GRID WORLD PROBLEM/GAME

$$\begin{bmatrix} \boxed{s_1} & \boxed{s_2} & \boxed{s_3} & \boxed{s_4} \\ \boxed{s_5} & \boxed{s_6} & \boxed{s_7} & \boxed{s_8} \\ \boxed{s_9} & \boxed{s_{10}} & \boxed{s_{11}} & \boxed{s_1} \end{bmatrix}$$

Bellman Eq

$$V_{i+1}(s) = \max_{a \in \mathbb{A}} \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V_i(s')]$$

Value iteration algorithm

Algorithm
Start with $V_0(s) = 0 \ \forall s \in \{s_1, s_2, \dots, s_{12}\}$ For i=1:1:H $\quad \forall s \in \{s_1, s_2, \dots, s_{12}\}$: $$V_{i+1}(s) \leftarrow \max_{a \in \mathbb{A}} \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V_i(s')]$$ $$\pi_{i+1}(s) \leftarrow arg \max_{a \in \mathbb{A}} \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V_i(s')]$$ This is called value-update or Bellman-update

$V_i^*(s)$ = expected sum of rewards accumulated starting from state s, acting optimally for i steps

$\pi_i^*(s)$ = optimal action when in state s and getting to act for i steps

THE GRID WORLD PROBLEM

$$\begin{bmatrix} \boxed{s_1} & \boxed{s_2} & \boxed{s_3} & \boxed{s_4} \\ \boxed{s_5} & \boxed{s_6} & \boxed{s_7} & \boxed{s_8} \\ \boxed{s_9} & \boxed{s_{10}} & \boxed{s_{11}} & \boxed{s_1} \end{bmatrix}$$

i	$V_i(s_1)$	$V_i(s_2)$	$V_i(s_3)$	$V_i(s_4)$	$V_i(s_5)$	$V_i(s_6)$	$V_i(s_7)$	$V_i(s_8)$	$V_i(s_9)$	$V_i(s_{10})$	$V_i(s_{11})$	$V_i(s_{12})$
1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0	-1	0	0	0	0

For i=2,3,…,H for the terminal states you write the rewards directly

(w/o computing the value using Bellman Eq)

$$R(s_4, a, s_4) = +1$$

$$R(s_8, a, s_8) = -1$$

When i=3:

i	$V_i(s_1)$	$V_i(s_2)$	$V_i(s_3)$	$V_i(s_4)$	$V_i(s_5)$	$V_i(s_6)$	$V_i(s_7)$	$V_i(s_8)$	$V_i(s_9)$	$V_i(s_{10})$	$V_i(s_{11})$	$V_i(s_{12})$
1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0	-1	0	0	0	0
3				1				-1				

$$V_{i+1}(s) \leftarrow \max_{a \in \mathbb{A}} \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V_i(s')]$$

$\boxed{\rightarrow}$ $T(s_3,a,s_4)[R(s_3,a,s_4)+(0.9)V_2(s_4)]+$ $\boxed{\rightarrow}$: $\boxed{\uparrow}$ $T(s_3,a,s_3)[R(s_3,a,s_3)+(0.9)V_2(s_3)]+$ $\boxed{\downarrow}$ $T(s_3,a,s_7)[R(s_3,a,s_7)+(0.9)V_2(s_7)]$	$\boxed{\rightarrow}$ $0.8[0+(0.9)(1)]+$ $\boxed{\rightarrow}$: $\boxed{\uparrow}$ $0.1[0+(0.9)(0)]+$ $\boxed{\downarrow}$ $0.1[0+(0.9)(0)]$	0.72
$\boxed{\uparrow}$ $T(s_3,a,s_3)[R(s_3,a,s_3)+(0.9)V_2(s_3)]+$ $\boxed{\uparrow}$: $\boxed{\rightarrow}$ $T(s_3,a,s_4)[R(s_3,a,s_4)+(0.9)V_2(s_4)]+$ $\boxed{\leftarrow}$ $T(s_3,a,s_2)[R(s_3,a,s_2)+(0.9)V_2(s_2)]$	$\boxed{\uparrow}$ $0.8[0+(0.9)(0)]+$ $\boxed{\uparrow}$: $\boxed{\rightarrow}$ $0.1[0+(0.9)(1)]+$ $\boxed{\leftarrow}$ $0.1[0+(0.9)(0)]$	0.09
$\boxed{\leftarrow}$ $T(s_3,a,s_2)[R(s_3,a,s_2)+(0.9)V_2(s_2)]+$ $\boxed{\leftarrow}$: $\boxed{\uparrow}$ $T(s_3,a,s_3)[R(s_3,a,s_3)+(0.9)V_2(s_3)]+$ $\boxed{\downarrow}$ $T(s_3,a,s_7)[R(s_3,a,s_7)+(0.9)V_2(s_7)]$	$\boxed{\leftarrow}$ $0.8[0+(0.9)(0)]+$ $\boxed{\leftarrow}$: $\boxed{\uparrow}$ $0.1[0+(0.9)(0)]+$ $\boxed{\downarrow}$ $0.1[0+(0.9)(0)]$	0
$\boxed{\downarrow}$ $T(s_3,a,s_7)[R(s_3,a,s_7)+(0.9)V_2(s_7)]+$ $\boxed{\downarrow}$: $\boxed{\leftarrow}$ $T(s_3,a,s_2)[R(s_3,a,s_2)+(0.9)V_2(s_2)]+$ $\boxed{\rightarrow}$ $T(s_3,a,s_4)[R(s_3,a,s_4)+(0.9)V_2(s_4)]$	$\boxed{\downarrow}$ $0.8[0+(0.9)(0)]+$ $\boxed{\downarrow}$: $\boxed{\leftarrow}$ $0.1[0+(0.9)(0)]+$ $\boxed{\rightarrow}$ $0.1[0+(0.9)(1)]$	0.09

And take my word for it, the values of the other states are zero.

When i=4:

i	$V_i(s_1)$	$V_i(s_2)$	$V_i(s_3)$	$V_i(s_4)$	$V_i(s_5)$	$V_i(s_6)$	$V_i(s_7)$	$V_i(s_8)$	$V_i(s_9)$	$V_i(s_{10})$	$V_i(s_{11})$	$V_i(s_{12})$
1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0	-1	0	0	0	0
3	0	0	0.72	1	0	0	0	-1	0	0	0	0
4				1		0		-1				

$$V_{i+1}(s) \leftarrow \max_{a\in\mathbb{A}} \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V_i(s')]$$

For s=2

$\boxed{\rightarrow}\ T(s_2,a,s_3)[R(s_2,a,s_3)+(0.9)V_3(s_3)]+$ $\boxed{\rightarrow}$: $\boxed{\uparrow}\ T(s_2,a,s_2)[R(s_2,a,s_2)+(0.9)V_3(s_2)]+$ $\boxed{\downarrow}\ T(s_2,a,s_2)[R(s_2,a,s_2)+(0.9)V_3(s_2)]$	$\boxed{\rightarrow}\ 0.8[0+(0.9)(0.72)]+$ $\boxed{\rightarrow}$: $\boxed{\uparrow}\ 0.1[0+(0.9)(0)]+$ $\boxed{\downarrow}\ 0.1[0+(0.9)(0)]$	0.5184 max
$\boxed{\uparrow}\ T(s_2,a,s_2)[R(s_2,a,s_2)+(0.9)V_3(s_2)]+$ $\boxed{\uparrow}$: $\boxed{\rightarrow}\ T(s_2,a,s_3)[R(s_2,a,s_3)+(0.9)V_3(s_3)]+$ $\boxed{\leftarrow}\ T(s_2,a,s_1)[R(s_2,a,s_1)+(0.9)V_3(s_1)]$	$\boxed{\uparrow}\ 0.8[0+(0.9)(0)]+$ $\boxed{\uparrow}$: $\boxed{\rightarrow}\ 0.1[0+(0.9)(0.72)]+$ $\boxed{\leftarrow}\ 0.1[0+(0.9)(0)]$	0.0648
$\boxed{\leftarrow}\ T(s_2,a,s_1)[R(s_2,a,s_1)+(0.9)V_3(s_1)]+$ $\boxed{\leftarrow}$: $\boxed{\uparrow}\ T(s_2,a,s_2)[R(s_2,a,s_2)+(0.9)V_3(s_2)]+$ $\boxed{\downarrow}\ T(s_2,a,s_2)[R(s_2,a,s_2)+(0.9)V_3(s_2)]$	$\boxed{\leftarrow}\ 0.8[0+(0.9)(0)]+$ $\boxed{\leftarrow}$: $\boxed{\uparrow}\ 0.1[0+(0.9)(0)]+$ $\boxed{\downarrow}\ 0.1[0+(0.9)(0)]$	0
$\boxed{\downarrow}\ T(s_2,a,s_2)[R(s_2,a,s_2)+(0.9)V_3(s_2)]+$ $\boxed{\downarrow}$: $\boxed{\leftarrow}\ T(s_2,a,s_1)[R(s_2,a,s_1)+(0.9)V_3(s_1)]+$ $\boxed{\rightarrow}\ T(s_2,a,s_3)[R(s_2,a,s_3)+(0.9)V_3(s_3)]$	$\boxed{\downarrow}\ 0.8[0+(0.9)(0)]+$ $\boxed{\downarrow}$: $\boxed{\leftarrow}\ 0.1[0+(0.9)(0)]+$ $\boxed{\rightarrow}\ 0.1[0+(0.9)(0.72)]$	0.0648

For s=3

$\boxed{\rightarrow}\ T(s_3,a,s_4)[R(s_3,a,s_4)+(0.9)V_3(s_4)]+$ $\boxed{\rightarrow}$: $\boxed{\uparrow}\ T(s_3,a,s_3)[R(s_3,a,s_3)+(0.9)V_3(s_3)]+$ $\boxed{\downarrow}\ T(s_3,a,s_7)[R(s_3,a,s_7)+(0.9)V_3(s_7)]$	$\boxed{\rightarrow}\ 0.8[0+(0.9)(1)]+$ $\boxed{\rightarrow}$: $\boxed{\uparrow}\ 0.1[0+(0.9)(0.72)]+$ $\boxed{\downarrow}\ 0.1[0+(0.9)(0)]$	0.7848 max
$\boxed{\uparrow}\ T(s_3,a,s_3)[R(s_3,a,s_3)+(0.9)V_3(s_3)]+$ $\boxed{\uparrow}$: $\boxed{\rightarrow}\ T(s_3,a,s_4)[R(s_3,a,s_4)+(0.9)V_3(s_4)]+$ $\boxed{\leftarrow}\ T(s_3,a,s_2)[R(s_3,a,s_2)+(0.9)V_3(s_2)]$	$\boxed{\uparrow}\ 0.8[0+(0.9)(0.72)]+$ $\boxed{\uparrow}$: $\boxed{\rightarrow}\ 0.1[0+(0.9)(1)]+$ $\boxed{\leftarrow}\ 0.1[0+(0.9)(0)]$	0.6084
$\boxed{\leftarrow}\ T(s_3,a,s_2)[R(s_3,a,s_2)+(0.9)V_3(s_2)]+$ $\boxed{\leftarrow}$: $\boxed{\uparrow}\ T(s_3,a,s_3)[R(s_3,a,s_3)+(0.9)V_3(s_3)]+$ $\boxed{\downarrow}\ T(s_3,a,s_7)[R(s_3,a,s_7)+(0.9)V_3(s_7)]$	$\boxed{\leftarrow}\ 0.8[0+(0.9)(0)]+$ $\boxed{\leftarrow}$: $\boxed{\uparrow}\ 0.1[0+(0.9)(0.72)]+$ $\boxed{\downarrow}\ 0.1[0+(0.9)(0)]$	0.0648
$\boxed{\downarrow}\ T(s_3,a,s_7)[R(s_3,a,s_7)+(0.9)V_3(s_7)]+$ $\boxed{\downarrow}$: $\boxed{\leftarrow}\ T(s_3,a,s_2)[R(s_3,a,s_2)+(0.9)V_3(s_2)]+$ $\boxed{\rightarrow}\ T(s_3,a,s_4)[R(s_3,a,s_4)+(0.9)V_3(s_4)]$	$\boxed{\downarrow}\ 0.8[0+(0.9)(0)]+$ $\boxed{\downarrow}$: $\boxed{\leftarrow}\ 0.1[0+(0.9)(0)]+$ $\boxed{\rightarrow}\ 0.1[0+(0.9)(1)]$	0.09

For s=7

$\boxed{\rightarrow}$: $\boxed{\rightarrow}\ T(s_7, a, s_8)[R(s_7, a, s_8) + (0.9)V_3(s_8)] +$ $\boxed{\uparrow}\ T(s_7, a, s_3)[R(s_7, a, s_3) + (0.9)V_3(s_3)] +$ $\boxed{\downarrow}\ T(s_7, a, s_{11})[R(s_7, a, s_{11}) + (0.9)V_3(s_{11})]$	$\boxed{\rightarrow}$: $\boxed{\rightarrow}\ 0.8[0 + (0.9)(-1)] +$ $\boxed{\uparrow}\ 0.1[0 + (0.9)(0.72)] +$ $\boxed{\downarrow}\ 0.1[0 + (0.9)(0)]$	-0.2016
$\boxed{\uparrow}$: $\boxed{\uparrow}\ T(s_7, a, s_3)[R(s_7, a, s_3) + (0.9)V_3(s_3)] +$ $\boxed{\rightarrow}\ T(s_7, a, s_8)[R(s_7, a, s_8) + (0.9)V_3(s_8)] +$ $\boxed{\leftarrow}\ T(s_7, a, s_7)[R(s_7, a, s_7) + (0.9)V_3(s_7)]$	$\boxed{\uparrow}$: $\boxed{\uparrow}\ 0.8[0 + (0.9)(0.72)] +$ $\boxed{\rightarrow}\ 0.1[0 + (0.9)(-1)] +$ $\boxed{\leftarrow}\ 0.1[0 + (0.9)(0)]$	0.4284 max
$\boxed{\leftarrow}$: $\boxed{\leftarrow}\ T(s_7, a, s_7)[R(s_7, a, s_7) + (0.9)V_3(s_7)] +$ $\boxed{\uparrow}\ T(s_7, a, s_3)[R(s_7, a, s_3) + (0.9)V_3(s_3)] +$ $\boxed{\downarrow}\ T(s_7, a, s_{11})[R(s_7, a, s_{11}) + (0.9)V_3(s_{11})]$	$\boxed{\leftarrow}$: $\boxed{\leftarrow}\ 0.8[0 + (0.9)(0)] +$ $\boxed{\uparrow}\ 0.1[0 + (0.9)(0.72)] +$ $\boxed{\downarrow}\ 0.1[0 + (0.9)(0)]$	0.0648
$\boxed{\downarrow}$: $\boxed{\downarrow}\ T(s_7, a, s_{11})[R(s_7, a, s_{11}) + (0.9)V_3(s_{11})] +$ $\boxed{\leftarrow}\ T(s_7, a, s_7)[R(s_7, a, s_7) + (0.9)V_3(s_7)] +$ $\boxed{\rightarrow}\ T(s_7, a, s_8)[R(s_7, a, s_8) + (0.9)V_3(s_8)]$	$\boxed{\downarrow}$: $\boxed{\downarrow}\ 0.8[0 + (0.9)(0)] +$ $\boxed{\leftarrow}\ 0.1[0 + (0.9)(0)] +$ $\boxed{\rightarrow}\ 0.1[0 + (0.9)(-1)]$	-0.09

i	$V_i(s_1)$	$V_i(s_2)$	$V_i(s_3)$	$V_i(s_4)$	$V_i(s_5)$	$V_i(s_6)$	$V_i(s_7)$	$V_i(s_8)$	$V_i(s_9)$	$V_i(s_{10})$	$V_i(s_{11})$	$V_i(s_{12})$
1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0	-1	0	0	0	0
3	0	0	0.72	1	0	0	0	-1	0	0	0	0
4	0	0.5184	0.7848	1	0	0	0.4284	-1	0	0	0	0

$$s\_prime\_list\_generator(state, action)$$

For each state-action pair there is a corresponding s-prime-list