ex1

January 3, 2025

Reinforcement Learning Assignment 1 - The Reinforcement Learning Framework

This notebook is a part of teaching material for ELEC-E8125

Sep 4, 2024 - Nov 30, 2024

Aalto University

1 Table of contents

Total Points: 100

Estimated runtime of all the cells: 30 minutes

2 1. Introduction

In this exercise we will take a first look at a reinforcement learning environment, its components and modify the reward function of a simple agent.

In this notebook two environments are used: Cartpole and Reacher. The cartpole environment is taken from OpenAI's Gym library. The reacher environment is custom made (and defined in `reacher.py`) but utilizes the Gym API.

2.1 1.1 Learning Objectives:

- To become familiar with assignment structure and the agent-environment relationshp
- To understand the effects of stochasticity
- To understand and explore the effects of task definition

2.2 1.2 Code Structure & Files

The `train.py` file instantiates the environment and the RL agent that acts in it. The `agent.py` file contains the implementation of a simple reinforcement learning agent; for the sake of this exercise, you can assume it to be a black box (you don't need to understand how it works, although you are encouraged to study it in more detail). You don't have to edit any other file other than `ex1.ipynb` to complete this exercise.

```
cfg                  # Config files for environments e.g. define the maximum number of steps
imgs                 # Images used in notebook
results
    logging          # Contains logged data
    model            # Contains the policies learned
    video            # Contains videos for each environment
        CartPole-v0
           test      # Videos saved during testing
           train     # Videos saved during training
        SpinningReacher-v0
            test
            train
 ex1.ipynb             # Main assignment file containing tasks <---------
 feedback.ipynb        # Please give feedback in here
 README.ipynb          # This file
 agent.py              # Contains functions that govern the policy
 reacher.py            # Defines the reacher environment
 train.py              # Contains training and testing functions
 utils.py              # Contains useful functions
```

Please consult `README.md` for more details the assignments.

2.3 Warnings:

- Don't copy and paste cells within a notebook. This will mess up the tracking metadata and prevent autograding from working.
- Only add new cells using the '+' button in the upper toolbar and do not split cells.
- Be cautious about things such as copying the whole notebook to Colab to work on it. This has sometimes resulted in removing all notebook metadata, making autograding impossible.

[41]:
```
skip_training = True  # Set this flag to True before validation and submission
```

[ ]:
```

```

[2]:
```
from pathlib import Path # to find directory
work_dir = Path().cwd()/'results'
import os

import train as t # for training
import utils as u # helper functions

import numpy as np # The numpy library can be used for math functions
import torch # Used to manage policy and learning
from IPython.display import Video, display, HTML # to display videos
```

3 2. Cartpole

The Cartpole environment consists of a cart and a pole mounted on top of it, as shown in Figure 1. The cart can move either to the left or to the right. The goal is to balance the pole in a vertical position in order to prevent it from falling down. The cart should also stay within limited distance from the center (trying to move outside screen boundaries is considered a failure).

Figure 1: The Cartpole environment

The state and the observation are four element vectors:

$$o = s = \begin{pmatrix} x \\ \dot{x} \\ \theta \\ \dot{\theta} \end{pmatrix},$$

where x is the position of the cart, $\dot{x}$ is its velocity, θ is the angle of the pole w.r.t. the vertical axis, and $\dot{\theta}$ is the angular velocity of the pole.

In the standard formulation, a reward of 1 is given for every timestep the pole remains balanced. Upon failing (the pole falls) or completing the task, an episode is finished.

The training script will record videos of the agent's learning progress during training, and the recorded videos are saved to `results/video/CartPole-v0/train`. By default, the training information is saved to `results/logging/CartPole-v0_{seed}.csv`. When the training is finished, the models are saved to `results/model/Cartpole-v0_params.pt`. Videos of the agent's behaviour during testing are saved to `results/video/CartPole-v0/test`.

`<h3><b>Student Task 1.</b> Training a Model for Simple Cartpole Environment (10 points) </h3>`

This task requires you to train a model for the cartpole environment with 100 timesteps per episode. Then test the model for 500 timesteps and report average reward. To do this, you can simply run the code in the cells below.

To see a full list of options that can be passed through `cfg_args` consult the configuation file found in `cfg/`.

- **1st** Run training over 100 steps per episode by using `t.train` function. See the cell below. The training will run for 500 episodes automatically.

- **2nd:** Export the training plot `episodesep_reward` from logged data (.csv format).
- **3rd:** Run testing over 500 steps by using `t.test` function. See the cell below. See the cell below. Notice `max_episode_steps` parameter.
- **4th: Manually** report the average reward after the cells have completed execution.

Table of Contents

[3]:
```
if not skip_training:
    t.train(cfg_path=Path().cwd()/'cfg'/'cartpole_v1.yaml',␣
 ↪cfg_args=dict(seed=1, max_episode_steps=100, model_name="CartPole-v1")) # <␣
 ↪5 mins
```

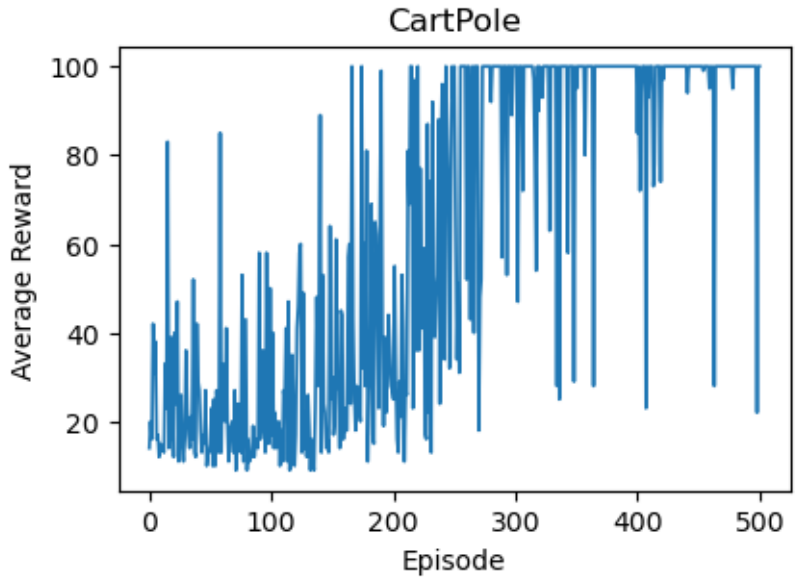
```
Numpy/Torch/Random Seed:  1
Configuration Settings: {'exp_name': 'ex1', 'seed': 1, 'env_name':
'CartPole-v1', 'model_name': 'CartPole-v1', 'max_episode_steps': 100,
'train_episodes': 500, 'batch_size': 64, 'min_update_samples': 2000, 'testing':
False, 'model_path': 'default', 'save_video': True, 'save_model': True,
'save_logging': True, 'silent': False, 'use_wandb': True, 'run_suffix': 0}
Training device: cpu
Observation space dimensions: 4
Action space dimensions: 2

Episode 0 finished. Total reward: 14.0 (14 timesteps)
Episode 5 finished. Total reward: 38.0 (38 timesteps)
Episode 10 finished. Total reward: 14.0 (14 timesteps)
Episode 15 finished. Total reward: 83.0 (83 timesteps)
Episode 20 finished. Total reward: 12.0 (12 timesteps)
Episode 25 finished. Total reward: 18.0 (18 timesteps)
Episode 30 finished. Total reward: 36.0 (36 timesteps)
Episode 35 finished. Total reward: 16.0 (16 timesteps)
Episode 40 finished. Total reward: 29.0 (29 timesteps)
Episode 45 finished. Total reward: 15.0 (15 timesteps)
Episode 50 finished. Total reward: 15.0 (15 timesteps)
Episode 55 finished. Total reward: 27.0 (27 timesteps)
Episode 60 finished. Total reward: 33.0 (33 timesteps)
Episode 65 finished. Total reward: 11.0 (11 timesteps)
Episode 70 finished. Total reward: 27.0 (27 timesteps)
Episode 75 finished. Total reward: 13.0 (13 timesteps)
Episode 80 finished. Total reward: 9.0 (9 timesteps)
Episode 85 finished. Total reward: 19.0 (19 timesteps)
Updating the policy…
Updating finished!
Episode 90 finished. Total reward: 58.0 (58 timesteps)
Episode 95 finished. Total reward: 13.0 (13 timesteps)
Episode 100 finished. Total reward: 17.0 (17 timesteps)
Episode 105 finished. Total reward: 17.0 (17 timesteps)
Episode 110 finished. Total reward: 27.0 (27 timesteps)
Episode 115 finished. Total reward: 9.0 (9 timesteps)
```

```
Episode 120 finished. Total reward: 17.0 (17 timesteps)
Episode 125 finished. Total reward: 13.0 (13 timesteps)
Episode 130 finished. Total reward: 26.0 (26 timesteps)
Episode 135 finished. Total reward: 9.0 (9 timesteps)
Episode 140 finished. Total reward: 89.0 (89 timesteps)
Episode 145 finished. Total reward: 14.0 (14 timesteps)
Episode 150 finished. Total reward: 30.0 (30 timesteps)
Episode 155 finished. Total reward: 27.0 (27 timesteps)
Episode 160 finished. Total reward: 23.0 (23 timesteps)
Updating the policy…
Updating finished!
Episode 165 finished. Total reward: 24.0 (24 timesteps)
Episode 170 finished. Total reward: 28.0 (28 timesteps)
Episode 175 finished. Total reward: 32.0 (32 timesteps)
Episode 180 finished. Total reward: 27.0 (27 timesteps)
Episode 185 finished. Total reward: 65.0 (65 timesteps)
Episode 190 finished. Total reward: 99.0 (99 timesteps)
Episode 195 finished. Total reward: 30.0 (30 timesteps)
Episode 200 finished. Total reward: 25.0 (25 timesteps)
Episode 205 finished. Total reward: 29.0 (29 timesteps)
Episode 210 finished. Total reward: 27.0 (27 timesteps)
Updating the policy…
Updating finished!
Episode 215 finished. Total reward: 100.0 (100 timesteps)
Episode 220 finished. Total reward: 100.0 (100 timesteps)
Episode 225 finished. Total reward: 59.0 (59 timesteps)
Episode 230 finished. Total reward: 74.0 (74 timesteps)
Episode 235 finished. Total reward: 49.0 (49 timesteps)
Episode 240 finished. Total reward: 96.0 (96 timesteps)
Episode 245 finished. Total reward: 73.0 (73 timesteps)
Updating the policy…
Updating finished!
Episode 250 finished. Total reward: 100.0 (100 timesteps)
Episode 255 finished. Total reward: 100.0 (100 timesteps)
Episode 260 finished. Total reward: 52.0 (52 timesteps)
Episode 265 finished. Total reward: 100.0 (100 timesteps)
Episode 270 finished. Total reward: 18.0 (18 timesteps)
Updating the policy…
Updating finished!
Episode 275 finished. Total reward: 100.0 (100 timesteps)
Episode 280 finished. Total reward: 92.0 (92 timesteps)
Episode 285 finished. Total reward: 100.0 (100 timesteps)
Episode 290 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 295 finished. Total reward: 100.0 (100 timesteps)
Episode 300 finished. Total reward: 100.0 (100 timesteps)
Episode 305 finished. Total reward: 100.0 (100 timesteps)
```

```
Episode 310 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 315 finished. Total reward: 97.0 (97 timesteps)
Episode 320 finished. Total reward: 100.0 (100 timesteps)
Episode 325 finished. Total reward: 100.0 (100 timesteps)
Episode 330 finished. Total reward: 100.0 (100 timesteps)
Episode 335 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 340 finished. Total reward: 100.0 (100 timesteps)
Episode 345 finished. Total reward: 100.0 (100 timesteps)
Episode 350 finished. Total reward: 95.0 (95 timesteps)
Episode 355 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 360 finished. Total reward: 100.0 (100 timesteps)
Episode 365 finished. Total reward: 100.0 (100 timesteps)
Episode 370 finished. Total reward: 100.0 (100 timesteps)
Episode 375 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 380 finished. Total reward: 100.0 (100 timesteps)
Episode 385 finished. Total reward: 100.0 (100 timesteps)
Episode 390 finished. Total reward: 100.0 (100 timesteps)
Episode 395 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 400 finished. Total reward: 85.0 (85 timesteps)
Episode 405 finished. Total reward: 100.0 (100 timesteps)
Episode 410 finished. Total reward: 100.0 (100 timesteps)
Episode 415 finished. Total reward: 100.0 (100 timesteps)
Episode 420 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 425 finished. Total reward: 100.0 (100 timesteps)
Episode 430 finished. Total reward: 100.0 (100 timesteps)
Episode 435 finished. Total reward: 100.0 (100 timesteps)
Episode 440 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 445 finished. Total reward: 100.0 (100 timesteps)
Episode 450 finished. Total reward: 100.0 (100 timesteps)
Episode 455 finished. Total reward: 100.0 (100 timesteps)
Episode 460 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 465 finished. Total reward: 100.0 (100 timesteps)
```

```
Episode 470 finished. Total reward: 100.0 (100 timesteps)
Episode 475 finished. Total reward: 100.0 (100 timesteps)
Episode 480 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 485 finished. Total reward: 100.0 (100 timesteps)
Episode 490 finished. Total reward: 100.0 (100 timesteps)
Episode 495 finished. Total reward: 100.0 (100 timesteps)
Episode 500 finished. Total reward: 100.0 (100 timesteps)
Model saved to /notebooks/rl2024/ex1/results/model/CartPole-v1_params.pt
------Training finished.------
```

[4]:
```
if not skip_training:
    u.plot_reward(Path().cwd()/'results'/'logging'/'CartPole-v1_1.csv',␣
 ↪'CartPole')
```



The command below will evaluate the trained model in 10 episodes and report the average reward (and episode length) for these 10 episodes. Do not delete the cell below as it is used for grading.

[5]:
```
%%capture --no-stdout
assert t.test(episodes=10, cfg_path=Path().cwd()/'cfg'/'cartpole_v1.yaml',
        cfg_args=dict(testing=True, save_video=(not skip_training),␣
 ↪save_model=False, seed=None,␣
 ↪max_episode_steps=500,model_name="CartPole-v1")) > 100
```

```
Numpy/Torch/Random Seed:  688
Loading model from /notebooks/rl2024/ex1/results/model/CartPole-v1_params.pt …
Testing…
```

```
Test ep reward: 123.0 seed: 434
Test ep reward: 144.0 seed: 308
Test ep reward: 128.0 seed: 922
Test ep reward: 127.0 seed: 231
Test ep reward: 121.0 seed: 961
Test ep reward: 140.0 seed: 363
Test ep reward: 128.0 seed: 626
Test ep reward: 130.0 seed: 905
Test ep reward: 137.0 seed: 497
Test ep reward: 134.0 seed: 816
Average test reward: 131.2 episode length: 131.2
```

Report below the average reward after testing the model

131.2

The agent acting in the environment can be seen using the following command. Change the `path` to pick the episode you want to visualize. Bear in mind by default video saving for training is taken every 50 episodes.

[6]:
```
if not skip_training:
    # Train Result
    video_dir = work_dir/'video'/'CartPole-v1'/'train'

    # List all MP4 files in the directory
    mp4_files = [file for file in os.listdir(video_dir) if file.endswith(".
 ↪mp4")]
    frame_colors = ['#FF5733', '#33FF57', '#5733FF', '#FFFF33', '#33FFFF',␣
 ↪'#FF33FF']
    # Display each MP4 file
    for i, mp4_file in enumerate(mp4_files):
        video_path = os.path.join(video_dir, mp4_file)
        video = Video(video_path, embed=True, html_attributes="loop autoplay",␣
 ↪width=200, height=100)
        frame_color = frame_colors[i % len(frame_colors)]
        video_frame = HTML(f'<div style="width: 200px; height: 100px;; border:␣
 ↪1px solid #FF5733;">{video._repr_html_()}</div>')
        print("test/",mp4_file)
        display(video_frame)
```

```
test/ ex1-episode-0.mp4
<IPython.core.display.HTML object>
test/ ex1-episode-50.mp4
<IPython.core.display.HTML object>
test/ ex1-episode-100.mp4
<IPython.core.display.HTML object>
```

```
test/ ex1-episode-150.mp4
<IPython.core.display.HTML object>
test/ ex1-episode-200.mp4
<IPython.core.display.HTML object>
test/ ex1-episode-250.mp4
<IPython.core.display.HTML object>
test/ ex1-episode-300.mp4
<IPython.core.display.HTML object>
test/ ex1-episode-350.mp4
<IPython.core.display.HTML object>
test/ ex1-episode-400.mp4
<IPython.core.display.HTML object>
test/ ex1-episode-450.mp4
<IPython.core.display.HTML object>
test/ ex1-episode-500.mp4
<IPython.core.display.HTML object>
```

[7]:
```python
if not skip_training:
    # Test Result

    video_dir = work_dir/'video'/'CartPole-v1'/'test'

    # List all MP4 files in the directory
    mp4_files = [file for file in os.listdir(video_dir) if file.endswith(".
↪mp4")]
    frame_colors = ['#FF5733', '#33FF57', '#5733FF', '#FFFF33', '#33FFFF',␣
↪'#FF33FF']
    # Display each MP4 file
    for i, mp4_file in enumerate(mp4_files):
        video_path = os.path.join(video_dir, mp4_file)
        video = Video(video_path, embed=True, html_attributes="loop autoplay",␣
↪width=200, height=100)
        frame_color = frame_colors[i % len(frame_colors)]
        video_frame = HTML(f'<div style="width: 200px; height: 100px;; border:␣
↪1px solid #5733FF;">{video._repr_html_()}</div>')
        print("test/",mp4_file)
        display(video_frame)
```

```
test/ ex1-episode-0.mp4
```

```
<IPython.core.display.HTML object>
test/ ex1-episode-1.mp4
<IPython.core.display.HTML object>
test/ ex1-episode-2.mp4
<IPython.core.display.HTML object>
test/ ex1-episode-3.mp4
<IPython.core.display.HTML object>
test/ ex1-episode-4.mp4
<IPython.core.display.HTML object>
test/ ex1-episode-5.mp4
<IPython.core.display.HTML object>
test/ ex1-episode-6.mp4
<IPython.core.display.HTML object>
test/ ex1-episode-7.mp4
<IPython.core.display.HTML object>
test/ ex1-episode-8.mp4
<IPython.core.display.HTML object>
test/ ex1-episode-9.mp4
<IPython.core.display.HTML object>
```

<h3>Student Question 1.1 Learning (10 points) </h3>

Test the trained model from Task 1 five times with different random seeds. Did the same model, trained to balance for 100 timesteps, learn to always balance the pole for 1000 timesteps? Why/why not?

Table of Contents

[8]:
```
#Evaluate the trained model by setting 5 different random seeds chosen by you
#Type 5 different  random seeds  in eval_seeds array
eval_seeds =  [1, 4, 5, 15, 98]
```

[9]:
```
assert len(set(eval_seeds)) == 5
```

[10]:
```
%%capture --no-stdout
def question1_1_evaluate1(eval_seeds):
    mean_reward = []

    for seed in eval_seeds:
```

```
        acc_rewards = t.test(episodes=1, cfg_path=Path().cwd()/'cfg'/
↪'cartpole_v1.yaml',
            cfg_args=dict(testing=True,save_video=False, save_logging=False,␣
↪save_model=False,
            seed=seed, max_episode_steps=1000, model_name="CartPole-v1"))
        mean_reward.append(acc_rewards)

    return np.mean(mean_reward)

assert question1_1_evaluate1(eval_seeds) > 100
```

```
Numpy/Torch/Random Seed:  1
Loading model from /notebooks/rl2024/ex1/results/model/CartPole-v1_params.pt …
Testing…
Test ep reward: 127.0 seed: 1
Average test reward: 127.0 episode length: 127.0
Numpy/Torch/Random Seed:  4
Loading model from /notebooks/rl2024/ex1/results/model/CartPole-v1_params.pt …
Testing…
Test ep reward: 148.0 seed: 4
Average test reward: 148.0 episode length: 148.0
Numpy/Torch/Random Seed:  5
Loading model from /notebooks/rl2024/ex1/results/model/CartPole-v1_params.pt …
Testing…
Test ep reward: 138.0 seed: 5
Average test reward: 138.0 episode length: 138.0
Numpy/Torch/Random Seed:  15
Loading model from /notebooks/rl2024/ex1/results/model/CartPole-v1_params.pt …
Testing…
Test ep reward: 130.0 seed: 15
Average test reward: 130.0 episode length: 130.0
Numpy/Torch/Random Seed:  98
Loading model from /notebooks/rl2024/ex1/results/model/CartPole-v1_params.pt …
Testing…
Test ep reward: 130.0 seed: 98
Average test reward: 130.0 episode length: 130.0
```

3.0.1 Question:

Did the same model consistently balance the pole for 1000 timesteps? Select the most appropriate answer.

Options: Choices: 1. No, the agent generally failed around 100-200 timesteps because it encountered unfamiliar states.

2. Yes, the agent balanced the pole for 1000 timesteps in every test due to a perfectly learned policy.
3. No, although the agent sometimes balanced the pole for 1000 timesteps, it typically failed

due to limited generalization to new states.

4. Yes, but only with some random seeds; it failed to generalize this behavior across all tests.
5. No, the agent frequently moved out of bounds or the pole tilted too much, showing poor adaptation.
6. Yes, the agent demonstrated perfect control in all tests due to effective training.
7. No, the agent's performance varied significantly with different random seeds, indicating inconsistency in learned behavior.

[11]:
```
sq1_1 = 3 #Answer question 1.1 with the appropriate answer number
```

[12]:
```
assert sq1_1 in range(1,8)
```

The following cells are used for grading

[ ]:

[ ]:

<h3>Student Task 2. Investigating Training Performance (10 points) </h3>

Repeat the experiment in Task 1 five times, each time training the model from scratch with 100 timesteps and testing it for 1000 timesteps. Use a different seed number for each training/testing cycle. You can use the box below to write a small script to do this. Use the result textbox below to report the average test reward for each repeat.

Table of Contents

Make sure to change the model_name for each agent to the ones given and and that the respective parameter file is saved in results/model/

In order to get the points for this exercise each of the agents must have a test accuracy of higher than 100.

[16]:
```
# %%capture
# uncomment if you want to skip the output

if not skip_training:
    model_names = ["CartPole-v1-model0", "CartPole-v1-model1",␣
 ↪"CartPole-v1-model2", "CartPole-v1-model3", "CartPole-v1-model4"]
    model_seeds = [1, 2, 4, 5, 15]

    assert len(set(model_seeds)) == 5
    '''
    TODO: Repeat the experiment in Task 1 five times
    '''
    ########## Your code starts here ##########
    for model_name, seed in zip(model_names, model_seeds):
        # Train the model
```

```
        t.train(cfg_path=Path().cwd() / 'cfg' / 'cartpole_v1.yaml',
                cfg_args=dict(seed=seed, max_episode_steps=100,␣
↪model_name=model_name))
    ########## Your code ends here ##########
```

```
Numpy/Torch/Random Seed:  1
Configuration Settings: {'exp_name': 'ex1', 'seed': 1, 'env_name':
'CartPole-v1', 'model_name': 'CartPole-v1-model0', 'max_episode_steps': 100,
'train_episodes': 500, 'batch_size': 64, 'min_update_samples': 2000, 'testing':
False, 'model_path': 'default', 'save_video': True, 'save_model': True,
'save_logging': True, 'silent': False, 'use_wandb': True, 'run_suffix': 0}
Training device: cpu
Observation space dimensions: 4
Action space dimensions: 2


/opt/software/lib/python3.10/site-
packages/gymnasium/wrappers/record_video.py:87: UserWarning: WARN:
Overwriting existing videos at
/notebooks/rl2024/ex1/results/video/CartPole-v1/train folder (try specifying a
different `video_folder` for the `RecordVideo` wrapper if this is not
desired)
  logger.warn(
Episode 0 finished. Total reward: 14.0 (14 timesteps)
Episode 5 finished. Total reward: 38.0 (38 timesteps)
Episode 10 finished. Total reward: 14.0 (14 timesteps)
Episode 15 finished. Total reward: 83.0 (83 timesteps)
Episode 20 finished. Total reward: 12.0 (12 timesteps)
Episode 25 finished. Total reward: 18.0 (18 timesteps)
Episode 30 finished. Total reward: 36.0 (36 timesteps)
Episode 35 finished. Total reward: 16.0 (16 timesteps)
Episode 40 finished. Total reward: 29.0 (29 timesteps)
Episode 45 finished. Total reward: 15.0 (15 timesteps)
Episode 50 finished. Total reward: 15.0 (15 timesteps)
Episode 55 finished. Total reward: 27.0 (27 timesteps)
Episode 60 finished. Total reward: 33.0 (33 timesteps)
Episode 65 finished. Total reward: 11.0 (11 timesteps)
Episode 70 finished. Total reward: 27.0 (27 timesteps)
Episode 75 finished. Total reward: 13.0 (13 timesteps)
Episode 80 finished. Total reward: 9.0 (9 timesteps)
Episode 85 finished. Total reward: 19.0 (19 timesteps)
Updating the policy…
Updating finished!
Episode 90 finished. Total reward: 58.0 (58 timesteps)
Episode 95 finished. Total reward: 13.0 (13 timesteps)
```

```
Episode 100 finished. Total reward: 17.0 (17 timesteps)
Episode 105 finished. Total reward: 17.0 (17 timesteps)
Episode 110 finished. Total reward: 27.0 (27 timesteps)
Episode 115 finished. Total reward: 9.0 (9 timesteps)
Episode 120 finished. Total reward: 17.0 (17 timesteps)
Episode 125 finished. Total reward: 13.0 (13 timesteps)
Episode 130 finished. Total reward: 26.0 (26 timesteps)
Episode 135 finished. Total reward: 9.0 (9 timesteps)
Episode 140 finished. Total reward: 89.0 (89 timesteps)
Episode 145 finished. Total reward: 14.0 (14 timesteps)
Episode 150 finished. Total reward: 30.0 (30 timesteps)
Episode 155 finished. Total reward: 27.0 (27 timesteps)
Episode 160 finished. Total reward: 23.0 (23 timesteps)
Updating the policy…
Updating finished!
Episode 165 finished. Total reward: 24.0 (24 timesteps)
Episode 170 finished. Total reward: 28.0 (28 timesteps)
Episode 175 finished. Total reward: 32.0 (32 timesteps)
Episode 180 finished. Total reward: 27.0 (27 timesteps)
Episode 185 finished. Total reward: 65.0 (65 timesteps)
Episode 190 finished. Total reward: 99.0 (99 timesteps)
Episode 195 finished. Total reward: 30.0 (30 timesteps)
Episode 200 finished. Total reward: 25.0 (25 timesteps)
Episode 205 finished. Total reward: 29.0 (29 timesteps)
Episode 210 finished. Total reward: 27.0 (27 timesteps)
Updating the policy…
Updating finished!
Episode 215 finished. Total reward: 100.0 (100 timesteps)
Episode 220 finished. Total reward: 100.0 (100 timesteps)
Episode 225 finished. Total reward: 59.0 (59 timesteps)
Episode 230 finished. Total reward: 74.0 (74 timesteps)
Episode 235 finished. Total reward: 49.0 (49 timesteps)
Episode 240 finished. Total reward: 96.0 (96 timesteps)
Episode 245 finished. Total reward: 73.0 (73 timesteps)
Updating the policy…
Updating finished!
Episode 250 finished. Total reward: 100.0 (100 timesteps)
Episode 255 finished. Total reward: 100.0 (100 timesteps)
Episode 260 finished. Total reward: 52.0 (52 timesteps)
Episode 265 finished. Total reward: 100.0 (100 timesteps)
Episode 270 finished. Total reward: 18.0 (18 timesteps)
Updating the policy…
Updating finished!
Episode 275 finished. Total reward: 100.0 (100 timesteps)
Episode 280 finished. Total reward: 92.0 (92 timesteps)
Episode 285 finished. Total reward: 100.0 (100 timesteps)
Episode 290 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
```

```
Updating finished!
Episode 295 finished. Total reward: 100.0 (100 timesteps)
Episode 300 finished. Total reward: 100.0 (100 timesteps)
Episode 305 finished. Total reward: 100.0 (100 timesteps)
Episode 310 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 315 finished. Total reward: 97.0 (97 timesteps)
Episode 320 finished. Total reward: 100.0 (100 timesteps)
Episode 325 finished. Total reward: 100.0 (100 timesteps)
Episode 330 finished. Total reward: 100.0 (100 timesteps)
Episode 335 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 340 finished. Total reward: 100.0 (100 timesteps)
Episode 345 finished. Total reward: 100.0 (100 timesteps)
Episode 350 finished. Total reward: 95.0 (95 timesteps)
Episode 355 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 360 finished. Total reward: 100.0 (100 timesteps)
Episode 365 finished. Total reward: 100.0 (100 timesteps)
Episode 370 finished. Total reward: 100.0 (100 timesteps)
Episode 375 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 380 finished. Total reward: 100.0 (100 timesteps)
Episode 385 finished. Total reward: 100.0 (100 timesteps)
Episode 390 finished. Total reward: 100.0 (100 timesteps)
Episode 395 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 400 finished. Total reward: 85.0 (85 timesteps)
Episode 405 finished. Total reward: 100.0 (100 timesteps)
Episode 410 finished. Total reward: 100.0 (100 timesteps)
Episode 415 finished. Total reward: 100.0 (100 timesteps)
Episode 420 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 425 finished. Total reward: 100.0 (100 timesteps)
Episode 430 finished. Total reward: 100.0 (100 timesteps)
Episode 435 finished. Total reward: 100.0 (100 timesteps)
Episode 440 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 445 finished. Total reward: 100.0 (100 timesteps)
Episode 450 finished. Total reward: 100.0 (100 timesteps)
Episode 455 finished. Total reward: 100.0 (100 timesteps)
```

```
Episode 460 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 465 finished. Total reward: 100.0 (100 timesteps)
Episode 470 finished. Total reward: 100.0 (100 timesteps)
Episode 475 finished. Total reward: 100.0 (100 timesteps)
Episode 480 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 485 finished. Total reward: 100.0 (100 timesteps)
Episode 490 finished. Total reward: 100.0 (100 timesteps)
Episode 495 finished. Total reward: 100.0 (100 timesteps)
Episode 500 finished. Total reward: 100.0 (100 timesteps)
Model saved to /notebooks/rl2024/ex1/results/model/CartPole-v1-model0_params.pt
------Training finished.------
Numpy/Torch/Random Seed:  2
Configuration Settings: {'exp_name': 'ex1', 'seed': 2, 'env_name':
'CartPole-v1', 'model_name': 'CartPole-v1-model1', 'max_episode_steps': 100,
'train_episodes': 500, 'batch_size': 64, 'min_update_samples': 2000, 'testing':
False, 'model_path': 'default', 'save_video': True, 'save_model': True,
'save_logging': True, 'silent': False, 'use_wandb': True, 'run_suffix': 0}
Training device: cpu
Observation space dimensions: 4
Action space dimensions: 2

Episode 0 finished. Total reward: 13.0 (13 timesteps)
Episode 5 finished. Total reward: 16.0 (16 timesteps)
Episode 10 finished. Total reward: 18.0 (18 timesteps)
Episode 15 finished. Total reward: 17.0 (17 timesteps)
Episode 20 finished. Total reward: 28.0 (28 timesteps)
Episode 25 finished. Total reward: 40.0 (40 timesteps)
Episode 30 finished. Total reward: 22.0 (22 timesteps)
Episode 35 finished. Total reward: 13.0 (13 timesteps)
Episode 40 finished. Total reward: 16.0 (16 timesteps)
Episode 45 finished. Total reward: 18.0 (18 timesteps)
Episode 50 finished. Total reward: 35.0 (35 timesteps)
Episode 55 finished. Total reward: 12.0 (12 timesteps)
Episode 60 finished. Total reward: 10.0 (10 timesteps)
Episode 65 finished. Total reward: 27.0 (27 timesteps)
Episode 70 finished. Total reward: 23.0 (23 timesteps)
Episode 75 finished. Total reward: 10.0 (10 timesteps)
Episode 80 finished. Total reward: 34.0 (34 timesteps)
Episode 85 finished. Total reward: 19.0 (19 timesteps)
Episode 90 finished. Total reward: 14.0 (14 timesteps)
Episode 95 finished. Total reward: 13.0 (13 timesteps)
Updating the policy…
Updating finished!
Episode 100 finished. Total reward: 19.0 (19 timesteps)
```

```
Episode 105 finished. Total reward: 14.0 (14 timesteps)
Episode 110 finished. Total reward: 19.0 (19 timesteps)
Episode 115 finished. Total reward: 16.0 (16 timesteps)
Episode 120 finished. Total reward: 15.0 (15 timesteps)
Episode 125 finished. Total reward: 28.0 (28 timesteps)
Episode 130 finished. Total reward: 38.0 (38 timesteps)
Episode 135 finished. Total reward: 52.0 (52 timesteps)
Episode 140 finished. Total reward: 35.0 (35 timesteps)
Episode 145 finished. Total reward: 20.0 (20 timesteps)
Episode 150 finished. Total reward: 44.0 (44 timesteps)
Episode 155 finished. Total reward: 14.0 (14 timesteps)
Episode 160 finished. Total reward: 21.0 (21 timesteps)
Episode 165 finished. Total reward: 15.0 (15 timesteps)
Updating the policy…
Updating finished!
Episode 170 finished. Total reward: 36.0 (36 timesteps)
Episode 175 finished. Total reward: 28.0 (28 timesteps)
Episode 180 finished. Total reward: 22.0 (22 timesteps)
Episode 185 finished. Total reward: 27.0 (27 timesteps)
Episode 190 finished. Total reward: 36.0 (36 timesteps)
Episode 195 finished. Total reward: 100.0 (100 timesteps)
Episode 200 finished. Total reward: 35.0 (35 timesteps)
Episode 205 finished. Total reward: 100.0 (100 timesteps)
Episode 210 finished. Total reward: 21.0 (21 timesteps)
Updating the policy…
Updating finished!
Episode 215 finished. Total reward: 55.0 (55 timesteps)
Episode 220 finished. Total reward: 26.0 (26 timesteps)
Episode 225 finished. Total reward: 100.0 (100 timesteps)
Episode 230 finished. Total reward: 100.0 (100 timesteps)
Episode 235 finished. Total reward: 61.0 (61 timesteps)
Updating the policy…
Updating finished!
Episode 240 finished. Total reward: 100.0 (100 timesteps)
Episode 245 finished. Total reward: 53.0 (53 timesteps)
Episode 250 finished. Total reward: 100.0 (100 timesteps)
Episode 255 finished. Total reward: 53.0 (53 timesteps)
Episode 260 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 265 finished. Total reward: 42.0 (42 timesteps)
Episode 270 finished. Total reward: 100.0 (100 timesteps)
Episode 275 finished. Total reward: 100.0 (100 timesteps)
Episode 280 finished. Total reward: 68.0 (68 timesteps)
Episode 285 finished. Total reward: 23.0 (23 timesteps)
Updating the policy…
Updating finished!
Episode 290 finished. Total reward: 36.0 (36 timesteps)
```

```
Episode 295 finished. Total reward: 100.0 (100 timesteps)
Episode 300 finished. Total reward: 100.0 (100 timesteps)
Episode 305 finished. Total reward: 83.0 (83 timesteps)
Updating the policy…
Updating finished!
Episode 310 finished. Total reward: 100.0 (100 timesteps)
Episode 315 finished. Total reward: 100.0 (100 timesteps)
Episode 320 finished. Total reward: 100.0 (100 timesteps)
Episode 325 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 330 finished. Total reward: 100.0 (100 timesteps)
Episode 335 finished. Total reward: 100.0 (100 timesteps)
Episode 340 finished. Total reward: 100.0 (100 timesteps)
Episode 345 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 350 finished. Total reward: 100.0 (100 timesteps)
Episode 355 finished. Total reward: 100.0 (100 timesteps)
Episode 360 finished. Total reward: 100.0 (100 timesteps)
Episode 365 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 370 finished. Total reward: 100.0 (100 timesteps)
Episode 375 finished. Total reward: 100.0 (100 timesteps)
Episode 380 finished. Total reward: 100.0 (100 timesteps)
Episode 385 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 390 finished. Total reward: 100.0 (100 timesteps)
Episode 395 finished. Total reward: 100.0 (100 timesteps)
Episode 400 finished. Total reward: 100.0 (100 timesteps)
Episode 405 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 410 finished. Total reward: 100.0 (100 timesteps)
Episode 415 finished. Total reward: 100.0 (100 timesteps)
Episode 420 finished. Total reward: 100.0 (100 timesteps)
Episode 425 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 430 finished. Total reward: 100.0 (100 timesteps)
Episode 435 finished. Total reward: 100.0 (100 timesteps)
Episode 440 finished. Total reward: 72.0 (72 timesteps)
Episode 445 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 450 finished. Total reward: 100.0 (100 timesteps)
```

```
Episode 455 finished. Total reward: 100.0 (100 timesteps)
Episode 460 finished. Total reward: 100.0 (100 timesteps)
Episode 465 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 470 finished. Total reward: 83.0 (83 timesteps)
Episode 475 finished. Total reward: 56.0 (56 timesteps)
Episode 480 finished. Total reward: 100.0 (100 timesteps)
Episode 485 finished. Total reward: 20.0 (20 timesteps)
Episode 490 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 495 finished. Total reward: 100.0 (100 timesteps)
Episode 500 finished. Total reward: 100.0 (100 timesteps)
Model saved to /notebooks/rl2024/ex1/results/model/CartPole-v1-model1_params.pt
------Training finished.------
Numpy/Torch/Random Seed:  4
Configuration Settings: {'exp_name': 'ex1', 'seed': 4, 'env_name':
'CartPole-v1', 'model_name': 'CartPole-v1-model2', 'max_episode_steps': 100,
'train_episodes': 500, 'batch_size': 64, 'min_update_samples': 2000, 'testing':
False, 'model_path': 'default', 'save_video': True, 'save_model': True,
'save_logging': True, 'silent': False, 'use_wandb': True, 'run_suffix': 0}
Training device: cpu
Observation space dimensions: 4
Action space dimensions: 2

Episode 0 finished. Total reward: 18.0 (18 timesteps)
Episode 5 finished. Total reward: 14.0 (14 timesteps)
Episode 10 finished. Total reward: 20.0 (20 timesteps)
Episode 15 finished. Total reward: 44.0 (44 timesteps)
Episode 20 finished. Total reward: 25.0 (25 timesteps)
Episode 25 finished. Total reward: 20.0 (20 timesteps)
Episode 30 finished. Total reward: 13.0 (13 timesteps)
Episode 35 finished. Total reward: 16.0 (16 timesteps)
Episode 40 finished. Total reward: 19.0 (19 timesteps)
Episode 45 finished. Total reward: 21.0 (21 timesteps)
Episode 50 finished. Total reward: 27.0 (27 timesteps)
Episode 55 finished. Total reward: 11.0 (11 timesteps)
Episode 60 finished. Total reward: 17.0 (17 timesteps)
Episode 65 finished. Total reward: 38.0 (38 timesteps)
Episode 70 finished. Total reward: 39.0 (39 timesteps)
Episode 75 finished. Total reward: 19.0 (19 timesteps)
Episode 80 finished. Total reward: 39.0 (39 timesteps)
Updating the policy…
Updating finished!
Episode 85 finished. Total reward: 40.0 (40 timesteps)
Episode 90 finished. Total reward: 46.0 (46 timesteps)
Episode 95 finished. Total reward: 48.0 (48 timesteps)
```

```
Episode 100 finished. Total reward: 42.0 (42 timesteps)
Episode 105 finished. Total reward: 24.0 (24 timesteps)
Episode 110 finished. Total reward: 11.0 (11 timesteps)
Episode 115 finished. Total reward: 29.0 (29 timesteps)
Episode 120 finished. Total reward: 38.0 (38 timesteps)
Episode 125 finished. Total reward: 28.0 (28 timesteps)
Episode 130 finished. Total reward: 48.0 (48 timesteps)
Episode 135 finished. Total reward: 14.0 (14 timesteps)
Episode 140 finished. Total reward: 20.0 (20 timesteps)
Episode 145 finished. Total reward: 27.0 (27 timesteps)
Episode 150 finished. Total reward: 13.0 (13 timesteps)
Updating the policy…
Updating finished!
Episode 155 finished. Total reward: 18.0 (18 timesteps)
Episode 160 finished. Total reward: 46.0 (46 timesteps)
Episode 165 finished. Total reward: 22.0 (22 timesteps)
Episode 170 finished. Total reward: 47.0 (47 timesteps)
Episode 175 finished. Total reward: 26.0 (26 timesteps)
Episode 180 finished. Total reward: 26.0 (26 timesteps)
Episode 185 finished. Total reward: 61.0 (61 timesteps)
Episode 190 finished. Total reward: 42.0 (42 timesteps)
Episode 195 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 200 finished. Total reward: 18.0 (18 timesteps)
Episode 205 finished. Total reward: 21.0 (21 timesteps)
Episode 210 finished. Total reward: 27.0 (27 timesteps)
Episode 215 finished. Total reward: 100.0 (100 timesteps)
Episode 220 finished. Total reward: 53.0 (53 timesteps)
Episode 225 finished. Total reward: 64.0 (64 timesteps)
Episode 230 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 235 finished. Total reward: 100.0 (100 timesteps)
Episode 240 finished. Total reward: 100.0 (100 timesteps)
Episode 245 finished. Total reward: 100.0 (100 timesteps)
Episode 250 finished. Total reward: 83.0 (83 timesteps)
Updating the policy…
Updating finished!
Episode 255 finished. Total reward: 100.0 (100 timesteps)
Episode 260 finished. Total reward: 100.0 (100 timesteps)
Episode 265 finished. Total reward: 53.0 (53 timesteps)
Episode 270 finished. Total reward: 95.0 (95 timesteps)
Episode 275 finished. Total reward: 40.0 (40 timesteps)
Updating the policy…
Updating finished!
Episode 280 finished. Total reward: 100.0 (100 timesteps)
Episode 285 finished. Total reward: 100.0 (100 timesteps)
```

```
Episode 290 finished. Total reward: 100.0 (100 timesteps)
Episode 295 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 300 finished. Total reward: 100.0 (100 timesteps)
Episode 305 finished. Total reward: 100.0 (100 timesteps)
Episode 310 finished. Total reward: 100.0 (100 timesteps)
Episode 315 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 320 finished. Total reward: 24.0 (24 timesteps)
Episode 325 finished. Total reward: 100.0 (100 timesteps)
Episode 330 finished. Total reward: 100.0 (100 timesteps)
Episode 335 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 340 finished. Total reward: 100.0 (100 timesteps)
Episode 345 finished. Total reward: 100.0 (100 timesteps)
Episode 350 finished. Total reward: 100.0 (100 timesteps)
Episode 355 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 360 finished. Total reward: 100.0 (100 timesteps)
Episode 365 finished. Total reward: 100.0 (100 timesteps)
Episode 370 finished. Total reward: 100.0 (100 timesteps)
Episode 375 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 380 finished. Total reward: 100.0 (100 timesteps)
Episode 385 finished. Total reward: 100.0 (100 timesteps)
Episode 390 finished. Total reward: 100.0 (100 timesteps)
Episode 395 finished. Total reward: 100.0 (100 timesteps)
Episode 400 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 405 finished. Total reward: 100.0 (100 timesteps)
Episode 410 finished. Total reward: 100.0 (100 timesteps)
Episode 415 finished. Total reward: 100.0 (100 timesteps)
Episode 420 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 425 finished. Total reward: 100.0 (100 timesteps)
Episode 430 finished. Total reward: 100.0 (100 timesteps)
Episode 435 finished. Total reward: 100.0 (100 timesteps)
Episode 440 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 445 finished. Total reward: 100.0 (100 timesteps)
```

```
Episode 450 finished. Total reward: 88.0 (88 timesteps)
Episode 455 finished. Total reward: 18.0 (18 timesteps)
Episode 460 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 465 finished. Total reward: 100.0 (100 timesteps)
Episode 470 finished. Total reward: 100.0 (100 timesteps)
Episode 475 finished. Total reward: 100.0 (100 timesteps)
Episode 480 finished. Total reward: 100.0 (100 timesteps)
Episode 485 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 490 finished. Total reward: 100.0 (100 timesteps)
Episode 495 finished. Total reward: 100.0 (100 timesteps)
Episode 500 finished. Total reward: 100.0 (100 timesteps)
Model saved to /notebooks/rl2024/ex1/results/model/CartPole-v1-model2_params.pt
------Training finished.------
Numpy/Torch/Random Seed:  5
Configuration Settings: {'exp_name': 'ex1', 'seed': 5, 'env_name':
'CartPole-v1', 'model_name': 'CartPole-v1-model3', 'max_episode_steps': 100,
'train_episodes': 500, 'batch_size': 64, 'min_update_samples': 2000, 'testing':
False, 'model_path': 'default', 'save_video': True, 'save_model': True,
'save_logging': True, 'silent': False, 'use_wandb': True, 'run_suffix': 0}
Training device: cpu
Observation space dimensions: 4
Action space dimensions: 2

Episode 0 finished. Total reward: 36.0 (36 timesteps)
Episode 5 finished. Total reward: 15.0 (15 timesteps)
Episode 10 finished. Total reward: 46.0 (46 timesteps)
Episode 15 finished. Total reward: 11.0 (11 timesteps)
Episode 20 finished. Total reward: 39.0 (39 timesteps)
Episode 25 finished. Total reward: 18.0 (18 timesteps)
Episode 30 finished. Total reward: 77.0 (77 timesteps)
Episode 35 finished. Total reward: 22.0 (22 timesteps)
Episode 40 finished. Total reward: 18.0 (18 timesteps)
Episode 45 finished. Total reward: 16.0 (16 timesteps)
Episode 50 finished. Total reward: 19.0 (19 timesteps)
Episode 55 finished. Total reward: 15.0 (15 timesteps)
Episode 60 finished. Total reward: 20.0 (20 timesteps)
Episode 65 finished. Total reward: 16.0 (16 timesteps)
Episode 70 finished. Total reward: 38.0 (38 timesteps)
Episode 75 finished. Total reward: 51.0 (51 timesteps)
Updating the policy…
Updating finished!
Episode 80 finished. Total reward: 30.0 (30 timesteps)
Episode 85 finished. Total reward: 24.0 (24 timesteps)
Episode 90 finished. Total reward: 50.0 (50 timesteps)
```

```
Episode 95 finished. Total reward: 23.0 (23 timesteps)
Episode 100 finished. Total reward: 16.0 (16 timesteps)
Episode 105 finished. Total reward: 24.0 (24 timesteps)
Episode 110 finished. Total reward: 30.0 (30 timesteps)
Episode 115 finished. Total reward: 24.0 (24 timesteps)
Episode 120 finished. Total reward: 42.0 (42 timesteps)
Episode 125 finished. Total reward: 16.0 (16 timesteps)
Episode 130 finished. Total reward: 25.0 (25 timesteps)
Episode 135 finished. Total reward: 10.0 (10 timesteps)
Episode 140 finished. Total reward: 22.0 (22 timesteps)
Episode 145 finished. Total reward: 53.0 (53 timesteps)
Updating the policy…
Updating finished!
Episode 150 finished. Total reward: 76.0 (76 timesteps)
Episode 155 finished. Total reward: 38.0 (38 timesteps)
Episode 160 finished. Total reward: 39.0 (39 timesteps)
Episode 165 finished. Total reward: 28.0 (28 timesteps)
Episode 170 finished. Total reward: 77.0 (77 timesteps)
Episode 175 finished. Total reward: 30.0 (30 timesteps)
Episode 180 finished. Total reward: 74.0 (74 timesteps)
Updating the policy…
Updating finished!
Episode 185 finished. Total reward: 66.0 (66 timesteps)
Episode 190 finished. Total reward: 100.0 (100 timesteps)
Episode 195 finished. Total reward: 25.0 (25 timesteps)
Episode 200 finished. Total reward: 72.0 (72 timesteps)
Episode 205 finished. Total reward: 100.0 (100 timesteps)
Episode 210 finished. Total reward: 17.0 (17 timesteps)
Updating the policy…
Updating finished!
Episode 215 finished. Total reward: 29.0 (29 timesteps)
Episode 220 finished. Total reward: 70.0 (70 timesteps)
Episode 225 finished. Total reward: 91.0 (91 timesteps)
Episode 230 finished. Total reward: 72.0 (72 timesteps)
Episode 235 finished. Total reward: 30.0 (30 timesteps)
Updating the policy…
Updating finished!
Episode 240 finished. Total reward: 73.0 (73 timesteps)
Episode 245 finished. Total reward: 96.0 (96 timesteps)
Episode 250 finished. Total reward: 100.0 (100 timesteps)
Episode 255 finished. Total reward: 66.0 (66 timesteps)
Episode 260 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 265 finished. Total reward: 100.0 (100 timesteps)
Episode 270 finished. Total reward: 100.0 (100 timesteps)
Episode 275 finished. Total reward: 100.0 (100 timesteps)
Episode 280 finished. Total reward: 100.0 (100 timesteps)
```

```
Updating the policy…
Updating finished!
Episode 285 finished. Total reward: 100.0 (100 timesteps)
Episode 290 finished. Total reward: 100.0 (100 timesteps)
Episode 295 finished. Total reward: 100.0 (100 timesteps)
Episode 300 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 305 finished. Total reward: 100.0 (100 timesteps)
Episode 310 finished. Total reward: 51.0 (51 timesteps)
Episode 315 finished. Total reward: 95.0 (95 timesteps)
Episode 320 finished. Total reward: 90.0 (90 timesteps)
Updating the policy…
Updating finished!
Episode 325 finished. Total reward: 100.0 (100 timesteps)
Episode 330 finished. Total reward: 100.0 (100 timesteps)
Episode 335 finished. Total reward: 68.0 (68 timesteps)
Episode 340 finished. Total reward: 100.0 (100 timesteps)
Episode 345 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 350 finished. Total reward: 100.0 (100 timesteps)
Episode 355 finished. Total reward: 100.0 (100 timesteps)
Episode 360 finished. Total reward: 100.0 (100 timesteps)
Episode 365 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 370 finished. Total reward: 66.0 (66 timesteps)
Episode 375 finished. Total reward: 100.0 (100 timesteps)
Episode 380 finished. Total reward: 99.0 (99 timesteps)
Episode 385 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 390 finished. Total reward: 100.0 (100 timesteps)
Episode 395 finished. Total reward: 100.0 (100 timesteps)
Episode 400 finished. Total reward: 93.0 (93 timesteps)
Episode 405 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 410 finished. Total reward: 100.0 (100 timesteps)
Episode 415 finished. Total reward: 100.0 (100 timesteps)
Episode 420 finished. Total reward: 100.0 (100 timesteps)
Episode 425 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 430 finished. Total reward: 100.0 (100 timesteps)
Episode 435 finished. Total reward: 100.0 (100 timesteps)
Episode 440 finished. Total reward: 100.0 (100 timesteps)
```

```
Episode 445 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 450 finished. Total reward: 100.0 (100 timesteps)
Episode 455 finished. Total reward: 100.0 (100 timesteps)
Episode 460 finished. Total reward: 100.0 (100 timesteps)
Episode 465 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 470 finished. Total reward: 100.0 (100 timesteps)
Episode 475 finished. Total reward: 100.0 (100 timesteps)
Episode 480 finished. Total reward: 100.0 (100 timesteps)
Episode 485 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 490 finished. Total reward: 100.0 (100 timesteps)
Episode 495 finished. Total reward: 100.0 (100 timesteps)
Episode 500 finished. Total reward: 100.0 (100 timesteps)
Model saved to /notebooks/rl2024/ex1/results/model/CartPole-v1-model3_params.pt
------Training finished.------
Numpy/Torch/Random Seed:  15
Configuration Settings: {'exp_name': 'ex1', 'seed': 15, 'env_name':
'CartPole-v1', 'model_name': 'CartPole-v1-model4', 'max_episode_steps': 100,
'train_episodes': 500, 'batch_size': 64, 'min_update_samples': 2000, 'testing':
False, 'model_path': 'default', 'save_video': True, 'save_model': True,
'save_logging': True, 'silent': False, 'use_wandb': True, 'run_suffix': 0}
Training device: cpu
Observation space dimensions: 4
Action space dimensions: 2

Episode 0 finished. Total reward: 17.0 (17 timesteps)
Episode 5 finished. Total reward: 35.0 (35 timesteps)
Episode 10 finished. Total reward: 24.0 (24 timesteps)
Episode 15 finished. Total reward: 32.0 (32 timesteps)
Episode 20 finished. Total reward: 13.0 (13 timesteps)
Episode 25 finished. Total reward: 12.0 (12 timesteps)
Episode 30 finished. Total reward: 19.0 (19 timesteps)
Episode 35 finished. Total reward: 53.0 (53 timesteps)
Episode 40 finished. Total reward: 41.0 (41 timesteps)
Episode 45 finished. Total reward: 27.0 (27 timesteps)
Episode 50 finished. Total reward: 60.0 (60 timesteps)
Episode 55 finished. Total reward: 12.0 (12 timesteps)
Episode 60 finished. Total reward: 18.0 (18 timesteps)
Episode 65 finished. Total reward: 27.0 (27 timesteps)
Episode 70 finished. Total reward: 34.0 (34 timesteps)
Episode 75 finished. Total reward: 24.0 (24 timesteps)
Episode 80 finished. Total reward: 19.0 (19 timesteps)
Episode 85 finished. Total reward: 24.0 (24 timesteps)
```

```
Updating the policy…
Updating finished!
Episode 90 finished. Total reward: 29.0 (29 timesteps)
Episode 95 finished. Total reward: 15.0 (15 timesteps)
Episode 100 finished. Total reward: 22.0 (22 timesteps)
Episode 105 finished. Total reward: 11.0 (11 timesteps)
Episode 110 finished. Total reward: 24.0 (24 timesteps)
Episode 115 finished. Total reward: 23.0 (23 timesteps)
Episode 120 finished. Total reward: 67.0 (67 timesteps)
Episode 125 finished. Total reward: 22.0 (22 timesteps)
Episode 130 finished. Total reward: 19.0 (19 timesteps)
Episode 135 finished. Total reward: 13.0 (13 timesteps)
Episode 140 finished. Total reward: 14.0 (14 timesteps)
Episode 145 finished. Total reward: 29.0 (29 timesteps)
Episode 150 finished. Total reward: 17.0 (17 timesteps)
Episode 155 finished. Total reward: 41.0 (41 timesteps)
Episode 160 finished. Total reward: 38.0 (38 timesteps)
Updating the policy…
Updating finished!
Episode 165 finished. Total reward: 67.0 (67 timesteps)
Episode 170 finished. Total reward: 52.0 (52 timesteps)
Episode 175 finished. Total reward: 16.0 (16 timesteps)
Episode 180 finished. Total reward: 82.0 (82 timesteps)
Episode 185 finished. Total reward: 55.0 (55 timesteps)
Episode 190 finished. Total reward: 32.0 (32 timesteps)
Episode 195 finished. Total reward: 39.0 (39 timesteps)
Episode 200 finished. Total reward: 31.0 (31 timesteps)
Episode 205 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 210 finished. Total reward: 100.0 (100 timesteps)
Episode 215 finished. Total reward: 100.0 (100 timesteps)
Episode 220 finished. Total reward: 100.0 (100 timesteps)
Episode 225 finished. Total reward: 84.0 (84 timesteps)
Episode 230 finished. Total reward: 58.0 (58 timesteps)
Episode 235 finished. Total reward: 25.0 (25 timesteps)
Episode 240 finished. Total reward: 63.0 (63 timesteps)
Updating the policy…
Updating finished!
Episode 245 finished. Total reward: 100.0 (100 timesteps)
Episode 250 finished. Total reward: 100.0 (100 timesteps)
Episode 255 finished. Total reward: 55.0 (55 timesteps)
Episode 260 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 265 finished. Total reward: 100.0 (100 timesteps)
Episode 270 finished. Total reward: 100.0 (100 timesteps)
Episode 275 finished. Total reward: 100.0 (100 timesteps)
```

```
Episode 280 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 285 finished. Total reward: 94.0 (94 timesteps)
Episode 290 finished. Total reward: 100.0 (100 timesteps)
Episode 295 finished. Total reward: 87.0 (87 timesteps)
Episode 300 finished. Total reward: 100.0 (100 timesteps)
Episode 305 finished. Total reward: 46.0 (46 timesteps)
Updating the policy…
Updating finished!
Episode 310 finished. Total reward: 100.0 (100 timesteps)
Episode 315 finished. Total reward: 100.0 (100 timesteps)
Episode 320 finished. Total reward: 100.0 (100 timesteps)
Episode 325 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 330 finished. Total reward: 88.0 (88 timesteps)
Episode 335 finished. Total reward: 100.0 (100 timesteps)
Episode 340 finished. Total reward: 53.0 (53 timesteps)
Episode 345 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 350 finished. Total reward: 100.0 (100 timesteps)
Episode 355 finished. Total reward: 100.0 (100 timesteps)
Episode 360 finished. Total reward: 100.0 (100 timesteps)
Episode 365 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 370 finished. Total reward: 100.0 (100 timesteps)
Episode 375 finished. Total reward: 100.0 (100 timesteps)
Episode 380 finished. Total reward: 100.0 (100 timesteps)
Episode 385 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 390 finished. Total reward: 100.0 (100 timesteps)
Episode 395 finished. Total reward: 100.0 (100 timesteps)
Episode 400 finished. Total reward: 100.0 (100 timesteps)
Episode 405 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 410 finished. Total reward: 100.0 (100 timesteps)
Episode 415 finished. Total reward: 100.0 (100 timesteps)
Episode 420 finished. Total reward: 100.0 (100 timesteps)
Episode 425 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 430 finished. Total reward: 100.0 (100 timesteps)
Episode 435 finished. Total reward: 100.0 (100 timesteps)
```

```
Episode 440 finished. Total reward: 100.0 (100 timesteps)
Episode 445 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 450 finished. Total reward: 100.0 (100 timesteps)
Episode 455 finished. Total reward: 100.0 (100 timesteps)
Episode 460 finished. Total reward: 100.0 (100 timesteps)
Episode 465 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 470 finished. Total reward: 100.0 (100 timesteps)
Episode 475 finished. Total reward: 100.0 (100 timesteps)
Episode 480 finished. Total reward: 100.0 (100 timesteps)
Episode 485 finished. Total reward: 100.0 (100 timesteps)
Updating the policy…
Updating finished!
Episode 490 finished. Total reward: 100.0 (100 timesteps)
Episode 495 finished. Total reward: 100.0 (100 timesteps)
Episode 500 finished. Total reward: 100.0 (100 timesteps)
Model saved to /notebooks/rl2024/ex1/results/model/CartPole-v1-model4_params.pt
------Training finished.------
```

Do not delete the cell below as it is used for grading.

[17]:
```
%%capture --no-stdout
test_seeds = [700, 800, 900, 1000, 2000]

for i in range(5):
    for seed in test_seeds:
        assert t.test(episodes=1, cfg_path=Path().cwd()/'cfg'/'cartpole_v1.
 ↪yaml',
            cfg_args=dict(testing=True, seed=seed,save_video=False,␣
 ↪save_logging=False,
            save_model=False, max_episode_steps=1000,␣
 ↪model_name=str("CartPole-v1-model"+str(i)))) > 100
```

```
Numpy/Torch/Random Seed:  700
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model0_params.pt …
Testing…
Test ep reward: 139.0 seed: 700
Average test reward: 139.0 episode length: 139.0
Numpy/Torch/Random Seed:  800
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model0_params.pt …
Testing…
Test ep reward: 129.0 seed: 800
Average test reward: 129.0 episode length: 129.0
```

```
Numpy/Torch/Random Seed:  900
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model0_params.pt …
Testing…
Test ep reward: 132.0 seed: 900
Average test reward: 132.0 episode length: 132.0
Numpy/Torch/Random Seed:  1000
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model0_params.pt …
Testing…
Test ep reward: 130.0 seed: 1000
Average test reward: 130.0 episode length: 130.0
Numpy/Torch/Random Seed:  2000
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model0_params.pt …
Testing…
Test ep reward: 123.0 seed: 2000
Average test reward: 123.0 episode length: 123.0
Numpy/Torch/Random Seed:  700
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model1_params.pt …
Testing…
Test ep reward: 133.0 seed: 700
Average test reward: 133.0 episode length: 133.0
Numpy/Torch/Random Seed:  800
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model1_params.pt …
Testing…
Test ep reward: 147.0 seed: 800
Average test reward: 147.0 episode length: 147.0
Numpy/Torch/Random Seed:  900
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model1_params.pt …
Testing…
Test ep reward: 137.0 seed: 900
Average test reward: 137.0 episode length: 137.0
Numpy/Torch/Random Seed:  1000
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model1_params.pt …
Testing…
Test ep reward: 149.0 seed: 1000
Average test reward: 149.0 episode length: 149.0
Numpy/Torch/Random Seed:  2000
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model1_params.pt …
Testing…
Test ep reward: 162.0 seed: 2000
Average test reward: 162.0 episode length: 162.0
```

```
Numpy/Torch/Random Seed:  700
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model2_params.pt …
Testing…
Test ep reward: 124.0 seed: 700
Average test reward: 124.0 episode length: 124.0
Numpy/Torch/Random Seed:  800
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model2_params.pt …
Testing…
Test ep reward: 129.0 seed: 800
Average test reward: 129.0 episode length: 129.0
Numpy/Torch/Random Seed:  900
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model2_params.pt …
Testing…
Test ep reward: 130.0 seed: 900
Average test reward: 130.0 episode length: 130.0
Numpy/Torch/Random Seed:  1000
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model2_params.pt …
Testing…
Test ep reward: 128.0 seed: 1000
Average test reward: 128.0 episode length: 128.0
Numpy/Torch/Random Seed:  2000
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model2_params.pt …
Testing…
Test ep reward: 138.0 seed: 2000
Average test reward: 138.0 episode length: 138.0
Numpy/Torch/Random Seed:  700
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model3_params.pt …
Testing…
Test ep reward: 164.0 seed: 700
Average test reward: 164.0 episode length: 164.0
Numpy/Torch/Random Seed:  800
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model3_params.pt …
Testing…
Test ep reward: 183.0 seed: 800
Average test reward: 183.0 episode length: 183.0
Numpy/Torch/Random Seed:  900
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model3_params.pt …
Testing…
Test ep reward: 171.0 seed: 900
Average test reward: 171.0 episode length: 171.0
```

```
Numpy/Torch/Random Seed:  1000
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model3_params.pt …
Testing…
Test ep reward: 181.0 seed: 1000
Average test reward: 181.0 episode length: 181.0
Numpy/Torch/Random Seed:  2000
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model3_params.pt …
Testing…
Test ep reward: 189.0 seed: 2000
Average test reward: 189.0 episode length: 189.0
Numpy/Torch/Random Seed:  700
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model4_params.pt …
Testing…
Test ep reward: 387.0 seed: 700
Average test reward: 387.0 episode length: 387.0
Numpy/Torch/Random Seed:  800
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model4_params.pt …
Testing…
Test ep reward: 1000.0 seed: 800
Average test reward: 1000.0 episode length: 1000.0
Numpy/Torch/Random Seed:  900
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model4_params.pt …
Testing…
Test ep reward: 316.0 seed: 900
Average test reward: 316.0 episode length: 316.0
Numpy/Torch/Random Seed:  1000
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model4_params.pt …
Testing…
Test ep reward: 390.0 seed: 1000
Average test reward: 390.0 episode length: 390.0
Numpy/Torch/Random Seed:  2000
Loading model from
/notebooks/rl2024/ex1/results/model/CartPole-v1-model4_params.pt …
Testing…
Test ep reward: 771.0 seed: 2000
Average test reward: 771.0 episode length: 771.0
```

DOUBLE CLICK HERE TO EDIT, CLEAR THIS TEXT AND REPORT HERE

`<h3><b>Student Question 2.1</b> Analysis of Training Performance (15 points) </h3>`

Are the behavior and performance of the trained models the same every time? Why/why not? Analyze the causes briefly. Hint: the random seed initializes the weights and the environment

settings randomly.

Table of Contents

Select all the correct answers. You can select no more than 6 answers, otherwise you lose all exercise points.

1. Yes, because the model has been optimized to always reach the highest possible reward regardless of random seeds or initial conditions.
2. No, because the training dynamics and the evaluating dynamics are not the same
3. No, because the trained model has stochastic behaviour which may perform differently in different environments
4. No, the behaviour and performance are not the same, as shown by the different average rewards in previous Task.
5. No, because there's a lot of stochasticity involved in the training process
6. No, because agent explores with random actions during training
7. No, because policy is initially randomly initialised with random weights, gradient updates of the policy are noisy.
8. No, because the environment may be stochastic: agent is randomly initialised in beginning of each episode, (transitions from one state to next one may follow a probability distribution)
9. Yes, all machine learning models perform consistently after training due to the deterministic nature of their algorithms.

[18]:
```
sq2_1 = [3, 4, 5, 6, 7, 8] #Answer question 2.1 with the appropriate answer␣
↪numbers
```

[19]:
```
assert 1 <= len(set(sq2_1)) <= 6
assert set(sq2_1) < set(range(1, 10))
```

Do not remove the following blocks as they are used for grading

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

<h3>Student Question 2.2 Stochasticity (10 points) </h3>

What are the implications of this stochasticity, when it comes to comparing reinforcement learning algorithms to each other? Please explain.

Table of Contents

Select the correct answer

1. It is only necessary to compare the maximum rewards achieved by each algorithm to determine the best one.
2. Stochasticity requires that algorithms be trained and evaluated multiple times to obtain a reliable estimate of performance, including measures of variance besides the average reward.
3. Algorithms should be trained once under identical conditions to ensure a fair comparison, focusing on the consistency of the training process.
4. The best approach is to evaluate algorithms based on their performance in a single, well-designed test to avoid the confounding effects of random variability.
5. To effectively compare algorithms, one should calculate the median reward from multiple trials, as this is the most robust measure against outliers.

[20]:
```
sq2_2 = 2 #Answer question 2.2 with the appropriate answer number
```

[21]:
```
assert sq2_2 in range(1,6)
```

The following cells are used for grading

[ ]:

[ ]:

4 Reacher

Now we will focus on designing a reward function for a different environment, the Reacher environment, where a two-joint manipulator needs to reach a goal (see Figure 2).

Figure 2: The Reacher environment

The Cartesian (x, y) position of the end-effector of the manipulator can be determined following the equation:

$$x = L_1 \sin(\theta_0) + L_2 \sin(\theta_0 + \theta_1) y = -L_1 \cos(\theta_0) - L_2 \cos(\theta_0 + \theta_1)$$

where $L1 = 1$, $L2 = 1$ are the lengths, and θ_0, θ_1 the joint angles of the first and second links respectively. The state (and observation) in this environment is the two element vector:

$$o = s = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix},$$

The action space now consists of 5 "options"; 4 correspond rotating the first/second joint left/right, and the final one performs no motion at all (the configuration doesn't change). The episode terminates when the agent reaches the target position, marked in red. Now, let us design a custom reward function and use it for training the RL agent.

```
<h3><b>Student Task 3.</b> Reward Functions (20 points) </h3>
```

Below two classes are shown that modify the reward function of the reacher function provided in `reacher.py`. Edit the function `get_reward` *below* (not in `reacher.py`) in both classes. For each class, write a reward function to incentivise the agent to learn the following behaviors:

Class 1) `SpinningReacherEnv`: Keep the manipulator rotating clockwise continuously (w.r.t. angle _0). You can use a lower number of training episodes for this, e.g. train(cfg_args=dict(env_name='SpinningReacher-v0', train_episodes=200), overrides=['env=reacher_v1'])

Class 2) `TargetReacherEnv`: Reach the goal point located in x = [1.0,1.0] (marked in red). Use at least 500 training episodes.

Train one model for each behavior.

Hint: Use the observation vector to get the quantities required to compute the new reward (such as the position of the manipulator). You can get the Cartesian position of the end-effector with `self.get_cartesian_pos(state)`.

Table of Contents

[22]:
```
from reacher import ReacherEnv
from typing import Optional
from gymnasium.envs.registration import register

class SpinningReacherEnv(ReacherEnv):
    def __init__(self, render_mode: Optional[str] = None,␣
 ↪max_episode_steps=200):
        super().__init__(render_mode=render_mode,␣
 ↪max_episode_steps=max_episode_steps)

    def get_reward(self, prev_state, action, next_state):
        '''
        TODO: Task 3: Implement and test the first reward function
        '''
        ########## Your code starts here ##########
         # Extract angles from previous and next states
        prev_theta0, _ = prev_state
        next_theta0, _ = next_state

        # Calculate the change in theta0
        delta_theta0 = next_theta0 - prev_theta0

        # Reward positive changes in theta0, indicating clockwise rotation
        reward = delta_theta0

        return reward
        ########## Your codes end here ##########


register("SpinningReacher-v0",
```

```
        entry_point="%s:SpinningReacherEnv"%__name__,
        max_episode_steps=200)

class TargetReacherEnv(ReacherEnv):
    def __init__(self, render_mode: Optional[str] = None,␣
 ↪max_episode_steps=200):
        super().__init__(render_mode=render_mode,␣
 ↪max_episode_steps=max_episode_steps)

    def get_reward(self, prev_state, action, next_state):
        '''
        # TODO: Task 3: Implement and test the second reward function
        '''
        ########## Your code starts here ##########
        # Extract angles from the state
        theta0, theta1 = next_state

        # Get the Cartesian position of the end-effector
        x, y = self.get_cartesian_pos(next_state)

        # Define the target position
        target_x, target_y = 1.0, 1.0

        # Calculate the Euclidean distance to the target
        distance = np.sqrt((x - target_x) ** 2 + (y - target_y) ** 2)

        # Reward is the negative distance (closer = higher reward)
        reward = -distance

        return reward
        ########## Your codes end here ##########

register("TargetReacher-v0",
        entry_point="%s:TargetReacherEnv"%__name__,
        max_episode_steps=200)
```

[23]:
```
if not skip_training:
      t.train(cfg_path=Path().cwd()/'cfg'/'reacher_v1.yaml',
    ␣
 ↪cfg_args=dict(env_name='SpinningReacher-v0',model_name='SpinningReacher-v0',␣
 ↪train_episodes=200, seed=1)) # < 5 mins
```

```
Numpy/Torch/Random Seed:  1
Configuration Settings: {'exp_name': 'ex1', 'seed': 1, 'env_name':
'SpinningReacher-v0', 'model_name': 'SpinningReacher-v0', 'max_episode_steps':
200, 'train_episodes': 200, 'batch_size': 64, 'min_update_samples': 2000,
'testing': False, 'model_path': 'default', 'save_video': True, 'save_model':
```

```
True, 'save_logging': True, 'silent': False, 'use_wandb': True, 'run_suffix': 0}
Training device: cpu
Observation space dimensions: 2
Action space dimensions: 5

Episode 0 finished. Total reward: 0.9999997168779373 (200 timesteps)
Episode 5 finished. Total reward: -4.199996635317802 (200 timesteps)
Episode 10 finished. Total reward: -0.20000001788139343 (200 timesteps)
Updating the policy…
Updating finished!
Episode 15 finished. Total reward: 0.2000000774860382 (200 timesteps)
Episode 20 finished. Total reward: 3.9999969750642776 (200 timesteps)
Updating the policy…
Updating finished!
Episode 25 finished. Total reward: 13.400116756558418 (200 timesteps)
Episode 30 finished. Total reward: 2.999997928738594 (121 timesteps)
Updating the policy…
Updating finished!
Episode 35 finished. Total reward: 20.600281551480293 (200 timesteps)
Episode 40 finished. Total reward: 1.5999991446733475 (33 timesteps)
Updating the policy…
Updating finished!
Episode 45 finished. Total reward: 14.200135067105293 (152 timesteps)
Episode 50 finished. Total reward: 14.200135067105293 (109 timesteps)
Episode 55 finished. Total reward: 15.600167110562325 (174 timesteps)
Updating the policy…
Updating finished!
Episode 60 finished. Total reward: 28.600464656949043 (200 timesteps)
Episode 65 finished. Total reward: 7.999993160367012 (77 timesteps)
Updating the policy…
Updating finished!
Episode 70 finished. Total reward: 28.000450924038887 (200 timesteps)
Episode 75 finished. Total reward: 26.20040972530842 (200 timesteps)
Updating the policy…
Updating finished!
Episode 80 finished. Total reward: 14.200135067105293 (99 timesteps)
Episode 85 finished. Total reward: 31.800537899136543 (200 timesteps)
Episode 90 finished. Total reward: 28.000450924038887 (200 timesteps)
Updating the policy…
Updating finished!
Episode 95 finished. Total reward: 31.400528743863106 (200 timesteps)
Episode 100 finished. Total reward: 31.800537899136543 (200 timesteps)
Updating the policy…
Updating finished!
Episode 105 finished. Total reward: 28.200455501675606 (164 timesteps)
Episode 110 finished. Total reward: 34.600601986050606 (200 timesteps)
Updating the policy…
Updating finished!
```

```
Episode 115 finished. Total reward: 35.40062029659748 (200 timesteps)
Episode 120 finished. Total reward: 35.6006248742342 (200 timesteps)
Updating the policy…
Updating finished!
Episode 125 finished. Total reward: 37.60067065060139 (200 timesteps)
Episode 130 finished. Total reward: 37.40066607296467 (200 timesteps)
Updating the policy…
Updating finished!
Episode 135 finished. Total reward: 38.000679805874825 (200 timesteps)
Episode 140 finished. Total reward: 38.20068438351154 (200 timesteps)
Updating the policy…
Updating finished!
Episode 145 finished. Total reward: 39.400711849331856 (200 timesteps)
Episode 150 finished. Total reward: 39.20070727169514 (200 timesteps)
Updating the policy…
Updating finished!
Episode 155 finished. Total reward: 39.400711849331856 (200 timesteps)
Episode 160 finished. Total reward: 39.20070727169514 (200 timesteps)
Updating the policy…
Updating finished!
Episode 165 finished. Total reward: 39.400711849331856 (200 timesteps)
Episode 170 finished. Total reward: 39.600716426968575 (200 timesteps)
Updating the policy…
Updating finished!
Episode 175 finished. Total reward: 39.80072100460529 (200 timesteps)
Episode 180 finished. Total reward: 39.600716426968575 (200 timesteps)
Updating the policy…
Updating finished!
Episode 185 finished. Total reward: 40.00072558224201 (200 timesteps)
Episode 190 finished. Total reward: 39.80072100460529 (200 timesteps)
Updating the policy…
Updating finished!
Episode 195 finished. Total reward: 40.00072558224201 (200 timesteps)
Episode 200 finished. Total reward: 39.600716426968575 (200 timesteps)
Model saved to /notebooks/rl2024/ex1/results/model/SpinningReacher-v0_params.pt
------Training finished.------
```

[24]:
```
if not skip_training:
       t.test(episodes=10, cfg_path=Path().cwd()/'cfg'/'reacher_v1.yaml',
      ␣
 ↪cfg_args=dict(env_name='SpinningReacher-v0',model_name='SpinningReacher-v0',␣
 ↪testing=True,seed=None))
```

```
Numpy/Torch/Random Seed:  60
Loading model from
/notebooks/rl2024/ex1/results/model/SpinningReacher-v0_params.pt …
Testing…
Test ep reward: 40.000725865364075 seed: 206
```

```
Test ep reward: 40.00072532892227 seed: 66
Test ep reward: 40.000723481178284 seed: 583
Test ep reward: 40.00072096288204 seed: 395
Test ep reward: 40.000724375247955 seed: 400
Test ep reward: 40.000725984573364 seed: 100
Test ep reward: 40.00072507560253 seed: 73
Test ep reward: 40.00072538852692 seed: 122
Test ep reward: 40.000723734498024 seed: 914
Test ep reward: 40.00072155892849 seed: 675
Average test reward: 40.000724175572394 episode length: 200.0
```

The agent acting in the environment can be seen using the following command. Change the `path` to pick the episode you want to visualize. Bear in mind by default video saving for training is taken every 50 episodes.

[27]:
```
if not skip_training:
      video = Video(work_dir/'video'/'SpinningReacher-v0'/'test'/
 ↪f'ex1-episode-0.mp4',
      embed=True, html_attributes="loop autoplay") # Set␣
 ↪html_attributes="controls" for video control
      display(video)
```

```
<IPython.core.display.Video object>
```

[28]:
```
if not skip_training:
      t.train(cfg_path=Path().cwd()/'cfg'/'reacher_v1.yaml',
      cfg_args=dict(env_name='TargetReacher-v0',model_name='TargetReacher-v0',␣
 ↪train_episodes=500, seed=1)) # < 5 mins
```

```
Numpy/Torch/Random Seed:  1
Configuration Settings: {'exp_name': 'ex1', 'seed': 1, 'env_name':
'TargetReacher-v0', 'model_name': 'TargetReacher-v0', 'max_episode_steps': 200,
'train_episodes': 500, 'batch_size': 64, 'min_update_samples': 2000, 'testing':
False, 'model_path': 'default', 'save_video': True, 'save_model': True,
'save_logging': True, 'silent': False, 'use_wandb': True, 'run_suffix': 0}
Training device: cpu
Observation space dimensions: 2
Action space dimensions: 5

Episode 0 finished. Total reward: -571.9537805008489 (200 timesteps)
Episode 5 finished. Total reward: -533.6843759806541 (200 timesteps)
Episode 10 finished. Total reward: -604.272767413936 (200 timesteps)
Updating the policy…
Updating finished!
Episode 15 finished. Total reward: -80.1796538766261 (40 timesteps)
Episode 20 finished. Total reward: -369.87731162214294 (200 timesteps)
Updating the policy…
Updating finished!
```

```
Episode 25 finished. Total reward: -394.4414388304313 (200 timesteps)
Episode 30 finished. Total reward: -378.48743626640504 (200 timesteps)
Updating the policy…
Updating finished!
Episode 35 finished. Total reward: -373.0086270680964 (200 timesteps)
Episode 40 finished. Total reward: -233.26818804176827 (134 timesteps)
Episode 45 finished. Total reward: -60.5237512490564 (31 timesteps)
Episode 50 finished. Total reward: -80.45341425616046 (44 timesteps)
Updating the policy…
Updating finished!
Episode 55 finished. Total reward: -35.80822392924983 (21 timesteps)
Episode 60 finished. Total reward: -151.32287389355596 (105 timesteps)
Episode 65 finished. Total reward: -82.79113477770156 (37 timesteps)
Episode 70 finished. Total reward: -325.3955444048953 (179 timesteps)
Updating the policy…
Updating finished!
Episode 75 finished. Total reward: -129.70303411433866 (85 timesteps)
Episode 80 finished. Total reward: -382.3229798029909 (189 timesteps)
Episode 85 finished. Total reward: -86.74025455218394 (38 timesteps)
Episode 90 finished. Total reward: -208.23570339057187 (111 timesteps)
Episode 95 finished. Total reward: -299.91780581086226 (164 timesteps)
Updating the policy…
Updating finished!
Episode 100 finished. Total reward: -367.3348305590615 (200 timesteps)
Episode 105 finished. Total reward: -34.85407608561298 (25 timesteps)
Episode 110 finished. Total reward: -357.7723130233212 (200 timesteps)
Episode 115 finished. Total reward: -91.7588256271416 (64 timesteps)
Episode 120 finished. Total reward: -273.353337928596 (169 timesteps)
Episode 125 finished. Total reward: -280.5299344373903 (163 timesteps)
Updating the policy…
Updating finished!
Episode 130 finished. Total reward: -222.27640157065375 (139 timesteps)
Episode 135 finished. Total reward: -365.1862383396505 (200 timesteps)
Episode 140 finished. Total reward: -51.60914469248095 (30 timesteps)
Episode 145 finished. Total reward: -338.49758369673265 (200 timesteps)
Episode 150 finished. Total reward: -39.6645531160769 (21 timesteps)
Episode 155 finished. Total reward: -35.86884124642807 (19 timesteps)
Updating the policy…
Updating finished!
Episode 160 finished. Total reward: -30.80251044942633 (17 timesteps)
Episode 165 finished. Total reward: -43.05902316159371 (25 timesteps)
Episode 170 finished. Total reward: -199.54354881468694 (112 timesteps)
Episode 175 finished. Total reward: -42.30615421536371 (21 timesteps)
Episode 180 finished. Total reward: -230.88535174481638 (142 timesteps)
Episode 185 finished. Total reward: -221.0472965146015 (126 timesteps)
Updating the policy…
Updating finished!
Episode 190 finished. Total reward: -35.838449292640334 (27 timesteps)
```

```
Episode 195 finished. Total reward: -47.84226076980664 (30 timesteps)
Episode 200 finished. Total reward: -298.46125389386214 (175 timesteps)
Episode 205 finished. Total reward: -30.488648940288062 (18 timesteps)
Episode 210 finished. Total reward: -32.43702402043547 (19 timesteps)
Episode 215 finished. Total reward: -61.46050258568872 (47 timesteps)
Updating the policy…
Updating finished!
Episode 220 finished. Total reward: -28.14286054331255 (16 timesteps)
Episode 225 finished. Total reward: -32.97550276825772 (19 timesteps)
Episode 230 finished. Total reward: -68.2774312735156 (52 timesteps)
Episode 235 finished. Total reward: -25.150060366048336 (15 timesteps)
Episode 240 finished. Total reward: -348.1143851654531 (200 timesteps)
Updating the policy…
Updating finished!
Episode 245 finished. Total reward: -78.39209810765745 (57 timesteps)
Episode 250 finished. Total reward: -355.72393519355114 (200 timesteps)
Episode 255 finished. Total reward: -45.29884072513697 (24 timesteps)
Episode 260 finished. Total reward: -26.612393535601136 (17 timesteps)
Episode 265 finished. Total reward: -27.73773619773191 (16 timesteps)
Updating the policy…
Updating finished!
Episode 270 finished. Total reward: -24.863364770094336 (16 timesteps)
Episode 275 finished. Total reward: -78.80020032225862 (62 timesteps)
Episode 280 finished. Total reward: -27.957475086126927 (16 timesteps)
Episode 285 finished. Total reward: -338.8562585659363 (200 timesteps)
Episode 290 finished. Total reward: -32.519009883618736 (21 timesteps)
Episode 295 finished. Total reward: -348.86823686322725 (200 timesteps)
Episode 300 finished. Total reward: -29.848933722007317 (19 timesteps)
Episode 305 finished. Total reward: -34.10132082024734 (21 timesteps)
Updating the policy…
Updating finished!
Episode 310 finished. Total reward: -24.5707829614306 (14 timesteps)
Episode 315 finished. Total reward: -67.85193572983357 (48 timesteps)
Episode 320 finished. Total reward: -28.76647837238569 (19 timesteps)
Episode 325 finished. Total reward: -37.43696515319926 (20 timesteps)
Episode 330 finished. Total reward: -28.279336483358463 (18 timesteps)
Episode 335 finished. Total reward: -40.156430425456065 (23 timesteps)
Episode 340 finished. Total reward: -157.4480321184227 (96 timesteps)
Episode 345 finished. Total reward: -26.602077847704916 (16 timesteps)
Episode 350 finished. Total reward: -25.18363684891801 (14 timesteps)
Updating the policy…
Updating finished!
Episode 355 finished. Total reward: -361.8265372343272 (200 timesteps)
Episode 360 finished. Total reward: -188.95992269848438 (100 timesteps)
Episode 365 finished. Total reward: -27.995677348774255 (18 timesteps)
Episode 370 finished. Total reward: -34.04309344852075 (18 timesteps)
Episode 375 finished. Total reward: -26.99371974246626 (15 timesteps)
Episode 380 finished. Total reward: -39.02818868254301 (20 timesteps)
```

```
Episode 385 finished. Total reward: -32.21459867668866 (18 timesteps)
Updating the policy…
Updating finished!
Episode 390 finished. Total reward: -27.44875967037614 (17 timesteps)
Episode 395 finished. Total reward: -197.8764987177071 (107 timesteps)
Episode 400 finished. Total reward: -26.872320754655696 (20 timesteps)
Episode 405 finished. Total reward: -334.41775388139894 (200 timesteps)
Episode 410 finished. Total reward: -32.260479806509785 (18 timesteps)
Episode 415 finished. Total reward: -32.507442517110384 (20 timesteps)
Episode 420 finished. Total reward: -24.796633858316344 (15 timesteps)
Episode 425 finished. Total reward: -198.9968210093196 (126 timesteps)
Episode 430 finished. Total reward: -28.215221855267405 (19 timesteps)
Episode 435 finished. Total reward: -347.24742005671413 (200 timesteps)
Episode 440 finished. Total reward: -33.60766783012935 (20 timesteps)
Updating the policy…
Updating finished!
Episode 445 finished. Total reward: -358.4500074227465 (200 timesteps)
Episode 450 finished. Total reward: -31.695111343516654 (18 timesteps)
Episode 455 finished. Total reward: -26.55483262913745 (16 timesteps)
Episode 460 finished. Total reward: -372.22721574912595 (200 timesteps)
Episode 465 finished. Total reward: -27.91654066424372 (17 timesteps)
Episode 470 finished. Total reward: -70.28367201682471 (62 timesteps)
Episode 475 finished. Total reward: -25.974489253201014 (15 timesteps)
Episode 480 finished. Total reward: -25.405948721391834 (16 timesteps)
Episode 485 finished. Total reward: -23.38993901562231 (14 timesteps)
Updating the policy…
Updating finished!
Episode 490 finished. Total reward: -30.086791020565375 (19 timesteps)
Episode 495 finished. Total reward: -30.35155274741469 (16 timesteps)
Episode 500 finished. Total reward: -356.0189181553706 (200 timesteps)
Model saved to /notebooks/rl2024/ex1/results/model/TargetReacher-v0_params.pt
------Training finished.------
```

[29]:
```
if not skip_training:
       t.test(episodes=10, cfg_path=Path().cwd()/'cfg'/'reacher_v1.yaml',
       cfg_args=dict(env_name='TargetReacher-v0',model_name='TargetReacher-v0',␣
 ↪seed=None, testing=True,))
```

```
Numpy/Torch/Random Seed:  519
Loading model from
/notebooks/rl2024/ex1/results/model/TargetReacher-v0_params.pt …
Testing…
Test ep reward: -27.82880207688945 seed: 25
Test ep reward: -26.783486480853362 seed: 466
Test ep reward: -26.184199947698463 seed: 374
Test ep reward: -24.769934864094974 seed: 517
Test ep reward: -24.320561223924084 seed: 875
Test ep reward: -24.613443126057277 seed: 180
```

```
Test ep reward: -26.138254807258836 seed: 20
Test ep reward: -27.78487429442449 seed: 519
Test ep reward: -25.025835976465842 seed: 873
Test ep reward: -23.337915984927395 seed: 768
Average test reward: -25.678730878259422 episode length: 15.3
```

[30]:
```
if not skip_training:
      video = Video(work_dir/'video'/'TargetReacher-v0'/'test'/f'ex1-episode-0.
 ↪mp4',
      embed=True, html_attributes="loop autoplay") # Set␣
 ↪html_attributes="controls" for video control
      display(video)
```

```
<IPython.core.display.Video object>
```

Do not delete the following cells as they are used for grading. Make sure to test that both agents work correctly for the selected set of seeds [860, 241, 73, 543, 582, 211, 524, 484, 954, 656].

[ ]:
```
%%capture --no-stdout
"TEST CELL"
```

[ ]:
```
%%capture --no-stdout
"TEST CELL"
```

```
<h3><b>Student Task 4.</b> Visualizing Behavior (10 points) </h3>
```

Now, let us visualize the reward function for the second behavior (reaching the goal [1,1]). Plot the values of the second reward function from Task 3 and the learned best action as a function of the state (the joint positions). Use the code below as a starting point. After plotting, answer the questions below.

Table of Contents

[31]:
```
import matplotlib.pyplot as plt
import seaborn as sns
import gymnasium as gym
from agent import Agent, Policy
```

[32]:
```
env_name = "TargetReacher-v0"
resolution = 101  # Resolution of the policy/reward image

# Load policy from default path to plot
policy_dir = Path().cwd()/'results'/'model'/f'{env_name}_params.pt'

sns.set()

# Create a gym environment
env = gym.make(env_name)
```

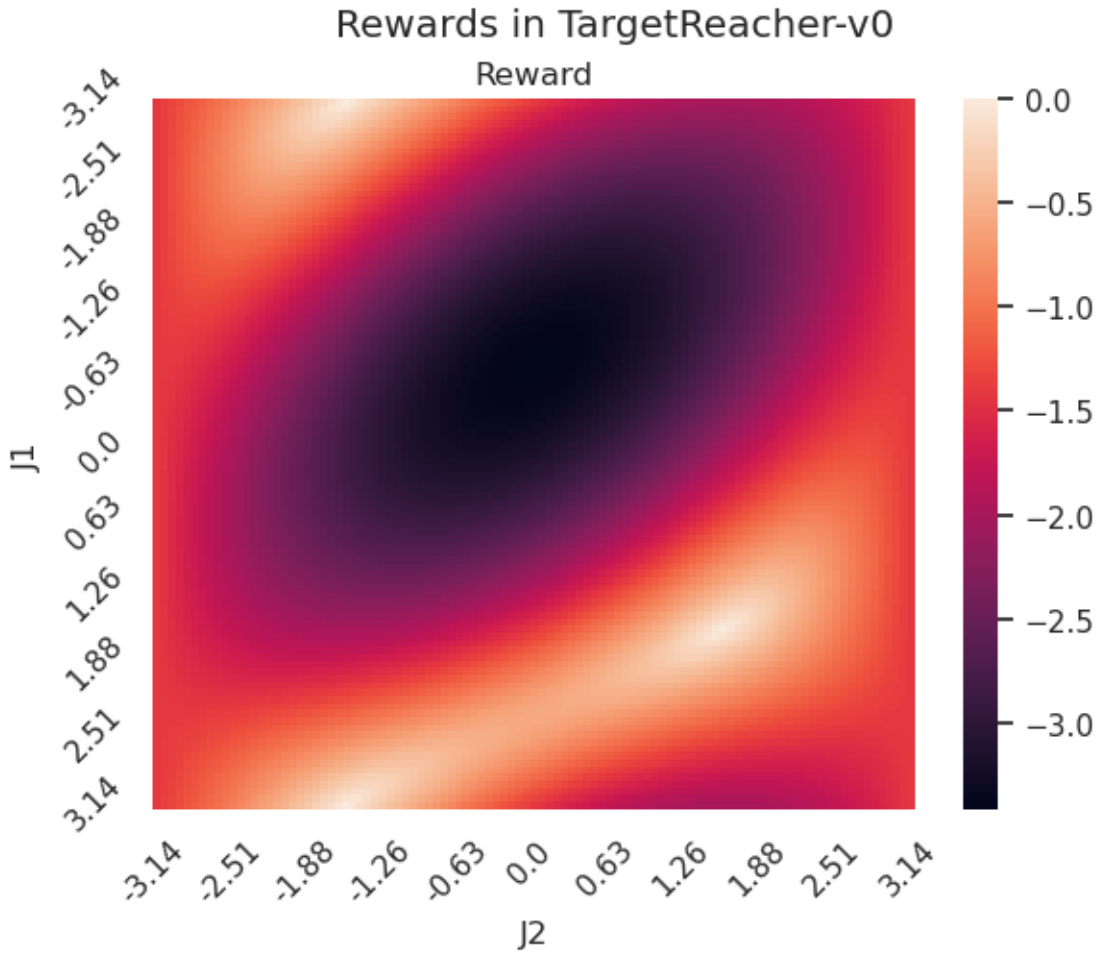
```
action_space_dim = u.get_space_dim(env.action_space)
observation_space_dim = u.get_space_dim(env.observation_space)
policy = Policy(observation_space_dim, action_space_dim)

if policy_dir:
    policy.load_state_dict(torch.load(policy_dir))
    print("Loading policy from", policy_dir)
else:
    print("Plotting a random policy")
```

```
Loading policy from
/notebooks/rl2024/ex1/results/model/TargetReacher-v0_params.pt
```

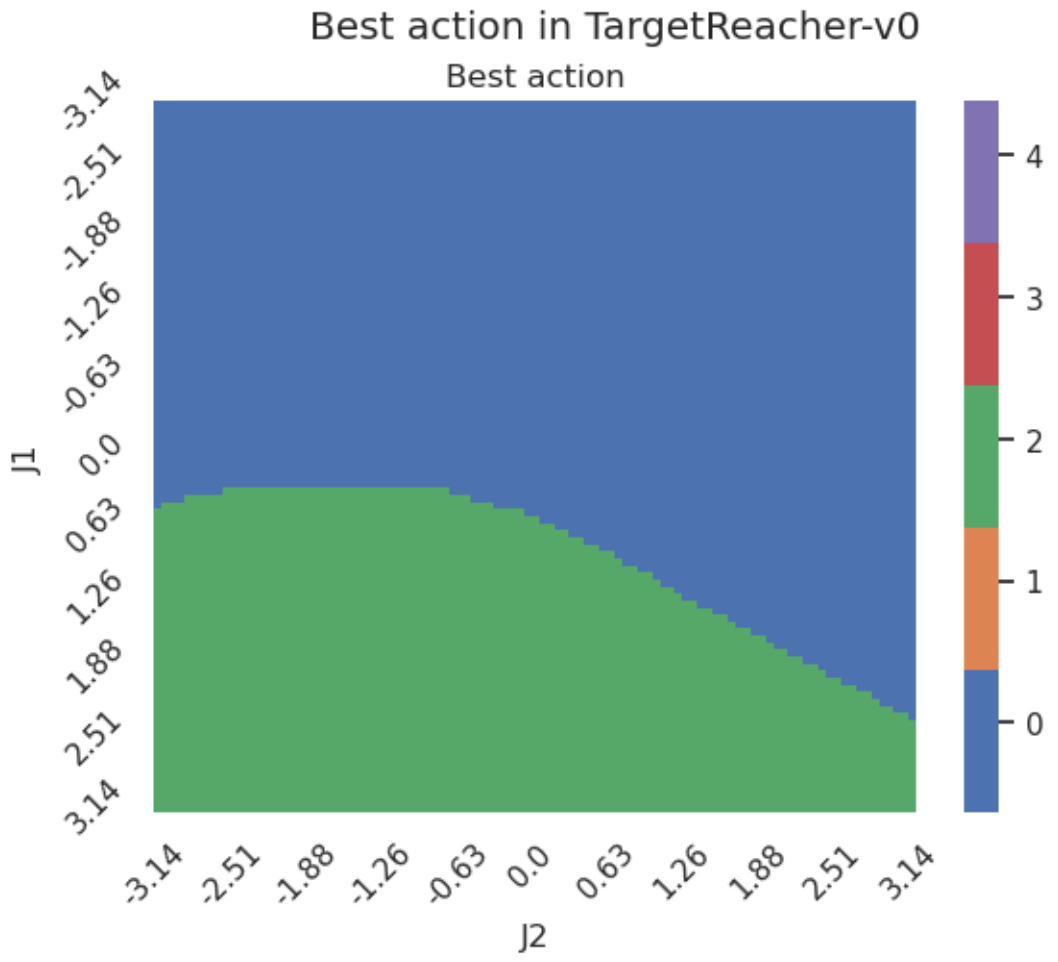
[33]:
```
# Create a grid and initialize arrays to store rewards and actions
npoints = resolution
state_range = np.linspace(-np.pi, np.pi, npoints)
rewards = np.zeros((npoints, npoints))
actions = np.zeros((npoints, npoints), dtype=np.int32)

# Loop through state[0] and state[1]
for i,th1 in enumerate(state_range):
    for j,th2 in enumerate(state_range):
        # Create the state vector from th1, th2
        state = np.array([th1, th2])

        # Query the policy and find the most probable action
        with torch.no_grad():
            action_dist, _ = policy(torch.from_numpy(state).float().
 ↪unsqueeze(0))
        action_probs = action_dist.probs.numpy()

        '''
        # TODO: Task 4:
        # 1. What's the best action, according to the policy?
        # 2. Compute the reward given state
        '''
        ########## Your code starts here ##########
        # Use the action probabilities in the action_probs vector
        # (it's a numpy array)
        # Find the action with the highest probability
        best_action = np.argmax(action_probs)
        actions[i, j] = best_action

        # Calculate the reward for the given state-action pair
        # Use the policy to get the reward for the state-action pair
        next_state = state  # For static reward calculation, we use the same␣
 ↪state
```

```
        reward = env.get_reward(state, best_action, next_state)
        rewards[i, j] = reward
        ########## Your code ends here ##########
```

[34]:
```
# Create the reward plot
num_ticks = 10
tick_skip = max(1, npoints // num_ticks)
tick_shift = 2*np.pi/npoints/2
tick_points = np.arange(npoints)[::tick_skip] + tick_shift
tick_labels = state_range.round(2)[::tick_skip]

sns.heatmap(rewards)
plt.xticks(tick_points, tick_labels, rotation=45)
plt.yticks(tick_points, tick_labels, rotation=45)
plt.xlabel("J2")
plt.ylabel("J1")
plt.title("Reward")
plt.suptitle("Rewards in %s" % env_name)
plt.show()
```



[35]:
```
# # Create the policy plot
cmap = sns.color_palette("deep", action_space_dim)
sns.heatmap(actions, cmap=cmap, vmin=0, vmax=action_space_dim-1)
plt.xticks(tick_points, tick_labels, rotation=45)
plt.yticks(tick_points, tick_labels, rotation=45)
colorbar = plt.gca().collections[0].colorbar
ticks = np.array(range(action_space_dim))*((action_space_dim-1)/
 ↪action_space_dim)+0.5
colorbar.set_ticks(ticks)
if env.spec.id == "Reacher-v1":
    # In Reacher, we can replace 0..4 with more readable labels
    labels = ["J1+", "J1-", "J2+", "J2-", "Stop"]
else:
    labels = list(map(str, range(action_space_dim)))
colorbar.set_ticklabels(labels)
plt.xlabel("J2")
plt.ylabel("J1")
plt.title("Best action")
plt.suptitle("Best action in %s" % env_name)
plt.show()
```



Do not remove this cell as it is used for grading

<h3>Student Question 4.1 Achieved Performance (5 points) </h3>

Where are the highest and lowest reward achieved?

Table of Contents

Highest Rewards: Located in areas with bright colors, closer to yellow. Specifically, the highest rewards are achieved when the joint angles (j1,j2) are around the coordinates where the color on the heatmap transitions to bright yellow.

Lowest Rewards: Located in areas with darker colors, closer to purple or black. The lowest rewards are achieved when the joint angles (j1,j2) are around the coordinates where the color on the heatmap transitions to dark purple or blue.

<h3>Student Question 4.2 Analysis of Behaviour (10 points) </h3>

Did the policy learn to reach the goal from every possible state (manipulator configuration) in an optimal way (i.e. with lowest possible number of steps)? Why/why not?

Table of Contents

Select all the correct answers.You can select no more than 4 answers, otherwise you lose all exercise points:

1. Yes, the policy efficiently learns to reach the goal from every state due to comprehensive state exploration.
2. No, the policy fails to explore all states adequately because the initial and goal states are nearly constant, leading to limited exploration ability.
3. Yes, the model is trained to always find the shortest path due to advanced algorithmic efficiency.
4. No, because the policy often chooses longer paths for certain initial states , not exploiting shorter alternatives.
5. Yes, continuous training ensures the policy adapts to find the best route from any state over time.
6. No, optimal paths are not always found because the exploration phase is limited, causing repeated suboptimal actions.
7. No, the policy sometimes achieves the goal optimally, but not consistently across all trials and configurations.
8. Yes, the training regime guarantees that all possible configurations are optimally addressed by the final model.
9. No, the policy cannot reach from all state because the policy has stochastic behavior.
10. No, bacause the policy overfits to the online it collects.
11. No, bacause the the reward function is not defined in an optimal way for this task.

[38]:
```
sq4_2 = [2, 4, 6, 7] #Answer question 4.2 with the appropriate answer numbers
```

[39]:
```
assert 1 <= len(set(sq4_2)) <= 4
assert set(sq4_2) < set(range(1, 12))
```

Do not remove the following blocks as they are used for grading

[ ]:

[ ]:

[ ]:

5 4. Submitting

Ensure all tasks and questions (in `ex1.ipynb`) are answered and the relevant plots are recorded in the relevant places. Details about attaching images and figures can be found below. The relevant

graphs to be included for this assignment are: - Task 1, CartPole `episodesep_reward` plot from logged csv file - x2 Task 4 reward plots

Ensure the correct model files are saved: - results/model/CartPole-v1_params.pt - results/model/CartPole-v1-model{0-4}_params.pt - results/model/SpinningReacher-v0_params.pt - results/model/TargetReacher-v0_params.pt

[42]:
```
# Make sure that skip training is set to True before submission
assert skip_training == True
```

5.1 4.1 Feedback

In order to help the staff of the course as well as the forthcoming students, it would be great if you could answer to the following questions in your submission:

1) How much time did you spend solving this exercise? (change the `hrs` variable below to a floating point number representing the number of hours taken e.g. 5.43)

[ ]:
```
hrs = 2.5
```

2) Difficulty of each task/question from 1-5 (int or float)

[ ]:
```
T1 = 1 # Student Task 1. Training a simple model for Cartpole environment
Q1_1 = 2 # Student Question 1.1 Learning
T2 = 4 # Student Task 2. Investigating training performance
Q2_1 = 3 # Student Question 2.1 Analyzing the training performance
Q2_2 = 3 # Stochasticity
T3 = 4 # Student Task 3. Reward function
T4 = 2 # Student Task 4. Visualize behavior
Q4_1 = 1 # Student Question 4.1 Achieved performance
Q4_2 = 3 # Student Question 4.2 Analysis of behavior
```

3) How well did you understand the content of the task/question from 1-5? (int or float)

[ ]:
```
T1 = 5 # Student Task 1. Training a simple model for Cartpole environment
Q1_1 = 3 # Student Question 1.1 Learning
T2 = 4 # Student Task 2. Investigating training performance
Q2_1 = 2 # Student Question 2.1 Analyzing the training performance
Q2_2 = 2 # Stochasticity
T3 = 4 # Student Task 3. Reward function
T4 = 4 # Student Task 4. Visualize behavior
Q4_1 = 5 # Student Question 4.1 Achieved performance
Q4_2 = 2 # Student Question 4.2 Analysis of behavior
```

4) General feedback. Consider questions like:

- Did the content of the lecture relate well with the assignment?
- To what extent did you find the material to be potentially useful for your research and studies?

And other feedback you think is worth including. Type in the box below

Please use the following section to record references. # References

[1] Sutton, Richard S., and Andrew G. Barto. "Reinforcement Learning: An Introduction (in progress)." London, England (2017). http://incompleteideas.net/book/RLbook2018.pdf