Reinforcement Learning

Lecture 2: Environments and bandits

Robert Lieck

Durham University

Lecture Overview

This lecture shows how to set up a modern RL environment ready for training using **Gymnasium**, **Brax** and a multi-armed bandits implementation as in chapter 2 of [1].

1. **Gym and Brax**
   - what are they?
   - how do they map to RL concepts?
   - gym.Env methods and attributes
   - nested spaces example
2. **Custom environments**
   - Gymnasium
   - Google brax
3. **Colab coding**
   - a random agent
   - exploring different environments
   - minimal REINFORCE example
   - multi-armed bandits

Gym and Brax what are they?

Gymnasium & Google brax

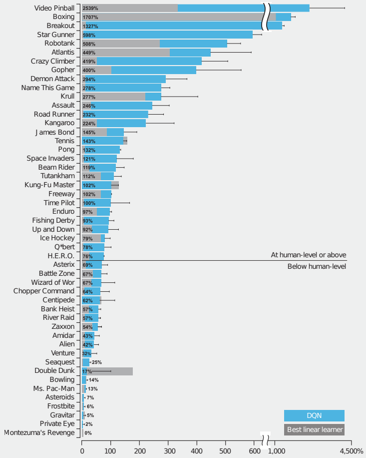
What is it?

- Collection of environments
- Most environments are freely available
- Standardised API makes comparisons and benchmarking easier
- Test same RL algorithms on many different problems

For example:

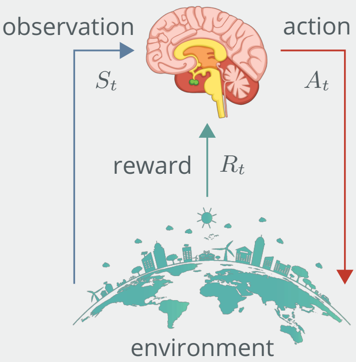
- Atari 2600 games
- Classic control problems
- Physics simulations

Environments and performance in [2]



Gym and Brax how does it map to RL concepts?

RL Agents



Algorithm: minimal example

```
import gymnasium as gym
env = gym.make('CartPole-v1')

agent = Agent()

# reset environment to base state
# and get initial observation and info
o, i = env.reset()

# loop until the episode is done
done = False
while not done:
    action = agent.sample_action(o)
    o, r, term, trunc, i = env.step(action)
    done = term or trunc

env.close()
```

Gymnasium gym.Env methods and attributes

Core methods

In gym/core.py:

- **obs, info = env.reset(seed)** - reset the environment, optionally seed randomness, return the initial observation and some diagnostic info
- **obs, rew, term, trunc, info = env.step(action)** - takes a single step and returns a tuple with the next observation, reward, whether the environment terminated normally or was truncated due to limits, and some diagnostic info
- **env.close()** - cleanup function
- **env.render()** - (optional) render the env state: print text, show an image...

Env attributes

- **env.action_space** - gym.Space object that describes the shape and type of actions
- **env.observation_space** - gym.Space object that describes the shape and type of observations
- **env.reward_range** - tuple of per-step reward range, defaults to $(-\infty, \infty)$

gym.Space in /gym/spaces/space.py:

- **Box, Discrete, Binary, ...**
- **Tuple, Dict** - can contain nested observation spaces

Example: nested observation space – gym/spaces/dict.py

```
env.nested_observation_space = spaces.Dict({
    ‘sensors’:  spaces.Dict({
        ‘position’: spaces.Box(low=−100, high=100, shape=(3,)),
        ‘velocity’: spaces.Box(low=−1, high=1, shape=(3,)),
        ‘front_cam’: spaces.Tuple((
            spaces.Box(low=0, high=1, shape=(10, 10, 3)),
            spaces.Box(low=0, high=1, shape=(10, 10, 3))
        )),
        ‘rear_cam’: spaces.Box(low=0, high=1, shape=(10, 10, 3)),
    }),
    ‘ext_controller’: spaces.MultiDiscrete((5, 2, 2))
})
```

Custom environments Gymnasium

Example: gym custom environment doc

```python
class CustomEnv(gym.Env):
    def __init__(self):
        ...

    def step(self, action):
        ...
        return observation, reward, terminated, truncated, info

    def reset(self, seed=None, options=None):
        ...
        return observation, info

    def render(self):
        ...

    def close(self):
        ...
```

Custom environments Google brax

Example: brax custom environment [3]

```
from brax.envs import env
class CustomEnv(env.Env):
  def __init__(self, **kwargs):
    super().__init__(config='dt: .02', **kwargs)

  def reset(self, rng: jnp.ndarray) -> env.State:
    zero = jnp.zeros(1)
    qp = brax.QP(pos=zero, vel=zero, rot=zero, ang=zero)
    obs, reward, done = jnp.zeros(2)
    return env.State(qp, obs, reward, done)

  def step(self, state: env.State, action: jnp.ndarray) -> env.State:
    vel = state.qp.vel + (action > 0) * self.sys.config.dt
    pos = state.qp.pos + vel * self.sys.config.dt
    qp = state.qp.replace(pos=pos, vel=vel)
    obs = jnp.array([pos[0], vel[0]])
    reward = pos[0]
    return state.replace(qp=qp, obs=obs, reward=reward)
```

Colab coding demo

Click this link:

Link to Colab code

In this demo we will cover:

- A random agent
- Exploring different environments
- Plotting and measuring baselines
- Minimal REINFORCE example
- Multi-armed Bandits – chapter two of [1]

References I

[1] Richard S Sutton and Andrew G Barto. Reinforcement learning: An introduction (second edition). Available online. MIT press, 2018.

[2] Volodymyr Mnih et al. “Human-level control through deep reinforcement learning”. In: nature 518.7540 (2015), pp. 529–533.

[3] C Daniel Freeman et al. “Brax–A Differentiable Physics Engine for Large Scale Rigid Body Simulation”. In: arXiv preprint arXiv:2106.13281 (2021).