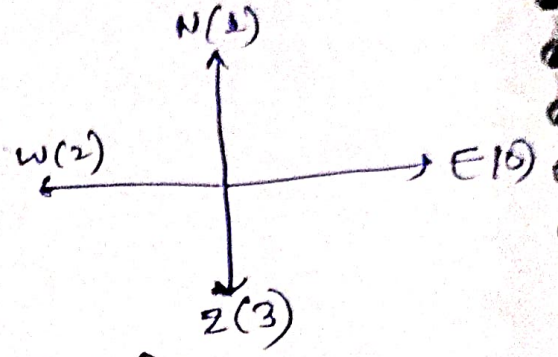
Ques 2: (b) Considering an example of value Iteration for 3 iterations.

Let take an example of 3×3 grid

(0,2)	(1,2)	(2,2)
(0,1)	(1,1)	(2,1)
(0,0)	(1,0)	(2,0)



Goal state = (2,2)
Obstacle state = (1,1)

East :- (x+1, y)
West :- (x-1, y)
North :- (x, y+1)
South → (x, y-1)

$\gamma = 0.9$

$\theta \in [0,1,2,3]$

rewards :-

step = -1
Collision = -100
goal = +50

Bellman Optimality Equation :-

$$V(s) = \max_{\pi} \sum_{a} \pi(a|s) \left(\sum_{r} p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) v(s') \right),$$

$$\forall s \in S$$

the value function is initialized $V_0(s) = 0$ for all state, this will help us an initial estimate and allow value iteration to iteratively update the value function using Bellman Optimality Equation.

Iteration 0:

$$V_0(s) = 0 \quad \forall s$$

Iteration 1:

since $V_0 = 0$;

$$V_1(s) = \max_{\pi} \sum_a \pi(a|s) \left(\sum_r p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) V_0(s') \right)$$

substituting $V_0(s') = 0$,

$$V_1(s) = \max_{\pi} \sum_a \pi(a|s) \left(\sum_r p(r|s,a) r \right)$$

$$= \max_a \sum_r p(r|s,a) r$$

{ best policy is deterministic & greedy,

$$\pi[a|s] = \begin{cases} 1, & a = a^* \\ 0, & \text{otherwise} \end{cases}$$

Case 1: near goal state

taking Example:

1) $(1, 2, 0)$

$(x, y, 0)$

$$V_1 = \underbrace{0.5 \times (50)}_{\text{goal}} + \overbrace{0.1(-100)}^{\text{boundary}} + \overbrace{0.1(-100)}^{\text{obstacle}}$$

$$= 40 - 10 - 10 = 20$$

] Forward Action

2) (illegible)

(illegible)

Reward = -1
Deterministic (prob 1)

$$\sum_{r} r\, P(r/s,a) = 1(-1) = -1$$

] Turn left

Reward = -1
Deterministic (prob 1)

$$\sum_{r} r\, P(r/s,a) = 1(-1) = -1$$

] Turn Right

$$\boxed{V_1(1,2,0) = \max(20,-1,-1) = 20}$$

for iteration 2:

$$V_2(s) = \max_a \left(\sum_r r\, p(r/s,a) + \gamma \sum_{s'} P(s'/s,a) V_1(s') \right)$$

$$V_1(1,2,0) = 20$$

⟹ for forward action:-

$$\sum_r p(r/s,a) = 0.8 \times 50 + 0.1 \times (-100) + 0.1(-100)$$
$$= 20$$

future reward term:

$$\gamma \sum_{s'} p(s'/s,a) V_1(s') = 0 \quad \text{[all outcomes are terminal]}$$

$$Q(s, \text{forward}) = 20$$

⇒ Turn left →

• turn left → next state = (1, 2, 1)

From iteration 1:

$$V_1(1,2,1) = -1$$

Immediate reward :—

$$\sum_r r\, p(r|s,a) = -2$$

future value :

$$\gamma \sum_{s'} p(s'|s,a) V_1(s') = 0.9 \times (-1) = 0.9$$

$$Q(s, \text{Turn left}) = -1 - 0.9 = -1.9$$

⇒ Similarly,

Turn Right →

Turn right → next state = (1, 2, 3)

Immediate reward :—

$$\sum_r r\, p(r|s,a) = -1$$

Future value:

$$\gamma \sum_{s'} p(s'|s,a) V_1(s') = 0.9 \times (-1) = 0.9$$

$$Q(s, \text{Turn Right}) = -1.9$$

$$V_2(1,2,0) = \max(20, -1.9, -1.9) = 20$$

Iteration 3:—

$V_2(1,2,0) = 20$

$$V_3(s) = \max_a \left(\sum_{s'} r\, p(s'|s,a) + \gamma \sum_{s'} p(s'|s,a)\, V_2(s') \right)$$

Ex:—

$s = (2,2,0)$ Action Forward

Immediate reward

$$\sum_{s'} r\, p(s'|s,a) = 0.8 \times 50 + 0.1 \times -100 + 0.1 \times -100$$

$$= 20$$

all outcomes are terminal, no future rewards

$$\boxed{Q(\text{forward}) = 20}$$

Turn Left

Turn left → new state = (1,2,1)

$V_2(1,2,1) = -1.9$

immediate reward

$$\sum_{r} r\, p(r|s,a) = -1$$

future rewards:—

$$\gamma \sum_{s'} p(s'|s,a) V_2(s') = 0.9 \times (-1.9)$$

$$= -1.71$$

$$\boxed{Q(\text{Turn left}) = -2.71}$$

Turn Right

Turn Right → new state = (1, 2, 3)

immediate reward:—

$$\sum r\,p(r|s,a) = -1$$

future reward

$$\gamma \sum_{s'} p(s'|s,a)\,V_2(s') = 0.9 \times (-1.9)$$
$$= -1.71$$

Total = $-1 - 1.71 = -2.71$

$$Q(\text{Turn Right}) = -2.71$$

$$V_3(1,2,0) = \max(20, -2.71, -2.71) = 20$$

Case II - distant corner state

for Example $s = (0,0,0)$

$V_0(s) = 0$ → Initialization & iteration 0

Iteration 1

$V_0 = 0$, future term disappears.

Forward Action

$$V_1(s) = \max_a \sum_r r\,p(r|s,a)$$

$$= 0.8 \times -1 + 0.1 \times -1 + 0.1 \times -100$$

$$= -0.8 - 0.1 - 10 = -10.9$$

$$Q(\text{forward Action}) = -10.9$$

Turn left: { Probability = 1, Reward = -1 } Deterministic

$$\sum_r r\,p(r|s,a) = 1(-1) = -1$$

$$q(s, \text{Turn left}) = -1$$

Turn right

similarly,

$$\sum_r r\,p(r|s,a) = 1(-1) = -1$$

$$q(s, \text{Turn Right}) = -1$$

Total maximum

$$V_2(0,0,0) = \max(-11.71, -1.9, -1.9)$$
$$= -1.9$$

Iteration 3

$$V_2(0,0,0) = -1.9$$
$$V_2(1,0,0) = -1.9$$
$$V_2(0,1,0) = -1.9$$

Immediate rewards :-

$$\sum_r r\,p(r|s,a) = 0.8 \times (-1) + 0.1 \times (-1) + 0.1 \times (-100)$$
$$= -10.9$$

Future rewards :-

$$\gamma \sum_{s'} p(s'|s,a)\, V_2(s') =$$

$$0.9 \times (0.8(-1.9) + 0.1(-1.9))$$

$$= -1.539$$

$Q(\text{forward}) = -10.9 - 1.539$
$= -12.439$

Turn Right :-

Immediate reward:

$$\sum_r r p(r|s,a) = -1$$

future rewards:

$$\gamma \sum_{s'} p(s'|s,a) V_2(s') = 0.9(-1.9) = -1.71$$

$Q(\text{Turn Right}) = -1 - 1.71 = -2.71$

Turn left :-

Immediate reward:

$$\sum_a r p(a|s,a) = -1$$

Future Rewards :-

$$\gamma \sum_{s'} p(s'|s,a) V_2(s') = 0.9(-1.9)$$
$$= -1.71$$

$Q(\text{Turn Left}) = -1.71 - 1$
$= -2.71$

Total maximum:

$$V_3(0,0,0) = \max(-12.439, -2.71, -2.71)$$

$$\boxed{V_3(0,0,0) = -2.71}$$

State	V_1	V_2	V_3
(1,2,0)	20	20	20
(0,0,0)	-1	-1.9	-2.71

from this, we can observe, that state who are near goal state are stable at value 20, but corner distant state become increasingly negative due to cumulative step cost and collision risk. the near goal state having positive stable value because terminal reward dominates.

Optimal Policy

again using same example, using the 3X3 guid:

Goal = (2,2)
Obstacle = (1,1)
$\gamma = 0.9$
step = -1

After performing value iterations until convergence, we obtain optimal value $V^*(s)$

optimal value function satisfies Bellman optimality equation:

$$V^*(s) = \max_a \left(\sum_r p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) V^*(s') \right)$$

$$Q^*(s,a) = \sum_r p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) V^*(s')$$

$$\boxed{V^*(s) = \max_a Q^*(s,a)}$$

Optimal policy:

$$\boxed{\pi^*(s) = \arg\max_a Q^*(s,a)}$$

this action maximizes the expected discounted return.

Ex In 3×3 grid:

after value iterations converges, we obtained:

$V^*(1,2,0) = 20$

$V^*(0,0,0) = -2.71$

Case 1: State $s = (1,2,0)$
(Position (1,2), facing East)

Optimal policy :-

$$\pi^*(s) = \arg\max_a Q^*(s,a)$$

where,

$$Q^*(s,a) = \sum_{r} p(r|s,a)\, r + \gamma \sum_{s'} p(s'|s,a)\, V^*(s')$$

Action: Forward

Possible transitions:

* 0.8 → (2,2) Goal → reward = 50
* 0.1 → Boundary → reward = −100
* 0.1 → Obstacle → reward = −100

Since goal is terminal, future value is 0

$$Q^*(1,2,0,\text{Forward}) = 0.8(50) + 0.1(-100) + 0.1(-100)$$

$$= 20$$

Action : TurnLeft

turning gives immediate reward

$$= -1$$

New state value :

$$V^*(1,2,1) = -2.71$$

$$Q^*(1,2,0,\text{TurnLeft}) = -1 + 0.9(-2.71)$$

$$= -1 - 2.439 = -3.439$$

Action: Turn Right

similarly:

$$Q^*(1,2,0,\text{Turn Right}) = -1 + 0.9(-2.71)$$
$$= -3.439$$

Compare:

$Q^*(\text{forward}) = 20$

$Q^*(\text{Turn Left}) = -3.439$

$Q^*(\text{Turn Right}) = -3.439$

$$\boxed{\pi^*(1,2,0) = \text{forward}}$$

Case 2 $S = (0,0,0)$

Position, $(0,0)$, facing East

$$V^*(0,0,0) = -2.71$$

Action: Forward

$$Q^*(0,0,0,\text{forward}) =$$

~~$Q^*(1,2,0,\text{East})$~~

$$0.8[-1 + 0.9(-2.71)] + 0.1[-1 + 0.9(-2.71)] + 0.1(-100)$$

$$= -2.7512 - 0.3439 - 10$$

$$= -13.0951$$

Action: Turn Left

$$Q^*(0,0,0, \text{Turn Left}) = -1 + 0.9V^*(0,0,1)$$
$$= -1 + 0.9(-2.71)$$
$$= -3.439$$

Action: Turn Right

$$Q^*(0,0,0, \text{Turn Right}) = -1 + 0.9V^*(0,0,3)$$
$$= -1 + 0.9 \times (-2.71)$$
$$= -3.439$$

Now $Q^*(\text{Turn left}) = Q^*(\text{Turn Right})$
$= -3.439$

$Q^*(\text{forward}) = -13.095$

$-3.439 > -13.095$

$\pi^*(0,0,0) =$ Turn left or Turn Right

For manual computation of first three iterations, a reduced 3X3 grid was used to clearly demonstrate the Bellman process, the same implementation generalizes to 10x10 grid Environment

c) Policy iteration consist 2 steps:—

(1) Policy Evaluation $\Rightarrow V_{\pi_k} = r_{\pi_k} + \gamma P_{\pi_k} V_{\pi_k}$

(2) Policy improvement $\Rightarrow \pi_{k+1} = \arg\max_{\pi} (r_\pi + \gamma P_\pi V_{\pi_k})$

$\Longrightarrow$ Initializing an arbitrary policy (assumption):

$\pi_0(s)$ = Forward $\forall s$

In every state, the agent choose Forward.

Now, policy Evaluation:

For fixed policy π, value function:

$$V^\pi(s) = \sum_r p(r|s,\pi(s))\, r + \gamma \sum_{s'} p(s'|s,\pi(s))\, V^\pi(s')$$

equation is solved Iteratively

Ex $s = (1,2,0)$ In grid size (3×3)

under policy π_0, action = Forward

transitions :-

0.8 → Goal → reward = 50

0.1 → Boundary → reward = -100

0.1 → Obstacle → reward = -100

Since, goal is terminal

$$V^{\pi_0}(1,2,0) =$$

$$= 0.8(50) + 0.1(-100) + 0.1(-100)$$

$$= 20$$

$$\boxed{V^{\pi_0}(1,2,0) = 20}$$

state $= (0,0,0)$

Transitions

$0.8 \rightarrow (1,0) \rightarrow$ reward $= -1$

$0.1 \rightarrow (1,0) \rightarrow$ reward $= -1$

$0.1 \rightarrow$ collision $\rightarrow$ reward $= -100$

$$V^{\pi_0}(0,0,0) = 0.8[-1 + 0.9V^{\pi_0}(1,0,0)] + 0.1[-1 + 0.9V^{\pi_0}(0,1,1)]$$

$$+ 0.1(-100)$$

$$= 0.8(-1) + 0.1(-1) + 0.1(-100)$$

$$= -0.8 - 0.1 - 10$$

$$= -10.9$$

$$V^{\pi_0}(0,0,0) = -10.9$$

Policy Improvement

$$Q^{\pi}(s,a) = \sum_{r} p(r/s,a)r + \gamma \sum_{s'} p(s'/s,a)V^{\pi}(s')$$

we compare all actions and choose the best.

we already computed,

$$Q^{\pi_0}(\text{Forward}) = -10.9$$

Turn-Left:

Turning gives, New state = (0,0,1)

Reward = -1

probability = 1

$$Q^{\pi_0}(\text{TurnLeft}) = -1 + 0.9V^{\pi_0}(0,0,1)$$

Initially, $V^{\pi_0}(0,0,1) = 0$

$$= -1 + 0.9(0) = -1$$

Similarly,

TurnRight:

$$Q^{\pi_0}(\text{TurnRight}) = -1 + 0.9(0) = -1$$

Now,

$$-1 > -10.9$$

policy improves,

$$\pi_1(0,0,0) = \text{TurnLeft (or TurnRight)}$$

second policy evaluation:

assuming, $\pi_1(0,0,0)$ = Turn left

$V^{\pi_1}(0,0,0) = -1 + 0.9\, V^{\pi}(0,0,1)$

& $(0,0,1)$ also turn similarly:

Eventually values propagate:

$V^{\pi_1}(0,0,0) = -1 + 0.9(-1 + 0.9(-1))$

$= -1 - 0.9 - 0.81$

$= -2.71$

It is:

$-1(-1 + 0.9 + 0.9^2)$

Final improvement check:

now encomputing Forward using updated values:

$Q(\text{Forward}) = 0.8[-1 + 0.9(-2.71)] + 0.1[-1 + 0.9(-2.71)] + 0.1(-100)$

earlier computed $= -13.095$

turning gives $= -1 + 0.9(-2.71) = -3.439$

Since:

$-3.439 > -13.095$

Policy remain same → stable

Final optimal policy:

$\pi^*(1,2,0)$ = Forward

$\pi^*(0,0,0)$ = Turn Right or Turn Left