Windy Gridworld

Alessandro Tenaglia

Machine and Reinforcement Learning in Control Applications

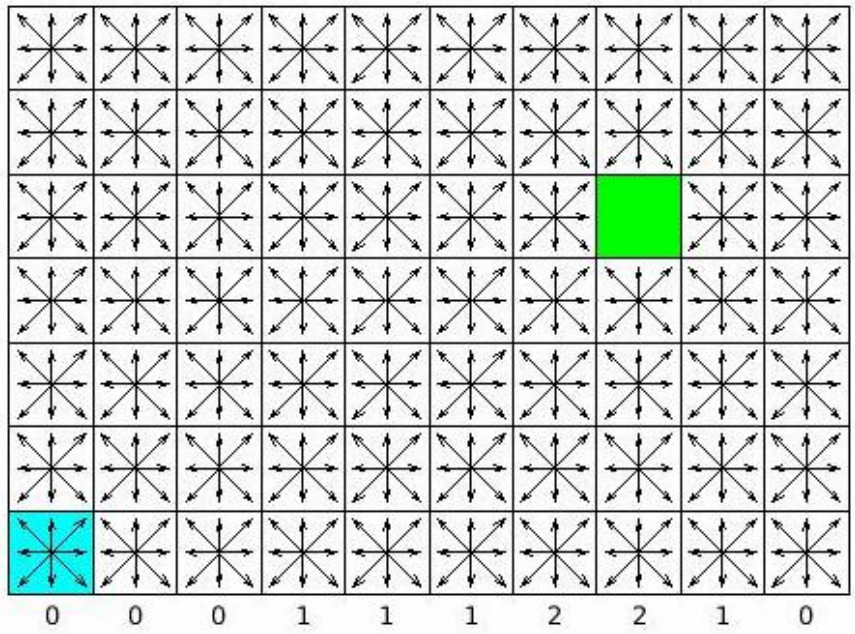
April 20, 2022

Problem

Learn to move in an unknown map with external disturbances

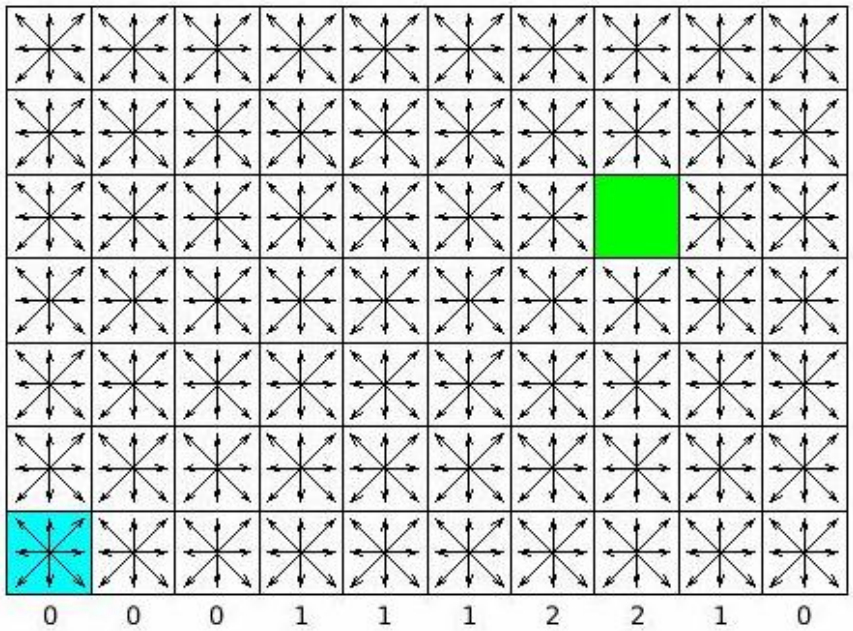
Problem formulation

- Consider the gridworld on the right;
- The goal is to reach the green box;
- There is a crosswind running upward through the grid;
  - its amount is shown below each column;
  - next position is shifted.



Problem formulation

- 8 possible directions: **N, S, E, W, NE, NW, SE, SW**;
- Reward:
  - -1 for each step



Model

- The **state** is the position in the Gridworld
  - we have $X \cdot Y$ states.
- The **action** is the direction of the movement
  - we have 8 actions.

Planning: Dynamic Programming

- Model the problem as an **MDP**;
- Compute the transition matrix **P**;
- Compute the reward matrix **R**;
- Find the optimal policy $\pi_\star$ using **Dynamic Programming** method (PI or VI).

Learning: Monte Carlo

- **Model-free**: no knowledge of MDP transitions and rewards;
- Simulate episodes:

$$S_0, A_0, R_0, S_1, A_1, R_1 ..., S_T, A_T, R_T$$

- Use the return G_t to update estimates:

$$V(S_t) \leftarrow V(S_t) + \alpha(G_t - V(S_t))$$

Learning: Temporal Difference

- **Model-free**: no knowledge of MDP transitions and rewards;
- **Bootstrap**: updates a guess towards a guess;
- Learns from incomplete episodes;
- Use the immediate reward R_t to update estimates:

$$V(S_t) \leftarrow V(S_t) + \alpha(R_t + \gamma V(S_{t+1}) - V(S_t))$$

MC vs TD

Monte Carlo

- Model-free
- Episodic tasks
- Full episode
- No bootstrap
- Zero bias
- Not very sensitive to initial value
- High variance

Temporal Difference

- Model-free
- Continuous and episodic tasks
- Just a single step (online)
- Bootstrap
- Non-zero bias
- Sensitive to initial value
- Low variance