FROZEN LAKE: MONTE CARLO METHOD

Irena Torosyan

TABLE OF CONTENT

01.

MONTE CARLO

model-free method for learning the state-value function

does not require a priori information about the state transition probabilities

DEFINITIONS

state value function

policy, state, time

$$v_\pi(s) = E_\pi[G_t | S_t = s]$$

weighted return

discount rate, reward

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots + \gamma^{T-t-1} R_T$$

FIRST VISIT MONTE CARLO METHOD

compute the return starting from the first visit of the state in the sequence from the time step $t = 2$

EPISODE

a single run or instance of a simulation or computation

EXAMPLE

start state: S0 = s1, S1 = s5,
S2 = s6, . . . , S11 = s12,
terminal state: S12 = s16

R1 = r5, R2 = r6, . . . , R12 = r16

the estimate (3 episodes)

$$\hat{v}(s_1) = \frac{G(s_1)^{\text{Episode 1}} + G(s_1)^{\text{Episode 2}} + G(s_1)^{\text{Episode 3}}}{3}$$

s_1	s_2 r_2	s_3 r_3	s_4 r_4
s_5 r_5	s_6 r_6	s_7 r_7	s_8 r_8
s_9 r_9	s_{10} r_{10}	s_{11} r_{11}	s_{12} r_{12}
s_{13} r_{13}	s_{14} r_{14}	s_{15} r_{15}	s_{16} r_{16}

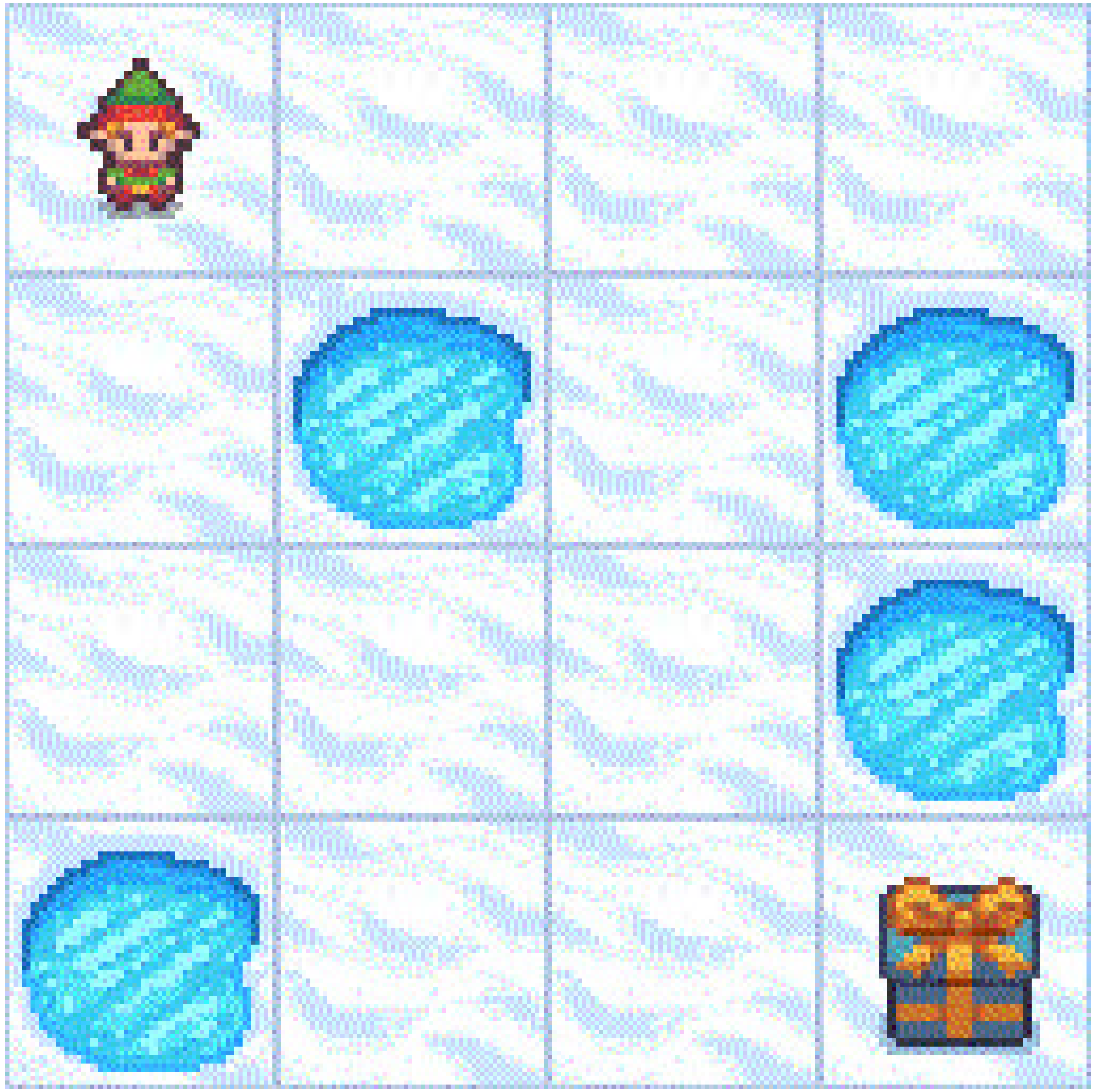
02.

FROZEN LAKE

The game setup and rules

The setup

- 4 x 4 board
- Fixed start and goal positions
- cross the frozen lake without falling into a hole
- hole positions are unknown
- slippery ice



action space

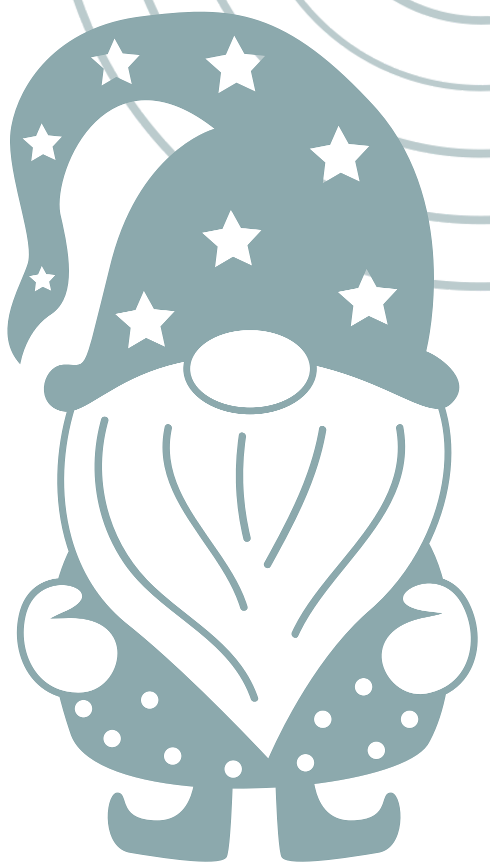
left, down, right, up

observation space

agent's current position

rewards

reach goal: +1
reach frozen: 0



03.

IMPLEMENTATION

Solving frozen lake

STRUCTURE OF THE CODE

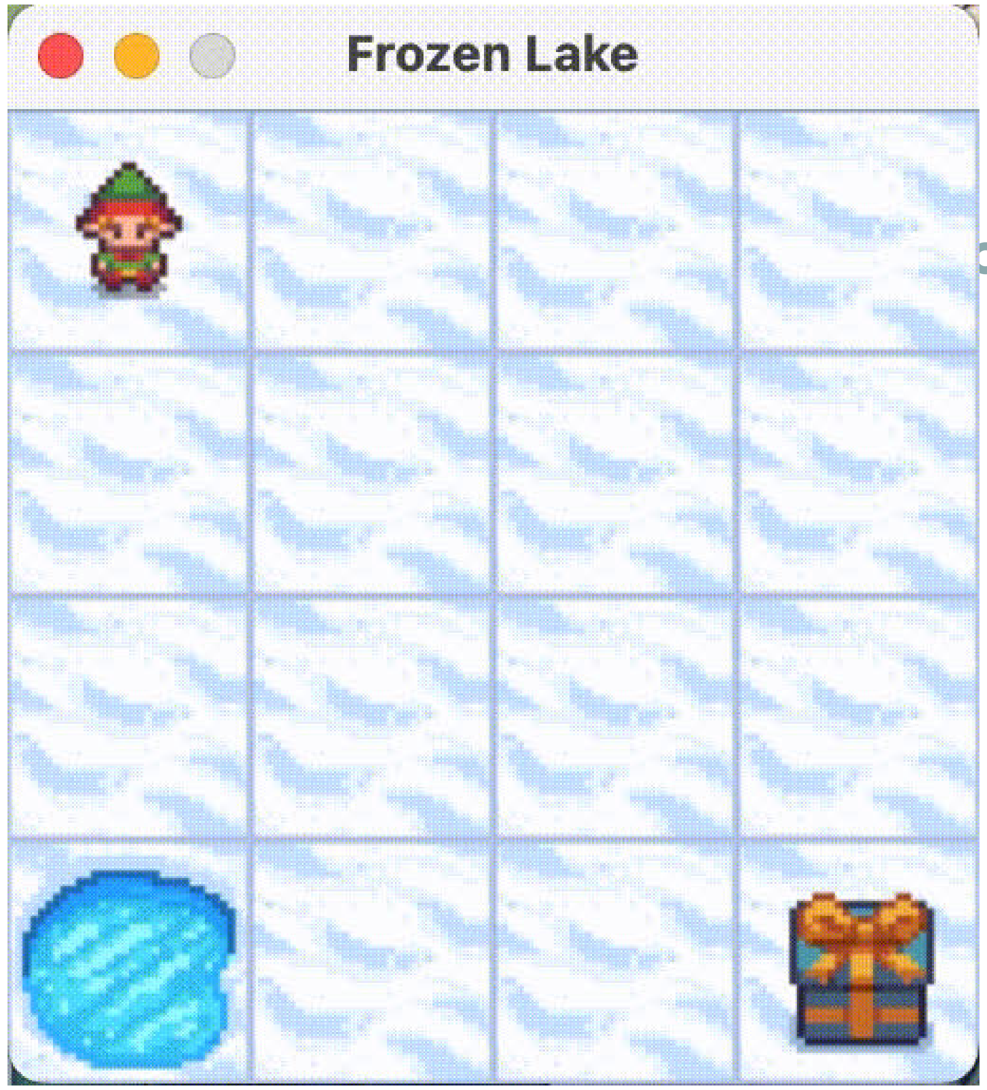
- Create a for loop that simulates episodes. Create vectors that store total returns and the total number of visits for every state during simulated episodes.
- In every episode simulation (iteration) compute the return from every visited state.
- In every iteration, update the vectors that store the total return and the total number of visits for every particular state.
- After the loop is completed, divide every entry of the vector storing total returns by the number of visits of a particular state.

03.

DEMONSTRATION

Code in action and metrics

DEMO

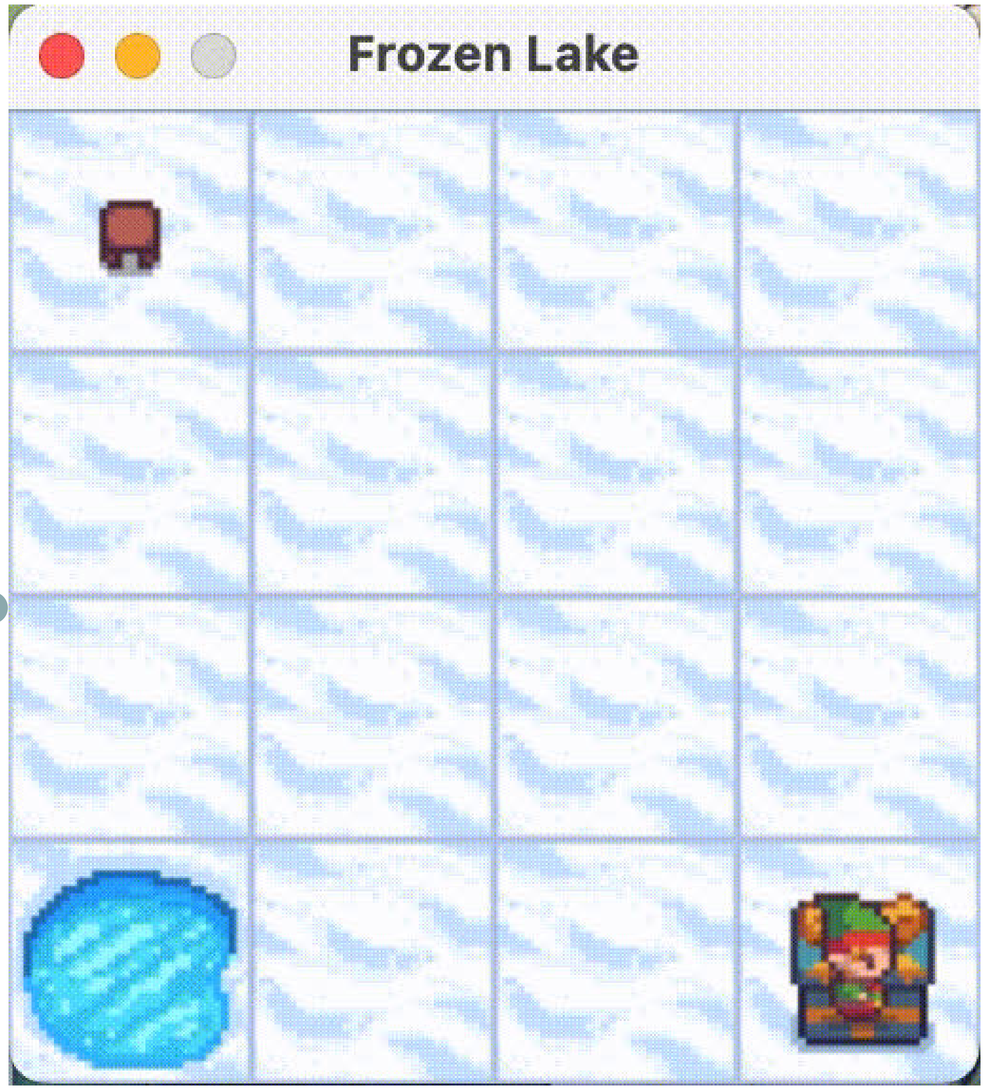


WIN

Not a very typical example, takes longer to reach the end usually

LOSS

Wondered around a lot, more typical



RESULTS

State value function estimates

Iterative Policy

0.43	0.47	0.53	0.57
0.38	0.45	0.55	0.62
0.27	0.41	0.59	0.73
0	0.35	0.65	0

Monte Carlo

0.42	0.47	0.53	0.59
0.38	0.46	0.55	0.62
0.26	0.42	0.59	0.74
0	0.36	0.66	0

THANK YOU