Online and Reinforcement Learning (2025)
Home Assignment 8

Davide Marchi 777881

Contents

1 MDP Classes

In this exercise we classify three MDPs—**RiverSwim**, **RiverSwim-2** (a modified version of RiverSwim), and the **4-room grid-world**—according to the following classes:

- **Ergodic MDPs**: Every stationary policy induces a Markov chain that has a unique recurrent class (i.e., it is a unichain).

- **Communicating MDPs**: For every pair of states $s, s' \in \mathcal{S}$, there exists some policy under which s' is reachable from s (and vice versa).

- **Weakly Communicating MDPs**: The state space can be partitioned into a communicating set of recurrent states and a set of transient states.

Since an ergodic MDP is a special case of a communicating MDP, and every communicating MDP is weakly communicating, the classes satisfy

$$\text{Ergodic} \subseteq \text{Communicating} \subseteq \text{Weakly Communicating}.$$

Below, I analyze each MDP and provide counterexamples where needed.

(i) RiverSwim

Ergodicity: **Answer:** `False`.

Explanation: Although the underlying transition graph of RiverSwim is (in principle) strongly connected (since for any two states one can design a policy to move from one to the other), there exist stationary policies that break the unichain property. For example, consider the policy that always selects the *left* action. Under this policy the agent remains confined to the left-hand bank (state 1) and never reaches any other state. This shows that not every stationary policy yields a single recurrent class; hence, the MDP is not ergodic.

Communicating: **Answer:** `True`.

Explanation: By the structural properties of RiverSwim, for any pair of states s and s' one can construct a policy that uses the *right* action (to move toward higher-indexed states) and the *left* action (to move back) appropriately such that s' is reachable from s. This mutual reachability implies that RiverSwim is a communicating MDP.

Weakly Communicating: **Answer:** `True`.

Explanation: Since being communicating implies that there exists at least one policy connecting every pair of states, the MDP is, in particular, weakly communicating.

(ii) RiverSwim-2

The modified MDP, RiverSwim-2, is defined as RiverSwim but with an extra state s_{extra} that has two actions:

- Under action a_1, the agent moves deterministically to state 1 (the left bank).

- Under action a_2, the agent moves to state 1 with probability 0.5 or to state 2 with probability 0.5.

All other transitions remain the same as in RiverSwim.

Ergodicity: Answer: `False`.

Explanation: Similar to the original RiverSwim, one can choose stationary policies (e.g., always taking an action that avoids exiting a subset of states) that do not visit all states. Thus, the unichain property does not hold.

Communicating: Answer: `False`.

Explanation: A key modification in RiverSwim-2 is that the extra state s_{extra} is not reachable from any of the original RiverSwim states. While once in s_{extra} the agent can move to state 1 (or 2), there is no action in the main chain that will take the agent to s_{extra}. Hence, there exist state pairs (namely, any state in the original set and s_{extra}) for which no policy can ensure mutual reachability. Therefore, RiverSwim-2 is not communicating.

Weakly Communicating: Answer: `True`.

Explanation: Even though the whole state space is not mutually reachable, the original states (the main chain) form a communicating recurrent class. The extra state s_{extra} is transient because whenever it is visited the agent will eventually be forced into the communicating set. This structure fits the definition of a weakly communicating MDP.

(iii) 4-room Grid-world

The 4-room grid-world (often also called the frozen lake MDP in this context) has the following characteristics:

- The agent has 4 actions (up, down, left, right) when not adjacent to walls.

- Due to a slippery floor, the chosen action results in moving in the intended direction with probability 0.7, staying in the same state with probability 0.1, or moving in one of the two perpendicular directions (each with probability 0.1).

- Walls act as reflectors (i.e., if an action leads into a wall, the agent remains in the current state).

- Upon reaching the rewarding state (highlighted in yellow), the agent is teleported back to the initial state.

Ergodicity: Answer: `False`.

Explanation: Even though the grid is fully connected, there exist stationary policies (for instance, those that consistently choose actions that lead to self-loops—by always hitting the wall) that confine the agent to a strict subset of states. Thus, not every stationary policy produces a Markov chain with a single recurrent class.

Communicating: Answer: `True`.

Explanation: The underlying structure of the grid is such that for any two states there exists a policy (using the stochastic nature of the transitions and the teleportation mechanism) that connects them. In other words, even if some policies lead to degenerate behavior, one can always design a policy that makes every state reachable from any other state.

Weakly Communicating: Answer: `True`.

Explanation: Since the grid-world is communicating (in the structural sense described above), it is trivially weakly communicating.

Summary Table

MDP	Ergodic	Communicating	Weakly Communicating
RiverSwim	False	True	True
RiverSwim-2	False	False	True
4-room grid-world	False	True	True

Table 1: Classification of the MDPs based on their connectivity properties.

2 Infinitely Many Choices for Optimal Bias in Average-Reward MDPs

3 Comparison between UCRL2 and UCRL2-L

4 Leveraging Prior Knowledge on the Graph of MDP to Reduce Regret