# On-policy Model-Based

# Rmax Algorithm

**Rmax Algorithm**

**Input:** $\mathbf{S}$, $\mathbf{A}$, $\gamma$, $m$, Rmax

**Output:** $\pi^*$

Initialize counter $c(s,a) \leftarrow 0$ for all $(s,a) \in \mathbf{S} \times \mathbf{A}$

Initialize the empirical known state-action MDP $\hat{M} = \langle \mathbf{S}, \mathbf{A}, \hat{\mathbf{T}}, \hat{\mathbf{R}}, \gamma \rangle$ :

$\hat{T}(s,a,s') = \mathbb{I}(s' = s), \quad \hat{R}(s,a) = Rmax$

**for** $t = 1, 2, 3, \ldots$ **do**

  Compute $\hat{\pi}_t^*$

  Execute $a_t = \hat{\pi}_t^*(s_t)$

  Act: receive reward $r_t$ and transition to the next state $s_{t+1}$.

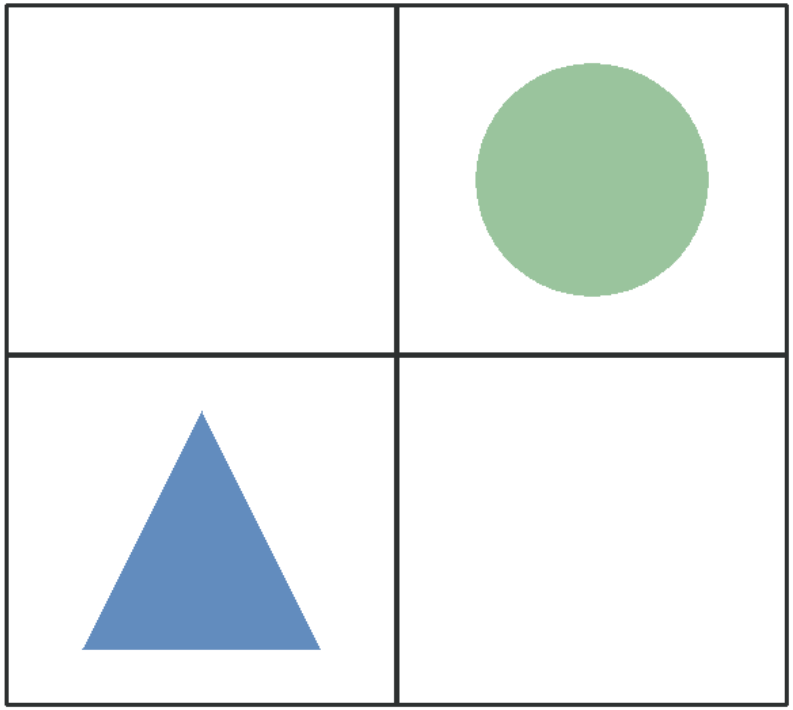
  $c(s_t, a_t) \leftarrow c(s_t, a_t) + 1$

  **if** $c(s_t, a_t) = m$ **then**

    Redefine: $\hat{T}(s,a,s') = \frac{count(s,a,s')}{count(s,a)}$ and $\hat{R}(s,a) = \frac{\sum_{(s,a,r,s')} r}{count(s,a)}$

# Example - grid world

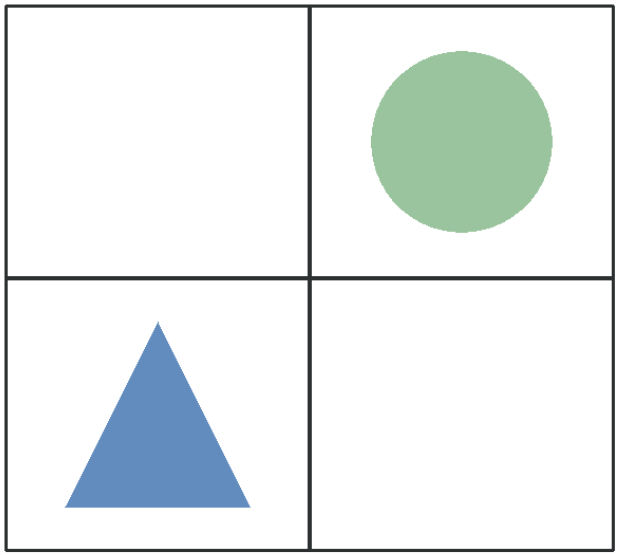
- 2×2 grid world
- The agent starts at location (1,1)
- the goal is at (2,2)
- can move [left, right, down, up]
- The reward is 0, unless we step into a goal state, and then it is 1



# Example - grid world

*MDP:*

- $S = \{(1,1), (1,2), (2,1), (2,2)\}$
- A=$\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.99$
- m= 2
- use Rmax to train an agent to solve the game



# Init (t=0)

Known status

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | U | U | U | U |
| → | U | U | U | U |
| ↓ | U | U | U | U |
| ← | U | U | U | U |

Reward counts

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | Rmax | Rmax | Rmax | Rmax |
| → | Rmax | Rmax | Rmax | Rmax |
| ↓ | Rmax | Rmax | Rmax | Rmax |
| ← | Rmax | Rmax | Rmax | Rmax |

Transition counts

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | 0 | 0 | 0 | 0 |
| → | 0 | 0 | 0 | 0 |
| ↓ | 0 | 0 | 0 | 0 |
| ← | 0 | 0 | 0 | 0 |

# $S_0 = (1,1)$ - agent moves left
# t=1

Known status

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | U | U | U | U |
| → | U | U | U | U |
| ↓ | U | U | U | U |
| ← | U | U | U | U |

Reward counts

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | Rmax | Rmax | Rmax | Rmax |
| → | Rmax | Rmax | Rmax | Rmax |
| ↓ | Rmax | Rmax | Rmax | Rmax |
| ← | Rmax [ 0 ] | Rmax | Rmax | Rmax |

Transition counts

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | 0 | 0 | 0 | 0 |
| → | 0 | 0 | 0 | 0 |
| ↓ | 0 | 0 | 0 | 0 |
| ← | 1 | 0 | 0 | 0 |

# $S_1 = (1,1)$ - agent moves left
# t=2

Known status

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | U | U | U | U |
| → | U | U | U | U |
| ↓ | U | U | U | U |
| ← | K | U | U | U |

Reward counts

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | Rmax | Rmax | Rmax | Rmax |
| → | Rmax | Rmax | Rmax | Rmax |
| ↓ | Rmax | Rmax | Rmax | Rmax |
| ← | 0 | Rmax | Rmax | Rmax |

Transition counts

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | 0 | 0 | 0 | 0 |
| → | 0 | 0 | 0 | 0 |
| ↓ | 0 | 0 | 0 | 0 |
| ← | 2 | 0 | 0 | 0 |

$\hat{T}((1,1), \leftarrow, s' \neq (1,1)) = 0$

$\hat{T}((1,1), \leftarrow, (1,1)) = 1$

# $S_2 =$ (1,1) - agent moves up
# t=3

Known status

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | U | U | U | U |
| → | U | U | U | U |
| ↓ | U | U | U | U |
| ← | K | U | U | U |

Reward counts

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | Rmax [ 0 ] | Rmax | Rmax | Rmax |
| → | Rmax | Rmax | Rmax | Rmax |
| ↓ | Rmax | Rmax | Rmax | Rmax |
| ← | 0 | Rmax | Rmax | Rmax |

Transition counts

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | 1 | 0 | 0 | 0 |
| → | 0 | 0 | 0 | 0 |
| ↓ | 0 | 0 | 0 | 0 |
| ← | 2 | 0 | 0 | 0 |

# $S_3 = (2,1)$ - agent moves down
# t=4

Known status

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | U | U | U | U |
| → | U | U | U | U |
| ↓ | U | U | U | U |
| ← | K | U | U | U |

Reward counts

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | Rmax [0] | Rmax | Rmax | Rmax |
| → | Rmax | Rmax | Rmax | Rmax |
| ↓ | Rmax | Rmax | Rmax [0] | Rmax |
| ← | 0 | Rmax | Rmax | Rmax |

Transition counts

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | 1 | 0 | 0 | 0 |
| → | 0 | 0 | 0 | 0 |
| ↓ | 0 | 0 | 1 | 0 |
| ← | 2 | 0 | 0 | 0 |

# $S_4 =$ (1,1) - agent moves up
# t=5

Known status

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | K | U | U | U |
| → | U | U | U | U |
| ↓ | U | U | U | U |
| ← | K | U | U | U |

Reward counts

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | 0 | Rmax | Rmax | Rmax |
| → | Rmax | Rmax | Rmax | Rmax |
| ↓ | Rmax | Rmax | Rmax [ 0 ] | Rmax |
| ← | 0 | Rmax | Rmax | Rmax |

Transition counts

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | 2 | 0 | 0 | 0 |
| → | 0 | 0 | 0 | 0 |
| ↓ | 0 | 0 | 1 | 0 |
| ← | 2 | 0 | 0 | 0 |

$\hat{T}((1,1), \uparrow, s' \neq (2,1)) = 0$

$\hat{T}((1,1), \uparrow, (2,1)) = 1$

# $S_5 = (2,1)$ - agent moves right
# t=6

Known status

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | K | U | U | U |
| → | U | U | U | U |
| ↓ | U | U | U | U |
| ← | K | U | U | U |

Reward counts

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | 0 | Rmax | Rmax | Rmax |
| → | Rmax | Rmax | Rmax [0] | Rmax |
| ↓ | Rmax | Rmax | Rmax [0] | Rmax |
| ← | 0 | Rmax | Rmax | Rmax |

Transition counts

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | 2 | 0 | 0 | 0 |
| → | 0 | 0 | 1 | 0 |
| ↓ | 0 | 0 | 1 | 0 |
| ← | 2 | 0 | 0 | 0 |

# After $|S| \cdot |A| = 16$ changes: (this is a DDP):

Known status

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | K | K | K | K |
| → | K | K | K | K |
| ↓ | K | K | K | K |
| ← | K | K | K | K |

Transition counts

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | 2 | 2 | 2 | 2 |
| → | 2 | 2 | 2 | 2 |
| ↓ | 2 | 2 | 2 | 2 |
| ← | 2 | 2 | 2 | 2 |

Reward counts

| | (1,1) | (1,2) | (2,1) | (2,2) |
|---|---|---|---|---|
| ↑ | 0 | 0 | 0 | 1 |
| → | 0 | 0 | 0 | 1 |
| ↓ | 0 | 0 | 0 | 1 |
| ← | 0 | 0 | 0 | 1 |

$\hat{T} = T, \hat{R} = R$