1091 + 0.5 * 3
4.2974 + 0.01 * -0.1406
----------
Finished episode: 286 Reward: -1273.5985 total_loss = 17.0593 = 0.0205 + 0.5 * 34.
0807 + 0.01 * -0.1491
----------
Finished episode: 287 Reward: -1280.8705 total_loss = 16.9087 = -0.0577 + 0.5 * 3
3.9360 + 0.01 * -0.1536
----------
Finished episode: 288 Reward: -1251.1106 total_loss = 17.7142 = -0.0213 + 0.5 * 3
5.4737 + 0.01 * -0.1329
----------
Finished episode: 289 Reward: -1297.5269 total_loss = 16.1443 = -0.0351 + 0.5 * 3
2.3619 + 0.01 * -0.1557
----------
Finished episode: 290 Reward: -1268.1941 total_loss = 16.7609 = -0.0124 + 0.5 * 3
3.5495 + 0.01 * -0.1396
----------
Finished episode: 291 Reward: -1267.2167 total_loss = 15.5407 = -0.0596 + 0.5 * 3
1.2035 + 0.01 * -0.1437
----------
Finished episode: 292 Reward: -1266.6675 total_loss = 17.5214 = -0.0638 + 0.5 * 3
5.1736 + 0.01 * -0.1594
----------
Finished episode: 293 Reward: -1222.3135 total_loss = 18.7801 = 0.0734 + 0.5 * 37.
4162 + 0.01 * -0.1405
----------
Finished episode: 294 Reward: -1224.1585 total_loss = 20.4819 = -0.1286 + 0.5 * 4
1.2238 + 0.01 * -0.1447
----------
Finished episode: 295 Reward: -1248.6983 total_loss = 17.6429 = -0.0772 + 0.5 * 3
5.4431 + 0.01 * -0.1366
----------
Finished episode: 296 Reward: -1247.1298 total_loss = 19.5197 = 0.0120 + 0.5 * 39.
0183 + 0.01 * -0.1384
----------
Finished episode: 297 Reward: -1227.0682 total_loss = 19.5947 = -0.0939 + 0.5 * 3
9.3795 + 0.01 * -0.1194
----------
Finished episode: 298 Reward: -1210.2238 total_loss = 19.5126 = -0.0523 + 0.5 * 3
9.1325 + 0.01 * -0.1316
----------
Finished episode: 299 Reward: -1256.6118 total_loss = 19.7696 = -0.0410 + 0.5 * 3
9.6238 + 0.01 * -0.1318
----------
Finished episode: 300 Reward: -1236.3916 total_loss = 17.0838 = -0.0211 + 0.5 * 3
4.2128 + 0.01 * -0.1437
----------
Finished episode: 301 Reward: -1265.0990 total_loss = 20.7059 = -0.0398 + 0.5 * 4
1.4940 + 0.01 * -0.1210
----------
Finished episode: 302 Reward: -1289.5345 total_loss = 17.0379 = 0.0651 + 0.5 * 33.
9482 + 0.01 * -0.1375
----------
Finished episode: 303 Reward: -1293.3729 total_loss = 16.3165 = -0.0724 + 0.5 * 3
2.7802 + 0.01 * -0.1170
----------
Finished episode: 304 Reward: -1264.6848 total_loss = 17.3535 = 0.1240 + 0.5 * 34.
4617 + 0.01 * -0.1383
----------
```

```
Finished episode: 305 Reward: -1258.0745 total_loss = 17.2845 = -0.0140 + 0.5 * 3
4.5998 + 0.01 * -0.1398
-----------------
Finished episode: 306 Reward: -1253.6618 total_loss = 17.4950 = 0.0260 + 0.5 * 34.
9406 + 0.01 * -0.1363
-----------------
Finished episode: 307 Reward: -1286.7196 total_loss = 16.7188 = -0.0118 + 0.5 * 3
3.4636 + 0.01 * -0.1162
-----------------
Finished episode: 308 Reward: -1258.2001 total_loss = 16.9078 = -0.0158 + 0.5 * 3
3.8497 + 0.01 * -0.1274
-----------------
Finished episode: 309 Reward: -1251.9669 total_loss = 16.5876 = -0.0214 + 0.5 * 3
3.2202 + 0.01 * -0.1068
-----------------
Finished episode: 310 Reward: -1269.5798 total_loss = 18.4002 = 0.0078 + 0.5 * 36.
7873 + 0.01 * -0.1219
-----------------
Finished episode: 311 Reward: -1264.1434 total_loss = 16.2578 = -0.0066 + 0.5 * 3
2.5314 + 0.01 * -0.1320
-----------------
Finished episode: 312 Reward: -1257.3419 total_loss = 17.7037 = 0.0813 + 0.5 * 35.
2467 + 0.01 * -0.1011
-----------------
Finished episode: 313 Reward: -1292.7572 total_loss = 16.2642 = 0.0206 + 0.5 * 32.
4897 + 0.01 * -0.1268
-----------------
Finished episode: 314 Reward: -1261.7685 total_loss = 17.6230 = -0.0723 + 0.5 * 3
5.3928 + 0.01 * -0.1094
-----------------
Finished episode: 315 Reward: -1271.6789 total_loss = 16.9991 = -0.0764 + 0.5 * 3
4.1538 + 0.01 * -0.1438
-----------------
Finished episode: 316 Reward: -1253.8473 total_loss = 14.9441 = 0.1273 + 0.5 * 29.
6362 + 0.01 * -0.1312
-----------------
Finished episode: 317 Reward: -1267.7131 total_loss = 17.2171 = -0.0796 + 0.5 * 3
4.5957 + 0.01 * -0.1186
-----------------
Finished episode: 318 Reward: -1240.4055 total_loss = 19.3914 = -0.1331 + 0.5 * 3
9.0509 + 0.01 * -0.0945
-----------------
Finished episode: 319 Reward: -1252.6365 total_loss = 17.0183 = -0.0149 + 0.5 * 3
4.0690 + 0.01 * -0.1299
-----------------
Finished episode: 320 Reward: -1265.1201 total_loss = 15.3032 = 0.0328 + 0.5 * 30.
5429 + 0.01 * -0.1114
-----------------
Finished episode: 321 Reward: -1235.2146 total_loss = 17.2524 = 0.1131 + 0.5 * 34.
2813 + 0.01 * -0.1315
-----------------
Finished episode: 322 Reward: -1218.7864 total_loss = 19.3709 = 0.0113 + 0.5 * 38.
7214 + 0.01 * -0.1157
-----------------
Finished episode: 323 Reward: -1282.4069 total_loss = 17.0304 = -0.0028 + 0.5 * 3
4.0690 + 0.01 * -0.1341
-----------------
Finished episode: 324 Reward: -1247.4307 total_loss = 17.1297 = -0.0551 + 0.5 * 3
4.3719 + 0.01 * -0.1197
-----------------
Finished episode: 325 Reward: -1212.0361 total_loss = 16.8050 = 0.0897 + 0.5 * 33.
```

```
4332 + 0.01 * -0.1280
----------------
Finished episode: 326 Reward: -1233.0561 total_loss = 18.1946 = 0.0498 + 0.5 * 36.
2919 + 0.01 * -0.1142
----------------
Finished episode: 327 Reward: -1250.2645 total_loss = 16.0131 = -0.0399 + 0.5 * 3
2.1086 + 0.01 * -0.1278
----------------
Finished episode: 328 Reward: -1245.8240 total_loss = 17.0351 = -0.0343 + 0.5 * 3
4.1407 + 0.01 * -0.0992
----------------
Finished episode: 329 Reward: -1235.2345 total_loss = 17.7976 = 0.0121 + 0.5 * 35.
5732 + 0.01 * -0.1122
----------------
Finished episode: 330 Reward: -1249.1603 total_loss = 15.6283 = 0.0287 + 0.5 * 31.
2015 + 0.01 * -0.1115
----------------
Finished episode: 331 Reward: -1237.2889 total_loss = 16.9819 = 0.1048 + 0.5 * 33.
7563 + 0.01 * -0.1121
----------------
Finished episode: 332 Reward: -1224.7735 total_loss = 17.7405 = -0.0068 + 0.5 * 3
5.4967 + 0.01 * -0.1053
----------------
Finished episode: 333 Reward: -1233.4901 total_loss = 18.0055 = 0.0466 + 0.5 * 35.
9195 + 0.01 * -0.0874
----------------
Finished episode: 334 Reward: -1224.2208 total_loss = 15.4881 = 0.1005 + 0.5 * 30.
7772 + 0.01 * -0.0973
----------------
Finished episode: 335 Reward: -1251.7206 total_loss = 16.8786 = -0.0318 + 0.5 * 3
3.8226 + 0.01 * -0.0852
----------------
Finished episode: 336 Reward: -1264.7611 total_loss = 13.3131 = 0.0940 + 0.5 * 26.
4400 + 0.01 * -0.0902
----------------
Finished episode: 337 Reward: -1249.0955 total_loss = 17.8562 = -0.1881 + 0.5 * 3
6.0903 + 0.01 * -0.0858
----------------
Finished episode: 338 Reward: -1258.3612 total_loss = 13.8933 = -0.0310 + 0.5 * 2
7.8507 + 0.01 * -0.1040
----------------
Finished episode: 339 Reward: -1230.1703 total_loss = 18.3815 = -0.0194 + 0.5 * 3
6.8034 + 0.01 * -0.0791
----------------
Finished episode: 340 Reward: -1217.7837 total_loss = 18.1074 = 0.0968 + 0.5 * 36.
0227 + 0.01 * -0.0815
----------------
Finished episode: 341 Reward: -1239.7191 total_loss = 15.6039 = -0.0125 + 0.5 * 3
1.2343 + 0.01 * -0.0754
----------------
Finished episode: 342 Reward: -1168.4050 total_loss = 16.2741 = 0.0270 + 0.5 * 32.
4961 + 0.01 * -0.0901
----------------
Finished episode: 343 Reward: -1245.5047 total_loss = 12.8206 = -0.0360 + 0.5 * 2
5.7153 + 0.01 * -0.1019
----------------
Finished episode: 344 Reward: -1252.2318 total_loss = 13.3642 = 0.0151 + 0.5 * 26.
6996 + 0.01 * -0.0657
----------------
Finished episode: 345 Reward: -1230.8300 total_loss = 15.1725 = -0.1413 + 0.5 * 3
0.6289 + 0.01 * -0.0645
```

```
----------------
Finished episode: 346 Reward: -1248.2628 total_loss = 16.4924 = -0.0207 + 0.5 * 3
3.0273 + 0.01 * -0.0593
----------------
Finished episode: 347 Reward: -1208.4548 total_loss = 15.4553 = -0.0278 + 0.5 * 3
0.9678 + 0.01 * -0.0804
----------------
Finished episode: 348 Reward: -1251.7304 total_loss = 15.1268 = -0.0028 + 0.5 * 3
0.2604 + 0.01 * -0.0604
----------------
Finished episode: 349 Reward: -1249.7498 total_loss = 15.6561 = -0.0228 + 0.5 * 3
1.3586 + 0.01 * -0.0423
----------------
Finished episode: 350 Reward: -1248.5646 total_loss = 15.3678 = 0.0312 + 0.5 * 30.
6744 + 0.01 * -0.0625
----------------
Finished episode: 351 Reward: -1234.9650 total_loss = 19.6096 = -0.0001 + 0.5 * 3
9.2206 + 0.01 * -0.0638
----------------
Finished episode: 352 Reward: -1243.7323 total_loss = 16.6184 = -0.0870 + 0.5 * 3
3.4117 + 0.01 * -0.0497
----------------
Finished episode: 353 Reward: -1208.4638 total_loss = 16.6763 = 0.0591 + 0.5 * 33.
2359 + 0.01 * -0.0739
----------------
Finished episode: 354 Reward: -1247.0259 total_loss = 16.2411 = -0.0430 + 0.5 * 3
2.5698 + 0.01 * -0.0822
----------------
Finished episode: 355 Reward: -1228.1970 total_loss = 17.5116 = -0.0164 + 0.5 * 3
5.0571 + 0.01 * -0.0586
----------------
Finished episode: 356 Reward: -1235.0845 total_loss = 15.7967 = -0.0223 + 0.5 * 3
1.6389 + 0.01 * -0.0471
----------------
Finished episode: 357 Reward: -1237.7962 total_loss = 17.2524 = -0.0869 + 0.5 * 3
4.6794 + 0.01 * -0.0416
----------------
Finished episode: 358 Reward: -1243.7595 total_loss = 15.4341 = -0.0153 + 0.5 * 3
0.9004 + 0.01 * -0.0728
----------------
Finished episode: 359 Reward: -1206.0019 total_loss = 17.6793 = 0.0152 + 0.5 * 35.
3293 + 0.01 * -0.0602
----------------
Finished episode: 360 Reward: -1248.8723 total_loss = 17.7736 = -0.0865 + 0.5 * 3
5.7213 + 0.01 * -0.0502
----------------
Finished episode: 361 Reward: -1244.8266 total_loss = 13.6335 = 0.0357 + 0.5 * 27.
1966 + 0.01 * -0.0483
----------------
Finished episode: 362 Reward: -1223.4509 total_loss = 14.2962 = 0.0015 + 0.5 * 28.
5910 + 0.01 * -0.0705
----------------
Finished episode: 363 Reward: -1222.7952 total_loss = 18.5309 = 0.1193 + 0.5 * 36.
8252 + 0.01 * -0.1016
----------------
Finished episode: 364 Reward: -1231.6768 total_loss = 16.5203 = -0.0609 + 0.5 * 3
3.1632 + 0.01 * -0.0375
----------------
Finished episode: 365 Reward: -1260.3683 total_loss = 15.4707 = 0.0803 + 0.5 * 30.
7817 + 0.01 * -0.0429
----------------
```

```
Finished episode: 366 Reward: -1247.3214 total_loss = 14.6629 = 0.1506 + 0.5 * 29.
0259 + 0.01 * -0.0640
-----------------
Finished episode: 367 Reward: -1250.4674 total_loss = 13.8508 = -0.0128 + 0.5 * 2
7.7282 + 0.01 * -0.0515
-----------------
Finished episode: 368 Reward: -1248.2333 total_loss = 14.1065 = -0.0118 + 0.5 * 2
8.2371 + 0.01 * -0.0301
-----------------
Finished episode: 369 Reward: -1247.9250 total_loss = 15.2854 = 0.0320 + 0.5 * 30.
5080 + 0.01 * -0.0645
-----------------
Finished episode: 370 Reward: -1225.5480 total_loss = 15.4349 = -0.0322 + 0.5 * 3
0.9354 + 0.01 * -0.0657
-----------------
Finished episode: 371 Reward: -1269.2711 total_loss = 15.5541 = -0.0729 + 0.5 * 3
1.2552 + 0.01 * -0.0525
-----------------
Finished episode: 372 Reward: -1235.6560 total_loss = 13.4923 = 0.0341 + 0.5 * 26.
9175 + 0.01 * -0.0534
-----------------
Finished episode: 373 Reward: -1250.6869 total_loss = 16.3940 = -0.0435 + 0.5 * 3
2.8761 + 0.01 * -0.0576
-----------------
Finished episode: 374 Reward: -1254.0721 total_loss = 15.1803 = -0.0036 + 0.5 * 3
0.3690 + 0.01 * -0.0562
-----------------
Finished episode: 375 Reward: -1256.2201 total_loss = 14.5797 = -0.0519 + 0.5 * 2
9.2641 + 0.01 * -0.0383
-----------------
Finished episode: 376 Reward: -1238.1141 total_loss = 15.2512 = -0.0917 + 0.5 * 3
0.6862 + 0.01 * -0.0236
-----------------
Finished episode: 377 Reward: -1239.6793 total_loss = 14.5864 = -0.0121 + 0.5 * 2
9.1979 + 0.01 * -0.0457
-----------------
Finished episode: 378 Reward: -1210.5270 total_loss = 18.9823 = -0.0331 + 0.5 * 3
8.0320 + 0.01 * -0.0618
-----------------
Finished episode: 379 Reward: -1234.9756 total_loss = 16.8877 = -0.0004 + 0.5 * 3
3.7773 + 0.01 * -0.0545
-----------------
Finished episode: 380 Reward: -1214.9894 total_loss = 15.8685 = -0.0543 + 0.5 * 3
1.8464 + 0.01 * -0.0450
-----------------
Finished episode: 381 Reward: -1220.5580 total_loss = 17.9609 = -0.0505 + 0.5 * 3
6.0241 + 0.01 * -0.0625
-----------------
Finished episode: 382 Reward: -1232.4985 total_loss = 13.1958 = -0.0809 + 0.5 * 2
6.5542 + 0.01 * -0.0387
-----------------
Finished episode: 383 Reward: -1249.8789 total_loss = 14.6133 = -0.0338 + 0.5 * 2
9.2949 + 0.01 * -0.0423
-----------------
Finished episode: 384 Reward: -1190.1612 total_loss = 13.1975 = 0.0832 + 0.5 * 26.
2292 + 0.01 * -0.0336
-----------------
Finished episode: 385 Reward: -1198.1155 total_loss = 17.0322 = 0.0382 + 0.5 * 33.
9891 + 0.01 * -0.0492
-----------------
Finished episode: 386 Reward: -1205.7040 total_loss = 16.2305 = -0.0024 + 0.5 * 3
```

```
2.4672 + 0.01 * -0.0719
----------------
Finished episode: 387 Reward: -1231.5509 total_loss = 14.2119 = 0.0533 + 0.5 * 28.
3171 + 0.01 * -0.0016
----------------
Finished episode: 388 Reward: -1243.4595 total_loss = 14.1500 = 0.0518 + 0.5 * 28.
1968 + 0.01 * -0.0246
----------------
Finished episode: 389 Reward: -1267.1400 total_loss = 12.9803 = 0.0428 + 0.5 * 25.
8760 + 0.01 * -0.0552
----------------
Finished episode: 390 Reward: -1246.4465 total_loss = 14.2064 = -0.0143 + 0.5 * 2
8.4428 + 0.01 * -0.0656
----------------
Finished episode: 391 Reward: -1261.9896 total_loss = 14.7179 = -0.0003 + 0.5 * 2
9.4366 + 0.01 * -0.0104
----------------
Finished episode: 392 Reward: -1235.1856 total_loss = 13.0970 = 0.0483 + 0.5 * 26.
0983 + 0.01 * -0.0356
----------------
Finished episode: 393 Reward: -1242.8237 total_loss = 17.6549 = 0.0022 + 0.5 * 35.
3062 + 0.01 * -0.0413
----------------
Finished episode: 394 Reward: -1257.8953 total_loss = 15.4012 = 0.0384 + 0.5 * 30.
7263 + 0.01 * -0.0396
----------------
Finished episode: 395 Reward: -1264.5743 total_loss = 15.7369 = -0.0949 + 0.5 * 3
1.6640 + 0.01 * -0.0206
----------------
Finished episode: 396 Reward: -1240.7984 total_loss = 16.1200 = -0.0808 + 0.5 * 3
2.4021 + 0.01 * -0.0193
----------------
Finished episode: 397 Reward: -1250.5307 total_loss = 14.5920 = 0.0068 + 0.5 * 29.
1712 + 0.01 * -0.0377
----------------
Finished episode: 398 Reward: -1255.1689 total_loss = 13.9170 = 0.0833 + 0.5 * 27.
6674 + 0.01 * 0.0100
----------------
Finished episode: 399 Reward: -1231.1084 total_loss = 15.5639 = -0.0974 + 0.5 * 3
1.3224 + 0.01 * 0.0141
----------------
Finished episode: 400 Reward: -1243.2967 total_loss = 17.3431 = -0.0211 + 0.5 * 3
4.7285 + 0.01 * -0.0097
----------------
Finished episode: 401 Reward: -1255.7069 total_loss = 14.5605 = -0.0060 + 0.5 * 2
9.1334 + 0.01 * -0.0157
----------------
Finished episode: 402 Reward: -1243.1557 total_loss = 14.3332 = -0.0749 + 0.5 * 2
8.8159 + 0.01 * 0.0087
----------------
Finished episode: 403 Reward: -1246.6050 total_loss = 15.3322 = 0.0301 + 0.5 * 30.
6040 + 0.01 * 0.0104
----------------
Finished episode: 404 Reward: -1230.1801 total_loss = 16.7914 = -0.0225 + 0.5 * 3
3.6275 + 0.01 * 0.0142
----------------
Finished episode: 405 Reward: -1231.7428 total_loss = 14.4253 = 0.0679 + 0.5 * 28.
7150 + 0.01 * -0.0152
----------------
Finished episode: 406 Reward: -1258.7949 total_loss = 15.0360 = -0.0046 + 0.5 * 3
0.0808 + 0.01 * 0.0204
```

```
----------------
Finished episode: 407 Reward: -1235.7652 total_loss = 17.7618 = 0.0535 + 0.5 * 35.
4162 + 0.01 * 0.0222
----------------
Finished episode: 408 Reward: -1228.2745 total_loss = 17.8064 = 0.0444 + 0.5 * 35.
5241 + 0.01 * 0.0003
----------------
Finished episode: 409 Reward: -1237.0934 total_loss = 13.3023 = 0.0561 + 0.5 * 26.
4915 + 0.01 * 0.0426
----------------
Finished episode: 410 Reward: -1270.6761 total_loss = 15.2783 = -0.0213 + 0.5 * 3
0.5983 + 0.01 * 0.0399
----------------
Finished episode: 411 Reward: -1249.5736 total_loss = 14.6263 = 0.0375 + 0.5 * 29.
1770 + 0.01 * 0.0341
----------------
Finished episode: 412 Reward: -1257.3611 total_loss = 15.6239 = 0.0585 + 0.5 * 31.
1303 + 0.01 * 0.0243
----------------
Finished episode: 413 Reward: -1219.4982 total_loss = 15.2404 = -0.0143 + 0.5 * 3
0.5088 + 0.01 * 0.0257
----------------
Finished episode: 414 Reward: -1254.2576 total_loss = 12.7097 = -0.0162 + 0.5 * 2
5.4514 + 0.01 * 0.0181
----------------
Finished episode: 415 Reward: -1226.8329 total_loss = 15.9464 = -0.0079 + 0.5 * 3
1.9083 + 0.01 * 0.0116
----------------
Finished episode: 416 Reward: -1262.3972 total_loss = 13.2747 = -0.0357 + 0.5 * 2
6.6202 + 0.01 * 0.0338
----------------
Finished episode: 417 Reward: -1249.3372 total_loss = 15.9358 = 0.0443 + 0.5 * 31.
7825 + 0.01 * 0.0301
----------------
Finished episode: 418 Reward: -1246.1432 total_loss = 15.3291 = -0.0216 + 0.5 * 3
0.7005 + 0.01 * 0.0428
----------------
Finished episode: 419 Reward: -1252.0427 total_loss = 14.0152 = -0.0309 + 0.5 * 2
8.0921 + 0.01 * 0.0033
----------------
Finished episode: 420 Reward: -1229.8673 total_loss = 15.8954 = 0.0148 + 0.5 * 31.
7602 + 0.01 * 0.0407
----------------
Finished episode: 421 Reward: -1228.3015 total_loss = 17.5992 = -0.0641 + 0.5 * 3
5.3260 + 0.01 * 0.0312
----------------
Finished episode: 422 Reward: -1263.1553 total_loss = 13.0119 = -0.1617 + 0.5 * 2
6.3463 + 0.01 * 0.0450
----------------
Finished episode: 423 Reward: -1270.1231 total_loss = 13.1884 = 0.1405 + 0.5 * 26.
0945 + 0.01 * 0.0676
----------------
Finished episode: 424 Reward: -1254.2709 total_loss = 16.5464 = 0.1555 + 0.5 * 32.
7809 + 0.01 * 0.0508
----------------
Finished episode: 425 Reward: -1252.2778 total_loss = 15.4172 = -0.0174 + 0.5 * 3
0.8679 + 0.01 * 0.0594
----------------
Finished episode: 426 Reward: -1273.2595 total_loss = 12.7282 = 0.0180 + 0.5 * 25.
4195 + 0.01 * 0.0445
----------------
```

```
Finished episode: 427 Reward: -1265.3948 total_loss = 13.0136 = -0.0717 + 0.5 * 2
6.1700 + 0.01 * 0.0319
-----------------
Finished episode: 428 Reward: -1277.4511 total_loss = 13.8170 = -0.0340 + 0.5 * 2
7.7009 + 0.01 * 0.0573
-----------------
Finished episode: 429 Reward: -1268.9065 total_loss = 13.1552 = -0.1417 + 0.5 * 2
6.5926 + 0.01 * 0.0663
-----------------
Finished episode: 430 Reward: -1255.3187 total_loss = 14.8137 = -0.0702 + 0.5 * 2
9.7672 + 0.01 * 0.0379
-----------------
Finished episode: 431 Reward: -1275.4162 total_loss = 14.5197 = -0.1261 + 0.5 * 2
9.2901 + 0.01 * 0.0747
-----------------
Finished episode: 432 Reward: -1271.3859 total_loss = 14.9009 = 0.0075 + 0.5 * 29.
7846 + 0.01 * 0.1074
-----------------
Finished episode: 433 Reward: -1275.3186 total_loss = 12.6689 = 0.0198 + 0.5 * 25.
2964 + 0.01 * 0.0900
-----------------
Finished episode: 434 Reward: -1246.7289 total_loss = 13.8816 = 0.0108 + 0.5 * 27.
7405 + 0.01 * 0.0592
-----------------
Finished episode: 435 Reward: -1239.4084 total_loss = 15.6375 = -0.0662 + 0.5 * 3
1.4056 + 0.01 * 0.0917
-----------------
Finished episode: 436 Reward: -1276.5116 total_loss = 14.7498 = 0.0203 + 0.5 * 29.
4570 + 0.01 * 0.0940
-----------------
Finished episode: 437 Reward: -1259.6340 total_loss = 12.5083 = -0.0049 + 0.5 * 2
5.0241 + 0.01 * 0.1207
-----------------
Finished episode: 438 Reward: -1257.2749 total_loss = 13.6579 = -0.0619 + 0.5 * 2
7.4376 + 0.01 * 0.1016
-----------------
Finished episode: 439 Reward: -1282.8781 total_loss = 13.0595 = -0.1081 + 0.5 * 2
6.3336 + 0.01 * 0.0862
-----------------
Finished episode: 440 Reward: -1251.1023 total_loss = 14.6163 = 0.0653 + 0.5 * 29.
1005 + 0.01 * 0.0720
-----------------
Finished episode: 441 Reward: -1253.4247 total_loss = 10.5597 = 0.0195 + 0.5 * 21.
0788 + 0.01 * 0.0736
-----------------
Finished episode: 442 Reward: -1265.1010 total_loss = 12.6028 = 0.0448 + 0.5 * 25.
1142 + 0.01 * 0.0908
-----------------
Finished episode: 443 Reward: -1237.3140 total_loss = 12.4029 = -0.0168 + 0.5 * 2
4.8378 + 0.01 * 0.0820
-----------------
Finished episode: 444 Reward: -1264.2959 total_loss = 13.5841 = -0.0930 + 0.5 * 2
7.3527 + 0.01 * 0.0715
-----------------
Finished episode: 445 Reward: -1235.7877 total_loss = 13.7962 = 0.1169 + 0.5 * 27.
3568 + 0.01 * 0.0879
-----------------
Finished episode: 446 Reward: -1249.2298 total_loss = 11.4619 = 0.0612 + 0.5 * 22.
7994 + 0.01 * 0.1001
-----------------
Finished episode: 447 Reward: -1242.8138 total_loss = 14.9677 = -0.0007 + 0.5 * 2
```

```
9.9350 + 0.01 * 0.0890
----------
Finished episode: 448 Reward: -1265.8292 total_loss = 12.0566 = -0.0478 + 0.5 * 2
4.2074 + 0.01 * 0.0722
----------
Finished episode: 449 Reward: -1254.9830 total_loss = 12.3635 = -0.1354 + 0.5 * 2
4.9959 + 0.01 * 0.0994
----------
Finished episode: 450 Reward: -1255.3730 total_loss = 13.0390 = 0.0842 + 0.5 * 25.
9083 + 0.01 * 0.0602
----------
Finished episode: 451 Reward: -1257.8662 total_loss = 11.4186 = -0.0028 + 0.5 * 2
2.8409 + 0.01 * 0.0960
----------
Finished episode: 452 Reward: -1242.2987 total_loss = 12.3589 = -0.0077 + 0.5 * 2
4.7322 + 0.01 * 0.0517
----------
Finished episode: 453 Reward: -1237.3138 total_loss = 11.4019 = 0.0218 + 0.5 * 22.
7581 + 0.01 * 0.1028
----------
Finished episode: 454 Reward: -1262.2290 total_loss = 11.7743 = -0.0099 + 0.5 * 2
3.5664 + 0.01 * 0.0970
----------
Finished episode: 455 Reward: -1262.5946 total_loss = 12.3156 = 0.0655 + 0.5 * 24.
4982 + 0.01 * 0.0923
----------
Finished episode: 456 Reward: -1250.3959 total_loss = 12.4335 = -0.0706 + 0.5 * 2
5.0065 + 0.01 * 0.0837
----------
Finished episode: 457 Reward: -1256.0609 total_loss = 13.2736 = -0.0046 + 0.5 * 2
6.5542 + 0.01 * 0.1016
----------
Finished episode: 458 Reward: -1271.0794 total_loss = 12.5018 = 0.0221 + 0.5 * 24.
9573 + 0.01 * 0.1033
----------
Finished episode: 459 Reward: -1265.2631 total_loss = 13.5377 = -0.0440 + 0.5 * 2
7.1613 + 0.01 * 0.1068
----------
Finished episode: 460 Reward: -1252.4990 total_loss = 14.5791 = 0.0632 + 0.5 * 29.
0299 + 0.01 * 0.0928
----------
Finished episode: 461 Reward: -1254.0652 total_loss = 17.1085 = -0.0061 + 0.5 * 3
4.2273 + 0.01 * 0.0958
----------
Finished episode: 462 Reward: -1263.5302 total_loss = 13.2132 = -0.0020 + 0.5 * 2
6.4283 + 0.01 * 0.1037
----------
Finished episode: 463 Reward: -1273.4512 total_loss = 13.2481 = -0.0036 + 0.5 * 2
6.5012 + 0.01 * 0.1149
----------
Finished episode: 464 Reward: -1271.7795 total_loss = 13.2728 = -0.0106 + 0.5 * 2
6.5648 + 0.01 * 0.1026
----------
Finished episode: 465 Reward: -1238.6400 total_loss = 12.9362 = -0.0023 + 0.5 * 2
5.8749 + 0.01 * 0.1089
----------
Finished episode: 466 Reward: -1229.9770 total_loss = 13.8760 = -0.0473 + 0.5 * 2
7.8447 + 0.01 * 0.0918
----------
Finished episode: 467 Reward: -1284.4639 total_loss = 14.4152 = -0.0312 + 0.5 * 2
8.8908 + 0.01 * 0.0962
```

```
----------
Finished episode: 468 Reward: -1262.6379 total_loss = 13.0299 = 0.1477 + 0.5 * 25.
7626 + 0.01 * 0.0912
----------
Finished episode: 469 Reward: -1281.8043 total_loss = 15.6933 = 0.1000 + 0.5 * 31.
1842 + 0.01 * 0.1212
----------
Finished episode: 470 Reward: -1252.5241 total_loss = 12.4703 = 0.0697 + 0.5 * 24.
7987 + 0.01 * 0.1353
----------
Finished episode: 471 Reward: -1263.5372 total_loss = 13.3977 = -0.0898 + 0.5 * 2
6.9730 + 0.01 * 0.1078
----------
Finished episode: 472 Reward: -1248.3623 total_loss = 15.1984 = -0.0096 + 0.5 * 3
0.4141 + 0.01 * 0.0934
----------
Finished episode: 473 Reward: -1259.1738 total_loss = 14.7698 = 0.1378 + 0.5 * 29.
2611 + 0.01 * 0.1407
----------
Finished episode: 474 Reward: -1246.1133 total_loss = 11.6287 = -0.0989 + 0.5 * 2
3.4528 + 0.01 * 0.1221
----------
Finished episode: 475 Reward: -1279.2278 total_loss = 14.4218 = 0.0139 + 0.5 * 28.
8132 + 0.01 * 0.1326
----------
Finished episode: 476 Reward: -1249.5804 total_loss = 11.6958 = 0.0637 + 0.5 * 23.
2606 + 0.01 * 0.1779
----------
Finished episode: 477 Reward: -1263.1219 total_loss = 13.4952 = 0.0963 + 0.5 * 26.
7954 + 0.01 * 0.1165
----------
Finished episode: 478 Reward: -1278.7364 total_loss = 14.5344 = 0.0283 + 0.5 * 29.
0098 + 0.01 * 0.1128
----------
Finished episode: 479 Reward: -1265.2144 total_loss = 12.0077 = -0.0698 + 0.5 * 2
4.1527 + 0.01 * 0.1092
----------
Finished episode: 480 Reward: -1236.3535 total_loss = 11.5118 = 0.0223 + 0.5 * 22.
9761 + 0.01 * 0.1457
----------
Finished episode: 481 Reward: -1272.4351 total_loss = 14.8821 = -0.0443 + 0.5 * 2
9.8498 + 0.01 * 0.1460
----------
Finished episode: 482 Reward: -1261.4836 total_loss = 13.4425 = 0.0486 + 0.5 * 26.
7844 + 0.01 * 0.1667
----------
Finished episode: 483 Reward: -1262.5003 total_loss = 13.3915 = 0.0046 + 0.5 * 26.
7708 + 0.01 * 0.1581
----------
Finished episode: 484 Reward: -1244.7074 total_loss = 14.8781 = 0.0104 + 0.5 * 29.
7322 + 0.01 * 0.1620
----------
Finished episode: 485 Reward: -1280.9973 total_loss = 12.2720 = 0.0336 + 0.5 * 24.
4741 + 0.01 * 0.1417
----------
Finished episode: 486 Reward: -1271.0817 total_loss = 12.0297 = -0.0152 + 0.5 * 2
4.0867 + 0.01 * 0.1522
----------
Finished episode: 487 Reward: -1283.3207 total_loss = 14.9391 = -0.0813 + 0.5 * 3
0.0385 + 0.01 * 0.1158
----------
```

```
Finished episode: 488 Reward: -1258.0734 total_loss = 14.5370 = 0.0023 + 0.5 * 29.
0664 + 0.01 * 0.1531
-----------------
Finished episode: 489 Reward: -1258.8676 total_loss = 14.9641 = -0.0955 + 0.5 * 3
0.1159 + 0.01 * 0.1597
-----------------
Finished episode: 490 Reward: -1272.5380 total_loss = 11.5423 = -0.0707 + 0.5 * 2
3.2225 + 0.01 * 0.1734
-----------------
Finished episode: 491 Reward: -1271.1947 total_loss = 14.2696 = -0.0130 + 0.5 * 2
8.5618 + 0.01 * 0.1802
-----------------
Finished episode: 492 Reward: -1274.7477 total_loss = 15.2640 = -0.0128 + 0.5 * 3
0.5509 + 0.01 * 0.1320
-----------------
Finished episode: 493 Reward: -1248.5310 total_loss = 14.8902 = -0.0205 + 0.5 * 2
9.8189 + 0.01 * 0.1290
-----------------
Finished episode: 494 Reward: -1259.6753 total_loss = 12.1268 = 0.0397 + 0.5 * 24.
1713 + 0.01 * 0.1393
-----------------
Finished episode: 495 Reward: -1282.8186 total_loss = 15.4237 = -0.0060 + 0.5 * 3
0.8555 + 0.01 * 0.1969
-----------------
Finished episode: 496 Reward: -1266.5331 total_loss = 12.2744 = 0.0543 + 0.5 * 24.
4368 + 0.01 * 0.1801
-----------------
Finished episode: 497 Reward: -1259.0506 total_loss = 14.0091 = 0.1205 + 0.5 * 27.
7735 + 0.01 * 0.1843
-----------------
Finished episode: 498 Reward: -1253.1630 total_loss = 12.4148 = 0.0654 + 0.5 * 24.
6949 + 0.01 * 0.1998
-----------------
Finished episode: 499 Reward: -1270.2699 total_loss = 13.5535 = 0.0694 + 0.5 * 26.
9645 + 0.01 * 0.1852
-----------------
Finished episode: 500 Reward: -1266.7354 total_loss = 14.2238 = 0.0334 + 0.5 * 28.
3769 + 0.01 * 0.1964
-----------------
Finished episode: 501 Reward: -1271.1188 total_loss = 12.8213 = -0.0241 + 0.5 * 2
5.6878 + 0.01 * 0.1549
-----------------
Finished episode: 502 Reward: -1278.7752 total_loss = 14.7725 = -0.1149 + 0.5 * 2
9.7719 + 0.01 * 0.1480
-----------------
Finished episode: 503 Reward: -1254.1510 total_loss = 16.2138 = 0.0283 + 0.5 * 32.
3671 + 0.01 * 0.1981
-----------------
Finished episode: 504 Reward: -1267.9023 total_loss = 13.7396 = 0.0032 + 0.5 * 27.
4694 + 0.01 * 0.1747
-----------------
Finished episode: 505 Reward: -1265.8072 total_loss = 13.5977 = -0.1162 + 0.5 * 2
7.4234 + 0.01 * 0.2168
-----------------
Finished episode: 506 Reward: -1262.8332 total_loss = 18.2974 = -0.1619 + 0.5 * 3
6.9154 + 0.01 * 0.1645
-----------------
Finished episode: 507 Reward: -1266.0570 total_loss = 13.6579 = -0.0416 + 0.5 * 2
7.3957 + 0.01 * 0.1761
-----------------
Finished episode: 508 Reward: -1258.2263 total_loss = 14.1404 = -0.0624 + 0.5 * 2
```

```
8.4027 + 0.01 * 0.1469
----------------
Finished episode: 509 Reward: -1282.8317 total_loss = 14.0297 = 0.1450 + 0.5 * 27.
7652 + 0.01 * 0.2016
----------------
Finished episode: 510 Reward: -1232.2599 total_loss = 17.3874 = -0.1328 + 0.5 * 3
5.0368 + 0.01 * 0.1805
----------------
Finished episode: 511 Reward: -1274.3719 total_loss = 13.7536 = 0.0187 + 0.5 * 27.
4659 + 0.01 * 0.1892
----------------
Finished episode: 512 Reward: -1274.9330 total_loss = 12.1380 = -0.0337 + 0.5 * 2
4.3398 + 0.01 * 0.1859
----------------
Finished episode: 513 Reward: -1260.9576 total_loss = 13.2036 = -0.0740 + 0.5 * 2
6.5515 + 0.01 * 0.1839
----------------
Finished episode: 514 Reward: -1248.5026 total_loss = 11.7104 = 0.0910 + 0.5 * 23.
2346 + 0.01 * 0.2125
----------------
Finished episode: 515 Reward: -1274.0317 total_loss = 15.5027 = -0.0452 + 0.5 * 3
1.0918 + 0.01 * 0.1977
----------------
Finished episode: 516 Reward: -1289.7460 total_loss = 13.4757 = 0.0015 + 0.5 * 26.
9453 + 0.01 * 0.1547
----------------
Finished episode: 517 Reward: -1277.2842 total_loss = 13.1722 = -0.0114 + 0.5 * 2
6.3626 + 0.01 * 0.2221
----------------
Finished episode: 518 Reward: -1274.6738 total_loss = 13.0111 = 0.0292 + 0.5 * 25.
9592 + 0.01 * 0.2241
----------------
Finished episode: 519 Reward: -1253.4809 total_loss = 12.6255 = -0.0436 + 0.5 * 2
5.3339 + 0.01 * 0.2179
----------------
Finished episode: 520 Reward: -1252.7183 total_loss = 12.3501 = -0.0118 + 0.5 * 2
4.7196 + 0.01 * 0.2140
----------------
Finished episode: 521 Reward: -1283.2486 total_loss = 14.5267 = -0.0110 + 0.5 * 2
9.0715 + 0.01 * 0.1978
----------------
Finished episode: 522 Reward: -1266.1529 total_loss = 13.3375 = 0.0286 + 0.5 * 26.
6146 + 0.01 * 0.1585
----------------
Finished episode: 523 Reward: -1254.0722 total_loss = 14.5013 = 0.0602 + 0.5 * 28.
8783 + 0.01 * 0.1893
----------------
Finished episode: 524 Reward: -1251.4503 total_loss = 12.4945 = -0.0612 + 0.5 * 2
5.1075 + 0.01 * 0.1970
----------------
Finished episode: 525 Reward: -1272.3285 total_loss = 11.8582 = -0.0681 + 0.5 * 2
3.8494 + 0.01 * 0.1612
----------------
Finished episode: 526 Reward: -1277.5539 total_loss = 12.2390 = -0.0279 + 0.5 * 2
4.5304 + 0.01 * 0.1646
----------------
Finished episode: 527 Reward: -1278.3649 total_loss = 13.0171 = 0.0519 + 0.5 * 25.
9273 + 0.01 * 0.1600
----------------
Finished episode: 528 Reward: -1244.2470 total_loss = 13.7368 = -0.0008 + 0.5 * 2
7.4702 + 0.01 * 0.2453
```

```
----------------
Finished episode: 529 Reward: -1275.0143 total_loss = 14.2056 = 0.0441 + 0.5 * 28.
3185 + 0.01 * 0.2237
----------------
Finished episode: 530 Reward: -1272.1085 total_loss = 14.7997 = -0.0693 + 0.5 * 2
9.7336 + 0.01 * 0.2263
----------------
Finished episode: 531 Reward: -1232.3811 total_loss = 13.3231 = 0.0992 + 0.5 * 26.
4426 + 0.01 * 0.2559
----------------
Finished episode: 532 Reward: -1247.0264 total_loss = 14.7030 = 0.0528 + 0.5 * 29.
2957 + 0.01 * 0.2264
----------------
Finished episode: 533 Reward: -1277.5206 total_loss = 15.7467 = -0.0904 + 0.5 * 3
1.6706 + 0.01 * 0.1796
----------------
Finished episode: 534 Reward: -1261.8941 total_loss = 10.6677 = 0.1274 + 0.5 * 21.
0767 + 0.01 * 0.2003
----------------
Finished episode: 535 Reward: -1277.1799 total_loss = 13.2894 = 0.0655 + 0.5 * 26.
4434 + 0.01 * 0.2142
----------------
Finished episode: 536 Reward: -1277.1531 total_loss = 12.3739 = -0.0354 + 0.5 * 2
4.8144 + 0.01 * 0.2065
----------------
Finished episode: 537 Reward: -1267.8118 total_loss = 13.3730 = 0.0396 + 0.5 * 26.
6631 + 0.01 * 0.1809
----------------
Finished episode: 538 Reward: -1252.8597 total_loss = 12.4299 = 0.0686 + 0.5 * 24.
7179 + 0.01 * 0.2300
----------------
Finished episode: 539 Reward: -1255.2314 total_loss = 13.8800 = -0.0387 + 0.5 * 2
7.8327 + 0.01 * 0.2333
----------------
Finished episode: 540 Reward: -1265.9342 total_loss = 12.2709 = -0.0306 + 0.5 * 2
4.5985 + 0.01 * 0.2252
----------------
Finished episode: 541 Reward: -1262.1176 total_loss = 14.1581 = 0.0545 + 0.5 * 28.
2033 + 0.01 * 0.1906
----------------
Finished episode: 542 Reward: -1262.1666 total_loss = 13.8070 = -0.0329 + 0.5 * 2
7.6757 + 0.01 * 0.2115
----------------
Finished episode: 543 Reward: -1279.3946 total_loss = 11.8462 = -0.0986 + 0.5 * 2
3.8861 + 0.01 * 0.1734
----------------
Finished episode: 544 Reward: -1275.7413 total_loss = 16.0968 = 0.1889 + 0.5 * 31.
8119 + 0.01 * 0.1876
----------------
Finished episode: 545 Reward: -1260.3372 total_loss = 14.4158 = 0.0259 + 0.5 * 28.
7762 + 0.01 * 0.1701
----------------
Finished episode: 546 Reward: -1283.4586 total_loss = 13.6677 = -0.0674 + 0.5 * 2
7.4665 + 0.01 * 0.1861
----------------
Finished episode: 547 Reward: -1244.4512 total_loss = 16.6617 = -0.0081 + 0.5 * 3
3.3361 + 0.01 * 0.1732
----------------
Finished episode: 548 Reward: -1250.6266 total_loss = 17.9916 = 0.0462 + 0.5 * 35.
8877 + 0.01 * 0.1508
----------------
```

```
Finished episode: 549 Reward: -1260.2183 total_loss = 13.8011 = -0.0253 + 0.5 * 2
7.6486 + 0.01 * 0.2028
-----------------
Finished episode: 550 Reward: -1271.3659 total_loss = 14.3800 = 0.0732 + 0.5 * 28.
6094 + 0.01 * 0.2081
-----------------
Finished episode: 551 Reward: -1229.5986 total_loss = 13.5188 = 0.0001 + 0.5 * 27.
0326 + 0.01 * 0.2287
-----------------
Finished episode: 552 Reward: -1271.3835 total_loss = 15.3645 = -0.0940 + 0.5 * 3
0.9131 + 0.01 * 0.1896
-----------------
Finished episode: 553 Reward: -1251.5379 total_loss = 14.6889 = 0.0249 + 0.5 * 29.
3236 + 0.01 * 0.2197
-----------------
Finished episode: 554 Reward: -1252.1226 total_loss = 11.3663 = -0.0642 + 0.5 * 2
2.8562 + 0.01 * 0.2339
-----------------
Finished episode: 555 Reward: -1228.4893 total_loss = 14.6898 = 0.0209 + 0.5 * 29.
3341 + 0.01 * 0.1909
-----------------
Finished episode: 556 Reward: -1259.6483 total_loss = 11.8845 = 0.1747 + 0.5 * 23.
4161 + 0.01 * 0.1808
-----------------
Finished episode: 557 Reward: -1274.3443 total_loss = 14.5347 = 0.0286 + 0.5 * 29.
0075 + 0.01 * 0.2350
-----------------
Finished episode: 558 Reward: -1252.3772 total_loss = 16.1295 = 0.0789 + 0.5 * 32.
0965 + 0.01 * 0.2329
-----------------
Finished episode: 559 Reward: -1278.1271 total_loss = 13.7630 = -0.0537 + 0.5 * 2
7.6303 + 0.01 * 0.1516
-----------------
Finished episode: 560 Reward: -1288.8055 total_loss = 14.1624 = -0.0037 + 0.5 * 2
8.3270 + 0.01 * 0.2535
-----------------
Finished episode: 561 Reward: -1255.7513 total_loss = 17.3707 = -0.0482 + 0.5 * 3
4.8333 + 0.01 * 0.2246
-----------------
Finished episode: 562 Reward: -1266.4042 total_loss = 16.1792 = 0.0070 + 0.5 * 32.
3398 + 0.01 * 0.2233
-----------------
Finished episode: 563 Reward: -1274.0370 total_loss = 15.6701 = 0.0574 + 0.5 * 31.
2214 + 0.01 * 0.1991
-----------------
Finished episode: 564 Reward: -1248.7894 total_loss = 14.7792 = -0.0319 + 0.5 * 2
9.6189 + 0.01 * 0.1691
-----------------
Finished episode: 565 Reward: -1285.5389 total_loss = 14.1599 = 0.0451 + 0.5 * 28.
2251 + 0.01 * 0.2268
-----------------
Finished episode: 566 Reward: -1285.6768 total_loss = 15.5569 = -0.0069 + 0.5 * 3
1.1221 + 0.01 * 0.2735
-----------------
Finished episode: 567 Reward: -1257.4953 total_loss = 15.0305 = 0.0812 + 0.5 * 29.
8934 + 0.01 * 0.2609
-----------------
Finished episode: 568 Reward: -1276.0100 total_loss = 13.4721 = 0.0646 + 0.5 * 26.
8108 + 0.01 * 0.2134
-----------------
Finished episode: 569 Reward: -1287.9326 total_loss = 12.7780 = -0.0207 + 0.5 * 2
```

```
5.5936 + 0.01 * 0.1949
----------------
Finished episode: 570 Reward: -1278.0609 total_loss = 14.4980 = 0.0073 + 0.5 * 28.
9776 + 0.01 * 0.1874
----------------
Finished episode: 571 Reward: -1242.7813 total_loss = 14.9917 = -0.0501 + 0.5 * 3
0.0802 + 0.01 * 0.1727
----------------
Finished episode: 572 Reward: -1265.2400 total_loss = 14.1661 = -0.0041 + 0.5 * 2
8.3355 + 0.01 * 0.2491
----------------
Finished episode: 573 Reward: -1260.7355 total_loss = 15.4154 = 0.0522 + 0.5 * 30.
7223 + 0.01 * 0.1988
----------------
Finished episode: 574 Reward: -1256.6771 total_loss = 15.2905 = 0.0051 + 0.5 * 30.
5663 + 0.01 * 0.2244
----------------
Finished episode: 575 Reward: -1268.8489 total_loss = 14.9640 = -0.0309 + 0.5 * 2
9.9860 + 0.01 * 0.1885
----------------
Finished episode: 576 Reward: -1250.3081 total_loss = 14.1636 = -0.0665 + 0.5 * 2
8.4563 + 0.01 * 0.1930
----------------
Finished episode: 577 Reward: -1264.4383 total_loss = 12.6114 = 0.0290 + 0.5 * 25.
1606 + 0.01 * 0.2056
----------------
Finished episode: 578 Reward: -1285.7707 total_loss = 12.9842 = 0.0077 + 0.5 * 25.
9494 + 0.01 * 0.1801
----------------
Finished episode: 579 Reward: -1262.8089 total_loss = 13.9177 = 0.0677 + 0.5 * 27.
6969 + 0.01 * 0.1624
----------------
Finished episode: 580 Reward: -1290.5574 total_loss = 13.1302 = -0.0705 + 0.5 * 2
6.3973 + 0.01 * 0.2058
----------------
Finished episode: 581 Reward: -1271.3020 total_loss = 15.2404 = -0.0779 + 0.5 * 3
0.6326 + 0.01 * 0.1966
----------------
Finished episode: 582 Reward: -1274.8912 total_loss = 13.3510 = 0.0000 + 0.5 * 26.
6972 + 0.01 * 0.2365
----------------
Finished episode: 583 Reward: -1259.6816 total_loss = 13.6790 = 0.0274 + 0.5 * 27.
2993 + 0.01 * 0.1956
----------------
Finished episode: 584 Reward: -1275.9600 total_loss = 13.4088 = -0.0699 + 0.5 * 2
6.9536 + 0.01 * 0.1858
----------------
Finished episode: 585 Reward: -1290.8394 total_loss = 15.1725 = -0.1799 + 0.5 * 3
0.7007 + 0.01 * 0.2028
----------------
Finished episode: 586 Reward: -1271.9130 total_loss = 13.9184 = 0.0792 + 0.5 * 27.
6750 + 0.01 * 0.1772
----------------
Finished episode: 587 Reward: -1278.4475 total_loss = 12.9359 = 0.0608 + 0.5 * 25.
7455 + 0.01 * 0.2320
----------------
Finished episode: 588 Reward: -1264.2429 total_loss = 11.9686 = -0.0145 + 0.5 * 2
3.9615 + 0.01 * 0.2342
----------------
Finished episode: 589 Reward: -1280.6202 total_loss = 14.2761 = 0.0025 + 0.5 * 28.
5425 + 0.01 * 0.2255
```

```
----------------
Finished episode: 590 Reward: -1263.8956 total_loss = 14.8372 = -0.1225 + 0.5 * 2
9.9150 + 0.01 * 0.2179
----------------
Finished episode: 591 Reward: -1265.8102 total_loss = 12.5086 = 0.0619 + 0.5 * 24.
8883 + 0.01 * 0.2515
----------------
Finished episode: 592 Reward: -1249.0987 total_loss = 13.5100 = 0.0654 + 0.5 * 26.
8852 + 0.01 * 0.2065
----------------
Finished episode: 593 Reward: -1267.9716 total_loss = 14.9141 = -0.0396 + 0.5 * 2
9.9035 + 0.01 * 0.1923
----------------
Finished episode: 594 Reward: -1276.6117 total_loss = 13.3524 = -0.0804 + 0.5 * 2
6.8619 + 0.01 * 0.1841
----------------
Finished episode: 595 Reward: -1247.8072 total_loss = 13.5203 = 0.0284 + 0.5 * 26.
9795 + 0.01 * 0.2175
----------------
Finished episode: 596 Reward: -1257.2917 total_loss = 12.5531 = 0.0296 + 0.5 * 25.
0420 + 0.01 * 0.2493
----------------
Finished episode: 597 Reward: -1259.2835 total_loss = 15.2944 = 0.0415 + 0.5 * 30.
5017 + 0.01 * 0.1989
----------------
Finished episode: 598 Reward: -1254.6604 total_loss = 12.0366 = 0.0310 + 0.5 * 24.
0070 + 0.01 * 0.2106
----------------
Finished episode: 599 Reward: -1252.6105 total_loss = 17.0804 = -0.0431 + 0.5 * 3
4.2432 + 0.01 * 0.1912
----------------
Finished episode: 600 Reward: -1271.4085 total_loss = 13.2551 = 0.0199 + 0.5 * 26.
4658 + 0.01 * 0.2281
----------------
Finished episode: 601 Reward: -1261.9284 total_loss = 13.0928 = -0.0301 + 0.5 * 2
6.2414 + 0.01 * 0.2208
----------------
Finished episode: 602 Reward: -1254.9760 total_loss = 15.8599 = -0.0650 + 0.5 * 3
1.8467 + 0.01 * 0.1516
----------------
Finished episode: 603 Reward: -1257.6148 total_loss = 11.6461 = -0.0656 + 0.5 * 2
3.4199 + 0.01 * 0.1743
----------------
Finished episode: 604 Reward: -1272.6573 total_loss = 14.5246 = -0.0524 + 0.5 * 2
9.1500 + 0.01 * 0.1991
----------------
Finished episode: 605 Reward: -1262.1637 total_loss = 13.8505 = -0.0294 + 0.5 * 2
7.7556 + 0.01 * 0.2107
----------------
Finished episode: 606 Reward: -1282.1933 total_loss = 11.4335 = -0.0882 + 0.5 * 2
3.0386 + 0.01 * 0.2400
----------------
Finished episode: 607 Reward: -1263.7887 total_loss = 12.9325 = 0.0577 + 0.5 * 25.
7445 + 0.01 * 0.2556
----------------
Finished episode: 608 Reward: -1245.4102 total_loss = 14.9338 = -0.0650 + 0.5 * 2
9.9930 + 0.01 * 0.2297
----------------
Finished episode: 609 Reward: -1280.6167 total_loss = 12.7847 = 0.0334 + 0.5 * 25.
4980 + 0.01 * 0.2310
----------------
```

```
Finished episode: 610 Reward: -1266.4328 total_loss = 12.6330 = -0.1407 + 0.5 * 2
5.5431 + 0.01 * 0.2216
-----------------
Finished episode: 611 Reward: -1269.6541 total_loss = 16.3114 = -0.0081 + 0.5 * 3
2.6352 + 0.01 * 0.1955
-----------------
Finished episode: 612 Reward: -1292.3161 total_loss = 14.1777 = -0.0355 + 0.5 * 2
8.4230 + 0.01 * 0.1628
-----------------
Finished episode: 613 Reward: -1249.7194 total_loss = 17.7656 = -0.0081 + 0.5 * 3
5.5431 + 0.01 * 0.2091
-----------------
Finished episode: 614 Reward: -1259.1213 total_loss = 16.7469 = -0.0200 + 0.5 * 3
3.5290 + 0.01 * 0.2379
-----------------
Finished episode: 615 Reward: -1265.5892 total_loss = 18.0555 = -0.0350 + 0.5 * 3
6.1771 + 0.01 * 0.1916
-----------------
Finished episode: 616 Reward: -1248.5488 total_loss = 17.1697 = -0.0513 + 0.5 * 3
4.4381 + 0.01 * 0.1929
-----------------
Finished episode: 617 Reward: -1286.6533 total_loss = 14.7636 = -0.0109 + 0.5 * 2
9.5449 + 0.01 * 0.2040
-----------------
Finished episode: 618 Reward: -1267.7392 total_loss = 14.0401 = -0.0568 + 0.5 * 2
8.1901 + 0.01 * 0.1949
-----------------
Finished episode: 619 Reward: -1293.0443 total_loss = 15.0601 = -0.0140 + 0.5 * 3
0.1438 + 0.01 * 0.2201
-----------------
Finished episode: 620 Reward: -1260.5651 total_loss = 15.5748 = 0.0642 + 0.5 * 31.
0176 + 0.01 * 0.1820
-----------------
Finished episode: 621 Reward: -1273.1778 total_loss = 13.6057 = 0.0749 + 0.5 * 27.
0566 + 0.01 * 0.2463
-----------------
Finished episode: 622 Reward: -1268.7105 total_loss = 16.3243 = 0.0420 + 0.5 * 32.
5607 + 0.01 * 0.1959
-----------------
Finished episode: 623 Reward: -1243.7812 total_loss = 16.2142 = -0.0298 + 0.5 * 3
2.4830 + 0.01 * 0.2480
-----------------
Finished episode: 624 Reward: -1266.9529 total_loss = 15.8157 = -0.0241 + 0.5 * 3
1.6755 + 0.01 * 0.2046
-----------------
Finished episode: 625 Reward: -1273.2953 total_loss = 11.7879 = -0.0250 + 0.5 * 2
3.6212 + 0.01 * 0.2284
-----------------
Finished episode: 626 Reward: -1253.0367 total_loss = 16.9240 = 0.0290 + 0.5 * 33.
7861 + 0.01 * 0.1868
-----------------
Finished episode: 627 Reward: -1280.3854 total_loss = 17.1172 = -0.1502 + 0.5 * 3
4.5309 + 0.01 * 0.1978
-----------------
Finished episode: 628 Reward: -1271.5664 total_loss = 13.9787 = 0.0293 + 0.5 * 27.
8952 + 0.01 * 0.1818
-----------------
Finished episode: 629 Reward: -1268.0015 total_loss = 13.7970 = 0.0834 + 0.5 * 27.
4232 + 0.01 * 0.2051
-----------------
Finished episode: 630 Reward: -1246.2496 total_loss = 14.9390 = 0.0025 + 0.5 * 29.
```

```
8679 + 0.01 * 0.2511
----------------
Finished episode: 631 Reward: -1254.9416 total_loss = 18.5634 = -0.0941 + 0.5 * 3
7.3108 + 0.01 * 0.2011
----------------
Finished episode: 632 Reward: -1273.5646 total_loss = 16.0273 = 0.0917 + 0.5 * 31.
8662 + 0.01 * 0.2474
----------------
Finished episode: 633 Reward: -1246.3573 total_loss = 15.3320 = 0.0425 + 0.5 * 30.
5753 + 0.01 * 0.1868
----------------
Finished episode: 634 Reward: -1261.7477 total_loss = 15.8903 = 0.0423 + 0.5 * 31.
6916 + 0.01 * 0.2263
----------------
Finished episode: 635 Reward: -1270.8340 total_loss = 14.3379 = -0.0913 + 0.5 * 2
8.8550 + 0.01 * 0.1769
----------------
Finished episode: 636 Reward: -1251.2230 total_loss = 16.4532 = 0.0203 + 0.5 * 32.
8615 + 0.01 * 0.2199
----------------
Finished episode: 637 Reward: -1281.7180 total_loss = 13.5724 = -0.1209 + 0.5 * 2
7.3825 + 0.01 * 0.2033
----------------
Finished episode: 638 Reward: -1252.6727 total_loss = 16.7873 = 0.0451 + 0.5 * 33.
4803 + 0.01 * 0.2109
----------------
Finished episode: 639 Reward: -1281.1863 total_loss = 14.5361 = 0.0268 + 0.5 * 29.
0145 + 0.01 * 0.1999
----------------
Finished episode: 640 Reward: -1270.4829 total_loss = 16.0241 = 0.0239 + 0.5 * 31.
9969 + 0.01 * 0.1780
----------------
Finished episode: 641 Reward: -1232.1504 total_loss = 17.8986 = 0.0195 + 0.5 * 35.
7542 + 0.01 * 0.1999
----------------
Finished episode: 642 Reward: -1275.7226 total_loss = 14.6986 = 0.1089 + 0.5 * 29.
1750 + 0.01 * 0.2208
----------------
Finished episode: 643 Reward: -1284.5066 total_loss = 15.2650 = 0.0712 + 0.5 * 30.
3826 + 0.01 * 0.2528
----------------
Finished episode: 644 Reward: -1266.0780 total_loss = 18.2751 = -0.0325 + 0.5 * 3
6.6121 + 0.01 * 0.1617
----------------
Finished episode: 645 Reward: -1276.5123 total_loss = 17.9692 = 0.0030 + 0.5 * 35.
9286 + 0.01 * 0.1880
----------------
Finished episode: 646 Reward: -1281.4631 total_loss = 17.2252 = 0.0817 + 0.5 * 34.
2829 + 0.01 * 0.2006
----------------
Finished episode: 647 Reward: -1275.4418 total_loss = 13.6908 = -0.0551 + 0.5 * 2
7.4878 + 0.01 * 0.2053
----------------
Finished episode: 648 Reward: -1293.9727 total_loss = 13.0326 = -0.0474 + 0.5 * 2
6.1556 + 0.01 * 0.2181
----------------
Finished episode: 649 Reward: -1274.6096 total_loss = 14.8256 = -0.0112 + 0.5 * 2
9.6689 + 0.01 * 0.2377
----------------
Finished episode: 650 Reward: -1290.8406 total_loss = 13.6372 = -0.1063 + 0.5 * 2
7.4832 + 0.01 * 0.1857
```

```
----------------
Finished episode: 651 Reward: -1289.9093 total_loss = 14.5339 = -0.0721 + 0.5 * 2
9.2072 + 0.01 * 0.2350
----------------
Finished episode: 652 Reward: -1258.5400 total_loss = 16.6954 = 0.0030 + 0.5 * 33.
3801 + 0.01 * 0.2360
----------------
Finished episode: 653 Reward: -1268.8572 total_loss = 17.6362 = 0.0071 + 0.5 * 35.
2540 + 0.01 * 0.2056
----------------
Finished episode: 654 Reward: -1298.7413 total_loss = 14.5201 = -0.0432 + 0.5 * 2
9.1218 + 0.01 * 0.2369
----------------
Finished episode: 655 Reward: -1249.0108 total_loss = 15.4748 = 0.0205 + 0.5 * 30.
9041 + 0.01 * 0.2217
----------------
Finished episode: 656 Reward: -1280.5671 total_loss = 17.3628 = -0.0240 + 0.5 * 3
4.7686 + 0.01 * 0.2445
----------------
Finished episode: 657 Reward: -1304.7523 total_loss = 14.4527 = -0.0545 + 0.5 * 2
9.0099 + 0.01 * 0.2297
----------------
Finished episode: 658 Reward: -1248.8905 total_loss = 16.0328 = 0.0863 + 0.5 * 31.
8887 + 0.01 * 0.2183
----------------
Finished episode: 659 Reward: -1296.5214 total_loss = 14.5965 = 0.0201 + 0.5 * 29.
1479 + 0.01 * 0.2393
----------------
Finished episode: 660 Reward: -1288.8503 total_loss = 13.4897 = 0.0978 + 0.5 * 26.
7796 + 0.01 * 0.2101
----------------
Finished episode: 661 Reward: -1304.2112 total_loss = 13.1965 = 0.0511 + 0.5 * 26.
2862 + 0.01 * 0.2342
----------------
Finished episode: 662 Reward: -1288.8150 total_loss = 14.5272 = -0.0226 + 0.5 * 2
9.0946 + 0.01 * 0.2472
----------------
Finished episode: 663 Reward: -1285.1162 total_loss = 12.9574 = -0.0162 + 0.5 * 2
5.9426 + 0.01 * 0.2268
----------------
Finished episode: 664 Reward: -1296.9036 total_loss = 13.3389 = 0.0056 + 0.5 * 26.
6631 + 0.01 * 0.1768
----------------
Finished episode: 665 Reward: -1297.8089 total_loss = 12.8427 = -0.0176 + 0.5 * 2
5.7157 + 0.01 * 0.2419
----------------
Finished episode: 666 Reward: -1300.5061 total_loss = 15.1833 = -0.0483 + 0.5 * 3
0.4589 + 0.01 * 0.2171
----------------
Finished episode: 667 Reward: -1280.7342 total_loss = 16.2035 = -0.1928 + 0.5 * 3
2.7882 + 0.01 * 0.2200
----------------
Finished episode: 668 Reward: -1293.3885 total_loss = 13.9986 = -0.0351 + 0.5 * 2
8.0633 + 0.01 * 0.2047
----------------
Finished episode: 669 Reward: -1278.1235 total_loss = 13.1154 = -0.0335 + 0.5 * 2
6.2935 + 0.01 * 0.2211
----------------
Finished episode: 670 Reward: -1289.4087 total_loss = 13.6221 = 0.0625 + 0.5 * 27.
1154 + 0.01 * 0.1915
----------------
```

```
Finished episode: 671 Reward: -1264.2772 total_loss = 13.5466 = -0.0131 + 0.5 * 2
7.1146 + 0.01 * 0.2344
-----------------
Finished episode: 672 Reward: -1299.0902 total_loss = 12.3738 = 0.0086 + 0.5 * 24.
7262 + 0.01 * 0.2130
-----------------
Finished episode: 673 Reward: -1283.3385 total_loss = 13.5349 = 0.1645 + 0.5 * 26.
7362 + 0.01 * 0.2255
-----------------
Finished episode: 674 Reward: -1303.9651 total_loss = 13.2442 = 0.0254 + 0.5 * 26.
4328 + 0.01 * 0.2396
-----------------
Finished episode: 675 Reward: -1284.4045 total_loss = 16.3383 = 0.0546 + 0.5 * 32.
5633 + 0.01 * 0.2052
-----------------
Finished episode: 676 Reward: -1258.0637 total_loss = 12.5021 = -0.0339 + 0.5 * 2
5.0679 + 0.01 * 0.2088
-----------------
Finished episode: 677 Reward: -1267.5519 total_loss = 16.4082 = 0.0352 + 0.5 * 32.
7409 + 0.01 * 0.2567
-----------------
Finished episode: 678 Reward: -1277.9532 total_loss = 16.8242 = 0.0607 + 0.5 * 33.
5222 + 0.01 * 0.2391
-----------------
Finished episode: 679 Reward: -1291.7457 total_loss = 14.4933 = 0.0187 + 0.5 * 28.
9449 + 0.01 * 0.2187
-----------------
Finished episode: 680 Reward: -1276.1230 total_loss = 14.2835 = 0.0789 + 0.5 * 28.
4046 + 0.01 * 0.2354
-----------------
Finished episode: 681 Reward: -1290.5657 total_loss = 15.0729 = -0.1280 + 0.5 * 3
0.3982 + 0.01 * 0.1846
-----------------
Finished episode: 682 Reward: -1305.9437 total_loss = 14.7007 = 0.0440 + 0.5 * 29.
3088 + 0.01 * 0.2318
-----------------
Finished episode: 683 Reward: -1297.6015 total_loss = 13.5464 = 0.0004 + 0.5 * 27.
0882 + 0.01 * 0.1854
-----------------
Finished episode: 684 Reward: -1295.8655 total_loss = 14.1004 = 0.0079 + 0.5 * 28.
1804 + 0.01 * 0.2270
-----------------
Finished episode: 685 Reward: -1268.4587 total_loss = 16.7174 = 0.0141 + 0.5 * 33.
4023 + 0.01 * 0.2170
-----------------
Finished episode: 686 Reward: -1294.4005 total_loss = 13.3648 = 0.0450 + 0.5 * 26.
6357 + 0.01 * 0.1946
-----------------
Finished episode: 687 Reward: -1300.4093 total_loss = 14.9552 = -0.0414 + 0.5 * 2
9.9887 + 0.01 * 0.2213
-----------------
Finished episode: 688 Reward: -1268.1444 total_loss = 15.8649 = 0.0230 + 0.5 * 31.
6800 + 0.01 * 0.1859
-----------------
Finished episode: 689 Reward: -1299.7172 total_loss = 13.8892 = 0.0573 + 0.5 * 27.
6595 + 0.01 * 0.2237
-----------------
Finished episode: 690 Reward: -1303.0538 total_loss = 15.2163 = -0.0600 + 0.5 * 3
0.5489 + 0.01 * 0.1927
-----------------
Finished episode: 691 Reward: -1308.1309 total_loss = 15.3904 = 0.0207 + 0.5 * 30.
```

```
7351 + 0.01 * 0.2154
----------
Finished episode: 692 Reward: -1308.0901 total_loss = 14.5042 = 0.0645 + 0.5 * 28.
8748 + 0.01 * 0.2233
----------
Finished episode: 693 Reward: -1285.3716 total_loss = 16.6230 = -0.0049 + 0.5 * 3
3.2517 + 0.01 * 0.2053
----------
Finished episode: 694 Reward: -1245.4002 total_loss = 18.7383 = -0.0354 + 0.5 * 3
7.5435 + 0.01 * 0.1941
----------
Finished episode: 695 Reward: -1281.9143 total_loss = 12.8457 = 0.0426 + 0.5 * 25.
6024 + 0.01 * 0.1940
----------
Finished episode: 696 Reward: -1284.4247 total_loss = 15.8610 = 0.0060 + 0.5 * 31.
7059 + 0.01 * 0.2022
----------
Finished episode: 697 Reward: -1274.3846 total_loss = 16.0923 = 0.0476 + 0.5 * 32.
0856 + 0.01 * 0.1875
----------
Finished episode: 698 Reward: -1273.0425 total_loss = 15.2578 = -0.0468 + 0.5 * 3
0.6052 + 0.01 * 0.2055
----------
Finished episode: 699 Reward: -1281.1714 total_loss = 15.4444 = 0.0672 + 0.5 * 30.
7508 + 0.01 * 0.1799
----------
Finished episode: 700 Reward: -1309.9646 total_loss = 15.4159 = 0.0202 + 0.5 * 30.
7865 + 0.01 * 0.2513
----------
Finished episode: 701 Reward: -1303.7179 total_loss = 15.1167 = -0.0922 + 0.5 * 3
0.4136 + 0.01 * 0.2073
----------
Finished episode: 702 Reward: -1315.1950 total_loss = 15.7389 = -0.0622 + 0.5 * 3
1.5986 + 0.01 * 0.1714
----------
Finished episode: 703 Reward: -1297.9688 total_loss = 14.7801 = -0.0021 + 0.5 * 2
9.5609 + 0.01 * 0.1808
----------
Finished episode: 704 Reward: -1302.3723 total_loss = 15.3308 = -0.0825 + 0.5 * 3
0.8227 + 0.01 * 0.1949
----------
Finished episode: 705 Reward: -1296.9773 total_loss = 14.9010 = -0.2077 + 0.5 * 3
0.2129 + 0.01 * 0.2241
----------
Finished episode: 706 Reward: -1242.1708 total_loss = 18.5154 = -0.0032 + 0.5 * 3
7.0334 + 0.01 * 0.1837
----------
Finished episode: 707 Reward: -1302.3037 total_loss = 15.5061 = -0.0076 + 0.5 * 3
1.0239 + 0.01 * 0.1808
----------
Finished episode: 708 Reward: -1305.6278 total_loss = 16.6977 = -0.0335 + 0.5 * 3
3.4582 + 0.01 * 0.2125
----------
Finished episode: 709 Reward: -1291.8995 total_loss = 13.8923 = 0.0509 + 0.5 * 27.
6790 + 0.01 * 0.1906
----------
Finished episode: 710 Reward: -1311.5105 total_loss = 14.3968 = -0.1606 + 0.5 * 2
9.1096 + 0.01 * 0.2633
----------
Finished episode: 711 Reward: -1315.0485 total_loss = 14.8703 = 0.0225 + 0.5 * 29.
6911 + 0.01 * 0.2233
```

```
----------------
Finished episode: 712 Reward: -1302.1812 total_loss = 13.7004 = 0.0714 + 0.5 * 27.
2545 + 0.01 * 0.1731
----------------
Finished episode: 713 Reward: -1283.7558 total_loss = 13.8925 = 0.0123 + 0.5 * 27.
7568 + 0.01 * 0.1870
----------------
Finished episode: 714 Reward: -1303.4517 total_loss = 14.1066 = 0.1508 + 0.5 * 27.
9081 + 0.01 * 0.1721
----------------
Finished episode: 715 Reward: -1296.2183 total_loss = 14.9505 = -0.0894 + 0.5 * 3
0.0757 + 0.01 * 0.2030
----------------
Finished episode: 716 Reward: -1292.9586 total_loss = 14.9127 = -0.0107 + 0.5 * 2
9.8425 + 0.01 * 0.2073
----------------
Finished episode: 717 Reward: -1273.3795 total_loss = 11.9991 = 0.0353 + 0.5 * 23.
9229 + 0.01 * 0.2391
----------------
Finished episode: 718 Reward: -1287.8594 total_loss = 13.4970 = -0.0816 + 0.5 * 2
7.1530 + 0.01 * 0.2026
----------------
Finished episode: 719 Reward: -1309.2029 total_loss = 14.2014 = -0.0848 + 0.5 * 2
8.5686 + 0.01 * 0.1897
----------------
Finished episode: 720 Reward: -1316.0396 total_loss = 13.4437 = -0.0642 + 0.5 * 2
7.0123 + 0.01 * 0.1691
----------------
Finished episode: 721 Reward: -1284.9777 total_loss = 14.2367 = -0.0474 + 0.5 * 2
8.5643 + 0.01 * 0.1925
----------------
Finished episode: 722 Reward: -1314.5470 total_loss = 14.2099 = 0.0555 + 0.5 * 28.
3053 + 0.01 * 0.1807
----------------
Finished episode: 723 Reward: -1314.6443 total_loss = 16.5929 = -0.0296 + 0.5 * 3
3.2415 + 0.01 * 0.1742
----------------
Finished episode: 724 Reward: -1291.3322 total_loss = 12.8007 = 0.0023 + 0.5 * 25.
5929 + 0.01 * 0.1894
----------------
Finished episode: 725 Reward: -1316.7529 total_loss = 14.5508 = -0.0632 + 0.5 * 2
9.2233 + 0.01 * 0.2288
----------------
Finished episode: 726 Reward: -1295.4222 total_loss = 17.1094 = -0.1468 + 0.5 * 3
4.5086 + 0.01 * 0.1934
----------------
Finished episode: 727 Reward: -1311.7935 total_loss = 13.8494 = -0.0507 + 0.5 * 2
7.7966 + 0.01 * 0.1775
----------------
Finished episode: 728 Reward: -1320.7772 total_loss = 15.9467 = 0.0963 + 0.5 * 31.
6976 + 0.01 * 0.1652
----------------
Finished episode: 729 Reward: -1307.9703 total_loss = 13.5966 = -0.0184 + 0.5 * 2
7.2260 + 0.01 * 0.1917
----------------
Finished episode: 730 Reward: -1310.2260 total_loss = 13.6178 = 0.0356 + 0.5 * 27.
1608 + 0.01 * 0.1800
----------------
Finished episode: 731 Reward: -1301.8085 total_loss = 14.5738 = -0.0055 + 0.5 * 2
9.1547 + 0.01 * 0.1935
----------------
```

```
Finished episode: 732 Reward: -1311.6887 total_loss = 13.4376 = 0.0288 + 0.5 * 26.
8137 + 0.01 * 0.1956
-----------------
Finished episode: 733 Reward: -1297.7057 total_loss = 16.9895 = -0.0318 + 0.5 * 3
4.0384 + 0.01 * 0.2059
-----------------
Finished episode: 734 Reward: -1320.1971 total_loss = 13.4552 = 0.0701 + 0.5 * 26.
7664 + 0.01 * 0.1926
-----------------
Finished episode: 735 Reward: -1306.0622 total_loss = 15.1994 = 0.0038 + 0.5 * 30.
3876 + 0.01 * 0.1888
-----------------
Finished episode: 736 Reward: -1286.7646 total_loss = 14.6468 = 0.0018 + 0.5 * 29.
2859 + 0.01 * 0.1972
-----------------
Finished episode: 737 Reward: -1315.9450 total_loss = 15.1077 = -0.0164 + 0.5 * 3
0.2445 + 0.01 * 0.1799
-----------------
Finished episode: 738 Reward: -1306.0015 total_loss = 14.7480 = -0.1575 + 0.5 * 2
9.8073 + 0.01 * 0.1898
-----------------
Finished episode: 739 Reward: -1309.9874 total_loss = 14.6487 = -0.0118 + 0.5 * 2
9.3170 + 0.01 * 0.2028
-----------------
Finished episode: 740 Reward: -1299.2587 total_loss = 11.7306 = -0.0333 + 0.5 * 2
3.5235 + 0.01 * 0.2108
-----------------
Finished episode: 741 Reward: -1288.7663 total_loss = 16.3327 = 0.0739 + 0.5 * 32.
5140 + 0.01 * 0.1832
-----------------
Finished episode: 742 Reward: -1287.2359 total_loss = 15.7154 = 0.0516 + 0.5 * 31.
3245 + 0.01 * 0.1615
-----------------
Finished episode: 743 Reward: -1283.0047 total_loss = 16.8555 = -0.0144 + 0.5 * 3
3.7363 + 0.01 * 0.1758
-----------------
Finished episode: 744 Reward: -1289.3629 total_loss = 15.3496 = 0.0286 + 0.5 * 30.
6378 + 0.01 * 0.2050
-----------------
Finished episode: 745 Reward: -1311.1153 total_loss = 14.2517 = -0.0901 + 0.5 * 2
8.6797 + 0.01 * 0.1975
-----------------
Finished episode: 746 Reward: -1279.1329 total_loss = 14.8795 = 0.0489 + 0.5 * 29.
6586 + 0.01 * 0.1344
-----------------
Finished episode: 747 Reward: -1302.7248 total_loss = 14.1271 = -0.0771 + 0.5 * 2
8.4047 + 0.01 * 0.1784
-----------------
Finished episode: 748 Reward: -1304.7724 total_loss = 14.4988 = -0.0785 + 0.5 * 2
9.1505 + 0.01 * 0.2059
-----------------
Finished episode: 749 Reward: -1309.3270 total_loss = 15.6618 = 0.0544 + 0.5 * 31.
2111 + 0.01 * 0.1892
-----------------
Finished episode: 750 Reward: -1307.3229 total_loss = 14.5830 = -0.0600 + 0.5 * 2
9.2818 + 0.01 * 0.2109
-----------------
Finished episode: 751 Reward: -1317.6484 total_loss = 14.9641 = -0.0052 + 0.5 * 2
9.9349 + 0.01 * 0.1925
-----------------
Finished episode: 752 Reward: -1318.2133 total_loss = 16.2013 = -0.0417 + 0.5 * 3
```

```
2.4823 + 0.01 * 0.1893
----------
Finished episode: 753 Reward: -1314.1054 total_loss = 15.8465 = 0.0188 + 0.5 * 31.
6523 + 0.01 * 0.1499
----------
Finished episode: 754 Reward: -1305.9936 total_loss = 13.8041 = 0.0550 + 0.5 * 27.
4950 + 0.01 * 0.1579
----------
Finished episode: 755 Reward: -1311.0287 total_loss = 14.5495 = 0.0353 + 0.5 * 29.
0249 + 0.01 * 0.1774
----------
Finished episode: 756 Reward: -1314.2989 total_loss = 14.7953 = -0.0103 + 0.5 * 2
9.6081 + 0.01 * 0.1491
----------
Finished episode: 757 Reward: -1309.5904 total_loss = 14.3860 = -0.0839 + 0.5 * 2
8.9363 + 0.01 * 0.1693
----------
Finished episode: 758 Reward: -1298.3270 total_loss = 17.1922 = 0.1217 + 0.5 * 34.
1373 + 0.01 * 0.1834
----------
Finished episode: 759 Reward: -1300.7104 total_loss = 12.2074 = -0.1015 + 0.5 * 2
4.6143 + 0.01 * 0.1697
----------
Finished episode: 760 Reward: -1309.9791 total_loss = 14.8905 = -0.0018 + 0.5 * 2
9.7794 + 0.01 * 0.2541
----------
Finished episode: 761 Reward: -1298.9179 total_loss = 14.8514 = 0.0543 + 0.5 * 29.
5900 + 0.01 * 0.2129
----------
Finished episode: 762 Reward: -1299.7614 total_loss = 14.3644 = -0.0259 + 0.5 * 2
8.7764 + 0.01 * 0.2117
----------
Finished episode: 763 Reward: -1307.9988 total_loss = 14.1454 = 0.0009 + 0.5 * 28.
2859 + 0.01 * 0.1570
----------
Finished episode: 764 Reward: -1290.0002 total_loss = 16.1015 = -0.0031 + 0.5 * 3
2.2058 + 0.01 * 0.1701
----------
Finished episode: 765 Reward: -1289.1400 total_loss = 15.7600 = 0.0243 + 0.5 * 31.
4678 + 0.01 * 0.1795
----------
Finished episode: 766 Reward: -1291.1194 total_loss = 14.1559 = -0.0161 + 0.5 * 2
8.3400 + 0.01 * 0.1926
----------
Finished episode: 767 Reward: -1303.2654 total_loss = 14.0970 = 0.0099 + 0.5 * 28.
1708 + 0.01 * 0.1732
----------
Finished episode: 768 Reward: -1306.3804 total_loss = 17.7397 = -0.0061 + 0.5 * 3
5.4882 + 0.01 * 0.1734
----------
Finished episode: 769 Reward: -1302.3329 total_loss = 12.6363 = 0.1221 + 0.5 * 25.
0250 + 0.01 * 0.1716
----------
Finished episode: 770 Reward: -1309.3514 total_loss = 14.2684 = -0.1173 + 0.5 * 2
8.7681 + 0.01 * 0.1706
----------
Finished episode: 771 Reward: -1291.5045 total_loss = 15.2643 = -0.0216 + 0.5 * 3
0.5681 + 0.01 * 0.1792
----------
Finished episode: 772 Reward: -1311.3178 total_loss = 14.1287 = 0.0478 + 0.5 * 28.
1574 + 0.01 * 0.2216
```

```
----------------
Finished episode: 773 Reward: -1306.3776 total_loss = 13.8888 = -0.0156 + 0.5 * 2
7.8053 + 0.01 * 0.1749
----------------
Finished episode: 774 Reward: -1316.7568 total_loss = 14.7261 = -0.0069 + 0.5 * 2
9.4629 + 0.01 * 0.1646
----------------
Finished episode: 775 Reward: -1298.9712 total_loss = 14.5683 = -0.1248 + 0.5 * 2
9.3828 + 0.01 * 0.1656
----------------
Finished episode: 776 Reward: -1301.4929 total_loss = 12.4954 = 0.1377 + 0.5 * 24.
7121 + 0.01 * 0.1602
----------------
Finished episode: 777 Reward: -1305.5240 total_loss = 13.2666 = 0.1174 + 0.5 * 26.
2959 + 0.01 * 0.1187
----------------
Finished episode: 778 Reward: -1311.5436 total_loss = 15.7073 = -0.0218 + 0.5 * 3
1.4545 + 0.01 * 0.1875
----------------
Finished episode: 779 Reward: -1309.6130 total_loss = 13.2155 = 0.0578 + 0.5 * 26.
3127 + 0.01 * 0.1354
----------------
Finished episode: 780 Reward: -1308.5577 total_loss = 13.4935 = 0.0225 + 0.5 * 26.
9379 + 0.01 * 0.2107
----------------
Finished episode: 781 Reward: -1303.3840 total_loss = 15.0948 = 0.0116 + 0.5 * 30.
1629 + 0.01 * 0.1698
----------------
Finished episode: 782 Reward: -1314.2900 total_loss = 15.3034 = 0.0629 + 0.5 * 30.
4772 + 0.01 * 0.1952
----------------
Finished episode: 783 Reward: -1307.7457 total_loss = 14.0039 = 0.0285 + 0.5 * 27.
9470 + 0.01 * 0.1926
----------------
Finished episode: 784 Reward: -1286.1046 total_loss = 14.7011 = -0.0055 + 0.5 * 2
9.4093 + 0.01 * 0.1861
----------------
Finished episode: 785 Reward: -1317.8936 total_loss = 13.8494 = 0.1230 + 0.5 * 27.
4490 + 0.01 * 0.1869
----------------
Finished episode: 786 Reward: -1305.4948 total_loss = 14.3128 = 0.0462 + 0.5 * 28.
5298 + 0.01 * 0.1714
----------------
Finished episode: 787 Reward: -1296.0572 total_loss = 16.4537 = 0.0340 + 0.5 * 32.
8360 + 0.01 * 0.1712
----------------
Finished episode: 788 Reward: -1320.3501 total_loss = 14.1215 = 0.0004 + 0.5 * 28.
2380 + 0.01 * 0.2000
----------------
Finished episode: 789 Reward: -1330.2403 total_loss = 15.7888 = -0.0826 + 0.5 * 3
1.7395 + 0.01 * 0.1691
----------------
Finished episode: 790 Reward: -1313.1565 total_loss = 12.4984 = 0.1353 + 0.5 * 24.
7225 + 0.01 * 0.1862
----------------
Finished episode: 791 Reward: -1313.6649 total_loss = 14.4948 = -0.0628 + 0.5 * 2
9.1120 + 0.01 * 0.1600
----------------
Finished episode: 792 Reward: -1306.1261 total_loss = 14.7564 = 0.0393 + 0.5 * 29.
4318 + 0.01 * 0.1236
----------------
```

```
Finished episode: 793 Reward: -1325.1825 total_loss = 12.3710 = 0.0232 + 0.5 * 24.
6923 + 0.01 * 0.1670
-----------------
Finished episode: 794 Reward: -1307.4578 total_loss = 13.7903 = -0.1666 + 0.5 * 2
7.9098 + 0.01 * 0.2012
-----------------
Finished episode: 795 Reward: -1285.9233 total_loss = 17.4107 = -0.0337 + 0.5 * 3
4.8852 + 0.01 * 0.1791
-----------------
Finished episode: 796 Reward: -1310.5464 total_loss = 15.5959 = -0.0173 + 0.5 * 3
1.2229 + 0.01 * 0.1779
-----------------
Finished episode: 797 Reward: -1300.3427 total_loss = 14.3730 = -0.0279 + 0.5 * 2
8.7989 + 0.01 * 0.1499
-----------------
Finished episode: 798 Reward: -1319.5249 total_loss = 15.9823 = 0.0142 + 0.5 * 31.
9339 + 0.01 * 0.1155
-----------------
Finished episode: 799 Reward: -1289.7568 total_loss = 14.6213 = -0.0969 + 0.5 * 2
9.4337 + 0.01 * 0.1339
-----------------
Finished episode: 800 Reward: -1309.5285 total_loss = 13.5422 = -0.0523 + 0.5 * 2
7.1858 + 0.01 * 0.1619
-----------------
Finished episode: 801 Reward: -1318.8314 total_loss = 14.8288 = 0.0164 + 0.5 * 29.
6216 + 0.01 * 0.1591
-----------------
Finished episode: 802 Reward: -1307.8104 total_loss = 14.2179 = -0.0583 + 0.5 * 2
8.5491 + 0.01 * 0.1597
-----------------
Finished episode: 803 Reward: -1303.3953 total_loss = 13.6504 = -0.0817 + 0.5 * 2
7.4608 + 0.01 * 0.1706
-----------------
Finished episode: 804 Reward: -1319.7702 total_loss = 14.9226 = 0.0641 + 0.5 * 29.
7142 + 0.01 * 0.1422
-----------------
Finished episode: 805 Reward: -1287.6171 total_loss = 14.2477 = 0.0608 + 0.5 * 28.
3701 + 0.01 * 0.1816
-----------------
Finished episode: 806 Reward: -1293.1656 total_loss = 15.1070 = -0.0739 + 0.5 * 3
0.3583 + 0.01 * 0.1689
-----------------
Finished episode: 807 Reward: -1276.5839 total_loss = 17.7571 = -0.0151 + 0.5 * 3
5.5406 + 0.01 * 0.1942
-----------------
Finished episode: 808 Reward: -1329.4453 total_loss = 14.8683 = -0.0043 + 0.5 * 2
9.7422 + 0.01 * 0.1546
-----------------
Finished episode: 809 Reward: -1303.3655 total_loss = 13.3803 = 0.0262 + 0.5 * 26.
7047 + 0.01 * 0.1781
-----------------
Finished episode: 810 Reward: -1318.7651 total_loss = 15.5573 = -0.0467 + 0.5 * 3
1.2051 + 0.01 * 0.1549
-----------------
Finished episode: 811 Reward: -1310.4851 total_loss = 13.5514 = 0.0050 + 0.5 * 27.
0888 + 0.01 * 0.2026
-----------------
Finished episode: 812 Reward: -1309.8155 total_loss = 14.4372 = -0.0223 + 0.5 * 2
8.9157 + 0.01 * 0.1581
-----------------
Finished episode: 813 Reward: -1290.9280 total_loss = 13.9025 = -0.0347 + 0.5 * 2
```

```
7.8706 + 0.01 * 0.1916
----------
Finished episode: 814 Reward: -1305.7769 total_loss = 14.0419 = -0.0664 + 0.5 * 2
8.2140 + 0.01 * 0.1244
----------
Finished episode: 815 Reward: -1314.0758 total_loss = 16.8393 = 0.0299 + 0.5 * 33.
6160 + 0.01 * 0.1350
----------
Finished episode: 816 Reward: -1316.4952 total_loss = 13.3247 = 0.0859 + 0.5 * 26.
4748 + 0.01 * 0.1429
----------
Finished episode: 817 Reward: -1301.4119 total_loss = 14.4597 = -0.0756 + 0.5 * 2
9.0676 + 0.01 * 0.1477
----------
Finished episode: 818 Reward: -1295.7403 total_loss = 14.2172 = 0.0235 + 0.5 * 28.
3842 + 0.01 * 0.1522
----------
Finished episode: 819 Reward: -1307.7058 total_loss = 13.1434 = -0.0267 + 0.5 * 2
6.3366 + 0.01 * 0.1816
----------
Finished episode: 820 Reward: -1328.8478 total_loss = 16.3403 = 0.1313 + 0.5 * 32.
4139 + 0.01 * 0.1973
----------
Finished episode: 821 Reward: -1325.3931 total_loss = 14.4256 = 0.0028 + 0.5 * 28.
8427 + 0.01 * 0.1460
----------
Finished episode: 822 Reward: -1315.4518 total_loss = 15.1931 = -0.0036 + 0.5 * 3
0.3907 + 0.01 * 0.1350
----------
Finished episode: 823 Reward: -1299.3483 total_loss = 18.0980 = -0.0441 + 0.5 * 3
6.2808 + 0.01 * 0.1660
----------
Finished episode: 824 Reward: -1311.9358 total_loss = 13.9613 = -0.0892 + 0.5 * 2
8.0981 + 0.01 * 0.1365
----------
Finished episode: 825 Reward: -1333.2059 total_loss = 16.2457 = -0.0093 + 0.5 * 3
2.5066 + 0.01 * 0.1632
----------
Finished episode: 826 Reward: -1307.6147 total_loss = 15.5257 = 0.1931 + 0.5 * 30.
6619 + 0.01 * 0.1574
----------
Finished episode: 827 Reward: -1309.4665 total_loss = 13.3832 = -0.0075 + 0.5 * 2
6.7784 + 0.01 * 0.1526
----------
Finished episode: 828 Reward: -1292.3328 total_loss = 14.9301 = 0.1052 + 0.5 * 29.
6466 + 0.01 * 0.1523
----------
Finished episode: 829 Reward: -1302.0164 total_loss = 14.7276 = -0.0434 + 0.5 * 2
9.5384 + 0.01 * 0.1842
----------
Finished episode: 830 Reward: -1305.0151 total_loss = 16.8830 = -0.0392 + 0.5 * 3
3.8416 + 0.01 * 0.1367
----------
Finished episode: 831 Reward: -1333.7564 total_loss = 15.9088 = -0.0728 + 0.5 * 3
1.9594 + 0.01 * 0.1986
----------
Finished episode: 832 Reward: -1323.3051 total_loss = 13.8130 = -0.0306 + 0.5 * 2
7.6836 + 0.01 * 0.1740
----------
Finished episode: 833 Reward: -1321.3777 total_loss = 13.7051 = 0.1044 + 0.5 * 27.
1983 + 0.01 * 0.1497
```

```
----------------
Finished episode: 834 Reward: -1329.8644 total_loss = 15.0087 = 0.0440 + 0.5 * 29.
9265 + 0.01 * 0.1485
----------------
Finished episode: 835 Reward: -1298.5599 total_loss = 13.7936 = -0.0626 + 0.5 * 2
7.7084 + 0.01 * 0.1943
----------------
Finished episode: 836 Reward: -1293.2551 total_loss = 14.9777 = -0.0016 + 0.5 * 2
9.9551 + 0.01 * 0.1854
----------------
Finished episode: 837 Reward: -1290.1652 total_loss = 18.2630 = 0.0254 + 0.5 * 36.
4717 + 0.01 * 0.1735
----------------
Finished episode: 838 Reward: -1294.7308 total_loss = 16.2881 = 0.0972 + 0.5 * 32.
3784 + 0.01 * 0.1664
----------------
Finished episode: 839 Reward: -1332.1538 total_loss = 15.2242 = 0.0558 + 0.5 * 30.
3334 + 0.01 * 0.1698
----------------
Finished episode: 840 Reward: -1319.0092 total_loss = 14.6268 = -0.0460 + 0.5 * 2
9.3416 + 0.01 * 0.1951
----------------
Finished episode: 841 Reward: -1317.9905 total_loss = 16.5981 = -0.1084 + 0.5 * 3
3.4098 + 0.01 * 0.1622
----------------
Finished episode: 842 Reward: -1326.9742 total_loss = 13.3108 = 0.1115 + 0.5 * 26.
3954 + 0.01 * 0.1611
----------------
Finished episode: 843 Reward: -1301.2076 total_loss = 15.4171 = 0.0395 + 0.5 * 30.
7524 + 0.01 * 0.1465
----------------
Finished episode: 844 Reward: -1300.3486 total_loss = 18.0161 = 0.1109 + 0.5 * 35.
8069 + 0.01 * 0.1765
----------------
Finished episode: 845 Reward: -1311.5854 total_loss = 12.2033 = 0.0879 + 0.5 * 24.
2276 + 0.01 * 0.1518
----------------
Finished episode: 846 Reward: -1332.4533 total_loss = 14.3371 = 0.0429 + 0.5 * 28.
5859 + 0.01 * 0.1227
----------------
Finished episode: 847 Reward: -1324.8918 total_loss = 15.3048 = -0.0014 + 0.5 * 3
0.6089 + 0.01 * 0.1726
----------------
Finished episode: 848 Reward: -1313.2322 total_loss = 13.8047 = 0.0108 + 0.5 * 27.
5841 + 0.01 * 0.1876
----------------
Finished episode: 849 Reward: -1315.2048 total_loss = 13.6268 = 0.0098 + 0.5 * 27.
2306 + 0.01 * 0.1614
----------------
Finished episode: 850 Reward: -1301.5023 total_loss = 14.0474 = 0.0528 + 0.5 * 27.
9859 + 0.01 * 0.1686
----------------
Finished episode: 851 Reward: -1331.3873 total_loss = 15.7716 = 0.0509 + 0.5 * 31.
4381 + 0.01 * 0.1755
----------------
Finished episode: 852 Reward: -1322.0182 total_loss = 15.8253 = -0.1109 + 0.5 * 3
1.8689 + 0.01 * 0.1768
----------------
Finished episode: 853 Reward: -1316.7492 total_loss = 13.8950 = 0.0154 + 0.5 * 27.
7558 + 0.01 * 0.1662
----------------
```

```
Finished episode: 854 Reward: -1318.8804 total_loss = 13.3090 = -0.1232 + 0.5 * 2
6.8612 + 0.01 * 0.1561
-----------------
Finished episode: 855 Reward: -1319.0581 total_loss = 14.3993 = -0.0777 + 0.5 * 2
8.9513 + 0.01 * 0.1375
-----------------
Finished episode: 856 Reward: -1304.1165 total_loss = 13.0478 = -0.0379 + 0.5 * 2
6.1682 + 0.01 * 0.1591
-----------------
Finished episode: 857 Reward: -1328.5103 total_loss = 15.9035 = 0.0583 + 0.5 * 31.
6869 + 0.01 * 0.1686
-----------------
Finished episode: 858 Reward: -1317.3794 total_loss = 15.1315 = -0.0265 + 0.5 * 3
0.3131 + 0.01 * 0.1511
-----------------
Finished episode: 859 Reward: -1309.9430 total_loss = 14.2595 = 0.0040 + 0.5 * 28.
5080 + 0.01 * 0.1479
-----------------
Finished episode: 860 Reward: -1313.5530 total_loss = 14.1521 = 0.0327 + 0.5 * 28.
2361 + 0.01 * 0.1401
-----------------
Finished episode: 861 Reward: -1303.0859 total_loss = 15.8374 = 0.1038 + 0.5 * 31.
4642 + 0.01 * 0.1530
-----------------
Finished episode: 862 Reward: -1293.7868 total_loss = 15.1169 = -0.0001 + 0.5 * 3
0.2311 + 0.01 * 0.1440
-----------------
Finished episode: 863 Reward: -1323.6219 total_loss = 15.1477 = 0.0890 + 0.5 * 30.
1141 + 0.01 * 0.1642
-----------------
Finished episode: 864 Reward: -1317.2953 total_loss = 14.3987 = 0.0166 + 0.5 * 28.
7612 + 0.01 * 0.1392
-----------------
Finished episode: 865 Reward: -1302.2359 total_loss = 16.5955 = 0.0891 + 0.5 * 33.
0095 + 0.01 * 0.1616
-----------------
Finished episode: 866 Reward: -1321.6912 total_loss = 14.4994 = 0.0226 + 0.5 * 28.
9501 + 0.01 * 0.1747
-----------------
Finished episode: 867 Reward: -1298.7566 total_loss = 17.2955 = 0.0626 + 0.5 * 34.
4629 + 0.01 * 0.1428
-----------------
Finished episode: 868 Reward: -1327.2547 total_loss = 15.5295 = -0.0275 + 0.5 * 3
1.1103 + 0.01 * 0.1796
-----------------
Finished episode: 869 Reward: -1321.3445 total_loss = 14.8388 = 0.0049 + 0.5 * 29.
6649 + 0.01 * 0.1447
-----------------
Finished episode: 870 Reward: -1309.5190 total_loss = 15.3308 = 0.0456 + 0.5 * 30.
5671 + 0.01 * 0.1709
-----------------
Finished episode: 871 Reward: -1299.1454 total_loss = 17.6456 = 0.0144 + 0.5 * 35.
2586 + 0.01 * 0.1830
-----------------
Finished episode: 872 Reward: -1311.9069 total_loss = 14.2650 = -0.0768 + 0.5 * 2
8.6805 + 0.01 * 0.1469
-----------------
Finished episode: 873 Reward: -1316.6646 total_loss = 15.9587 = -0.0313 + 0.5 * 3
1.9768 + 0.01 * 0.1570
-----------------
Finished episode: 874 Reward: -1295.5868 total_loss = 15.0591 = -0.1021 + 0.5 * 3
```

```
0.3201 + 0.01 * 0.1199
----------
Finished episode: 875 Reward: -1314.6556 total_loss = 18.3432 = -0.0206 + 0.5 * 3
6.7246 + 0.01 * 0.1587
----------
Finished episode: 876 Reward: -1329.9932 total_loss = 14.3508 = -0.0006 + 0.5 * 2
8.6999 + 0.01 * 0.1458
----------
Finished episode: 877 Reward: -1321.7124 total_loss = 15.4859 = 0.0318 + 0.5 * 30.
9053 + 0.01 * 0.1394
----------
Finished episode: 878 Reward: -1308.0790 total_loss = 15.8123 = 0.0205 + 0.5 * 31.
5803 + 0.01 * 0.1633
----------
Finished episode: 879 Reward: -1320.1695 total_loss = 14.5206 = 0.0447 + 0.5 * 28.
9484 + 0.01 * 0.1695
----------
Finished episode: 880 Reward: -1324.8891 total_loss = 14.2452 = 0.1745 + 0.5 * 28.
1383 + 0.01 * 0.1490
----------
Finished episode: 881 Reward: -1324.4008 total_loss = 16.1836 = -0.0043 + 0.5 * 3
2.3734 + 0.01 * 0.1180
----------
Finished episode: 882 Reward: -1313.1239 total_loss = 14.4643 = 0.0663 + 0.5 * 28.
7926 + 0.01 * 0.1765
----------
Finished episode: 883 Reward: -1289.0203 total_loss = 15.3527 = 0.1255 + 0.5 * 30.
4506 + 0.01 * 0.1851
----------
Finished episode: 884 Reward: -1311.7906 total_loss = 13.1098 = 0.0560 + 0.5 * 26.
1043 + 0.01 * 0.1647
----------
Finished episode: 885 Reward: -1306.4385 total_loss = 14.9921 = -0.0627 + 0.5 * 3
0.1070 + 0.01 * 0.1305
----------
Finished episode: 886 Reward: -1313.2255 total_loss = 13.5025 = 0.0523 + 0.5 * 26.
8973 + 0.01 * 0.1571
----------
Finished episode: 887 Reward: -1327.3574 total_loss = 15.6642 = -0.0142 + 0.5 * 3
1.3539 + 0.01 * 0.1489
----------
Finished episode: 888 Reward: -1314.7682 total_loss = 14.5156 = -0.1022 + 0.5 * 2
9.2337 + 0.01 * 0.0911
----------
Finished episode: 889 Reward: -1326.5953 total_loss = 13.4113 = -0.0250 + 0.5 * 2
6.8702 + 0.01 * 0.1251
----------
Finished episode: 890 Reward: -1295.7064 total_loss = 16.2103 = -0.0359 + 0.5 * 3
2.4894 + 0.01 * 0.1428
----------
Finished episode: 891 Reward: -1328.6892 total_loss = 14.5678 = 0.0504 + 0.5 * 29.
0321 + 0.01 * 0.1394
----------
Finished episode: 892 Reward: -1321.8291 total_loss = 15.2143 = 0.0297 + 0.5 * 30.
3660 + 0.01 * 0.1608
----------
Finished episode: 893 Reward: -1333.7987 total_loss = 15.8660 = -0.0134 + 0.5 * 3
1.7559 + 0.01 * 0.1528
----------
Finished episode: 894 Reward: -1318.7550 total_loss = 14.0671 = 0.0939 + 0.5 * 27.
9436 + 0.01 * 0.1330
```

```
----------------
Finished episode: 895 Reward: -1330.7772 total_loss = 17.1432 = -0.0363 + 0.5 * 3
4.3564 + 0.01 * 0.1287
----------------
Finished episode: 896 Reward: -1296.7073 total_loss = 14.7598 = -0.0149 + 0.5 * 2
9.5470 + 0.01 * 0.1222
----------------
Finished episode: 897 Reward: -1319.3743 total_loss = 14.8753 = -0.0026 + 0.5 * 2
9.7535 + 0.01 * 0.1221
----------------
Finished episode: 898 Reward: -1332.5768 total_loss = 15.4282 = -0.0551 + 0.5 * 3
0.9639 + 0.01 * 0.1429
----------------
Finished episode: 899 Reward: -1318.6306 total_loss = 15.4314 = 0.0258 + 0.5 * 30.
8081 + 0.01 * 0.1624
----------------
Finished episode: 900 Reward: -1333.8132 total_loss = 14.3369 = 0.0048 + 0.5 * 28.
6611 + 0.01 * 0.1532
----------------
Finished episode: 901 Reward: -1324.2326 total_loss = 14.6512 = -0.0214 + 0.5 * 2
9.3428 + 0.01 * 0.1206
----------------
Finished episode: 902 Reward: -1320.7314 total_loss = 16.6491 = -0.1210 + 0.5 * 3
3.5370 + 0.01 * 0.1551
----------------
Finished episode: 903 Reward: -1323.8504 total_loss = 15.1769 = 0.0286 + 0.5 * 30.
2933 + 0.01 * 0.1686
----------------
Finished episode: 904 Reward: -1320.4018 total_loss = 17.8610 = -0.0626 + 0.5 * 3
5.8441 + 0.01 * 0.1615
----------------
Finished episode: 905 Reward: -1331.9733 total_loss = 15.0640 = -0.0378 + 0.5 * 3
0.2005 + 0.01 * 0.1526
----------------
Finished episode: 906 Reward: -1317.3628 total_loss = 14.0126 = 0.0956 + 0.5 * 27.
8315 + 0.01 * 0.1269
----------------
Finished episode: 907 Reward: -1324.9884 total_loss = 15.9781 = -0.0575 + 0.5 * 3
2.0681 + 0.01 * 0.1517
----------------
Finished episode: 908 Reward: -1314.7547 total_loss = 15.1488 = -0.0764 + 0.5 * 3
0.4477 + 0.01 * 0.1388
----------------
Finished episode: 909 Reward: -1309.2247 total_loss = 14.7583 = 0.0826 + 0.5 * 29.
3486 + 0.01 * 0.1447
----------------
Finished episode: 910 Reward: -1318.1614 total_loss = 13.5911 = 0.0158 + 0.5 * 27.
1478 + 0.01 * 0.1423
----------------
Finished episode: 911 Reward: -1301.4085 total_loss = 13.9258 = -0.0373 + 0.5 * 2
7.9237 + 0.01 * 0.1244
----------------
Finished episode: 912 Reward: -1317.4465 total_loss = 12.6945 = 0.0583 + 0.5 * 25.
2692 + 0.01 * 0.1655
----------------
Finished episode: 913 Reward: -1311.2210 total_loss = 15.7057 = 0.0449 + 0.5 * 31.
3188 + 0.01 * 0.1445
----------------
Finished episode: 914 Reward: -1322.7088 total_loss = 13.0713 = -0.0179 + 0.5 * 2
6.1754 + 0.01 * 0.1560
----------------
```

```
Finished episode: 915 Reward: -1301.1733 total_loss = 15.6939 = 0.0935 + 0.5 * 31.
1981 + 0.01 * 0.1362
-----------------
Finished episode: 916 Reward: -1309.0567 total_loss = 14.0792 = -0.0010 + 0.5 * 2
8.1575 + 0.01 * 0.1462
-----------------
Finished episode: 917 Reward: -1315.2237 total_loss = 14.4374 = -0.0820 + 0.5 * 2
9.0360 + 0.01 * 0.1396
-----------------
Finished episode: 918 Reward: -1314.9539 total_loss = 13.5021 = 0.0567 + 0.5 * 26.
8878 + 0.01 * 0.1541
-----------------
Finished episode: 919 Reward: -1313.5416 total_loss = 13.4123 = -0.0022 + 0.5 * 2
6.8264 + 0.01 * 0.1267
-----------------
Finished episode: 920 Reward: -1323.9989 total_loss = 15.5125 = 0.0314 + 0.5 * 30.
9595 + 0.01 * 0.1333
-----------------
Finished episode: 921 Reward: -1314.9803 total_loss = 14.7210 = 0.0327 + 0.5 * 29.
3736 + 0.01 * 0.1547
-----------------
Finished episode: 922 Reward: -1323.7227 total_loss = 13.8116 = -0.2060 + 0.5 * 2
8.0319 + 0.01 * 0.1607
-----------------
Finished episode: 923 Reward: -1333.8108 total_loss = 15.8774 = 0.0073 + 0.5 * 31.
7373 + 0.01 * 0.1434
-----------------
Finished episode: 924 Reward: -1303.3763 total_loss = 14.8574 = 0.0137 + 0.5 * 29.
6848 + 0.01 * 0.1287
-----------------
Finished episode: 925 Reward: -1313.8553 total_loss = 16.0633 = -0.0098 + 0.5 * 3
2.1437 + 0.01 * 0.1299
-----------------
Finished episode: 926 Reward: -1322.2405 total_loss = 15.0559 = -0.1071 + 0.5 * 3
0.3238 + 0.01 * 0.1171
-----------------
Finished episode: 927 Reward: -1323.6369 total_loss = 15.4093 = -0.0251 + 0.5 * 3
0.8665 + 0.01 * 0.1067
-----------------
Finished episode: 928 Reward: -1324.1743 total_loss = 13.7948 = -0.0706 + 0.5 * 2
7.7283 + 0.01 * 0.1283
-----------------
Finished episode: 929 Reward: -1320.5436 total_loss = 15.4415 = -0.0339 + 0.5 * 3
0.9478 + 0.01 * 0.1534
-----------------
Finished episode: 930 Reward: -1315.6464 total_loss = 14.3020 = -0.0417 + 0.5 * 2
8.6841 + 0.01 * 0.1706
-----------------
Finished episode: 931 Reward: -1294.1646 total_loss = 16.3980 = -0.0839 + 0.5 * 3
2.9612 + 0.01 * 0.1276
-----------------
Finished episode: 932 Reward: -1321.9395 total_loss = 14.2786 = 0.0069 + 0.5 * 28.
5402 + 0.01 * 0.1636
-----------------
Finished episode: 933 Reward: -1330.3926 total_loss = 14.9546 = 0.0873 + 0.5 * 29.
7309 + 0.01 * 0.1854
-----------------
Finished episode: 934 Reward: -1302.3300 total_loss = 14.0509 = -0.0451 + 0.5 * 2
8.1890 + 0.01 * 0.1454
-----------------
Finished episode: 935 Reward: -1328.1410 total_loss = 14.7088 = 0.0250 + 0.5 * 29.
```

```
3650 + 0.01 * 0.1370
----------------
Finished episode: 936 Reward: -1322.9408 total_loss = 14.4306 = -0.0209 + 0.5 * 2
8.9002 + 0.01 * 0.1368
----------------
Finished episode: 937 Reward: -1331.3095 total_loss = 15.1769 = -0.0552 + 0.5 * 3
0.4609 + 0.01 * 0.1681
----------------
Finished episode: 938 Reward: -1311.7950 total_loss = 13.8994 = 0.0249 + 0.5 * 27.
7457 + 0.01 * 0.1687
----------------
Finished episode: 939 Reward: -1310.5945 total_loss = 15.1188 = -0.1040 + 0.5 * 3
0.4429 + 0.01 * 0.1314
----------------
Finished episode: 940 Reward: -1329.7109 total_loss = 14.9179 = -0.0514 + 0.5 * 2
9.9354 + 0.01 * 0.1642
----------------
Finished episode: 941 Reward: -1314.3484 total_loss = 15.8276 = 0.0131 + 0.5 * 31.
6262 + 0.01 * 0.1359
----------------
Finished episode: 942 Reward: -1312.8713 total_loss = 15.5510 = -0.0257 + 0.5 * 3
1.1506 + 0.01 * 0.1379
----------------
Finished episode: 943 Reward: -1313.0124 total_loss = 14.7341 = 0.0665 + 0.5 * 29.
3335 + 0.01 * 0.0844
----------------
Finished episode: 944 Reward: -1300.9719 total_loss = 14.1501 = -0.0055 + 0.5 * 2
8.3090 + 0.01 * 0.1029
----------------
Finished episode: 945 Reward: -1332.0293 total_loss = 16.1203 = 0.0040 + 0.5 * 32.
2297 + 0.01 * 0.1491
----------------
Finished episode: 946 Reward: -1319.0111 total_loss = 15.4465 = -0.0209 + 0.5 * 3
0.9319 + 0.01 * 0.1372
----------------
Finished episode: 947 Reward: -1333.9508 total_loss = 13.5358 = 0.0044 + 0.5 * 27.
0606 + 0.01 * 0.1079
----------------
Finished episode: 948 Reward: -1303.3444 total_loss = 15.2131 = -0.0251 + 0.5 * 3
0.4738 + 0.01 * 0.1314
----------------
Finished episode: 949 Reward: -1320.2620 total_loss = 16.4434 = -0.0469 + 0.5 * 3
2.9781 + 0.01 * 0.1213
----------------
Finished episode: 950 Reward: -1323.0546 total_loss = 15.0382 = -0.0315 + 0.5 * 3
0.1365 + 0.01 * 0.1430
----------------
Finished episode: 951 Reward: -1307.4536 total_loss = 16.8503 = 0.1194 + 0.5 * 33.
4584 + 0.01 * 0.1657
----------------
Finished episode: 952 Reward: -1327.1748 total_loss = 14.3617 = -0.1165 + 0.5 * 2
8.9538 + 0.01 * 0.1270
----------------
Finished episode: 953 Reward: -1333.3840 total_loss = 15.1607 = -0.0025 + 0.5 * 3
0.3226 + 0.01 * 0.1879
----------------
Finished episode: 954 Reward: -1320.4888 total_loss = 15.7126 = 0.0016 + 0.5 * 31.
4189 + 0.01 * 0.1551
----------------
Finished episode: 955 Reward: -1332.2426 total_loss = 14.1041 = 0.0409 + 0.5 * 28.
1231 + 0.01 * 0.1614
```

```
----------------
Finished episode: 956 Reward: -1334.1301 total_loss = 15.3078 = -0.0192 + 0.5 * 3
0.6507 + 0.01 * 0.1734
----------------
Finished episode: 957 Reward: -1289.9421 total_loss = 16.9415 = 0.1304 + 0.5 * 33.
6191 + 0.01 * 0.1566
----------------
Finished episode: 958 Reward: -1326.8964 total_loss = 14.5926 = -0.0542 + 0.5 * 2
9.2908 + 0.01 * 0.1391
----------------
Finished episode: 959 Reward: -1310.9921 total_loss = 17.6218 = -0.0734 + 0.5 * 3
5.3871 + 0.01 * 0.1579
----------------
Finished episode: 960 Reward: -1311.1043 total_loss = 14.5450 = 0.1000 + 0.5 * 28.
8872 + 0.01 * 0.1396
----------------
Finished episode: 961 Reward: -1330.7481 total_loss = 14.8404 = -0.0459 + 0.5 * 2
9.7698 + 0.01 * 0.1479
----------------
Finished episode: 962 Reward: -1305.0847 total_loss = 12.8884 = 0.0013 + 0.5 * 25.
7719 + 0.01 * 0.1228
----------------
Finished episode: 963 Reward: -1326.1957 total_loss = 13.4446 = 0.1173 + 0.5 * 26.
6520 + 0.01 * 0.1278
----------------
Finished episode: 964 Reward: -1288.8218 total_loss = 15.5833 = -0.0179 + 0.5 * 3
1.1992 + 0.01 * 0.1534
----------------
Finished episode: 965 Reward: -1320.6571 total_loss = 13.4144 = -0.1348 + 0.5 * 2
7.0955 + 0.01 * 0.1497
----------------
Finished episode: 966 Reward: -1320.0739 total_loss = 16.1119 = 0.0417 + 0.5 * 32.
1376 + 0.01 * 0.1340
----------------
Finished episode: 967 Reward: -1325.0382 total_loss = 15.3253 = 0.0418 + 0.5 * 30.
5639 + 0.01 * 0.1594
----------------
Finished episode: 968 Reward: -1325.1343 total_loss = 14.1250 = 0.0829 + 0.5 * 28.
0813 + 0.01 * 0.1471
----------------
Finished episode: 969 Reward: -1328.9379 total_loss = 14.9572 = 0.0018 + 0.5 * 29.
9078 + 0.01 * 0.1454
----------------
Finished episode: 970 Reward: -1315.1430 total_loss = 14.4960 = -0.0922 + 0.5 * 2
9.1733 + 0.01 * 0.1611
----------------
Finished episode: 971 Reward: -1319.6126 total_loss = 16.3294 = 0.0101 + 0.5 * 32.
6356 + 0.01 * 0.1489
----------------
Finished episode: 972 Reward: -1307.1488 total_loss = 14.9296 = 0.0376 + 0.5 * 29.
7817 + 0.01 * 0.1164
----------------
Finished episode: 973 Reward: -1324.6073 total_loss = 14.2593 = -0.0065 + 0.5 * 2
8.5282 + 0.01 * 0.1693
----------------
Finished episode: 974 Reward: -1330.8930 total_loss = 15.0425 = 0.0208 + 0.5 * 30.
0407 + 0.01 * 0.1387
----------------
Finished episode: 975 Reward: -1314.6330 total_loss = 12.8887 = -0.0056 + 0.5 * 2
5.7859 + 0.01 * 0.1424
----------------
```

```
Finished episode: 976 Reward: -1312.4090 total_loss = 14.7114 = -0.0094 + 0.5 * 2
9.4389 + 0.01 * 0.1294
-----------------
Finished episode: 977 Reward: -1324.0674 total_loss = 15.6413 = -0.0260 + 0.5 * 3
1.3324 + 0.01 * 0.1178
-----------------
Finished episode: 978 Reward: -1327.8508 total_loss = 13.8263 = -0.0489 + 0.5 * 2
7.7480 + 0.01 * 0.1148
-----------------
Finished episode: 979 Reward: -1332.2642 total_loss = 13.9112 = -0.0574 + 0.5 * 2
7.9342 + 0.01 * 0.1536
-----------------
Finished episode: 980 Reward: -1331.1797 total_loss = 13.8982 = 0.0079 + 0.5 * 27.
7767 + 0.01 * 0.1901
-----------------
Finished episode: 981 Reward: -1332.1814 total_loss = 15.8933 = 0.0492 + 0.5 * 31.
6856 + 0.01 * 0.1292
-----------------
Finished episode: 982 Reward: -1333.9112 total_loss = 13.1791 = 0.0762 + 0.5 * 26.
2023 + 0.01 * 0.1778
-----------------
Finished episode: 983 Reward: -1325.9332 total_loss = 15.6024 = 0.1325 + 0.5 * 30.
9375 + 0.01 * 0.1141
-----------------
Finished episode: 984 Reward: -1322.3346 total_loss = 15.3658 = 0.0605 + 0.5 * 30.
6079 + 0.01 * 0.1405
-----------------
Finished episode: 985 Reward: -1314.4285 total_loss = 12.8199 = -0.0561 + 0.5 * 2
5.7489 + 0.01 * 0.1478
-----------------
Finished episode: 986 Reward: -1325.6319 total_loss = 13.0016 = -0.0009 + 0.5 * 2
6.0018 + 0.01 * 0.1595
-----------------
Finished episode: 987 Reward: -1314.6818 total_loss = 13.8866 = 0.0560 + 0.5 * 27.
6582 + 0.01 * 0.1600
-----------------
Finished episode: 988 Reward: -1314.5201 total_loss = 15.9895 = 0.1312 + 0.5 * 31.
7135 + 0.01 * 0.1553
-----------------
Finished episode: 989 Reward: -1332.1650 total_loss = 13.9364 = -0.1014 + 0.5 * 2
8.0724 + 0.01 * 0.1650
-----------------
Finished episode: 990 Reward: -1315.8172 total_loss = 12.7319 = 0.0451 + 0.5 * 25.
3718 + 0.01 * 0.0933
-----------------
Finished episode: 991 Reward: -1312.7142 total_loss = 16.4893 = -0.0031 + 0.5 * 3
2.9816 + 0.01 * 0.1600
-----------------
Finished episode: 992 Reward: -1309.5735 total_loss = 14.9015 = 0.0384 + 0.5 * 29.
7231 + 0.01 * 0.1549
-----------------
Finished episode: 993 Reward: -1319.2275 total_loss = 16.4035 = -0.0034 + 0.5 * 3
2.8108 + 0.01 * 0.1454
-----------------
Finished episode: 994 Reward: -1324.8889 total_loss = 16.7029 = -0.0101 + 0.5 * 3
3.4228 + 0.01 * 0.1524
-----------------
Finished episode: 995 Reward: -1319.8733 total_loss = 15.6083 = -0.0100 + 0.5 * 3
1.2341 + 0.01 * 0.1292
-----------------
Finished episode: 996 Reward: -1328.3687 total_loss = 15.5365 = 0.0108 + 0.5 * 31.
```

```
0489 + 0.01 * 0.1292
----------------
Finished episode: 997 Reward: -1320.7008 total_loss = 14.3026 = 0.0434 + 0.5 * 28.
5150 + 0.01 * 0.1691
----------------
Finished episode: 998 Reward: -1323.6432 total_loss = 14.7596 = -0.0206 + 0.5 * 2
9.5574 + 0.01 * 0.1471
----------------
Finished episode: 999 Reward: -1322.7430 total_loss = 14.6683 = 0.0240 + 0.5 * 29.
2861 + 0.01 * 0.1298
----------------
```

DDPG and TD3

The Deterministic Policy Gradient method was proposed by Silver et. al. 2014 (http://proceedings.mlr.press/v32/silver14.pdf (http://proceedings.mlr.press/v32/silver14.pdf)), and DDPG is its deep version.

The DPG also uses the actor-critic paradigm, but maitains a deterministic version of policy. It optimizes the critic through the Bellman Equation, and optimize the actor through the chain rule.

In this assignment, you may need to import some python files like DDPG.py and TD3.py to insert the method into training. Here are some solutions from stackoverflow:
https://stackoverflow.com/questions/48905127/importing-py-files-in-google-colab (https://stackoverflow.com/questions/48905127/importing-py-files-in-google-colab).

It is easier to just copy it from Drive than upload it.

1. Store MYLIB.py in your Drive. (for this assignment, it will be the utils.py, DDPG.py and TD3.py)
2. Open the Colab.
3. Open the left side pane, select Files view (the file icon).
4. Click Mount Drive then Connect to Google Drive (the folder with google drive icon).
5. Copy it by running "! cp drive/My\ Drive/MYLIB.py . " in your Colab file code line.
6. import MYLIB

TODOs for You (Please write down the answer in this block)

The TD3 is short for *Twin Delayed Deep Deterministic Policy Gradient*, their official open-source implementation is extremely clear and easy to follow! So I believe there is no need for you to build up the wheels one more time.

However, you really need to know about how this method works! TD3 proposes several improvements based on the method of DDPG to improve its sample efficiency.

- Q6. In this part, your task is to read the paper, and read the code of the official implementation of TD3 and DDPG at:

https://github.com/sfujim/TD3/blob/master/DDPG.py (https://github.com/sfujim/TD3/blob/master/DDPG.py)

https://github.com/sfujim/TD3/blob/master/TD3.py (https://github.com/sfujim/TD3/blob/master/TD3.py)

Then, please try to find the proposed improvements in TD3 over DDPG and summary them HERE:

1. It is indeed the DDPG has achieved better performance sometimes. However,it is brittle regarding with hyperparameters and other kinds of tuning. One of the challenging issue of the DDPG is about the overestimates of the Q-values, resulting in the policy breaking since it exploits the errors in the Q function. Twin Delayed DDPG (TD3) is an algorithm that addresses this issue by introducing three critical tricks.
2. Clipped Double-Q Learning. TD3 learns two Q-functions rather than one (hence “twin”), and utilizes the smaller of the two Q-values to form the targets in the Bellman error loss functions. For example:

```
target_Q1, target_Q2 = self.critic_target(next_state, next_action)
target_Q = torch.min(target_Q1, target_Q2)
```

3. “Delayed” Policy Updates. TD3 updates the policy (and target networks) less frequently than the Q-function. For example: if self.total_it % self.policy_freq == 0:

```
actor_loss = -self.critic.Q1(state, self.actor(state)).mean()
self.actor_optimizer.zero_grad()
actor_loss.backward()
self.actor_optimizer.step()
for param, target_param in zip(self.critic.parameters(), self.critic_target.parame
ters()):
    target_param.data.copy_(self.tau * param.data + (1 - self.tau) * target_param.
data)
for param, target_param in zip(self.actor.parameters(), self.actor_target.paramete
rs()):
    target_param.data.copy_(self.tau * param.data + (1 - self.tau) * target_param.
data)
```

4. Target Policy Smoothing. It is necessary to note that the TD3 adds noise to the target action for the purpose of making it harder for the policy to exploit Q-function errors by smoothing out Q along changes in action. For example:

```
noise = (torch.randn_like(action) * self.policy_noise).clamp(-self.noise_clip, sel
f.noise_clip)
next_action = (self.actor_target(next_state) + noise).clamp(-self.max_action, sel
f.max_action)
```

- Q7. Among all those improvements, which do you believe is the most important one? You may take some ablation studies to support your claim. (i.e., draw some learning curves with different settings together and draw your conclusions)

  1.The improvement of two Clipped Critic networks is the most important.

  2.It is necessary to notice that the maximum value of the Q-value is required to be calculated for the next state. This in fact means that the best action needs to be chosen in the next state. However, the Q-values in the early stages are evolving and therefore we lack enough information for getting the best action in the next state. Considering that in early stages, not a lot of states have been explored and actions been tried, the issue of the overestimate of the target Q-values may appear. Therefore, it is crucial for using double Q-learning where the actions are derived from the Target Q Network in the Target calculation. Since target Q network is stationary for a while, it can be used to derive action which can be used in the target calculation. Hence, two functions are used to calculate the Q value and the algorithm uses the smaller of the two Q values to update the Target.

- Q8. What is the difference between TD3(DDPG) and PPO in the OPTIMIZATION step (including but not restricted in terms of the sampling-training proportion)? Actually the improvements of PPO over TRPO was pointed as a benefit of more training iterations, can you further improve the sample efficiency of TD3?

  1.Difference between TD3(DDPG) and PPO in the OPTIMIZATION step

  For the PPO algorithm:
  1.1. PPO is an on-policy algorithm.
  1.2. PPO methods are simpler to implement.
  1.3. There are two variants of PPO including PPO-Penalty and PPO-Clip.
  For the DDPG algorithm:
  1.1. DDPG is an off-policy algorithm.
  1.2. DDPG can be thought of as being deep Q-learning for continuous action spaces.
  1.3. It uses off-policy data and the Bellman equation to learn the Q-function and uses the Q-function to learn the policy DDPG can only be used for environments with continuous action spaces.
  For the TD3 algorithm:
  1.1. TD3 is an off-policy algorithm.
  1.2. TD3 can only be used for environments with continuous action spaces.
  1.3. It is frequently brittle with respect to hyper-parameters and other kinds of tuning.
  1.4. TD3 learns two Q-functions instead of one and uses the smaller of the two Q-values to form the targets in the loss functions.
  1.5. TD3 updates the policy (and target networks) less frequently than Q-function.
  1.6. TD3 adds noise to target action to exploit Q-function errors by smoothing out Q along with changes in action.

  2.Further improve the sample efficiency of TD3

2.1. Reinforcement learning algorithms are notoriously known for the amount of samples they need for training.Typically, on-policy algorithms are much less sample efficient compared to off-policy algorithms. But there are other algorithmic features that allow improving the sample efficiency even more, like using a DND in NEC, or usingHindsight Experience Replay.

2.2. In addition, a variant of Hindsight Experience Replay, Curious and Aggressive Hindsight Experience Replay, is also beneficial for improving the sample efficiency of reinforcement learning methods.

2.3. In the following reference,a novel method for improving the sample efficiency is a population-based automated RL (AutoRL). In the framework of the AutoRl, the hyperparameters and also the neural architecture are optimized while simultaneously training the agent. By the way of sharing the collected experience across the population, the sample efficiency of the meta-optimization is substantially increased. [Ref.] D. Steckelmacher & H. Plisnier & M. Diederik et.al. Sample-efficient modelfree reinforcement learning with off-policy critics. European Conference on Machine Learning, 2019.

2.4. GAEL (Generalized Advantage Estimate learning) helps the RL agent learn from an expert player or an expert/pre-learned trajectory. This helps in providing imitation based learning and improves convergence and sample efficiency.

- Q9. (i) Please describe the difference of the exploration strategies between PPO, DDPG and TD3. (ii) Provide a comparison between the exploration strategies of those continuous control algorithms and DQN.

1. Please describe the difference of the exploration strategies between PPO, DDPG and TD3.

1.1. In RL models with continuous action spaces, instead of greedy mechanism undirected exploration is applied. This method is used in DDPG, PPO and other continuous control. Authors of DDPG constructed undirected exploration policy by adding noise sampled from a noise process to the actor policy.

1.2. The TD3 stands for Twin Delayed Deep Deterministic. TD3 retains the Actor-Critic architecture used in DDPG, and adds 3 new properties that greatly help to overcome overestimation:

1.2.1 TD3 maintains a pair of critics Q1 amd Q2 (hence the name “twin”) along with a single actor. For each time step, TD3 uses the smaller of the two Q-values.

1.2.2. TD3 updates the policy (and target networks) less frequently than the Q-function updates (one policy update (actor) for every two Q-function (critic) updates) TD3 adds exploration noise to the target action.

1.2.3. TD3 uses Gaussian noise, not Ornstein-Uhlenbeck noise as in DDPG.

1.3. PPO trains a stochastic policy in an on-policy way. This means that it explores by sampling actions according to the latest version of its stochastic policy. The amount of randomness in action selection depends on both initial conditions and the training procedure. Over the course of training, the policy typically becomes progressively less random, as the update rule encourages it to exploit rewards that it has already found. This may cause the policy to get trapped in local optima.

2. Provide a comparison between the exploration strategies of those continuous control and DQN.

One way to ensure adequate exploration in DQN and Double DQN is to use the annealing greedy mechanism. For the first episodes, exploitation is selected with a small probability, for example, 0.02 (i.e., the action will be chosen very randomly) and the exploration is selected with a probability 0.98. Starting from certain number of episode, the exploration will be performed with a minimal probability, for example,0.0

The following four blocks download the code in official implementation to your google drive so that the following script can run them. Note that the downloaded files may disappear due to some colab mechansim.

In [17]:

```
from os import makedirs as mkdir
mkdir('results', exist_ok=True)
```

In [18]:

```
# The following scripts run the DDPG algorithm.

alias = 'ddpg' # an alias of your experiment, used as a label


import matplotlib.pyplot as plt
import numpy as np
import torch
import gym
import argparse
import os
import torch.nn.functional as F
import utils
import TD3
import DDPG

def eval_policy(policy, eval_episodes=10):
    eval_env = gym.make(ENV_NAME)

    avg_reward = 0.
    for _ in range(eval_episodes):
        state, done = eval_env.reset(), False
        while not done:
            action = policy.select_action(np.array(state))
            state, reward, done, _ = eval_env.step(action)
            avg_reward += reward

    avg_reward /= eval_episodes
    #print("---------------------------------------")
    #print(f"Evaluation over {eval_episodes} episodes: {avg_reward:.3f}")
    #print("---------------------------------------")
    return avg_reward

env = gym.make(ENV_NAME)
torch.manual_seed(0)
np.random.seed(0)

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

args_policy_noise = 0.2
args_noise_clip = 0.5
args_policy_freq = 2
args_max_timesteps = 100000
args_expl_noise = 0.1
args_batch_size = 25
args_eval_freq = 1000
args_start_timesteps = 0

kwargs = {
    "state_dim": state_dim,
    "action_dim": action_dim,
    "max_action": max_action,
    "discount": 0.99,
    "tau": 0.005
}
```

```
args_policy = 'DDPG'

if args_policy == "TD3":
    # Target policy smoothing is scaled wrt the action scale
    kwargs["policy_noise"] = args_policy_noise * max_action
    kwargs["noise_clip"] = args_noise_clip * max_action
    kwargs["policy_freq"] = args_policy_freq
    policy = TD3.TD3(**kwargs)
elif args_policy == "DDPG":
    policy = DDPG.DDPG(**kwargs)
replay_buffer = utils.ReplayBuffer(state_dim, action_dim)

# Evaluate untrained policy
evaluations = [eval_policy(policy)]

state, done = env.reset(), False
episode_reward = 0
episode_timesteps = 0
episode_num = 0
counter = 0
msk_list = []
temp_curve = [eval_policy(policy)]
temp_val = []
for t in range(int(args_max_timesteps)):
    episode_timesteps += 1
    counter += 1
    # Select action randomly or according to policy
    if t < args_start_timesteps:
        action = np.random.uniform(-max_action,max_action,action_dim)
    else:
        if np.random.uniform(0,1) < 0.1:
            action = np.random.uniform(-max_action,max_action,action_dim)
        else:
            action = (
                policy.select_action(np.array(state))
                + np.random.normal(0, max_action * args_expl_noise, size=action_dim)
            ).clip(-max_action, max_action)

    # Perform action
    next_state, reward, done, _ = env.step(action)
    done_bool = float(done) if episode_timesteps < env._max_episode_steps else 0

    replay_buffer.add(state, action, next_state, reward, done_bool)

    state = next_state
    episode_reward += reward

    if t >= args_start_timesteps:
        '''TD3'''
        last_val = 999.
        patient = 5
        for i in range(1):
            policy.train(replay_buffer, args_batch_size)


    # Train agent after collecting sufficient data
    if done:
        print(f"Total T: {t+1} Episode Num: {episode_num+1} Episode T: {episode_timesteps} Re
ward: {episode_reward:.3f}")
        msk_list = []
        state, done = env.reset(), False
```

```
        episode_reward = 0
        episode_timesteps = 0
        episode_num += 1

    # Evaluate episode
    if (t + 1) % args_eval_freq == 0:
        evaluations.append(eval_policy(policy))
        print('recent Evaluation:',evaluations[-1])
        np.save('results/evaluations_alias{}_ENV{}'.format(alias,ENV_NAME),evaluations)
```

```
Total T: 200 Episode Num: 1 Episode T: 200 Reward: -1452.225
Total T: 400 Episode Num: 2 Episode T: 200 Reward: -1421.314
Total T: 600 Episode Num: 3 Episode T: 200 Reward: -1331.717
Total T: 800 Episode Num: 4 Episode T: 200 Reward: -1197.341
Total T: 1000 Episode Num: 5 Episode T: 200 Reward: -1542.474
recent Evaluation: -1453.3473573186948
Total T: 1200 Episode Num: 6 Episode T: 200 Reward: -1525.906
Total T: 1400 Episode Num: 7 Episode T: 200 Reward: -1204.637
Total T: 1600 Episode Num: 8 Episode T: 200 Reward: -1343.908
Total T: 1800 Episode Num: 9 Episode T: 200 Reward: -1268.647
Total T: 2000 Episode Num: 10 Episode T: 200 Reward: -1472.785
recent Evaluation: -1312.350163403058
Total T: 2200 Episode Num: 11 Episode T: 200 Reward: -1255.025
Total T: 2400 Episode Num: 12 Episode T: 200 Reward: -1477.864
Total T: 2600 Episode Num: 13 Episode T: 200 Reward: -1513.019
Total T: 2800 Episode Num: 14 Episode T: 200 Reward: -948.967
Total T: 3000 Episode Num: 15 Episode T: 200 Reward: -1548.249
recent Evaluation: -1494.0800492857502
Total T: 3200 Episode Num: 16 Episode T: 200 Reward: -1375.211
Total T: 3400 Episode Num: 17 Episode T: 200 Reward: -1428.271
Total T: 3600 Episode Num: 18 Episode T: 200 Reward: -956.032
Total T: 3800 Episode Num: 19 Episode T: 200 Reward: -1477.895
Total T: 4000 Episode Num: 20 Episode T: 200 Reward: -1174.887
recent Evaluation: -1420.829600343822
Total T: 4200 Episode Num: 21 Episode T: 200 Reward: -1581.850
Total T: 4400 Episode Num: 22 Episode T: 200 Reward: -1061.968
Total T: 4600 Episode Num: 23 Episode T: 200 Reward: -1525.369
Total T: 4800 Episode Num: 24 Episode T: 200 Reward: -1550.577
Total T: 5000 Episode Num: 25 Episode T: 200 Reward: -1555.735
recent Evaluation: -1455.4930837489057
Total T: 5200 Episode Num: 26 Episode T: 200 Reward: -1377.232
Total T: 5400 Episode Num: 27 Episode T: 200 Reward: -1390.264
Total T: 5600 Episode Num: 28 Episode T: 200 Reward: -1376.348
Total T: 5800 Episode Num: 29 Episode T: 200 Reward: -1613.844
Total T: 6000 Episode Num: 30 Episode T: 200 Reward: -1427.677
recent Evaluation: -1388.375429669532
Total T: 6200 Episode Num: 31 Episode T: 200 Reward: -1213.466
Total T: 6400 Episode Num: 32 Episode T: 200 Reward: -1295.735
Total T: 6600 Episode Num: 33 Episode T: 200 Reward: -1624.679
Total T: 6800 Episode Num: 34 Episode T: 200 Reward: -1549.925
Total T: 7000 Episode Num: 35 Episode T: 200 Reward: -1377.719
recent Evaluation: -1432.8983968093535
Total T: 7200 Episode Num: 36 Episode T: 200 Reward: -1489.854
Total T: 7400 Episode Num: 37 Episode T: 200 Reward: -1269.967
Total T: 7600 Episode Num: 38 Episode T: 200 Reward: -1522.082
Total T: 7800 Episode Num: 39 Episode T: 200 Reward: -1336.153
Total T: 8000 Episode Num: 40 Episode T: 200 Reward: -1546.706
recent Evaluation: -1372.0029146826712
Total T: 8200 Episode Num: 41 Episode T: 200 Reward: -1480.860
Total T: 8400 Episode Num: 42 Episode T: 200 Reward: -993.127
Total T: 8600 Episode Num: 43 Episode T: 200 Reward: -1606.080
Total T: 8800 Episode Num: 44 Episode T: 200 Reward: -1574.129
Total T: 9000 Episode Num: 45 Episode T: 200 Reward: -1124.130
recent Evaluation: -1538.923955231508
Total T: 9200 Episode Num: 46 Episode T: 200 Reward: -1406.059
Total T: 9400 Episode Num: 47 Episode T: 200 Reward: -1400.721
Total T: 9600 Episode Num: 48 Episode T: 200 Reward: -1605.685
Total T: 9800 Episode Num: 49 Episode T: 200 Reward: -1068.792
Total T: 10000 Episode Num: 50 Episode T: 200 Reward: -1558.733
recent Evaluation: -1523.5087839815549
Total T: 10200 Episode Num: 51 Episode T: 200 Reward: -1377.639
```

```
Total T: 10400 Episode Num: 52 Episode T: 200 Reward: -1079.933
Total T: 10600 Episode Num: 53 Episode T: 200 Reward: -1176.986
Total T: 10800 Episode Num: 54 Episode T: 200 Reward: -1537.831
Total T: 11000 Episode Num: 55 Episode T: 200 Reward: -1579.477
recent Evaluation: -1433.128287384317
Total T: 11200 Episode Num: 56 Episode T: 200 Reward: -1293.690
Total T: 11400 Episode Num: 57 Episode T: 200 Reward: -1407.105
Total T: 11600 Episode Num: 58 Episode T: 200 Reward: -1553.166
Total T: 11800 Episode Num: 59 Episode T: 200 Reward: -1528.172
Total T: 12000 Episode Num: 60 Episode T: 200 Reward: -1551.863
recent Evaluation: -1452.587829964126
Total T: 12200 Episode Num: 61 Episode T: 200 Reward: -1303.634
Total T: 12400 Episode Num: 62 Episode T: 200 Reward: -1545.977
Total T: 12600 Episode Num: 63 Episode T: 200 Reward: -1560.872
Total T: 12800 Episode Num: 64 Episode T: 200 Reward: -1538.999
Total T: 13000 Episode Num: 65 Episode T: 200 Reward: -1181.898
recent Evaluation: -1437.7471821726672
Total T: 13200 Episode Num: 66 Episode T: 200 Reward: -1519.574
Total T: 13400 Episode Num: 67 Episode T: 200 Reward: -1565.840
Total T: 13600 Episode Num: 68 Episode T: 200 Reward: -1569.086
Total T: 13800 Episode Num: 69 Episode T: 200 Reward: -1353.221
Total T: 14000 Episode Num: 70 Episode T: 200 Reward: -1596.770
recent Evaluation: -1501.728524691613
Total T: 14200 Episode Num: 71 Episode T: 200 Reward: -1637.349
Total T: 14400 Episode Num: 72 Episode T: 200 Reward: -1504.919
Total T: 14600 Episode Num: 73 Episode T: 200 Reward: -1246.986
Total T: 14800 Episode Num: 74 Episode T: 200 Reward: -1576.876
Total T: 15000 Episode Num: 75 Episode T: 200 Reward: -968.891
recent Evaluation: -1489.6202395706819
Total T: 15200 Episode Num: 76 Episode T: 200 Reward: -1389.588
Total T: 15400 Episode Num: 77 Episode T: 200 Reward: -1605.987
Total T: 15600 Episode Num: 78 Episode T: 200 Reward: -1461.591
Total T: 15800 Episode Num: 79 Episode T: 200 Reward: -1246.755
Total T: 16000 Episode Num: 80 Episode T: 200 Reward: -1537.938
recent Evaluation: -1399.0652371358526
Total T: 16200 Episode Num: 81 Episode T: 200 Reward: -1491.734
Total T: 16400 Episode Num: 82 Episode T: 200 Reward: -1070.933
Total T: 16600 Episode Num: 83 Episode T: 200 Reward: -1192.353
Total T: 16800 Episode Num: 84 Episode T: 200 Reward: -1130.225
Total T: 17000 Episode Num: 85 Episode T: 200 Reward: -1336.540
recent Evaluation: -1558.2891938904042
Total T: 17200 Episode Num: 86 Episode T: 200 Reward: -1481.335
Total T: 17400 Episode Num: 87 Episode T: 200 Reward: -1547.939
Total T: 17600 Episode Num: 88 Episode T: 200 Reward: -1542.869
Total T: 17800 Episode Num: 89 Episode T: 200 Reward: -1559.725
Total T: 18000 Episode Num: 90 Episode T: 200 Reward: -1527.590
recent Evaluation: -1419.1378692444587
Total T: 18200 Episode Num: 91 Episode T: 200 Reward: -1523.063
Total T: 18400 Episode Num: 92 Episode T: 200 Reward: -1286.768
Total T: 18600 Episode Num: 93 Episode T: 200 Reward: -1037.437
Total T: 18800 Episode Num: 94 Episode T: 200 Reward: -1108.331
Total T: 19000 Episode Num: 95 Episode T: 200 Reward: -988.953
recent Evaluation: -339.7839939089142
Total T: 19200 Episode Num: 96 Episode T: 200 Reward: -970.746
Total T: 19400 Episode Num: 97 Episode T: 200 Reward: -694.838
Total T: 19600 Episode Num: 98 Episode T: 200 Reward: -864.985
Total T: 19800 Episode Num: 99 Episode T: 200 Reward: -611.151
Total T: 20000 Episode Num: 100 Episode T: 200 Reward: -377.919
recent Evaluation: -179.4970095998468
Total T: 20200 Episode Num: 101 Episode T: 200 Reward: -895.622
Total T: 20400 Episode Num: 102 Episode T: 200 Reward: -978.701
```

```
Total T: 20600 Episode Num: 103 Episode T: 200 Reward: -510.493
Total T: 20800 Episode Num: 104 Episode T: 200 Reward: -381.877
Total T: 21000 Episode Num: 105 Episode T: 200 Reward: -619.775
recent Evaluation: -178.29206987316275
Total T: 21200 Episode Num: 106 Episode T: 200 Reward: -337.160
Total T: 21400 Episode Num: 107 Episode T: 200 Reward: -500.426
Total T: 21600 Episode Num: 108 Episode T: 200 Reward: -750.902
Total T: 21800 Episode Num: 109 Episode T: 200 Reward: -388.284
Total T: 22000 Episode Num: 110 Episode T: 200 Reward: -619.960
recent Evaluation: -86.03722055105368
Total T: 22200 Episode Num: 111 Episode T: 200 Reward: -257.240
Total T: 22400 Episode Num: 112 Episode T: 200 Reward: -5.387
Total T: 22600 Episode Num: 113 Episode T: 200 Reward: -470.061
Total T: 22800 Episode Num: 114 Episode T: 200 Reward: -256.609
Total T: 23000 Episode Num: 115 Episode T: 200 Reward: -239.922
recent Evaluation: -262.88415257023496
Total T: 23200 Episode Num: 116 Episode T: 200 Reward: -1.322
Total T: 23400 Episode Num: 117 Episode T: 200 Reward: -1.904
Total T: 23600 Episode Num: 118 Episode T: 200 Reward: -126.752
Total T: 23800 Episode Num: 119 Episode T: 200 Reward: -2.288
Total T: 24000 Episode Num: 120 Episode T: 200 Reward: -124.418
recent Evaluation: -170.21401046632448
Total T: 24200 Episode Num: 121 Episode T: 200 Reward: -124.915
Total T: 24400 Episode Num: 122 Episode T: 200 Reward: -242.312
Total T: 24600 Episode Num: 123 Episode T: 200 Reward: -128.483
Total T: 24800 Episode Num: 124 Episode T: 200 Reward: -2.873
Total T: 25000 Episode Num: 125 Episode T: 200 Reward: -131.081
recent Evaluation: -234.0506619199363
Total T: 25200 Episode Num: 126 Episode T: 200 Reward: -249.918
Total T: 25400 Episode Num: 127 Episode T: 200 Reward: -130.393
Total T: 25600 Episode Num: 128 Episode T: 200 Reward: -259.260
Total T: 25800 Episode Num: 129 Episode T: 200 Reward: -125.087
Total T: 26000 Episode Num: 130 Episode T: 200 Reward: -248.723
recent Evaluation: -189.8568771237772
Total T: 26200 Episode Num: 131 Episode T: 200 Reward: -366.425
Total T: 26400 Episode Num: 132 Episode T: 200 Reward: -258.759
Total T: 26600 Episode Num: 133 Episode T: 200 Reward: -130.501
Total T: 26800 Episode Num: 134 Episode T: 200 Reward: -129.745
Total T: 27000 Episode Num: 135 Episode T: 200 Reward: -231.888
recent Evaluation: -150.10456249198555
Total T: 27200 Episode Num: 136 Episode T: 200 Reward: -252.049
Total T: 27400 Episode Num: 137 Episode T: 200 Reward: -124.597
Total T: 27600 Episode Num: 138 Episode T: 200 Reward: -246.286
Total T: 27800 Episode Num: 139 Episode T: 200 Reward: -13.844
Total T: 28000 Episode Num: 140 Episode T: 200 Reward: -382.004
recent Evaluation: -183.11672710346875
Total T: 28200 Episode Num: 141 Episode T: 200 Reward: -124.673
Total T: 28400 Episode Num: 142 Episode T: 200 Reward: -128.477
Total T: 28600 Episode Num: 143 Episode T: 200 Reward: -116.678
Total T: 28800 Episode Num: 144 Episode T: 200 Reward: -369.384
Total T: 29000 Episode Num: 145 Episode T: 200 Reward: -261.093
recent Evaluation: -113.26880928284777
Total T: 29200 Episode Num: 146 Episode T: 200 Reward: -232.198
Total T: 29400 Episode Num: 147 Episode T: 200 Reward: -239.399
Total T: 29600 Episode Num: 148 Episode T: 200 Reward: -4.300
Total T: 29800 Episode Num: 149 Episode T: 200 Reward: -130.095
Total T: 30000 Episode Num: 150 Episode T: 200 Reward: -127.471
recent Evaluation: -86.02925489836053
Total T: 30200 Episode Num: 151 Episode T: 200 Reward: -240.493
Total T: 30400 Episode Num: 152 Episode T: 200 Reward: -392.675
Total T: 30600 Episode Num: 153 Episode T: 200 Reward: -125.831
```

```
Total T: 30800 Episode Num: 154 Episode T: 200 Reward: -238.080
Total T: 31000 Episode Num: 155 Episode T: 200 Reward: -118.050
recent Evaluation: -139.0066219639282
Total T: 31200 Episode Num: 156 Episode T: 200 Reward: -238.065
Total T: 31400 Episode Num: 157 Episode T: 200 Reward: -128.125
Total T: 31600 Episode Num: 158 Episode T: 200 Reward: -130.299
Total T: 31800 Episode Num: 159 Episode T: 200 Reward: -125.932
Total T: 32000 Episode Num: 160 Episode T: 200 Reward: -122.091
recent Evaluation: -122.33931269035
Total T: 32200 Episode Num: 161 Episode T: 200 Reward: -130.447
Total T: 32400 Episode Num: 162 Episode T: 200 Reward: -242.669
Total T: 32600 Episode Num: 163 Episode T: 200 Reward: -339.462
Total T: 32800 Episode Num: 164 Episode T: 200 Reward: -331.560
Total T: 33000 Episode Num: 165 Episode T: 200 Reward: -131.529
recent Evaluation: -134.89290892918208
Total T: 33200 Episode Num: 166 Episode T: 200 Reward: -125.482
Total T: 33400 Episode Num: 167 Episode T: 200 Reward: -227.037
Total T: 33600 Episode Num: 168 Episode T: 200 Reward: -249.733
Total T: 33800 Episode Num: 169 Episode T: 200 Reward: -4.383
Total T: 34000 Episode Num: 170 Episode T: 200 Reward: -122.770
recent Evaluation: -146.34184974685857
Total T: 34200 Episode Num: 171 Episode T: 200 Reward: -253.649
Total T: 34400 Episode Num: 172 Episode T: 200 Reward: -4.153
Total T: 34600 Episode Num: 173 Episode T: 200 Reward: -128.418
Total T: 34800 Episode Num: 174 Episode T: 200 Reward: -256.884
Total T: 35000 Episode Num: 175 Episode T: 200 Reward: -249.875
recent Evaluation: -135.68361983105677
Total T: 35200 Episode Num: 176 Episode T: 200 Reward: -124.129
Total T: 35400 Episode Num: 177 Episode T: 200 Reward: -249.984
Total T: 35600 Episode Num: 178 Episode T: 200 Reward: -247.825
Total T: 35800 Episode Num: 179 Episode T: 200 Reward: -120.880
Total T: 36000 Episode Num: 180 Episode T: 200 Reward: -5.502
recent Evaluation: -149.27108305990328
Total T: 36200 Episode Num: 181 Episode T: 200 Reward: -251.703
Total T: 36400 Episode Num: 182 Episode T: 200 Reward: -367.517
Total T: 36600 Episode Num: 183 Episode T: 200 Reward: -127.952
Total T: 36800 Episode Num: 184 Episode T: 200 Reward: -244.529
Total T: 37000 Episode Num: 185 Episode T: 200 Reward: -127.201
recent Evaluation: -148.34091783184073
Total T: 37200 Episode Num: 186 Episode T: 200 Reward: -359.146
Total T: 37400 Episode Num: 187 Episode T: 200 Reward: -123.739
Total T: 37600 Episode Num: 188 Episode T: 200 Reward: -131.857
Total T: 37800 Episode Num: 189 Episode T: 200 Reward: -250.349
Total T: 38000 Episode Num: 190 Episode T: 200 Reward: -242.292
recent Evaluation: -189.0984997806494
Total T: 38200 Episode Num: 191 Episode T: 200 Reward: -360.691
Total T: 38400 Episode Num: 192 Episode T: 200 Reward: -5.882
Total T: 38600 Episode Num: 193 Episode T: 200 Reward: -126.659
Total T: 38800 Episode Num: 194 Episode T: 200 Reward: -126.916
Total T: 39000 Episode Num: 195 Episode T: 200 Reward: -237.618
recent Evaluation: -121.73158853201224
Total T: 39200 Episode Num: 196 Episode T: 200 Reward: -121.427
Total T: 39400 Episode Num: 197 Episode T: 200 Reward: -249.927
Total T: 39600 Episode Num: 198 Episode T: 200 Reward: -374.575
Total T: 39800 Episode Num: 199 Episode T: 200 Reward: -122.087
Total T: 40000 Episode Num: 200 Episode T: 200 Reward: -241.615
recent Evaluation: -132.61272400830094
Total T: 40200 Episode Num: 201 Episode T: 200 Reward: -7.282
Total T: 40400 Episode Num: 202 Episode T: 200 Reward: -251.545
Total T: 40600 Episode Num: 203 Episode T: 200 Reward: -5.763
Total T: 40800 Episode Num: 204 Episode T: 200 Reward: -120.066
```

```
Total T: 41000 Episode Num: 205 Episode T: 200 Reward: -258.320
recent Evaluation: -126.43025980787289
Total T: 41200 Episode Num: 206 Episode T: 200 Reward: -4.415
Total T: 41400 Episode Num: 207 Episode T: 200 Reward: -3.546
Total T: 41600 Episode Num: 208 Episode T: 200 Reward: -118.715
Total T: 41800 Episode Num: 209 Episode T: 200 Reward: -238.851
Total T: 42000 Episode Num: 210 Episode T: 200 Reward: -135.137
recent Evaluation: -170.6425947308261
Total T: 42200 Episode Num: 211 Episode T: 200 Reward: -128.664
Total T: 42400 Episode Num: 212 Episode T: 200 Reward: -230.065
Total T: 42600 Episode Num: 213 Episode T: 200 Reward: -229.782
Total T: 42800 Episode Num: 214 Episode T: 200 Reward: -231.178
Total T: 43000 Episode Num: 215 Episode T: 200 Reward: -480.055
recent Evaluation: -125.19420559091957
Total T: 43200 Episode Num: 216 Episode T: 200 Reward: -234.661
Total T: 43400 Episode Num: 217 Episode T: 200 Reward: -247.579
Total T: 43600 Episode Num: 218 Episode T: 200 Reward: -129.540
Total T: 43800 Episode Num: 219 Episode T: 200 Reward: -130.145
Total T: 44000 Episode Num: 220 Episode T: 200 Reward: -133.606
recent Evaluation: -147.54046455617186
Total T: 44200 Episode Num: 221 Episode T: 200 Reward: -373.724
Total T: 44400 Episode Num: 222 Episode T: 200 Reward: -4.997
Total T: 44600 Episode Num: 223 Episode T: 200 Reward: -134.320
Total T: 44800 Episode Num: 224 Episode T: 200 Reward: -126.873
Total T: 45000 Episode Num: 225 Episode T: 200 Reward: -7.898
recent Evaluation: -186.27362025635057
Total T: 45200 Episode Num: 226 Episode T: 200 Reward: -378.974
Total T: 45400 Episode Num: 227 Episode T: 200 Reward: -249.327
Total T: 45600 Episode Num: 228 Episode T: 200 Reward: -129.820
Total T: 45800 Episode Num: 229 Episode T: 200 Reward: -130.129
Total T: 46000 Episode Num: 230 Episode T: 200 Reward: -239.217
recent Evaluation: -152.35511345000265
Total T: 46200 Episode Num: 231 Episode T: 200 Reward: -243.120
Total T: 46400 Episode Num: 232 Episode T: 200 Reward: -336.429
Total T: 46600 Episode Num: 233 Episode T: 200 Reward: -258.045
Total T: 46800 Episode Num: 234 Episode T: 200 Reward: -130.951
Total T: 47000 Episode Num: 235 Episode T: 200 Reward: -235.802
recent Evaluation: -207.93132617575174
Total T: 47200 Episode Num: 236 Episode T: 200 Reward: -234.576
Total T: 47400 Episode Num: 237 Episode T: 200 Reward: -240.840
Total T: 47600 Episode Num: 238 Episode T: 200 Reward: -129.833
Total T: 47800 Episode Num: 239 Episode T: 200 Reward: -229.798
Total T: 48000 Episode Num: 240 Episode T: 200 Reward: -133.060
recent Evaluation: -137.35053593067988
Total T: 48200 Episode Num: 241 Episode T: 200 Reward: -243.922
Total T: 48400 Episode Num: 242 Episode T: 200 Reward: -122.404
Total T: 48600 Episode Num: 243 Episode T: 200 Reward: -126.823
Total T: 48800 Episode Num: 244 Episode T: 200 Reward: -134.095
Total T: 49000 Episode Num: 245 Episode T: 200 Reward: -337.892
recent Evaluation: -144.21206333460043
Total T: 49200 Episode Num: 246 Episode T: 200 Reward: -126.064
Total T: 49400 Episode Num: 247 Episode T: 200 Reward: -376.888
Total T: 49600 Episode Num: 248 Episode T: 200 Reward: -128.900
Total T: 49800 Episode Num: 249 Episode T: 200 Reward: -4.966
Total T: 50000 Episode Num: 250 Episode T: 200 Reward: -121.515
recent Evaluation: -175.85816333949288
Total T: 50200 Episode Num: 251 Episode T: 200 Reward: -228.212
Total T: 50400 Episode Num: 252 Episode T: 200 Reward: -414.640
Total T: 50600 Episode Num: 253 Episode T: 200 Reward: -254.063
Total T: 50800 Episode Num: 254 Episode T: 200 Reward: -137.948
Total T: 51000 Episode Num: 255 Episode T: 200 Reward: -123.208
```

```
recent Evaluation: -112.4145310825098
Total T: 51200 Episode Num: 256 Episode T: 200 Reward: -134.590
Total T: 51400 Episode Num: 257 Episode T: 200 Reward: -121.074
Total T: 51600 Episode Num: 258 Episode T: 200 Reward: -121.510
Total T: 51800 Episode Num: 259 Episode T: 200 Reward: -117.710
Total T: 52000 Episode Num: 260 Episode T: 200 Reward: -132.233
recent Evaluation: -122.9523726904337
Total T: 52200 Episode Num: 261 Episode T: 200 Reward: -130.913
Total T: 52400 Episode Num: 262 Episode T: 200 Reward: -8.475
Total T: 52600 Episode Num: 263 Episode T: 200 Reward: -349.427
Total T: 52800 Episode Num: 264 Episode T: 200 Reward: -3.589
Total T: 53000 Episode Num: 265 Episode T: 200 Reward: -245.009
recent Evaluation: -163.93528403638132
Total T: 53200 Episode Num: 266 Episode T: 200 Reward: -120.207
Total T: 53400 Episode Num: 267 Episode T: 200 Reward: -4.438
Total T: 53600 Episode Num: 268 Episode T: 200 Reward: -379.826
Total T: 53800 Episode Num: 269 Episode T: 200 Reward: -132.125
Total T: 54000 Episode Num: 270 Episode T: 200 Reward: -376.528
recent Evaluation: -158.71509333456177
Total T: 54200 Episode Num: 271 Episode T: 200 Reward: -247.745
Total T: 54400 Episode Num: 272 Episode T: 200 Reward: -124.327
Total T: 54600 Episode Num: 273 Episode T: 200 Reward: -130.542
Total T: 54800 Episode Num: 274 Episode T: 200 Reward: -329.512
Total T: 55000 Episode Num: 275 Episode T: 200 Reward: -245.885
recent Evaluation: -89.7208396179299
Total T: 55200 Episode Num: 276 Episode T: 200 Reward: -252.542
Total T: 55400 Episode Num: 277 Episode T: 200 Reward: -118.706
Total T: 55600 Episode Num: 278 Episode T: 200 Reward: -226.869
Total T: 55800 Episode Num: 279 Episode T: 200 Reward: -483.662
Total T: 56000 Episode Num: 280 Episode T: 200 Reward: -132.334
recent Evaluation: -135.73087786933243
Total T: 56200 Episode Num: 281 Episode T: 200 Reward: -234.317
Total T: 56400 Episode Num: 282 Episode T: 200 Reward: -3.215
Total T: 56600 Episode Num: 283 Episode T: 200 Reward: -130.549
Total T: 56800 Episode Num: 284 Episode T: 200 Reward: -126.554
Total T: 57000 Episode Num: 285 Episode T: 200 Reward: -122.200
recent Evaluation: -111.5109575825087
Total T: 57200 Episode Num: 286 Episode T: 200 Reward: -129.430
Total T: 57400 Episode Num: 287 Episode T: 200 Reward: -120.091
Total T: 57600 Episode Num: 288 Episode T: 200 Reward: -122.145
Total T: 57800 Episode Num: 289 Episode T: 200 Reward: -229.057
Total T: 58000 Episode Num: 290 Episode T: 200 Reward: -259.216
recent Evaluation: -90.63621845333083
Total T: 58200 Episode Num: 291 Episode T: 200 Reward: -122.001
Total T: 58400 Episode Num: 292 Episode T: 200 Reward: -134.049
Total T: 58600 Episode Num: 293 Episode T: 200 Reward: -135.658
Total T: 58800 Episode Num: 294 Episode T: 200 Reward: -133.939
Total T: 59000 Episode Num: 295 Episode T: 200 Reward: -130.494
recent Evaluation: -139.3865992448481
Total T: 59200 Episode Num: 296 Episode T: 200 Reward: -130.314
Total T: 59400 Episode Num: 297 Episode T: 200 Reward: -122.712
Total T: 59600 Episode Num: 298 Episode T: 200 Reward: -128.581
Total T: 59800 Episode Num: 299 Episode T: 200 Reward: -131.993
Total T: 60000 Episode Num: 300 Episode T: 200 Reward: -225.503
recent Evaluation: -181.00040586853504
Total T: 60200 Episode Num: 301 Episode T: 200 Reward: -253.649
Total T: 60400 Episode Num: 302 Episode T: 200 Reward: -246.891
Total T: 60600 Episode Num: 303 Episode T: 200 Reward: -375.246
Total T: 60800 Episode Num: 304 Episode T: 200 Reward: -135.295
Total T: 61000 Episode Num: 305 Episode T: 200 Reward: -119.084
recent Evaluation: -124.28234473844111
```

```
Total T: 61200 Episode Num: 306 Episode T: 200 Reward: -256.428
Total T: 61400 Episode Num: 307 Episode T: 200 Reward: -376.144
Total T: 61600 Episode Num: 308 Episode T: 200 Reward: -4.891
Total T: 61800 Episode Num: 309 Episode T: 200 Reward: -369.480
Total T: 62000 Episode Num: 310 Episode T: 200 Reward: -241.903
recent Evaluation: -181.5658094605411
Total T: 62200 Episode Num: 311 Episode T: 200 Reward: -371.758
Total T: 62400 Episode Num: 312 Episode T: 200 Reward: -117.837
Total T: 62600 Episode Num: 313 Episode T: 200 Reward: -123.128
Total T: 62800 Episode Num: 314 Episode T: 200 Reward: -244.447
Total T: 63000 Episode Num: 315 Episode T: 200 Reward: -4.070
recent Evaluation: -207.40179016360918
Total T: 63200 Episode Num: 316 Episode T: 200 Reward: -351.516
Total T: 63400 Episode Num: 317 Episode T: 200 Reward: -246.603
Total T: 63600 Episode Num: 318 Episode T: 200 Reward: -123.514
Total T: 63800 Episode Num: 319 Episode T: 200 Reward: -367.224
Total T: 64000 Episode Num: 320 Episode T: 200 Reward: -117.991
recent Evaluation: -124.63315363495857
Total T: 64200 Episode Num: 321 Episode T: 200 Reward: -128.638
Total T: 64400 Episode Num: 322 Episode T: 200 Reward: -4.333
Total T: 64600 Episode Num: 323 Episode T: 200 Reward: -128.868
Total T: 64800 Episode Num: 324 Episode T: 200 Reward: -357.837
Total T: 65000 Episode Num: 325 Episode T: 200 Reward: -129.811
recent Evaluation: -179.82251181110166
Total T: 65200 Episode Num: 326 Episode T: 200 Reward: -234.545
Total T: 65400 Episode Num: 327 Episode T: 200 Reward: -124.908
Total T: 65600 Episode Num: 328 Episode T: 200 Reward: -327.394
Total T: 65800 Episode Num: 329 Episode T: 200 Reward: -246.826
Total T: 66000 Episode Num: 330 Episode T: 200 Reward: -122.766
recent Evaluation: -138.70729209544916
Total T: 66200 Episode Num: 331 Episode T: 200 Reward: -249.551
Total T: 66400 Episode Num: 332 Episode T: 200 Reward: -122.440
Total T: 66600 Episode Num: 333 Episode T: 200 Reward: -248.220
Total T: 66800 Episode Num: 334 Episode T: 200 Reward: -129.466
Total T: 67000 Episode Num: 335 Episode T: 200 Reward: -361.809
recent Evaluation: -136.64039765473126
Total T: 67200 Episode Num: 336 Episode T: 200 Reward: -248.307
Total T: 67400 Episode Num: 337 Episode T: 200 Reward: -246.318
Total T: 67600 Episode Num: 338 Episode T: 200 Reward: -4.954
Total T: 67800 Episode Num: 339 Episode T: 200 Reward: -245.007
Total T: 68000 Episode Num: 340 Episode T: 200 Reward: -377.718
recent Evaluation: -137.45939003134177
Total T: 68200 Episode Num: 341 Episode T: 200 Reward: -4.915
Total T: 68400 Episode Num: 342 Episode T: 200 Reward: -232.874
Total T: 68600 Episode Num: 343 Episode T: 200 Reward: -134.834
Total T: 68800 Episode Num: 344 Episode T: 200 Reward: -127.351
Total T: 69000 Episode Num: 345 Episode T: 200 Reward: -253.447
recent Evaluation: -127.39702975390735
Total T: 69200 Episode Num: 346 Episode T: 200 Reward: -120.018
Total T: 69400 Episode Num: 347 Episode T: 200 Reward: -235.306
Total T: 69600 Episode Num: 348 Episode T: 200 Reward: -125.696
Total T: 69800 Episode Num: 349 Episode T: 200 Reward: -308.724
Total T: 70000 Episode Num: 350 Episode T: 200 Reward: -375.302
recent Evaluation: -146.40689450329995
Total T: 70200 Episode Num: 351 Episode T: 200 Reward: -119.221
Total T: 70400 Episode Num: 352 Episode T: 200 Reward: -249.644
Total T: 70600 Episode Num: 353 Episode T: 200 Reward: -121.089
Total T: 70800 Episode Num: 354 Episode T: 200 Reward: -2.826
Total T: 71000 Episode Num: 355 Episode T: 200 Reward: -118.743
recent Evaluation: -163.00808180374082
Total T: 71200 Episode Num: 356 Episode T: 200 Reward: -125.236
```

```
Total T: 71400 Episode Num: 357 Episode T: 200 Reward: -121.156
Total T: 71600 Episode Num: 358 Episode T: 200 Reward: -132.963
Total T: 71800 Episode Num: 359 Episode T: 200 Reward: -350.974
Total T: 72000 Episode Num: 360 Episode T: 200 Reward: -115.764
recent Evaluation: -134.3408420439079
Total T: 72200 Episode Num: 361 Episode T: 200 Reward: -118.833
Total T: 72400 Episode Num: 362 Episode T: 200 Reward: -366.873
Total T: 72600 Episode Num: 363 Episode T: 200 Reward: -242.563
Total T: 72800 Episode Num: 364 Episode T: 200 Reward: -120.112
Total T: 73000 Episode Num: 365 Episode T: 200 Reward: -124.487
recent Evaluation: -101.24472583416339
Total T: 73200 Episode Num: 366 Episode T: 200 Reward: -125.209
Total T: 73400 Episode Num: 367 Episode T: 200 Reward: -3.706
Total T: 73600 Episode Num: 368 Episode T: 200 Reward: -127.210
Total T: 73800 Episode Num: 369 Episode T: 200 Reward: -134.375
Total T: 74000 Episode Num: 370 Episode T: 200 Reward: -372.330
recent Evaluation: -110.42967638227294
Total T: 74200 Episode Num: 371 Episode T: 200 Reward: -119.511
Total T: 74400 Episode Num: 372 Episode T: 200 Reward: -243.497
Total T: 74600 Episode Num: 373 Episode T: 200 Reward: -134.386
Total T: 74800 Episode Num: 374 Episode T: 200 Reward: -4.213
Total T: 75000 Episode Num: 375 Episode T: 200 Reward: -248.662
recent Evaluation: -163.0091929091089
Total T: 75200 Episode Num: 376 Episode T: 200 Reward: -352.938
Total T: 75400 Episode Num: 377 Episode T: 200 Reward: -124.436
Total T: 75600 Episode Num: 378 Episode T: 200 Reward: -234.116
Total T: 75800 Episode Num: 379 Episode T: 200 Reward: -257.334
Total T: 76000 Episode Num: 380 Episode T: 200 Reward: -2.199
recent Evaluation: -161.15904244791207
Total T: 76200 Episode Num: 381 Episode T: 200 Reward: -378.158
Total T: 76400 Episode Num: 382 Episode T: 200 Reward: -246.671
Total T: 76600 Episode Num: 383 Episode T: 200 Reward: -119.428
Total T: 76800 Episode Num: 384 Episode T: 200 Reward: -361.939
Total T: 77000 Episode Num: 385 Episode T: 200 Reward: -132.935
recent Evaluation: -74.40362746612065
Total T: 77200 Episode Num: 386 Episode T: 200 Reward: -1.516
Total T: 77400 Episode Num: 387 Episode T: 200 Reward: -342.312
Total T: 77600 Episode Num: 388 Episode T: 200 Reward: -125.704
Total T: 77800 Episode Num: 389 Episode T: 200 Reward: -371.746
Total T: 78000 Episode Num: 390 Episode T: 200 Reward: -124.905
recent Evaluation: -137.71092668123842
Total T: 78200 Episode Num: 391 Episode T: 200 Reward: -1.954
Total T: 78400 Episode Num: 392 Episode T: 200 Reward: -2.932
Total T: 78600 Episode Num: 393 Episode T: 200 Reward: -249.394
Total T: 78800 Episode Num: 394 Episode T: 200 Reward: -124.251
Total T: 79000 Episode Num: 395 Episode T: 200 Reward: -119.569
recent Evaluation: -137.79922278117186
Total T: 79200 Episode Num: 396 Episode T: 200 Reward: -128.801
Total T: 79400 Episode Num: 397 Episode T: 200 Reward: -122.607
Total T: 79600 Episode Num: 398 Episode T: 200 Reward: -236.598
Total T: 79800 Episode Num: 399 Episode T: 200 Reward: -132.041
Total T: 80000 Episode Num: 400 Episode T: 200 Reward: -116.126
recent Evaluation: -171.93369118848776
Total T: 80200 Episode Num: 401 Episode T: 200 Reward: -251.693
Total T: 80400 Episode Num: 402 Episode T: 200 Reward: -345.699
Total T: 80600 Episode Num: 403 Episode T: 200 Reward: -121.244
Total T: 80800 Episode Num: 404 Episode T: 200 Reward: -119.521
Total T: 81000 Episode Num: 405 Episode T: 200 Reward: -3.213
recent Evaluation: -147.57352047840428
Total T: 81200 Episode Num: 406 Episode T: 200 Reward: -4.501
Total T: 81400 Episode Num: 407 Episode T: 200 Reward: -4.444
```

```
Total T: 81600 Episode Num: 408 Episode T: 200 Reward: -371.259
Total T: 81800 Episode Num: 409 Episode T: 200 Reward: -362.641
Total T: 82000 Episode Num: 410 Episode T: 200 Reward: -242.871
recent Evaluation: -150.77355652910634
Total T: 82200 Episode Num: 411 Episode T: 200 Reward: -128.437
Total T: 82400 Episode Num: 412 Episode T: 200 Reward: -250.428
Total T: 82600 Episode Num: 413 Episode T: 200 Reward: -125.826
Total T: 82800 Episode Num: 414 Episode T: 200 Reward: -1.908
Total T: 83000 Episode Num: 415 Episode T: 200 Reward: -233.073
recent Evaluation: -174.99088798624535
Total T: 83200 Episode Num: 416 Episode T: 200 Reward: -123.452
Total T: 83400 Episode Num: 417 Episode T: 200 Reward: -126.710
Total T: 83600 Episode Num: 418 Episode T: 200 Reward: -246.277
Total T: 83800 Episode Num: 419 Episode T: 200 Reward: -2.015
Total T: 84000 Episode Num: 420 Episode T: 200 Reward: -118.469
recent Evaluation: -145.80784926181695
Total T: 84200 Episode Num: 421 Episode T: 200 Reward: -231.666
Total T: 84400 Episode Num: 422 Episode T: 200 Reward: -368.430
Total T: 84600 Episode Num: 423 Episode T: 200 Reward: -240.641
Total T: 84800 Episode Num: 424 Episode T: 200 Reward: -364.111
Total T: 85000 Episode Num: 425 Episode T: 200 Reward: -133.497
recent Evaluation: -154.47999530540878
Total T: 85200 Episode Num: 426 Episode T: 200 Reward: -257.272
Total T: 85400 Episode Num: 427 Episode T: 200 Reward: -122.390
Total T: 85600 Episode Num: 428 Episode T: 200 Reward: -3.492
Total T: 85800 Episode Num: 429 Episode T: 200 Reward: -123.724
Total T: 86000 Episode Num: 430 Episode T: 200 Reward: -243.292
recent Evaluation: -138.66162599400133
Total T: 86200 Episode Num: 431 Episode T: 200 Reward: -133.126
Total T: 86400 Episode Num: 432 Episode T: 200 Reward: -251.877
Total T: 86600 Episode Num: 433 Episode T: 200 Reward: -116.724
Total T: 86800 Episode Num: 434 Episode T: 200 Reward: -128.666
Total T: 87000 Episode Num: 435 Episode T: 200 Reward: -368.187
recent Evaluation: -133.2005644093549
Total T: 87200 Episode Num: 436 Episode T: 200 Reward: -244.549
Total T: 87400 Episode Num: 437 Episode T: 200 Reward: -252.119
Total T: 87600 Episode Num: 438 Episode T: 200 Reward: -3.165
Total T: 87800 Episode Num: 439 Episode T: 200 Reward: -122.892
Total T: 88000 Episode Num: 440 Episode T: 200 Reward: -373.227
recent Evaluation: -187.6686184815257
Total T: 88200 Episode Num: 441 Episode T: 200 Reward: -234.220
Total T: 88400 Episode Num: 442 Episode T: 200 Reward: -251.749
Total T: 88600 Episode Num: 443 Episode T: 200 Reward: -237.609
Total T: 88800 Episode Num: 444 Episode T: 200 Reward: -124.847
Total T: 89000 Episode Num: 445 Episode T: 200 Reward: -250.467
recent Evaluation: -115.57907447366513
Total T: 89200 Episode Num: 446 Episode T: 200 Reward: -244.817
Total T: 89400 Episode Num: 447 Episode T: 200 Reward: -133.819
Total T: 89600 Episode Num: 448 Episode T: 200 Reward: -243.386
Total T: 89800 Episode Num: 449 Episode T: 200 Reward: -227.859
Total T: 90000 Episode Num: 450 Episode T: 200 Reward: -491.463
recent Evaluation: -101.43491774230519
Total T: 90200 Episode Num: 451 Episode T: 200 Reward: -121.137
Total T: 90400 Episode Num: 452 Episode T: 200 Reward: -2.931
Total T: 90600 Episode Num: 453 Episode T: 200 Reward: -235.202
Total T: 90800 Episode Num: 454 Episode T: 200 Reward: -122.206
Total T: 91000 Episode Num: 455 Episode T: 200 Reward: -3.478
recent Evaluation: -179.68051645107766
Total T: 91200 Episode Num: 456 Episode T: 200 Reward: -235.642
Total T: 91400 Episode Num: 457 Episode T: 200 Reward: -464.724
Total T: 91600 Episode Num: 458 Episode T: 200 Reward: -354.277
```

```
Total T: 91800 Episode Num: 459 Episode T: 200 Reward: -132.823
Total T: 92000 Episode Num: 460 Episode T: 200 Reward: -125.733
recent Evaluation: -151.68596028309986
Total T: 92200 Episode Num: 461 Episode T: 200 Reward: -364.730
Total T: 92400 Episode Num: 462 Episode T: 200 Reward: -243.386
Total T: 92600 Episode Num: 463 Episode T: 200 Reward: -268.744
Total T: 92800 Episode Num: 464 Episode T: 200 Reward: -229.170
Total T: 93000 Episode Num: 465 Episode T: 200 Reward: -3.490
recent Evaluation: -110.14629456330519
Total T: 93200 Episode Num: 466 Episode T: 200 Reward: -359.088
Total T: 93400 Episode Num: 467 Episode T: 200 Reward: -118.127
Total T: 93600 Episode Num: 468 Episode T: 200 Reward: -249.673
Total T: 93800 Episode Num: 469 Episode T: 200 Reward: -2.312
Total T: 94000 Episode Num: 470 Episode T: 200 Reward: -360.971
recent Evaluation: -147.45591042977395
Total T: 94200 Episode Num: 471 Episode T: 200 Reward: -231.792
Total T: 94400 Episode Num: 472 Episode T: 200 Reward: -232.854
Total T: 94600 Episode Num: 473 Episode T: 200 Reward: -128.043
Total T: 94800 Episode Num: 474 Episode T: 200 Reward: -122.837
Total T: 95000 Episode Num: 475 Episode T: 200 Reward: -3.249
recent Evaluation: -100.7290875225244
Total T: 95200 Episode Num: 476 Episode T: 200 Reward: -3.092
Total T: 95400 Episode Num: 477 Episode T: 200 Reward: -320.857
Total T: 95600 Episode Num: 478 Episode T: 200 Reward: -239.707
Total T: 95800 Episode Num: 479 Episode T: 200 Reward: -123.822
Total T: 96000 Episode Num: 480 Episode T: 200 Reward: -117.995
recent Evaluation: -101.31036025855965
Total T: 96200 Episode Num: 481 Episode T: 200 Reward: -126.084
Total T: 96400 Episode Num: 482 Episode T: 200 Reward: -229.523
Total T: 96600 Episode Num: 483 Episode T: 200 Reward: -2.358
Total T: 96800 Episode Num: 484 Episode T: 200 Reward: -245.686
Total T: 97000 Episode Num: 485 Episode T: 200 Reward: -120.552
recent Evaluation: -135.0514678587004
Total T: 97200 Episode Num: 486 Episode T: 200 Reward: -242.593
Total T: 97400 Episode Num: 487 Episode T: 200 Reward: -118.544
Total T: 97600 Episode Num: 488 Episode T: 200 Reward: -245.656
Total T: 97800 Episode Num: 489 Episode T: 200 Reward: -120.172
Total T: 98000 Episode Num: 490 Episode T: 200 Reward: -116.501
recent Evaluation: -146.75852587152733
Total T: 98200 Episode Num: 491 Episode T: 200 Reward: -350.271
Total T: 98400 Episode Num: 492 Episode T: 200 Reward: -361.240
Total T: 98600 Episode Num: 493 Episode T: 200 Reward: -128.999
Total T: 98800 Episode Num: 494 Episode T: 200 Reward: -129.969
Total T: 99000 Episode Num: 495 Episode T: 200 Reward: -235.959
recent Evaluation: -145.9810141507658
Total T: 99200 Episode Num: 496 Episode T: 200 Reward: -365.984
Total T: 99400 Episode Num: 497 Episode T: 200 Reward: -123.502
Total T: 99600 Episode Num: 498 Episode T: 200 Reward: -128.639
Total T: 99800 Episode Num: 499 Episode T: 200 Reward: -1.718
Total T: 100000 Episode Num: 500 Episode T: 200 Reward: -116.618
recent Evaluation: -186.7883984198052
```

In [19]:

```
# The following scripts run the TD3 algorithm.

alias = 'td3'


import matplotlib.pyplot as plt
import numpy as np
import torch
import gym
import argparse
import os
import torch.nn.functional as F
import utils
import TD3
import DDPG

def eval_policy(policy, eval_episodes=10):
    eval_env = gym.make(ENV_NAME)

    avg_reward = 0.
    for _ in range(eval_episodes):
        state, done = eval_env.reset(), False
        while not done:
            action = policy.select_action(np.array(state))
            state, reward, done, _ = eval_env.step(action)
            avg_reward += reward

    avg_reward /= eval_episodes
    #print("---------------------------------------")
    #print(f"Evaluation over {eval_episodes} episodes: {avg_reward:.3f}")
    #print("---------------------------------------")
    return avg_reward

env = gym.make(ENV_NAME)
torch.manual_seed(0)
np.random.seed(0)

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

args_policy_noise = 0.2
args_noise_clip = 0.5
args_policy_freq = 2
args_max_timesteps = 100000
args_expl_noise = 0.1
args_batch_size = 25
args_eval_freq = 1000
args_start_timesteps = 0

kwargs = {
    "state_dim": state_dim,
    "action_dim": action_dim,
    "max_action": max_action,
    "discount": 0.99,
    "tau": 0.005
}
```

```
args_policy = 'TD3'

if args_policy == "TD3":
    # Target policy smoothing is scaled wrt the action scale
    kwargs["policy_noise"] = args_policy_noise * max_action
    kwargs["noise_clip"] = args_noise_clip * max_action
    kwargs["policy_freq"] = args_policy_freq
    policy = TD3.TD3(**kwargs)
elif args_policy == "OurDDPG":
    policy = OurDDPG.DDPG(**kwargs)
elif args_policy == "DDPG":
    policy = DDPG.DDPG(**kwargs)
replay_buffer = utils.ReplayBuffer(state_dim, action_dim)

# Evaluate untrained policy
evaluations = [eval_policy(policy)]

state, done = env.reset(), False
episode_reward = 0
episode_timesteps = 0
episode_num = 0
counter = 0
msk_list = []
temp_curve = [eval_policy(policy)]
temp_val = []
for t in range(int(args_max_timesteps)):
    episode_timesteps += 1
    counter += 1
    # Select action randomly or according to policy
    if t < args_start_timesteps:
        action = np.random.uniform(-max_action,max_action,action_dim)
    else:
        if np.random.uniform(0,1) < 0.1:
            action = np.random.uniform(-max_action,max_action,action_dim)
        else:
            action = (
                policy.select_action(np.array(state))
                + np.random.normal(0, max_action * args_expl_noise, size=action_dim)
            ).clip(-max_action, max_action)

    # Perform action
    next_state, reward, done, _ = env.step(action)
    done_bool = float(done) if episode_timesteps < env._max_episode_steps else 0

    replay_buffer.add(state, action, next_state, reward, done_bool)

    state = next_state
    episode_reward += reward

    if t >= args_start_timesteps:
        '''TD3'''
        last_val = 999.
        patient = 5
        for i in range(1):
            policy.train(replay_buffer, args_batch_size)


    # Train agent after collecting sufficient data
    if done:
        print(f"Total T: {t+1} Episode Num: {episode_num+1} Episode T: {episode_timesteps} Re
ward: {episode_reward:.3f}")
```

```
        msk_list = []
        state, done = env.reset(), False
        episode_reward = 0
        episode_timesteps = 0
        episode_num += 1

    # Evaluate episode
    if (t + 1) % args_eval_freq == 0:
        evaluations.append(eval_policy(policy))
        print('recent Evaluation:',evaluations[-1])
        np.save('results/evaluations_alias{}_ENV{}'.format(alias,ENV_NAME),evaluations)
```

```
Total T: 200 Episode Num: 1 Episode T: 200 Reward: -1258.078
Total T: 400 Episode Num: 2 Episode T: 200 Reward: -1666.855
Total T: 600 Episode Num: 3 Episode T: 200 Reward: -1706.568
Total T: 800 Episode Num: 4 Episode T: 200 Reward: -1743.297
Total T: 1000 Episode Num: 5 Episode T: 200 Reward: -1702.652
recent Evaluation: -1510.0795264051367
Total T: 1200 Episode Num: 6 Episode T: 200 Reward: -1708.091
Total T: 1400 Episode Num: 7 Episode T: 200 Reward: -1540.797
Total T: 1600 Episode Num: 8 Episode T: 200 Reward: -1596.404
Total T: 1800 Episode Num: 9 Episode T: 200 Reward: -1601.232
Total T: 2000 Episode Num: 10 Episode T: 200 Reward: -1571.199
recent Evaluation: -1560.8328332634014
Total T: 2200 Episode Num: 11 Episode T: 200 Reward: -1658.157
Total T: 2400 Episode Num: 12 Episode T: 200 Reward: -1514.381
Total T: 2600 Episode Num: 13 Episode T: 200 Reward: -1448.979
Total T: 2800 Episode Num: 14 Episode T: 200 Reward: -1430.510
Total T: 3000 Episode Num: 15 Episode T: 200 Reward: -1575.445
recent Evaluation: -1447.2521821861687
Total T: 3200 Episode Num: 16 Episode T: 200 Reward: -1167.695
Total T: 3400 Episode Num: 17 Episode T: 200 Reward: -1308.763
Total T: 3600 Episode Num: 18 Episode T: 200 Reward: -1407.998
Total T: 3800 Episode Num: 19 Episode T: 200 Reward: -1277.058
Total T: 4000 Episode Num: 20 Episode T: 200 Reward: -1268.961
recent Evaluation: -1356.130884241889
Total T: 4200 Episode Num: 21 Episode T: 200 Reward: -1395.528
Total T: 4400 Episode Num: 22 Episode T: 200 Reward: -1043.777
Total T: 4600 Episode Num: 23 Episode T: 200 Reward: -885.901
Total T: 4800 Episode Num: 24 Episode T: 200 Reward: -1226.807
Total T: 5000 Episode Num: 25 Episode T: 200 Reward: -920.007
recent Evaluation: -1182.3816913198493
Total T: 5200 Episode Num: 26 Episode T: 200 Reward: -1021.339
Total T: 5400 Episode Num: 27 Episode T: 200 Reward: -1052.884
Total T: 5600 Episode Num: 28 Episode T: 200 Reward: -871.984
Total T: 5800 Episode Num: 29 Episode T: 200 Reward: -1093.239
Total T: 6000 Episode Num: 30 Episode T: 200 Reward: -1207.517
recent Evaluation: -862.6562552077942
Total T: 6200 Episode Num: 31 Episode T: 200 Reward: -882.075
Total T: 6400 Episode Num: 32 Episode T: 200 Reward: -1003.490
Total T: 6600 Episode Num: 33 Episode T: 200 Reward: -877.956
Total T: 6800 Episode Num: 34 Episode T: 200 Reward: -883.286
Total T: 7000 Episode Num: 35 Episode T: 200 Reward: -1031.384
recent Evaluation: -839.5246017674284
Total T: 7200 Episode Num: 36 Episode T: 200 Reward: -771.413
Total T: 7400 Episode Num: 37 Episode T: 200 Reward: -799.409
Total T: 7600 Episode Num: 38 Episode T: 200 Reward: -860.044
Total T: 7800 Episode Num: 39 Episode T: 200 Reward: -643.163
Total T: 8000 Episode Num: 40 Episode T: 200 Reward: -759.550
recent Evaluation: -565.3844291474832
Total T: 8200 Episode Num: 41 Episode T: 200 Reward: -751.449
Total T: 8400 Episode Num: 42 Episode T: 200 Reward: -517.579
Total T: 8600 Episode Num: 43 Episode T: 200 Reward: -635.052
Total T: 8800 Episode Num: 44 Episode T: 200 Reward: -637.300
Total T: 9000 Episode Num: 45 Episode T: 200 Reward: -637.188
recent Evaluation: -562.9453309885416
Total T: 9200 Episode Num: 46 Episode T: 200 Reward: -496.115
Total T: 9400 Episode Num: 47 Episode T: 200 Reward: -264.840
Total T: 9600 Episode Num: 48 Episode T: 200 Reward: -257.647
Total T: 9800 Episode Num: 49 Episode T: 200 Reward: -129.660
Total T: 10000 Episode Num: 50 Episode T: 200 Reward: -125.916
recent Evaluation: -228.41285485569716
Total T: 10200 Episode Num: 51 Episode T: 200 Reward: -389.401
```

```
Total T: 10400 Episode Num: 52 Episode T: 200 Reward: -120.228
Total T: 10600 Episode Num: 53 Episode T: 200 Reward: -126.277
Total T: 10800 Episode Num: 54 Episode T: 200 Reward: -127.021
Total T: 11000 Episode Num: 55 Episode T: 200 Reward: -127.579
recent Evaluation: -213.66972506805755
Total T: 11200 Episode Num: 56 Episode T: 200 Reward: -315.076
Total T: 11400 Episode Num: 57 Episode T: 200 Reward: -244.320
Total T: 11600 Episode Num: 58 Episode T: 200 Reward: -126.947
Total T: 11800 Episode Num: 59 Episode T: 200 Reward: -240.018
Total T: 12000 Episode Num: 60 Episode T: 200 Reward: -116.833
recent Evaluation: -160.2151654800682
Total T: 12200 Episode Num: 61 Episode T: 200 Reward: -128.752
Total T: 12400 Episode Num: 62 Episode T: 200 Reward: -128.656
Total T: 12600 Episode Num: 63 Episode T: 200 Reward: -2.701
Total T: 12800 Episode Num: 64 Episode T: 200 Reward: -122.780
Total T: 13000 Episode Num: 65 Episode T: 200 Reward: -240.682
recent Evaluation: -149.20420641586358
Total T: 13200 Episode Num: 66 Episode T: 200 Reward: -240.976
Total T: 13400 Episode Num: 67 Episode T: 200 Reward: -330.661
Total T: 13600 Episode Num: 68 Episode T: 200 Reward: -122.715
Total T: 13800 Episode Num: 69 Episode T: 200 Reward: -116.231
Total T: 14000 Episode Num: 70 Episode T: 200 Reward: -248.123
recent Evaluation: -148.44977357333505
Total T: 14200 Episode Num: 71 Episode T: 200 Reward: -125.965
Total T: 14400 Episode Num: 72 Episode T: 200 Reward: -355.370
Total T: 14600 Episode Num: 73 Episode T: 200 Reward: -250.399
Total T: 14800 Episode Num: 74 Episode T: 200 Reward: -119.406
Total T: 15000 Episode Num: 75 Episode T: 200 Reward: -362.762
recent Evaluation: -167.76233675171244
Total T: 15200 Episode Num: 76 Episode T: 200 Reward: -3.826
Total T: 15400 Episode Num: 77 Episode T: 200 Reward: -123.531
Total T: 15600 Episode Num: 78 Episode T: 200 Reward: -346.246
Total T: 15800 Episode Num: 79 Episode T: 200 Reward: -114.982
Total T: 16000 Episode Num: 80 Episode T: 200 Reward: -116.230
recent Evaluation: -155.90691787976215
Total T: 16200 Episode Num: 81 Episode T: 200 Reward: -124.485
Total T: 16400 Episode Num: 82 Episode T: 200 Reward: -114.858
Total T: 16600 Episode Num: 83 Episode T: 200 Reward: -123.498
Total T: 16800 Episode Num: 84 Episode T: 200 Reward: -0.955
Total T: 17000 Episode Num: 85 Episode T: 200 Reward: -0.619
recent Evaluation: -174.3222954197024
Total T: 17200 Episode Num: 86 Episode T: 200 Reward: -245.157
Total T: 17400 Episode Num: 87 Episode T: 200 Reward: -245.614
Total T: 17600 Episode Num: 88 Episode T: 200 Reward: -126.705
Total T: 17800 Episode Num: 89 Episode T: 200 Reward: -123.135
Total T: 18000 Episode Num: 90 Episode T: 200 Reward: -127.763
recent Evaluation: -157.50202917148255
Total T: 18200 Episode Num: 91 Episode T: 200 Reward: -282.134
Total T: 18400 Episode Num: 92 Episode T: 200 Reward: -242.920
Total T: 18600 Episode Num: 93 Episode T: 200 Reward: -231.453
Total T: 18800 Episode Num: 94 Episode T: 200 Reward: -123.104
Total T: 19000 Episode Num: 95 Episode T: 200 Reward: -117.079
recent Evaluation: -145.06484118907173
Total T: 19200 Episode Num: 96 Episode T: 200 Reward: -120.627
Total T: 19400 Episode Num: 97 Episode T: 200 Reward: -114.761
Total T: 19600 Episode Num: 98 Episode T: 200 Reward: -124.510
Total T: 19800 Episode Num: 99 Episode T: 200 Reward: -116.408
Total T: 20000 Episode Num: 100 Episode T: 200 Reward: -123.214
recent Evaluation: -174.2771841888361
Total T: 20200 Episode Num: 101 Episode T: 200 Reward: -117.264
Total T: 20400 Episode Num: 102 Episode T: 200 Reward: -246.288
```

```
Total T: 20600 Episode Num: 103 Episode T: 200 Reward: -124.360
Total T: 20800 Episode Num: 104 Episode T: 200 Reward: -116.633
Total T: 21000 Episode Num: 105 Episode T: 200 Reward: -120.901
recent Evaluation: -165.09774222414543
Total T: 21200 Episode Num: 106 Episode T: 200 Reward: -1.755
Total T: 21400 Episode Num: 107 Episode T: 200 Reward: -234.214
Total T: 21600 Episode Num: 108 Episode T: 200 Reward: -120.036
Total T: 21800 Episode Num: 109 Episode T: 200 Reward: -240.455
Total T: 22000 Episode Num: 110 Episode T: 200 Reward: -119.211
recent Evaluation: -188.69160704640655
Total T: 22200 Episode Num: 111 Episode T: 200 Reward: -237.429
Total T: 22400 Episode Num: 112 Episode T: 200 Reward: -127.858
Total T: 22600 Episode Num: 113 Episode T: 200 Reward: -122.824
Total T: 22800 Episode Num: 114 Episode T: 200 Reward: -3.446
Total T: 23000 Episode Num: 115 Episode T: 200 Reward: -315.261
recent Evaluation: -130.94529990597442
Total T: 23200 Episode Num: 116 Episode T: 200 Reward: -119.229
Total T: 23400 Episode Num: 117 Episode T: 200 Reward: -366.063
Total T: 23600 Episode Num: 118 Episode T: 200 Reward: -0.442
Total T: 23800 Episode Num: 119 Episode T: 200 Reward: -119.445
Total T: 24000 Episode Num: 120 Episode T: 200 Reward: -235.875
recent Evaluation: -119.51683078510173
Total T: 24200 Episode Num: 121 Episode T: 200 Reward: -511.085
Total T: 24400 Episode Num: 122 Episode T: 200 Reward: -120.573
Total T: 24600 Episode Num: 123 Episode T: 200 Reward: -125.258
Total T: 24800 Episode Num: 124 Episode T: 200 Reward: -126.551
Total T: 25000 Episode Num: 125 Episode T: 200 Reward: -120.972
recent Evaluation: -140.85774224601917
Total T: 25200 Episode Num: 126 Episode T: 200 Reward: -0.647
Total T: 25400 Episode Num: 127 Episode T: 200 Reward: -118.275
Total T: 25600 Episode Num: 128 Episode T: 200 Reward: -116.625
Total T: 25800 Episode Num: 129 Episode T: 200 Reward: -244.057
Total T: 26000 Episode Num: 130 Episode T: 200 Reward: -244.869
recent Evaluation: -141.99194603699544
Total T: 26200 Episode Num: 131 Episode T: 200 Reward: -117.254
Total T: 26400 Episode Num: 132 Episode T: 200 Reward: -125.453
Total T: 26600 Episode Num: 133 Episode T: 200 Reward: -122.648
Total T: 26800 Episode Num: 134 Episode T: 200 Reward: -120.364
Total T: 27000 Episode Num: 135 Episode T: 200 Reward: -124.144
recent Evaluation: -141.56259366637715
Total T: 27200 Episode Num: 136 Episode T: 200 Reward: -114.906
Total T: 27400 Episode Num: 137 Episode T: 200 Reward: -123.911
Total T: 27600 Episode Num: 138 Episode T: 200 Reward: -117.308
Total T: 27800 Episode Num: 139 Episode T: 200 Reward: -247.600
Total T: 28000 Episode Num: 140 Episode T: 200 Reward: -123.454
recent Evaluation: -139.92313473809014
Total T: 28200 Episode Num: 141 Episode T: 200 Reward: -115.216
Total T: 28400 Episode Num: 142 Episode T: 200 Reward: -227.312
Total T: 28600 Episode Num: 143 Episode T: 200 Reward: -1.110
Total T: 28800 Episode Num: 144 Episode T: 200 Reward: -345.542
Total T: 29000 Episode Num: 145 Episode T: 200 Reward: -119.008
recent Evaluation: -155.33206825007193
Total T: 29200 Episode Num: 146 Episode T: 200 Reward: -483.260
Total T: 29400 Episode Num: 147 Episode T: 200 Reward: -121.101
Total T: 29600 Episode Num: 148 Episode T: 200 Reward: -117.841
Total T: 29800 Episode Num: 149 Episode T: 200 Reward: -122.041
Total T: 30000 Episode Num: 150 Episode T: 200 Reward: -124.546
recent Evaluation: -155.25337078406045
Total T: 30200 Episode Num: 151 Episode T: 200 Reward: -239.486
Total T: 30400 Episode Num: 152 Episode T: 200 Reward: -0.988
Total T: 30600 Episode Num: 153 Episode T: 200 Reward: -123.134
```

```
Total T: 30800 Episode Num: 154 Episode T: 200 Reward: -362.701
Total T: 31000 Episode Num: 155 Episode T: 200 Reward: -1.368
recent Evaluation: -118.502209145689
Total T: 31200 Episode Num: 156 Episode T: 200 Reward: -243.405
Total T: 31400 Episode Num: 157 Episode T: 200 Reward: -358.641
Total T: 31600 Episode Num: 158 Episode T: 200 Reward: -123.767
Total T: 31800 Episode Num: 159 Episode T: 200 Reward: -357.279
Total T: 32000 Episode Num: 160 Episode T: 200 Reward: -346.154
recent Evaluation: -130.30462808263945
Total T: 32200 Episode Num: 161 Episode T: 200 Reward: -124.204
Total T: 32400 Episode Num: 162 Episode T: 200 Reward: -117.024
Total T: 32600 Episode Num: 163 Episode T: 200 Reward: -118.627
Total T: 32800 Episode Num: 164 Episode T: 200 Reward: -122.560
Total T: 33000 Episode Num: 165 Episode T: 200 Reward: -1.608
recent Evaluation: -144.5007847483684
Total T: 33200 Episode Num: 166 Episode T: 200 Reward: -126.294
Total T: 33400 Episode Num: 167 Episode T: 200 Reward: -248.524
Total T: 33600 Episode Num: 168 Episode T: 200 Reward: -125.716
Total T: 33800 Episode Num: 169 Episode T: 200 Reward: -232.871
Total T: 34000 Episode Num: 170 Episode T: 200 Reward: -121.208
recent Evaluation: -164.79796559346195
Total T: 34200 Episode Num: 171 Episode T: 200 Reward: -124.156
Total T: 34400 Episode Num: 172 Episode T: 200 Reward: -122.794
Total T: 34600 Episode Num: 173 Episode T: 200 Reward: -237.066
Total T: 34800 Episode Num: 174 Episode T: 200 Reward: -237.257
Total T: 35000 Episode Num: 175 Episode T: 200 Reward: -118.843
recent Evaluation: -141.1217127546846
Total T: 35200 Episode Num: 176 Episode T: 200 Reward: -115.742
Total T: 35400 Episode Num: 177 Episode T: 200 Reward: -238.816
Total T: 35600 Episode Num: 178 Episode T: 200 Reward: -237.990
Total T: 35800 Episode Num: 179 Episode T: 200 Reward: -1.362
Total T: 36000 Episode Num: 180 Episode T: 200 Reward: -127.863
recent Evaluation: -156.68104750140554
Total T: 36200 Episode Num: 181 Episode T: 200 Reward: -120.466
Total T: 36400 Episode Num: 182 Episode T: 200 Reward: -1.420
Total T: 36600 Episode Num: 183 Episode T: 200 Reward: -124.532
Total T: 36800 Episode Num: 184 Episode T: 200 Reward: -117.205
Total T: 37000 Episode Num: 185 Episode T: 200 Reward: -234.087
recent Evaluation: -162.36902832678354
Total T: 37200 Episode Num: 186 Episode T: 200 Reward: -115.942
Total T: 37400 Episode Num: 187 Episode T: 200 Reward: -254.695
Total T: 37600 Episode Num: 188 Episode T: 200 Reward: -364.600
Total T: 37800 Episode Num: 189 Episode T: 200 Reward: -122.274
Total T: 38000 Episode Num: 190 Episode T: 200 Reward: -124.084
recent Evaluation: -155.86657151499088
Total T: 38200 Episode Num: 191 Episode T: 200 Reward: -239.412
Total T: 38400 Episode Num: 192 Episode T: 200 Reward: -121.615
Total T: 38600 Episode Num: 193 Episode T: 200 Reward: -230.103
Total T: 38800 Episode Num: 194 Episode T: 200 Reward: -357.312
Total T: 39000 Episode Num: 195 Episode T: 200 Reward: -239.602
recent Evaluation: -176.382730467056
Total T: 39200 Episode Num: 196 Episode T: 200 Reward: -0.688
Total T: 39400 Episode Num: 197 Episode T: 200 Reward: -122.151
Total T: 39600 Episode Num: 198 Episode T: 200 Reward: -125.426
Total T: 39800 Episode Num: 199 Episode T: 200 Reward: -119.411
Total T: 40000 Episode Num: 200 Episode T: 200 Reward: -234.406
recent Evaluation: -143.2322498222447
Total T: 40200 Episode Num: 201 Episode T: 200 Reward: -2.965
Total T: 40400 Episode Num: 202 Episode T: 200 Reward: -119.954
Total T: 40600 Episode Num: 203 Episode T: 200 Reward: -121.605
Total T: 40800 Episode Num: 204 Episode T: 200 Reward: -0.935
```

```
Total T: 41000 Episode Num: 205 Episode T: 200 Reward: -355.667
recent Evaluation: -140.78896802209582
Total T: 41200 Episode Num: 206 Episode T: 200 Reward: -317.202
Total T: 41400 Episode Num: 207 Episode T: 200 Reward: -123.062
Total T: 41600 Episode Num: 208 Episode T: 200 Reward: -243.447
Total T: 41800 Episode Num: 209 Episode T: 200 Reward: -121.423
Total T: 42000 Episode Num: 210 Episode T: 200 Reward: -122.485
recent Evaluation: -154.8226425633274
Total T: 42200 Episode Num: 211 Episode T: 200 Reward: -0.541
Total T: 42400 Episode Num: 212 Episode T: 200 Reward: -123.013
Total T: 42600 Episode Num: 213 Episode T: 200 Reward: -123.731
Total T: 42800 Episode Num: 214 Episode T: 200 Reward: -124.964
Total T: 43000 Episode Num: 215 Episode T: 200 Reward: -338.364
recent Evaluation: -116.66788581394913
Total T: 43200 Episode Num: 216 Episode T: 200 Reward: -118.733
Total T: 43400 Episode Num: 217 Episode T: 200 Reward: -120.329
Total T: 43600 Episode Num: 218 Episode T: 200 Reward: -123.703
Total T: 43800 Episode Num: 219 Episode T: 200 Reward: -232.967
Total T: 44000 Episode Num: 220 Episode T: 200 Reward: -1.978
recent Evaluation: -176.817775222965
Total T: 44200 Episode Num: 221 Episode T: 200 Reward: -241.322
Total T: 44400 Episode Num: 222 Episode T: 200 Reward: -0.887
Total T: 44600 Episode Num: 223 Episode T: 200 Reward: -230.920
Total T: 44800 Episode Num: 224 Episode T: 200 Reward: -115.021
Total T: 45000 Episode Num: 225 Episode T: 200 Reward: -242.997
recent Evaluation: -104.69454693203139
Total T: 45200 Episode Num: 226 Episode T: 200 Reward: -237.170
Total T: 45400 Episode Num: 227 Episode T: 200 Reward: -238.534
Total T: 45600 Episode Num: 228 Episode T: 200 Reward: -120.104
Total T: 45800 Episode Num: 229 Episode T: 200 Reward: -119.918
Total T: 46000 Episode Num: 230 Episode T: 200 Reward: -239.756
recent Evaluation: -183.1742808370805
Total T: 46200 Episode Num: 231 Episode T: 200 Reward: -121.327
Total T: 46400 Episode Num: 232 Episode T: 200 Reward: -119.822
Total T: 46600 Episode Num: 233 Episode T: 200 Reward: -229.544
Total T: 46800 Episode Num: 234 Episode T: 200 Reward: -118.295
Total T: 47000 Episode Num: 235 Episode T: 200 Reward: -235.554
recent Evaluation: -83.50890215638378
Total T: 47200 Episode Num: 236 Episode T: 200 Reward: -123.673
Total T: 47400 Episode Num: 237 Episode T: 200 Reward: -231.883
Total T: 47600 Episode Num: 238 Episode T: 200 Reward: -116.252
Total T: 47800 Episode Num: 239 Episode T: 200 Reward: -230.091
Total T: 48000 Episode Num: 240 Episode T: 200 Reward: -360.138
recent Evaluation: -154.0485401060037
Total T: 48200 Episode Num: 241 Episode T: 200 Reward: -121.034
Total T: 48400 Episode Num: 242 Episode T: 200 Reward: -332.553
Total T: 48600 Episode Num: 243 Episode T: 200 Reward: -230.920
Total T: 48800 Episode Num: 244 Episode T: 200 Reward: -123.838
Total T: 49000 Episode Num: 245 Episode T: 200 Reward: -123.130
recent Evaluation: -120.6687768377197
Total T: 49200 Episode Num: 246 Episode T: 200 Reward: -119.381
Total T: 49400 Episode Num: 247 Episode T: 200 Reward: -233.573
Total T: 49600 Episode Num: 248 Episode T: 200 Reward: -236.661
Total T: 49800 Episode Num: 249 Episode T: 200 Reward: -120.259
Total T: 50000 Episode Num: 250 Episode T: 200 Reward: -357.776
recent Evaluation: -118.06793834114085
Total T: 50200 Episode Num: 251 Episode T: 200 Reward: -239.729
Total T: 50400 Episode Num: 252 Episode T: 200 Reward: -229.366
Total T: 50600 Episode Num: 253 Episode T: 200 Reward: -3.390
Total T: 50800 Episode Num: 254 Episode T: 200 Reward: -239.731
Total T: 51000 Episode Num: 255 Episode T: 200 Reward: -115.338
```

```
recent Evaluation: -179.4772885574052
Total T: 51200 Episode Num: 256 Episode T: 200 Reward: -343.458
Total T: 51400 Episode Num: 257 Episode T: 200 Reward: -116.549
Total T: 51600 Episode Num: 258 Episode T: 200 Reward: -123.443
Total T: 51800 Episode Num: 259 Episode T: 200 Reward: -120.299
Total T: 52000 Episode Num: 260 Episode T: 200 Reward: -121.053
recent Evaluation: -153.15976079606145
Total T: 52200 Episode Num: 261 Episode T: 200 Reward: -239.842
Total T: 52400 Episode Num: 262 Episode T: 200 Reward: -117.914
Total T: 52600 Episode Num: 263 Episode T: 200 Reward: -244.587
Total T: 52800 Episode Num: 264 Episode T: 200 Reward: -334.682
Total T: 53000 Episode Num: 265 Episode T: 200 Reward: -114.057
recent Evaluation: -119.53982616276521
Total T: 53200 Episode Num: 266 Episode T: 200 Reward: -117.229
Total T: 53400 Episode Num: 267 Episode T: 200 Reward: -118.542
Total T: 53600 Episode Num: 268 Episode T: 200 Reward: -353.153
Total T: 53800 Episode Num: 269 Episode T: 200 Reward: -313.550
Total T: 54000 Episode Num: 270 Episode T: 200 Reward: -115.765
recent Evaluation: -154.19814259321475
Total T: 54200 Episode Num: 271 Episode T: 200 Reward: -238.113
Total T: 54400 Episode Num: 272 Episode T: 200 Reward: -123.600
Total T: 54600 Episode Num: 273 Episode T: 200 Reward: -343.270
Total T: 54800 Episode Num: 274 Episode T: 200 Reward: -232.068
Total T: 55000 Episode Num: 275 Episode T: 200 Reward: -232.157
recent Evaluation: -106.49350096047681
Total T: 55200 Episode Num: 276 Episode T: 200 Reward: -238.627
Total T: 55400 Episode Num: 277 Episode T: 200 Reward: -115.815
Total T: 55600 Episode Num: 278 Episode T: 200 Reward: -115.084
Total T: 55800 Episode Num: 279 Episode T: 200 Reward: -244.562
Total T: 56000 Episode Num: 280 Episode T: 200 Reward: -0.868
recent Evaluation: -163.19831821806042
Total T: 56200 Episode Num: 281 Episode T: 200 Reward: -120.972
Total T: 56400 Episode Num: 282 Episode T: 200 Reward: -120.457
Total T: 56600 Episode Num: 283 Episode T: 200 Reward: -243.780
Total T: 56800 Episode Num: 284 Episode T: 200 Reward: -124.038
Total T: 57000 Episode Num: 285 Episode T: 200 Reward: -122.175
recent Evaluation: -141.9874471199026
Total T: 57200 Episode Num: 286 Episode T: 200 Reward: -0.635
Total T: 57400 Episode Num: 287 Episode T: 200 Reward: -124.390
Total T: 57600 Episode Num: 288 Episode T: 200 Reward: -117.027
Total T: 57800 Episode Num: 289 Episode T: 200 Reward: -125.555
Total T: 58000 Episode Num: 290 Episode T: 200 Reward: -234.830
recent Evaluation: -143.89655595679835
Total T: 58200 Episode Num: 291 Episode T: 200 Reward: -229.623
Total T: 58400 Episode Num: 292 Episode T: 200 Reward: -250.918
Total T: 58600 Episode Num: 293 Episode T: 200 Reward: -0.328
Total T: 58800 Episode Num: 294 Episode T: 200 Reward: -3.546
Total T: 59000 Episode Num: 295 Episode T: 200 Reward: -116.697
recent Evaluation: -157.99117168127833
Total T: 59200 Episode Num: 296 Episode T: 200 Reward: -119.732
Total T: 59400 Episode Num: 297 Episode T: 200 Reward: -120.365
Total T: 59600 Episode Num: 298 Episode T: 200 Reward: -124.403
Total T: 59800 Episode Num: 299 Episode T: 200 Reward: -280.430
Total T: 60000 Episode Num: 300 Episode T: 200 Reward: -348.101
recent Evaluation: -131.35558850946703
Total T: 60200 Episode Num: 301 Episode T: 200 Reward: -118.447
Total T: 60400 Episode Num: 302 Episode T: 200 Reward: -117.572
Total T: 60600 Episode Num: 303 Episode T: 200 Reward: -0.835
Total T: 60800 Episode Num: 304 Episode T: 200 Reward: -123.695
Total T: 61000 Episode Num: 305 Episode T: 200 Reward: -246.845
recent Evaluation: -119.16812161071307
```

```
Total T: 61200 Episode Num: 306 Episode T: 200 Reward: -120.142
Total T: 61400 Episode Num: 307 Episode T: 200 Reward: -119.798
Total T: 61600 Episode Num: 308 Episode T: 200 Reward: -121.072
Total T: 61800 Episode Num: 309 Episode T: 200 Reward: -362.743
Total T: 62000 Episode Num: 310 Episode T: 200 Reward: -116.383
recent Evaluation: -157.12126381743644
Total T: 62200 Episode Num: 311 Episode T: 200 Reward: -351.949
Total T: 62400 Episode Num: 312 Episode T: 200 Reward: -239.835
Total T: 62600 Episode Num: 313 Episode T: 200 Reward: -118.526
Total T: 62800 Episode Num: 314 Episode T: 200 Reward: -118.750
Total T: 63000 Episode Num: 315 Episode T: 200 Reward: -122.559
recent Evaluation: -130.81004321472594
Total T: 63200 Episode Num: 316 Episode T: 200 Reward: -123.962
Total T: 63400 Episode Num: 317 Episode T: 200 Reward: -127.943
Total T: 63600 Episode Num: 318 Episode T: 200 Reward: -230.406
Total T: 63800 Episode Num: 319 Episode T: 200 Reward: -124.630
Total T: 64000 Episode Num: 320 Episode T: 200 Reward: -121.528
recent Evaluation: -119.15472271032468
Total T: 64200 Episode Num: 321 Episode T: 200 Reward: -239.636
Total T: 64400 Episode Num: 322 Episode T: 200 Reward: -236.425
Total T: 64600 Episode Num: 323 Episode T: 200 Reward: -124.657
Total T: 64800 Episode Num: 324 Episode T: 200 Reward: -240.044
Total T: 65000 Episode Num: 325 Episode T: 200 Reward: -116.287
recent Evaluation: -130.44993631589193
Total T: 65200 Episode Num: 326 Episode T: 200 Reward: -122.479
Total T: 65400 Episode Num: 327 Episode T: 200 Reward: -124.539
Total T: 65600 Episode Num: 328 Episode T: 200 Reward: -121.803
Total T: 65800 Episode Num: 329 Episode T: 200 Reward: -239.867
Total T: 66000 Episode Num: 330 Episode T: 200 Reward: -123.160
recent Evaluation: -119.42551126613486
Total T: 66200 Episode Num: 331 Episode T: 200 Reward: -119.193
Total T: 66400 Episode Num: 332 Episode T: 200 Reward: -241.576
Total T: 66600 Episode Num: 333 Episode T: 200 Reward: -228.551
Total T: 66800 Episode Num: 334 Episode T: 200 Reward: -122.221
Total T: 67000 Episode Num: 335 Episode T: 200 Reward: -123.504
recent Evaluation: -118.57835777487819
Total T: 67200 Episode Num: 336 Episode T: 200 Reward: -1.090
Total T: 67400 Episode Num: 337 Episode T: 200 Reward: -121.664
Total T: 67600 Episode Num: 338 Episode T: 200 Reward: -246.118
Total T: 67800 Episode Num: 339 Episode T: 200 Reward: -244.511
Total T: 68000 Episode Num: 340 Episode T: 200 Reward: -117.871
recent Evaluation: -117.80829799955877
Total T: 68200 Episode Num: 341 Episode T: 200 Reward: -125.173
Total T: 68400 Episode Num: 342 Episode T: 200 Reward: -246.886
Total T: 68600 Episode Num: 343 Episode T: 200 Reward: -124.055
Total T: 68800 Episode Num: 344 Episode T: 200 Reward: -363.576
Total T: 69000 Episode Num: 345 Episode T: 200 Reward: -121.703
recent Evaluation: -164.7042714278249
Total T: 69200 Episode Num: 346 Episode T: 200 Reward: -242.386
Total T: 69400 Episode Num: 347 Episode T: 200 Reward: -122.161
Total T: 69600 Episode Num: 348 Episode T: 200 Reward: -119.462
Total T: 69800 Episode Num: 349 Episode T: 200 Reward: -116.912
Total T: 70000 Episode Num: 350 Episode T: 200 Reward: -243.632
recent Evaluation: -94.78337140501192
Total T: 70200 Episode Num: 351 Episode T: 200 Reward: -237.947
Total T: 70400 Episode Num: 352 Episode T: 200 Reward: -119.764
Total T: 70600 Episode Num: 353 Episode T: 200 Reward: -126.459
Total T: 70800 Episode Num: 354 Episode T: 200 Reward: -373.974
Total T: 71000 Episode Num: 355 Episode T: 200 Reward: -229.830
recent Evaluation: -144.83828660032714
Total T: 71200 Episode Num: 356 Episode T: 200 Reward: -242.920
```

```
Total T: 71400 Episode Num: 357 Episode T: 200 Reward: -120.870
Total T: 71600 Episode Num: 358 Episode T: 200 Reward: -122.738
Total T: 71800 Episode Num: 359 Episode T: 200 Reward: -369.017
Total T: 72000 Episode Num: 360 Episode T: 200 Reward: -123.320
recent Evaluation: -107.129791199219
Total T: 72200 Episode Num: 361 Episode T: 200 Reward: -127.303
Total T: 72400 Episode Num: 362 Episode T: 200 Reward: -235.044
Total T: 72600 Episode Num: 363 Episode T: 200 Reward: -116.179
Total T: 72800 Episode Num: 364 Episode T: 200 Reward: -354.851
Total T: 73000 Episode Num: 365 Episode T: 200 Reward: -118.956
recent Evaluation: -129.94527896155054
Total T: 73200 Episode Num: 366 Episode T: 200 Reward: -117.020
Total T: 73400 Episode Num: 367 Episode T: 200 Reward: -327.747
Total T: 73600 Episode Num: 368 Episode T: 200 Reward: -121.978
Total T: 73800 Episode Num: 369 Episode T: 200 Reward: -235.836
Total T: 74000 Episode Num: 370 Episode T: 200 Reward: -119.504
recent Evaluation: -131.44712519795783
Total T: 74200 Episode Num: 371 Episode T: 200 Reward: -4.063
Total T: 74400 Episode Num: 372 Episode T: 200 Reward: -114.820
Total T: 74600 Episode Num: 373 Episode T: 200 Reward: -124.384
Total T: 74800 Episode Num: 374 Episode T: 200 Reward: -123.265
Total T: 75000 Episode Num: 375 Episode T: 200 Reward: -250.773
recent Evaluation: -130.70854618984436
Total T: 75200 Episode Num: 376 Episode T: 200 Reward: -232.885
Total T: 75400 Episode Num: 377 Episode T: 200 Reward: -122.761
Total T: 75600 Episode Num: 378 Episode T: 200 Reward: -235.871
Total T: 75800 Episode Num: 379 Episode T: 200 Reward: -364.623
Total T: 76000 Episode Num: 380 Episode T: 200 Reward: -1.055
recent Evaluation: -156.10259432384277
Total T: 76200 Episode Num: 381 Episode T: 200 Reward: -239.259
Total T: 76400 Episode Num: 382 Episode T: 200 Reward: -355.669
Total T: 76600 Episode Num: 383 Episode T: 200 Reward: -127.135
Total T: 76800 Episode Num: 384 Episode T: 200 Reward: -118.681
Total T: 77000 Episode Num: 385 Episode T: 200 Reward: -117.931
recent Evaluation: -120.94782529601783
Total T: 77200 Episode Num: 386 Episode T: 200 Reward: -121.701
Total T: 77400 Episode Num: 387 Episode T: 200 Reward: -2.221
Total T: 77600 Episode Num: 388 Episode T: 200 Reward: -233.778
Total T: 77800 Episode Num: 389 Episode T: 200 Reward: -124.314
Total T: 78000 Episode Num: 390 Episode T: 200 Reward: -121.047
recent Evaluation: -145.33418776842666
Total T: 78200 Episode Num: 391 Episode T: 200 Reward: -120.620
Total T: 78400 Episode Num: 392 Episode T: 200 Reward: -242.514
Total T: 78600 Episode Num: 393 Episode T: 200 Reward: -231.053
Total T: 78800 Episode Num: 394 Episode T: 200 Reward: -120.715
Total T: 79000 Episode Num: 395 Episode T: 200 Reward: -249.834
recent Evaluation: -145.39320309766038
Total T: 79200 Episode Num: 396 Episode T: 200 Reward: -120.232
Total T: 79400 Episode Num: 397 Episode T: 200 Reward: -118.388
Total T: 79600 Episode Num: 398 Episode T: 200 Reward: -119.853
Total T: 79800 Episode Num: 399 Episode T: 200 Reward: -124.874
Total T: 80000 Episode Num: 400 Episode T: 200 Reward: -122.966
recent Evaluation: -138.8747178822827
Total T: 80200 Episode Num: 401 Episode T: 200 Reward: -347.893
Total T: 80400 Episode Num: 402 Episode T: 200 Reward: -336.751
Total T: 80600 Episode Num: 403 Episode T: 200 Reward: -239.443
Total T: 80800 Episode Num: 404 Episode T: 200 Reward: -238.828
Total T: 81000 Episode Num: 405 Episode T: 200 Reward: -246.688
recent Evaluation: -98.33863467645779
Total T: 81200 Episode Num: 406 Episode T: 200 Reward: -127.008
Total T: 81400 Episode Num: 407 Episode T: 200 Reward: -342.731
```

```
Total T: 81600 Episode Num: 408 Episode T: 200 Reward: -120.504
Total T: 81800 Episode Num: 409 Episode T: 200 Reward: -127.492
Total T: 82000 Episode Num: 410 Episode T: 200 Reward: -125.212
recent Evaluation: -143.73456530163944
Total T: 82200 Episode Num: 411 Episode T: 200 Reward: -121.796
Total T: 82400 Episode Num: 412 Episode T: 200 Reward: -1.322
Total T: 82600 Episode Num: 413 Episode T: 200 Reward: -126.786
Total T: 82800 Episode Num: 414 Episode T: 200 Reward: -250.959
Total T: 83000 Episode Num: 415 Episode T: 200 Reward: -348.336
recent Evaluation: -173.5677952145989
Total T: 83200 Episode Num: 416 Episode T: 200 Reward: -0.581
Total T: 83400 Episode Num: 417 Episode T: 200 Reward: -228.434
Total T: 83600 Episode Num: 418 Episode T: 200 Reward: -127.938
Total T: 83800 Episode Num: 419 Episode T: 200 Reward: -115.875
Total T: 84000 Episode Num: 420 Episode T: 200 Reward: -252.561
recent Evaluation: -143.16261155462456
Total T: 84200 Episode Num: 421 Episode T: 200 Reward: -229.550
Total T: 84400 Episode Num: 422 Episode T: 200 Reward: -121.150
Total T: 84600 Episode Num: 423 Episode T: 200 Reward: -121.795
Total T: 84800 Episode Num: 424 Episode T: 200 Reward: -127.102
Total T: 85000 Episode Num: 425 Episode T: 200 Reward: -118.921
recent Evaluation: -157.03547584427568
Total T: 85200 Episode Num: 426 Episode T: 200 Reward: -290.433
Total T: 85400 Episode Num: 427 Episode T: 200 Reward: -120.204
Total T: 85600 Episode Num: 428 Episode T: 200 Reward: -239.938
Total T: 85800 Episode Num: 429 Episode T: 200 Reward: -122.537
Total T: 86000 Episode Num: 430 Episode T: 200 Reward: -348.644
recent Evaluation: -155.56694691503745
Total T: 86200 Episode Num: 431 Episode T: 200 Reward: -118.407
Total T: 86400 Episode Num: 432 Episode T: 200 Reward: -349.440
Total T: 86600 Episode Num: 433 Episode T: 200 Reward: -124.701
Total T: 86800 Episode Num: 434 Episode T: 200 Reward: -236.436
Total T: 87000 Episode Num: 435 Episode T: 200 Reward: -118.175
recent Evaluation: -197.18022866658356
Total T: 87200 Episode Num: 436 Episode T: 200 Reward: -235.194
Total T: 87400 Episode Num: 437 Episode T: 200 Reward: -376.718
Total T: 87600 Episode Num: 438 Episode T: 200 Reward: -124.875
Total T: 87800 Episode Num: 439 Episode T: 200 Reward: -239.162
Total T: 88000 Episode Num: 440 Episode T: 200 Reward: -255.155
recent Evaluation: -153.67763206098817
Total T: 88200 Episode Num: 441 Episode T: 200 Reward: -122.176
Total T: 88400 Episode Num: 442 Episode T: 200 Reward: -124.986
Total T: 88600 Episode Num: 443 Episode T: 200 Reward: -116.759
Total T: 88800 Episode Num: 444 Episode T: 200 Reward: -352.506
Total T: 89000 Episode Num: 445 Episode T: 200 Reward: -242.041
recent Evaluation: -118.94552296763102
Total T: 89200 Episode Num: 446 Episode T: 200 Reward: -121.842
Total T: 89400 Episode Num: 447 Episode T: 200 Reward: -121.501
Total T: 89600 Episode Num: 448 Episode T: 200 Reward: -118.297
Total T: 89800 Episode Num: 449 Episode T: 200 Reward: -122.749
Total T: 90000 Episode Num: 450 Episode T: 200 Reward: -116.012
recent Evaluation: -154.4059096214927
Total T: 90200 Episode Num: 451 Episode T: 200 Reward: -289.894
Total T: 90400 Episode Num: 452 Episode T: 200 Reward: -230.184
Total T: 90600 Episode Num: 453 Episode T: 200 Reward: -0.958
Total T: 90800 Episode Num: 454 Episode T: 200 Reward: -242.609
Total T: 91000 Episode Num: 455 Episode T: 200 Reward: -120.954
recent Evaluation: -121.08941083783472
Total T: 91200 Episode Num: 456 Episode T: 200 Reward: -120.777
Total T: 91400 Episode Num: 457 Episode T: 200 Reward: -124.115
Total T: 91600 Episode Num: 458 Episode T: 200 Reward: -240.273
```

```
Total T: 91800 Episode Num: 459 Episode T: 200 Reward: -114.847
Total T: 92000 Episode Num: 460 Episode T: 200 Reward: -120.206
recent Evaluation: -132.2730255031158
Total T: 92200 Episode Num: 461 Episode T: 200 Reward: -245.671
Total T: 92400 Episode Num: 462 Episode T: 200 Reward: -126.519
Total T: 92600 Episode Num: 463 Episode T: 200 Reward: -235.893
Total T: 92800 Episode Num: 464 Episode T: 200 Reward: -3.778
Total T: 93000 Episode Num: 465 Episode T: 200 Reward: -117.232
recent Evaluation: -157.89732228429637
Total T: 93200 Episode Num: 466 Episode T: 200 Reward: -129.568
Total T: 93400 Episode Num: 467 Episode T: 200 Reward: -245.127
Total T: 93600 Episode Num: 468 Episode T: 200 Reward: -122.942
Total T: 93800 Episode Num: 469 Episode T: 200 Reward: -353.429
Total T: 94000 Episode Num: 470 Episode T: 200 Reward: -3.876
recent Evaluation: -133.69442869217775
Total T: 94200 Episode Num: 471 Episode T: 200 Reward: -238.317
Total T: 94400 Episode Num: 472 Episode T: 200 Reward: -119.774
Total T: 94600 Episode Num: 473 Episode T: 200 Reward: -246.843
Total T: 94800 Episode Num: 474 Episode T: 200 Reward: -128.236
Total T: 95000 Episode Num: 475 Episode T: 200 Reward: -123.193
recent Evaluation: -135.94789433334734
Total T: 95200 Episode Num: 476 Episode T: 200 Reward: -230.345
Total T: 95400 Episode Num: 477 Episode T: 200 Reward: -119.919
Total T: 95600 Episode Num: 478 Episode T: 200 Reward: -123.574
Total T: 95800 Episode Num: 479 Episode T: 200 Reward: -130.933
Total T: 96000 Episode Num: 480 Episode T: 200 Reward: -332.566
recent Evaluation: -133.69789246846724
Total T: 96200 Episode Num: 481 Episode T: 200 Reward: -6.095
Total T: 96400 Episode Num: 482 Episode T: 200 Reward: -130.134
Total T: 96600 Episode Num: 483 Episode T: 200 Reward: -244.483
Total T: 96800 Episode Num: 484 Episode T: 200 Reward: -121.961
Total T: 97000 Episode Num: 485 Episode T: 200 Reward: -6.212
recent Evaluation: -148.8327526547953
Total T: 97200 Episode Num: 486 Episode T: 200 Reward: -125.763
Total T: 97400 Episode Num: 487 Episode T: 200 Reward: -125.946
Total T: 97600 Episode Num: 488 Episode T: 200 Reward: -130.526
Total T: 97800 Episode Num: 489 Episode T: 200 Reward: -133.272
Total T: 98000 Episode Num: 490 Episode T: 200 Reward: -123.833
recent Evaluation: -182.754982266662
Total T: 98200 Episode Num: 491 Episode T: 200 Reward: -363.517
Total T: 98400 Episode Num: 492 Episode T: 200 Reward: -342.048
Total T: 98600 Episode Num: 493 Episode T: 200 Reward: -6.785
Total T: 98800 Episode Num: 494 Episode T: 200 Reward: -369.604
Total T: 99000 Episode Num: 495 Episode T: 200 Reward: -7.685
recent Evaluation: -113.14168954679644
Total T: 99200 Episode Num: 496 Episode T: 200 Reward: -358.571
Total T: 99400 Episode Num: 497 Episode T: 200 Reward: -337.519
Total T: 99600 Episode Num: 498 Episode T: 200 Reward: -7.656
Total T: 99800 Episode Num: 499 Episode T: 200 Reward: -238.573
Total T: 100000 Episode Num: 500 Episode T: 200 Reward: -120.427
recent Evaluation: -161.97815428314306
```

Four-Solution-Maze Environment (optional)

TODOs for you:

- Q11. (bonus) In this section, another environment named Four-Solution-Maze is provided for you to evaluate your algorithms.

The task is quite simple, yet never easy for even PPO/TD3.

The default size of the maze is 64x64, and in each game (espisode), the agent is initialized randomly in the maze. There are 4 positions in the maze that has non-trivial reward of +10, while reaching other region will recieve only a tiny punishment of -0.1. An optimal policy should be able to find the shortest path to the most recent reward region (i.e., one of the four high-reward regions.).

The action space is continuous with range [-1,1], larger actions will be clipped.

In [20]:

```
import numpy as np
import matplotlib.pyplot as plt
from pylab import *
from numpy import *
import copy

class FourWayGridWorld:
    def __init__(self, N=17,left = 10,right = 10, up=10, down = 10):
        self.N = N
        self.left = left
        self.right = right
        self.up = up
        self.down = down
        self.map = np.ones((N,N))*(-0.1)
        self.map[int((N-1)/2),0] = self.left
        self.map[0,int((N-1)/2)] = self.up
        self.map[N-1,int((N-1)/2)] = self.down
        self.map[int((N-1)/2),N-1] = self.right
        self.loc = np.asarray([np.random.randint(N),np.random.randint(N)])
        self.step_num = 0
    def step(self,action):
        action = np.clip(action,-1,1)
        new_loc = np.clip(self.loc + action,0,self.N-1)
        self.loc = new_loc
        reward = self.map[int(round(self.loc[0])),int(round(self.loc[1]))]
        self.step_num+=1
        return self.loc,reward,self.ifdone()
    def ifdone(self):
        if self.step_num >= 2*self.N:
            return True
        else:
            return False
    def render(self):
        map_self = copy.deepcopy(self.map)
        map_self[int(self.loc[0]),int(self.loc[1])] = -5
        plt.imshow(map_self)
    def reset(self):
        self.map = np.ones((self.N,self.N))*(-0.1)
        self.map[int((self.N-1)/2),0] = self.left
        self.map[0,int((self.N-1)/2)] = self.up
        self.map[self.N-1,int((self.N-1)/2)] = self.down
        self.map[int((self.N-1)/2),self.N-1] = self.right
        self.loc = np.asarray([np.random.randint(self.N),np.random.randint(self.N)])
        self.step_num = 0
        return self.loc
```

In [21]:

```
env = FourWayGridWorld(33)
```

In [22]:

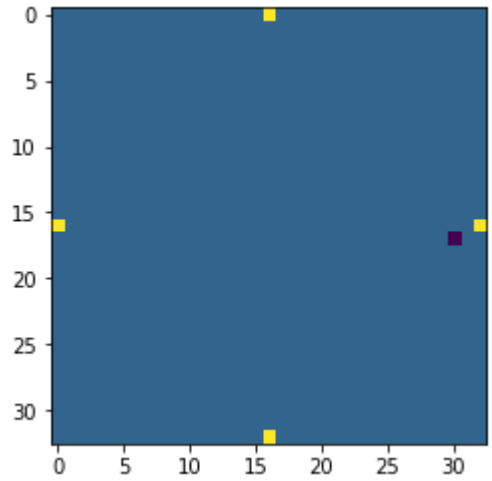
```
env.reset()
```

Out[22]:

```
array([17, 30])
```

In [23]:

```
env.render()
```



In [24]:

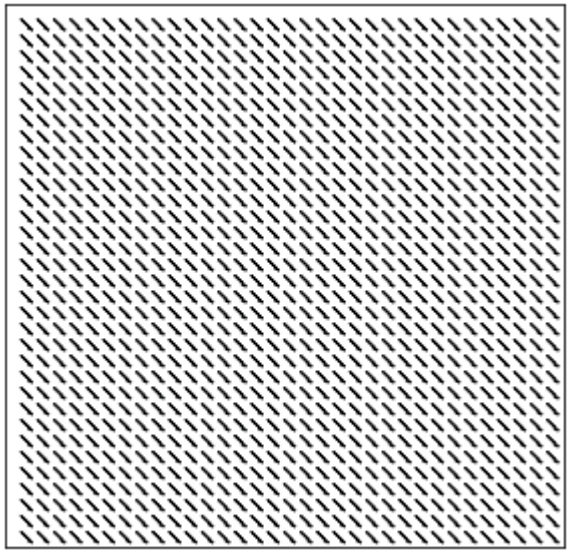
```
# This section is used to visualize your learned policy
from torch import Tensor
output_i = np.zeros((33,33))
output_j = np.zeros((33,33))
output_i_m = np.zeros((33,33))
output_j_m = np.zeros((33,33))
value_ij = np.zeros((33,33))
for i in range(33):
    for j in range(33):
        states = Tensor(np.asarray([i,j])).float().unsqueeze(0)

        '''
        you need to revise the following line,
        to fit your policy/network outputs
        '''
        action, value =[0.5,0.5],1
        output_i[i,j] = action[0]
        output_j[i,j] = action[1]
        value_ij[i,j] = value

plt.figure(figsize= (5,5))
for i in range(33):
    for j in range(33):
        plt.arrow(j,-i,output_j[i,j],-output_i[i,j],head_width=0.2,shape='left')
xlim(-1,33)
ylim(-33,1)
yticks([2*i-32 for i in range(17)],[2*i for i in range(17)])
plt.xticks([])
plt.yticks([])
```

Out[24]:

```
([], <a list of 0 Text yticklabel objects>)
```



In [ ]:

```
```