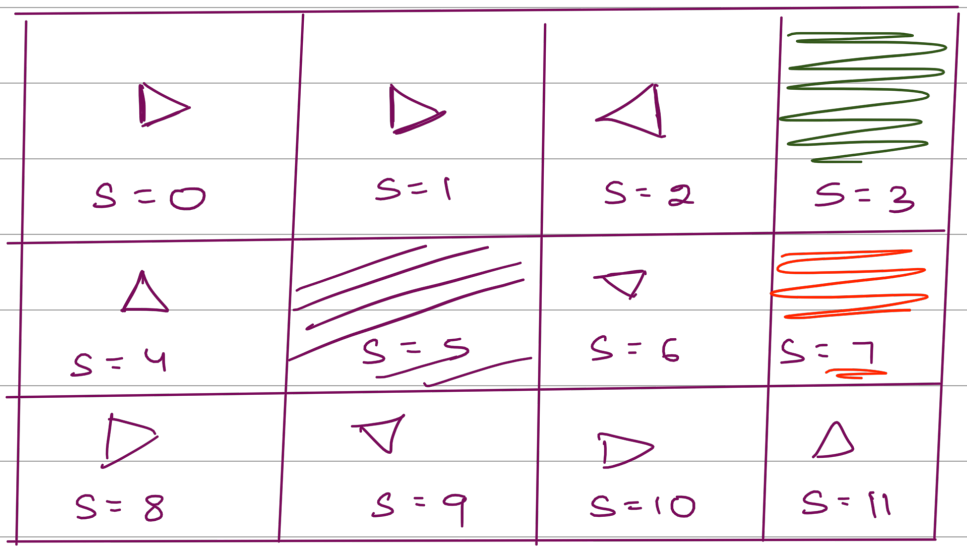
Policy Iteration



Given policy

$v(6)$ = rewards following policy $[6, 10, 11, 7]$

$= r(6) + r(10) + r(11) + r(7)$

For any state s at time $t = 0$, utility can be written as:-

$v(S_{t=0})$ = rewards following policy $[S_{t=0}, S_{t=1}, S_{t=2}, S_{t=3} \ldots]$

$= r(S_{t=0}) + r(S_{t=1}) + r(S_{t=2}) + \ldots$

$= \sum_t r(S_t)$

$v(6) = r(6) + r(10) + r(11) + r(7)$

utility of $S = 10$ $= r(10) + r(11) + r(7)$

$v(6) = r(6) + v(10)$

$v(S_{t=i}) = r(S_{t=i}) + v(S_{t=i+1})$

Utility of a state = immediate reward of that state + utility of its successor state following the given policy.

$$V(S) = \gamma(S) + V(S')$$

Discounted rewards:

$s = 0$

By following the policy, the future states are [0, 1, 2, 1, 2, 1, 2, 1, 2 ...] which is a loop of 1 and 2 forever.

If following the definition of utility

$$V(0) = \gamma(0) + \gamma(1) + \gamma(2) + \gamma(1) + \ldots \gamma(2)$$
$$= -\infty$$

Address the concept of Discounted Rewards

" Reward received in the future is less valuable than reward received right now." — Introduce discount factor γ

Additive rewards:-

$$v(s_{t=0}) = \sum_t r(s_t)$$

Using discounted rewards:

$$v(s_{t=0}) = \gamma^t \sum_t r(s_t)$$

$0 < \gamma < 1$ utility value is guaranteed to be bounded

finite no. of states = finite no. of rewards

$v(s_{t=0})$ = discounted rewards following policy $[s_{t=0}, s_{t=1}, s_{t=2}, s_{t=3}, \cdots]$

$$= r(s_{t=0}) + \gamma(s_{t=1}) + \gamma^2(s_{t=2}) \cdots$$

$$\boxed{v(s) = r(s) + \gamma v(s')}$$

$v(s_0)$ = discounted rewards following policy $[s_{t=0}, s_{t=1}, s_{t=2}, s_{t=3}, \cdots]$

$$= r(s_{t=0}) + \gamma r(s_{t=1}) + \gamma^2 r(s_{t=2}) + \gamma^3 r(s_{t=3}) + \cdots$$

$$\leq r_{max} + \gamma r_{max} + \gamma^2 r_{max} + \gamma^3 r_{max} + \cdots$$

$$= r_{max}\left[\frac{1}{1-\gamma}\right]$$

Add Back Stochastic :-

Deterministic Env

$s = 6$ determined path $[6, 10, 11, 7]$

$V(6) = r(6) + r(10) + r(11) + r(7)$

In Stochastic Env : not guaranteed to follow path $[6, 10, 11, 7]$

prob :- $(0.8)^3 = 0.512$

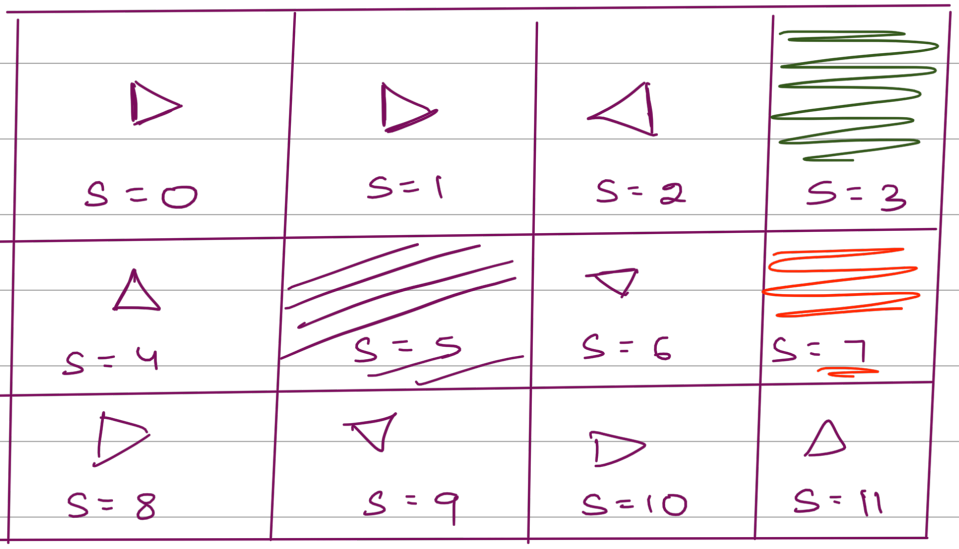
Define Utility in a stochastic Environment :-

$$V(s_0) = E\left[\sum_t \gamma^t r(s_t)\right]$$

E is the expectation w.r.t probability distribution over possible paths following a certain policy

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

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



Given policy

$S=6$:- to reach $S'=10$ Probability $=0.8$
to reach $S'=7$ Probability $=0.1$
to stay in 6 Probability $=0.1$

Therefore, the utility of $S=6$ is :-

$$V(6) = \gamma(6) + \beta\left[0.8\,V(10) + 0.1\,V(6) + 0.1\,V(7)\right]$$

$$V(0) = \gamma(0) + \beta\left[0.8\,V(1) + 0.1\,V(4) + 0.1\,V(0)\right]$$

$$V(1) = \gamma(1) + \beta\left[0.8\,V(2) + 0.1\,V(1) + 0.1\,V(1)\right]$$

$$V(2) = \gamma(2) + \beta\left[0.8\,V(1) + 0.1\,V(6) + 0.1\,V(2)\right]$$

$$V(3) = \gamma(3) + \beta\left[V(3)\right]$$

$$V(4) = \gamma(4) + \beta\left[0.8\,V(0) + 0.1\,V(4) + 0.1\,V(4)\right]$$

$$V(6) = \gamma(6) + \beta\left[0.8\,V(10) + 0.1\,V(6) + 0.1\,V(7)\right]$$

$$V(7) = \gamma(7) + \beta\left[V(7)\right]$$

$$V(8) = \gamma(8) + \beta\left[0.8\,V(9) + 0.1\,V(4) + 0.1\,V(8)\right]$$

$$V(9) = \gamma(9) + \beta\left[0.8\,V(9) + 0.1\,V(8) + 0.1\,V(10)\right]$$

$$V(10) = R(10) + \gamma\,[0.8V(11) + 0.1V(6) + 0.1V(10)]$$

$$V(11) = R(11) + \gamma\,[0.8V(7) + 0.1V(10) + 0.1V(11)]$$

11 Equations, 11 unknowns.

Policy Evaluation

Determine the utility of each state

DP approach – iterative policy evaluation or sweep.

① Initialize the utility of each state = 0

② Loop through the states

③ For each state, update the utility using the equation

$$V(S) = r(S) + \gamma \sum_{S'} P(S' \mid S, a = \pi(S))\, V(S')$$

Eg: set $\gamma = 0.5$

Loop through $S = 0$ to $S = 11$, we get

$V(0) = -0.04$

$V(1) = -0.04$

$V(2) = -0.056$

$V(3) = 1$

$V(4) = -0.056$

$V(6) = -0.04$

$V(7) = -1$

$V(8) = -0.0428$

$V(9) = -0.04214$

$V(10) = -0.042$

$V(11) = -0.4421$

in-place sweep use updated utility in the same sweep

$$V(1) = r(1) + \gamma [0.8\, V(2) + 0.1\, V(1) + 0.1\, V(1)]$$

$$= -0.04 + 0.5 [0.8 \times 0 + 0.1 \times 0 + 0.1 \times 0]$$

$$= -0.04$$

in-place sweep

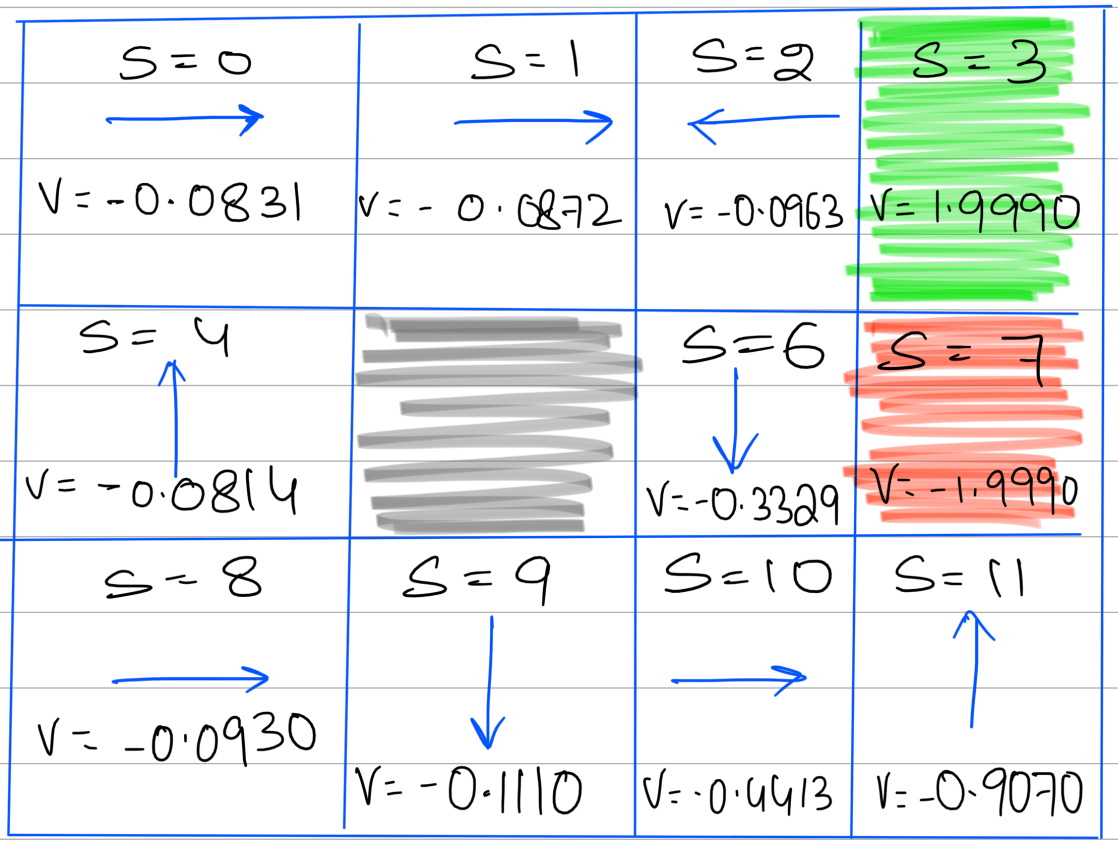
$$V(2) = r(2) + \gamma [0.8\, V(1) + 0.1\, V(6) + 0.1\, V(2)]$$

$$= -0.04 + 0.5 [0.8 (-0.04) + 0.1 \times 0 + 0.1 \times 0] = -0.056$$

We repeat sweeping for several times until the changes of utility values between consecutive sweeps

are marginal

Utility after 11 Sweeps :-



Here :- For $S = 6$

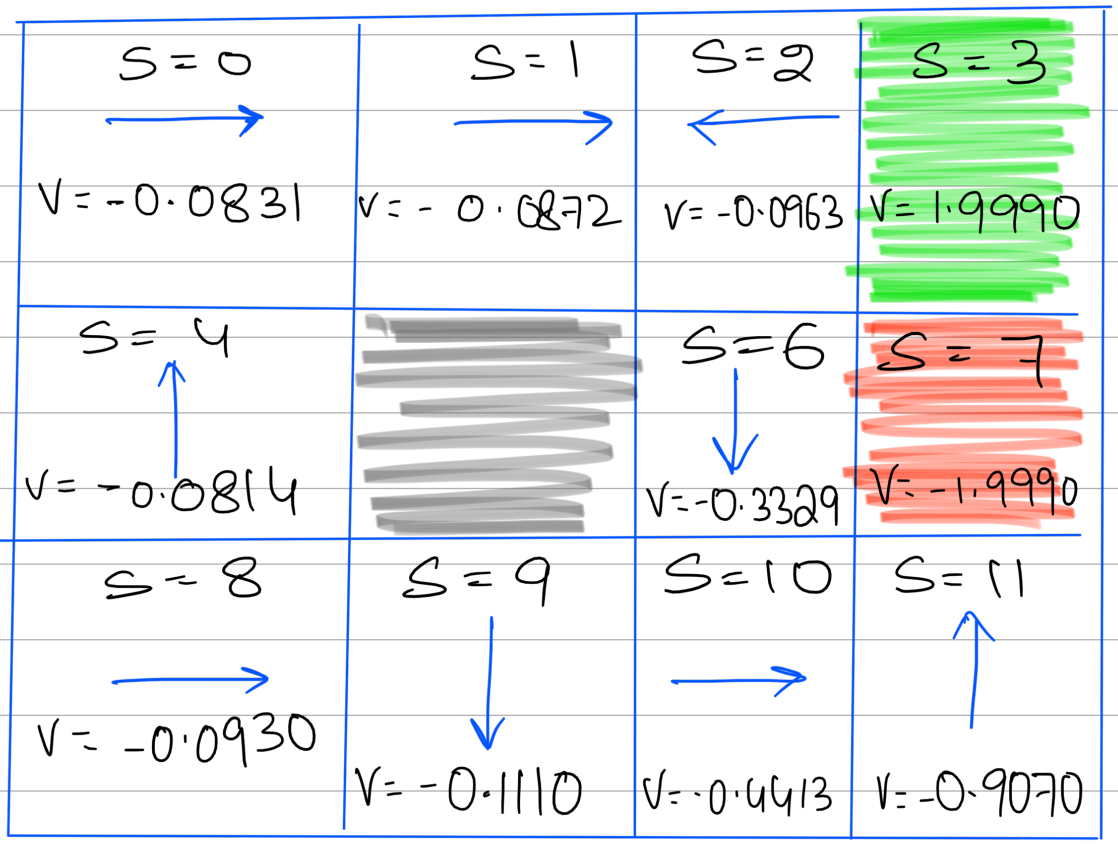
$$V(6) = r(6) + \beta \left[0.8\, V(10) + 0.1\, V(6) + 0.1\, V(7)\right]$$

$$V(6) = -0.04 + 0.5 \left[0.8\,(-0.4413) + 0.1\,(-0.3329) + 0.1\,(-1.9990)\right]$$

$$\simeq -0.3329$$

Sweep approach is alternate method to solve linear equations

Policy Improvement



① After we get the utility for this policy, it is time to improve the policy.

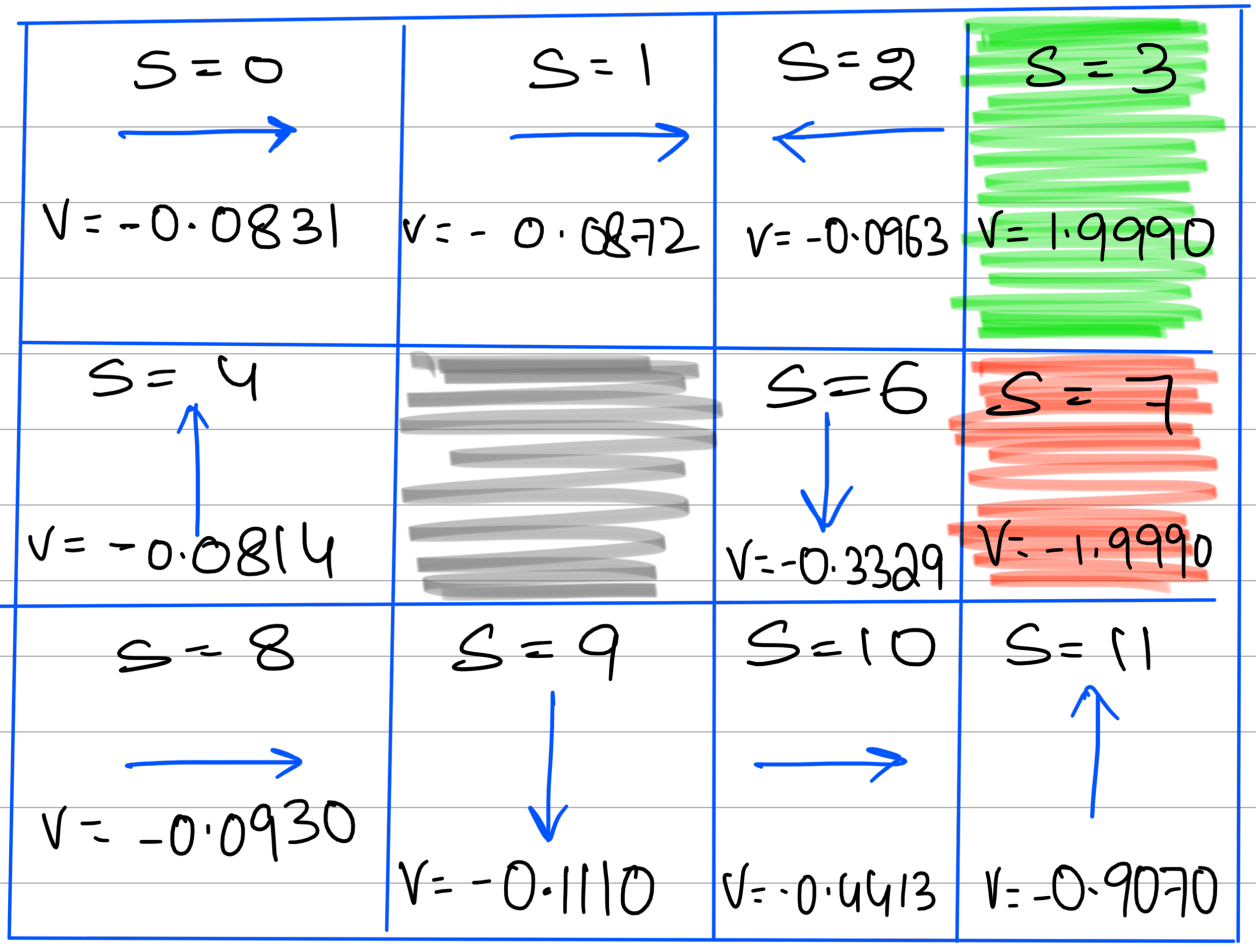
* Utility represents how good a state is.

② When choosing ACTIONS for a state, prefer successor states with higher utility

③ when $s=2$

potential successor state include 1,3,6

Clearly $s=3$ is best choice since $V=1.9990$



Since MDP is stochastic, selecting a preferable successor state does not guarantee we will reach it.

Rather than successor states, compare actions.

For S=6 [actions: UP, DOWN, RIGHT, LEFT]

choosing UP → the reward is

$0.8\,v(2) + 0.1\,v(6) + 0.1\,v(7)$

$= 0.8(-0.0963) + 0.1(-0.3329) + 0.1(-1.990) = -0.3093$ best choice

choosing LEFT → the reward is

$0.8\,v(6) + 0.1\,v(10) + 0.1\,v(2) = -0.3201$

$0.8(-0.3329) + 0.1(-0.4413) + 0.1(-0.0963)$

DOWN :-

$0.8\ v(10) + 0.1(v(6)) + 0.1\,v(7) = -0.5862$

$= 0.8[-0.4413) + 0.1(-0.3329) + 0.1(-1.9990)$

RIGHT:

$0.8\ v(7) + 0.1(v(2)) + 0.1\,v(10)$

$= 0.8(-1.9990) + 0.1(-0.0963) + 0.1(-0.4413)$

$= -1.6530$

Hence update policy of state $S=6$ to UP

Perform this process for each state

$$\pi(s) \leftarrow \text{argmax}_a \left[\sum_{s'} p(s'|s,a)\, v(s') \right]$$

After policy improvement, the policy looks like :-

S = 0 ↓ V = -0.0831	S = 1 ← V = -0.0872	S = 2 → V = -0.0963	S = 3 V = 1.9990
S = 4 → V = -0.0814		S = 6 ↑ V = -0.3329	S = 7 V = -1.9990
S = 8 ↑ V = -0.0930	S = 9 ← V = -0.1110	S = 10 ← V = -0.4413	S = 11 ← V = -0.9070

This is a better policy than the previous one. But, this is not the best one.

↳ Repeat process of policy evaluation and policy improvement again and again.

↳ Do not need to initialize utility to zero continue with the utility values, we got. This helps to converge faster

policy evaluation

* update utility values while keeping policy constant

from $\pi[i], V[i-1]$

↓

$\pi[i], V[i]$

policy improvement

* update policy while keeping utility values constant

from $\pi[i], V[i]$

↓

$\pi[i+1], V[i]$

When $\pi[i] == \pi[i+1]$

break; # convergence occurs

In example from:

Randomly generated policy $\pi[1]$

optimal policy π^* $\pi[6]$

Effects of discounted factor

Discounted factor γ is used to discount the rewards of future states when adding them together

$$V(S_{t=0}) = r(S_{t=0}) + \gamma\, r(S_{t=1}) + \gamma^2 r(S_{t=2}) \cdots$$

$$= \gamma^t \sum_t r(S_t)$$

If we choose 0.1 as γ, we have :-

$$V(S_{t=0}) = r(S_{t=0}) + 0.1\, r(S_{t=1}) + 0.1^2 r(S_{t=2}) \cdots$$

If we choose 0.9 as γ, we have :-

$$V(S_{t=0}) = r(S_{t=0}) + 0.9\, r(S_{t=1}) + 0.9^2 r(S_{t=2}) \cdots$$

Smaller γ (0.1) emphasizes immediate reward of current state

Larger γ (0.9) emphasizes long term reward in the future