Project proposal

This document details the proposal for the RLDMU project, 2023.

Organisation

The project is done in Python and the work is shared across the 2 team members. The project’s source code can always be monitored on GitHub: <https://github.com/maximewel/rldmuu_project>

The work is done in an iterative way. Firstly, a working discrete environment is created and tested. Then the environment can be rendered with a 2D GUI. Lastly, the environment can be passed to continuous observations in order to test the differences between Q-table based methods and Deep learning ones.

Environment overview

Arrived on the moon with the Lunar Lander, it is now time to explore. For this purpose, the Lunar Explorer was developed. Its objective is to get to a specific location (green cell). On the way there, there are potential interesting ores that could be a source of yet unknown metals. The Lunar Explorer aims to collect as many other precious metals as possible (blue cell). However, the Lunar Explorer has a limited drilling capacity (e.g. 3 drillings). It will therefore not be able to drill all the available resources but will have to select the most profitable ones. Furthermore, its landing point has been precisely defined (black cell). Overall, the Lunar Explorer should explore and discover the optimal exploitation of the current map.

However, no gas station is yet available on the moon. The Lunar Explorer can only use the initially available fuel and avoid useless trips – translating to a penalty of “” for each iteration. The lunar surface is also not very friendly. This is why, depending on the field, the Lunar Explorer can suffer from unexpected behaviours and must adapt its movements.

The goal of this environment is to have an agent balance between the two main mechanics – minerals and terrain type – in order to exploit the map.

Example of a standard map, no movement tile is reflected:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **1** | **…** | **…** | **…** | **…** | **…** | **…** | **…** | **…** | **…** | **n** |
| **1** |  |  |  |  |  |  |  |  |  |  |  |
| **…** |  |  |  |  |  |  |  |  |  |  |  |
| **…** |  |  |  |  |  |  |  |  |  |  |  |
| **…** |  |  |  |  |  |  |  |  |  |  |  |
| **…** |  |  |  |  |  |  |  |  |  |  |  |
| **…** |  |  |  |  |  |  |  |  |  |  |  |
| **…** |  |  |  |  |  |  |  |  |  |  |  |
| **…** |  |  |  |  |  |  |  |  |  |  |  |
| **…** |  |  |  |  |  |  |  |  |  |  |  |
| **n** |  |  |  |  |  |  |  |  |  |  |  |

TABLE 1 Example of the Lunar Explorer environment

*Legend*

* Start location
* Potentially interesting mineral
* End location, stops simulation

Discrete environment

The first iteration aims to create a basic environment, with discrete observations and discrete states. This is typically a problem that can be solved by a Qtable-based algorithm.

Observation space

The observation space has 4 components, each of them having a limited range of valid values.

State space: , with

* being the horizontal position
* being the vertical position
* being the horizontal speed,
* being the vertical speed,

The type of tile is not included in the observations, as it is useless for the Q-table based algorithms. They will learn a map by heart and will not learn to recognize and adapt to a tile type. Furthermore, as only the current tile is returned as an observation, no generalization ability should be developed by an algorithm running on this env.

More explanation about the speed mechanism is detailed on the next section.

A little verification/analysis of the Q-table resulting of that:

* For a 30\*30 map, this represents 8100 possible states.

Action space:

* On the 30\*30 map, this represents a Qtable of size 48600.

Movements / Speed

The agent can move around the map by going in either direction or stopping. The speed can be seen as a 2D vector of two integer values . In discrete mode, the speed can be either “moving” or “not moving” . Only one speed component can be activated at a time (no diagonal movements). The speed behaves as follow:

* The speed encodes the last movement performed. Going right (or left) sets to (resp. ) the horizontal speed, down / up set to (resp. ) the vertical speed for the next state.
* Doing nothing means the explorer stops and resets the speed to 0.
* No cost in turning: When going from horizontal to vertical or opposite, the absolute value of the speed is carried, only the orientation changes. This is required to keep the speed consistent across all Manhattan paths and avoid shortcuts on slow tiles.
  + This means that going right then up results in the speed to go from
* Going into an opposite direction resets the speed to (left to right, top to bottom, and inverse)

Surface tile types

The surface of the moon is not uniform. The Lunar Explorer has many challenges on its path.

The next state is relative to the absolute value of the speed of the rover, the current and next tile type, and the chosen action:

1. Speed type “slow”: This terrain is frail and can break under the explorer’s weight: advance carefully.
   * If the speed is when exiting the tile, the explorer can safely travel. If the speed is not , then the rover has a chance to fall, and the simulation stops. This means that the explorer has to stop before exiting every slow tile it wants to go through.
2. Speed type “random”: This terrain is covered in craters, don’t get lost! There is a chance to be moved to adjacent tiles when exiting it, regardless of the direction taken.
3. Speed type “normal”: Regular moon terrain. The agent will always move to the juxtaposed state, according to the action.
4. Speed type “fast”: This road-like terrain is particularly adapted to the explorer’s locomotion. When exiting with , The agent has probability to move to the juxtaposed state and probability to move two states away from the current state, according to the action.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Current state with speed type** | | **Absolute speed** | **Action (Left, right, top, down)** | **New state coordinates** | **Prob.** |
|  |  | 0 |  |  |  |
|  |  | 1 |  |  |  |
|  |  | 1 |  |  |  |
|  |  | any |  |  |  |
|  |  | any |  |  |  |
|  |  | any |  |  |  |
|  |  | 0 |  |  |  |
|  |  | 1 |  |  |  |
|  |  | 1 |  |  |  |

Solution proposal

We will use a plethora of different algorithms developed during the course and lab of the RDLMU course. Namely, we have Q-table based algorithms:

* Qlearning
* Sarsa
* DynaQ
* Eventually in combination with Eligibility traces

And Deep Learning based ones:

* DQN in PyTorch, soft update
* DQN in TensorFlow, full update every N epochs

The test process can be:

1. Decade an epsilon value and a decaying methodology, explore different algorithms; and
2. see how they react on the same map, how fast / stable they are in learning, their maximum performance, etc.

From discrete to continuous

If time allows it, the “position” and “speed” observations can be moved to continuous. In this case, similar to the Lunar Lander[[1]](#footnote-1), the actions increment the speed by a set amount of acceleration each time they are called to remain discrete actions. The changes would then be:

* Some tiles can slightly change behaviour (e.g. speed tile “propulsion” may add speed, only activated above a set amount of speed; craters having moving probabilities of being lost depending on speed value, etc…). The speed would stay limited to the bound [-1; 1] in order to have a single tile detection in the board management.

In order to fit the continuous space and to allow for the testing of generalisation, the observation space would be:

State space: with

* “Activation grid”: Simple array given in observation with the same dimensions as the grid. The value is for the tile where the agent is, where the agent has never been, and a decreasing value between and for each tile previously passed through. This allows the agent to easily identify its position and path to reach it.
* FOV: tile type of close tiles in order to be able to work on new maps and not only train on one. The agents must be able to react to their surroundings, i.e. for deep agents, not Q-table based ones. To avoid giving the whole grid at each observation, the FOV gives the tile type of tiles that are close to the player in an surrounding.
* being the horizontal speed,
* being the vertical speed,

In this form, the seed can be changed during training to teach the agent to react to the observation and not only find the best path for a given, single map.

1. See <https://gymnasium.farama.org/environments/box2d/lunar_lander/> [↑](#footnote-ref-1)