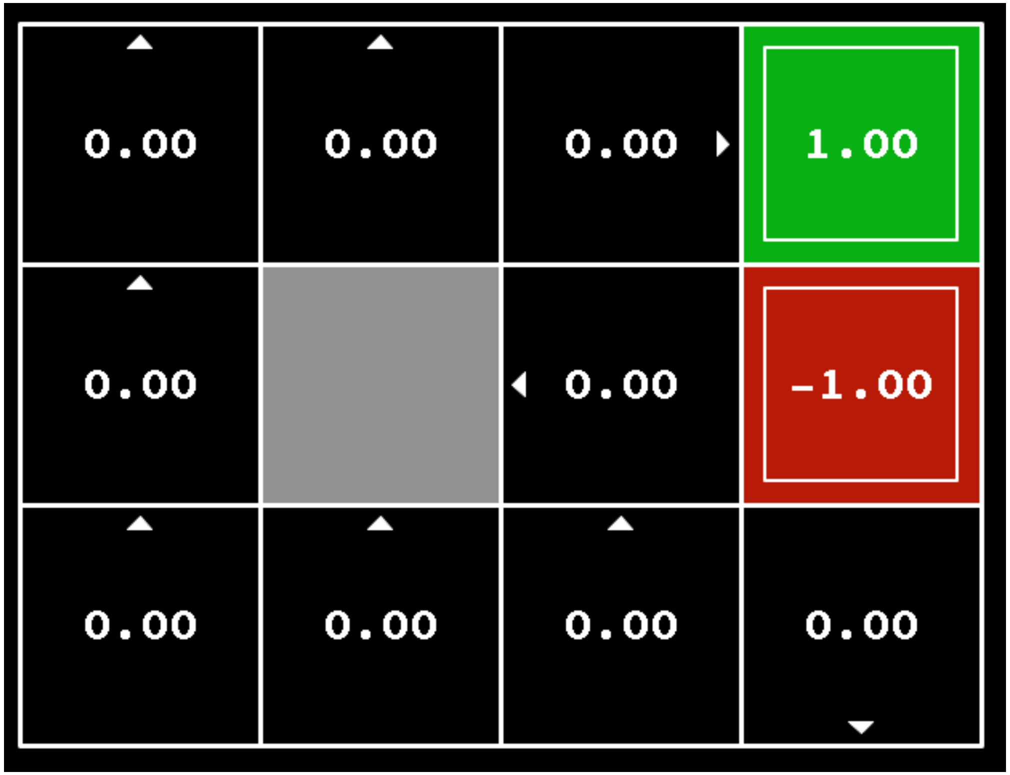
Outline

- Reinforcement Learning for Gomoku
- RL for News Recommendation
- RL for Text Generation
- Lab3 (In-class)
- Project3 – Blackjack (Homework)

Lab3 – Grid World

- Recall: What is Grid World problem?
  - States
  - Actions
  - Rewards
  - …



▲ 0.00	▲ 0.00	0.00 ▶	1.00
▲ 0.00		◀ 0.00	-1.00
▲ 0.00	▲ 0.00	▲ 0.00	0.00 ▼

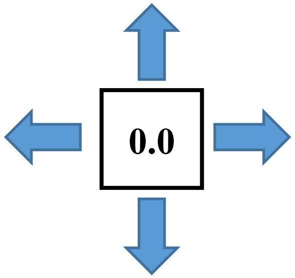
Lab3 – Grid World

- How about today’ s Grid World problem?
  - States
  - Actions:

    Up, Down, Left, Right
    Deterministically goes to next state
  - Rewards

0.0	A	0.0	B	0.0
0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	B_{TO}	0.0
0.0	0.0	0.0	0.0	0.0
0.0	A_{TO}	0.0	0.0	0.0

$$R(s,a) = \begin{cases} 0.0, & taking\ a\ will\ stay\ in\ the\ grid\ world \\ -1.0, & taking\ a\ will\ jump\ out\ of\ the\ grid\ world \\ r, & current\ state\ is\ special \end{cases}$$



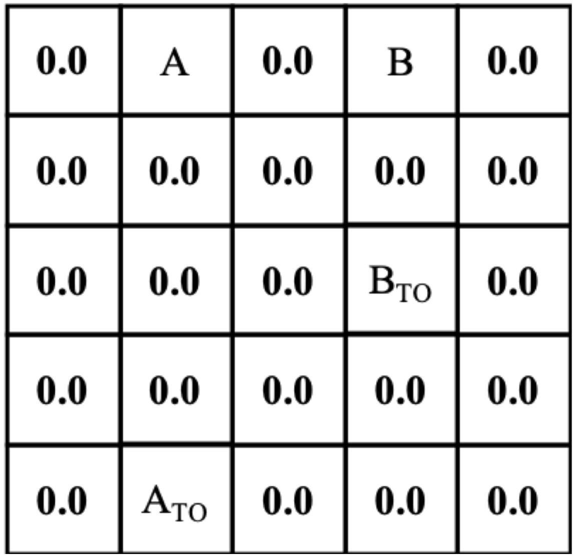
Lab3 – Grid World

- How about today’s Grid World problem?
  - Value Iteration

$$V^*(s) = \max_{a\in A(s)} R(s,a) + \gamma \sum_{s'} P(s'|s,a) * V^*(s')$$

- ➢ Synchronous Update
- ➢ Asynchronous Update

Make sure you use synchronous update in this Lab.



0.0	A	0.0	B	0.0
0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	B_{TO}	0.0
0.0	0.0	0.0	0.0	0.0
0.0	A_{TO}	0.0	0.0	0.0

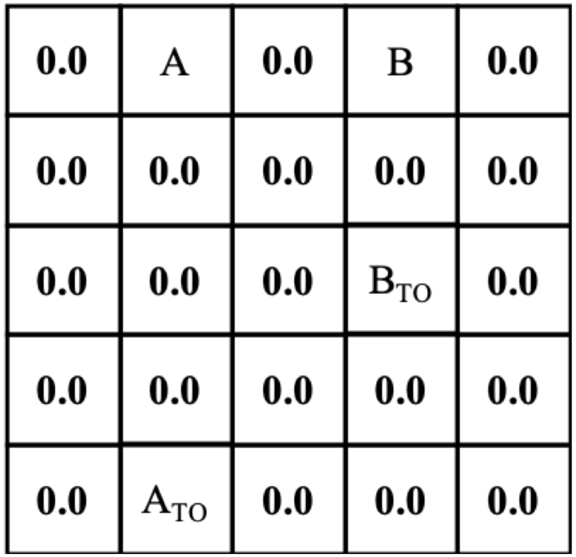
Lab3 – Grid World

- How about today’ s Grid World problem?
  - Policy Evaluation

$$V^{\pi}(s) = R(s,\pi(s)) + \gamma \sum_{s'} P(s'|s,\pi(s)) * V^{\pi}\ (s')$$

  - Policy Improvement

$$\pi(s) = \arg \max_{a\ \in A(s)} Q(s,a)$$



0.0	A	0.0	B	0.0
0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	B_{TO}	0.0
0.0	0.0	0.0	0.0	0.0
0.0	A_{TO}	0.0	0.0	0.0

Lab3

- Problem:
  - Solve the Grid World Problem based on MDP
- Requirement:
  - Print the iteration numbers and optimal values of all states using value iteration and policy iteration
- Address: http://10.192.9.85/contest/5/problem/01

Lab3

- Value Iteration

```
function VALUE-ITERATION($mdp$, $\epsilon$) returns a utility function
    inputs: $mdp$, an MDP with states $S$, actions $A(s)$, transition model $P(s'|s,a)$,
                rewards $R(s)$, discount $\gamma$ .
              $\epsilon$, the maximum error allowed  in the utility of all states
    local variables: $U, U'$, vectors of utilities for states in $S$, initially zero
                          $\delta$, the maximum change in the utility of any stage in an iteration

    repeat
        $U \leftarrow U';\ \delta \leftarrow 0$
        for each state $s$ in $S$ do
            $U'[s] \leftarrow \max_{a \in A(s)} R(s,a) + \gamma \sum_{s'} P(s'|s,a)\ U[s']$
            $\delta \leftarrow \delta + |U'[s] - U[s]|$
    until $\delta < \epsilon$
    return $U$
```

Lab3

- Policy Iteration

```
function POLICY-ITERATION($mdp$) returns a policy

  inputs: $mdp$, an MDP with states $S$, actions $A(s)$, transition model $P(s'|s,a)$
  local variables: $U$, a vector of utilities for states in $S$, initially zero
                   $\pi$, a policy vector indexed by state, initially random

  repeat
      $U \leftarrow$ POLICY_EVALUATION($\pi, U, mdp$)
      $unchanged? \leftarrow$ true
      for each state $s$ in $S$ do
      if $\max_{a \in A(s)} Q(s,a) > Q(s,\pi[s])$ then do
          $\pi[s] \leftarrow \arg\max_{a \in A(s)} Q(s,a)$
          $unchanged? \leftarrow$ false
  until $unchanged?$
  return $\pi$
```

Think by yourself :
How to compute the
Q-value in this problem?

Outline

- Reinforcement Learning for Gomoku
- RL for News Recommendation
- RL for Text Generation
- Lab3 (In-class)
- Project3 – Blackjack (Homework)

Blackjack

- You need to submit your own version of code.
- You are encouraged to discuss with your group members.

  It might take some time to get familiar with all the supportive codes.
- Homework 3 is due on **23:55 pm, 09 Dec, 2020**