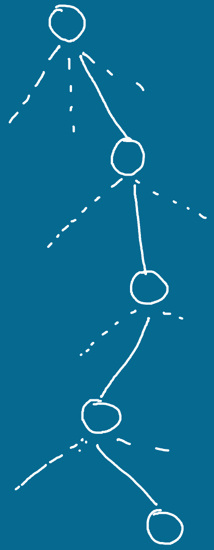
Value Backup



opponent

My move

TERMINATES

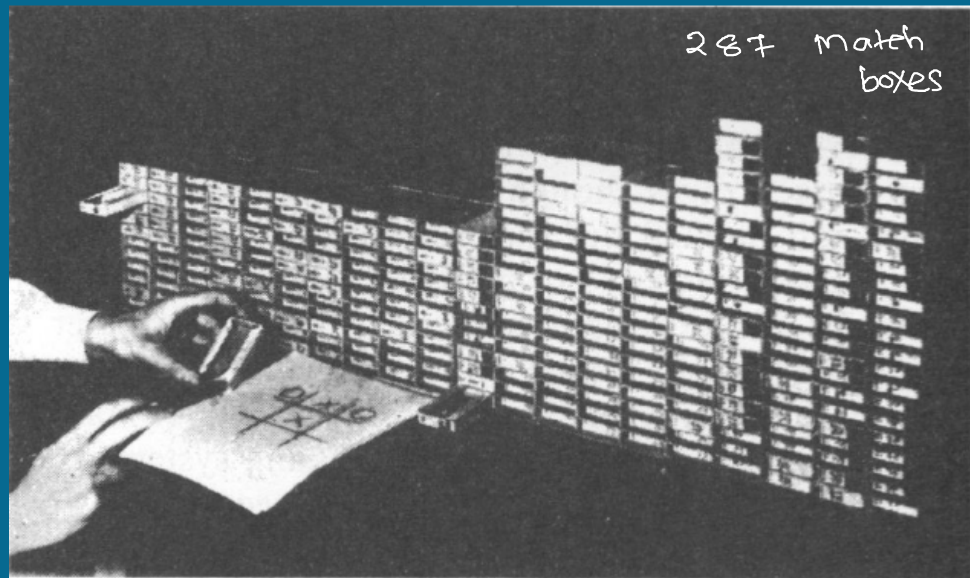
- Fully informed
- Partially informed

History

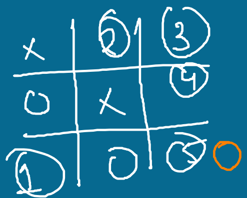
MENACE
(1961-62)

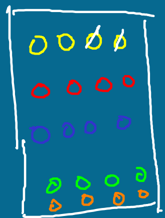
trial-and-error learning

(Donald Michie)



Matchbox Educable Noughts & Crosses Engine.





Menace — 'O'
I am — 'x'

N-Arm bandit :

Choices (Actions)

t	1	2	3	...	N
1		2			
2					
3					
4					
5					
⋮					
T					

R_t ← Reward from Environment

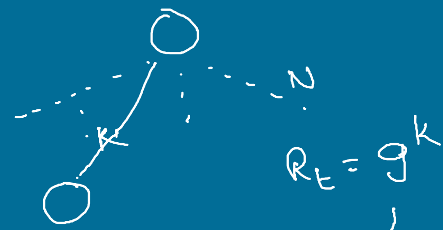
- Stationary distribution
- "Explore-Exploit" Time period

Expected total reward

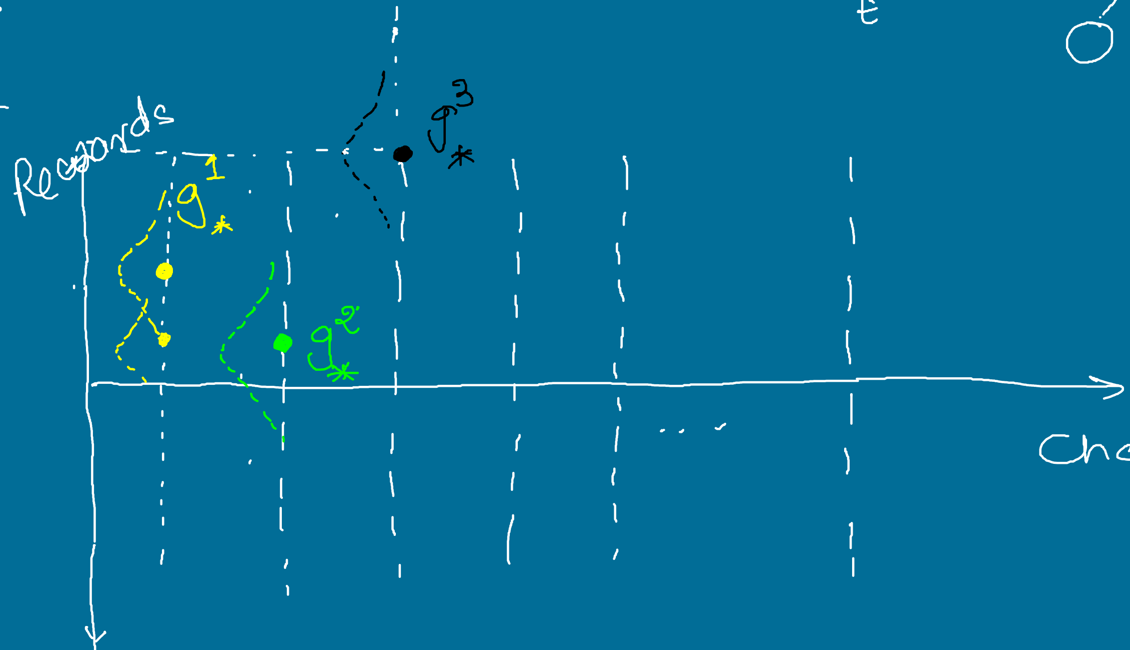
$$\mathbb{E}\left[R_1 + R_2 + \dots + R_T\right]$$

$t = 1$

$a_t = k$



Stationary distribution



Action value method :

$Q_1(a) = 0$

value ↑ action ↑

A_1	A_2	A_3	A_4	A_5	$A_6 \cdots$
a_2	$\boxed{a_1}$	$\boxed{a_1}$	a_2	$\boxed{a_1}$	
$R_1 = r(a_2)$	R_2	R_3	R_4	R_5	R_6

$a \in \{a_1, a_2\}$

μ_1, σ_1

μ_2, σ_2

$$Q_6(a_1) = \frac{\text{sum of rewards when } a_1 \text{ was taken prior to 't'}}{\#\ \text{action } a_1 \text{ was taken}}$$

at $t = 6$ $\longrightarrow$ $\boxed{A_6 = \underset{a \in \{a_1, a_2\}}{\operatorname{argmax}} Q_6(a)}$

$$Q_t(a) = \sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i = a} \Big/ \sum_{i=1}^{t-1} \mathbb{1}_{A_i = a}$$

$\mathbb{1}_{\text{predicate}} = \begin{cases} 1 & \text{True} \\ 0 & \text{False} \end{cases}$

Explore - Exploit : ε - greedy approach $\varepsilon > 0$

$$A_t \leftarrow \begin{cases} \arg\max_a Q_t(a) & \text{with prob } 1-\varepsilon \\ \text{a random action} & \text{with prob } \varepsilon \end{cases}$$

Incremental Implementation :

$$Q_n(a) \doteq \frac{R_1 + R_2 + \dots + R_{n-1}}{n-1}$$

R_i denotes the reward received after i^{th} selection of action a.

$$Q_{n+1}(a) = \frac{1}{n} \sum_{i=1}^{n} R_i$$

$$= \frac{1}{n} \left(R_n + \sum_{i=1}^{n-1} R_i \right)$$

$$= \frac{1}{n} \left(R_n + (n-1) Q_n(a) \right)$$

$$\boxed{Q_{n+1} = Q_n + \frac{1}{n}\left[R_n - Q_n\right]}$$

A simple Bandit Algorithm :

- stationary
- Non-stationary

↓

- $Q_{n+1} = Q_n + \alpha\left[R_n - Q_n\right]$

fixed

(weighted average)

$$Q_{n+1} = \underbrace{(1-\alpha)^n}\, Q_1 + \sum_{i=1}^{n} \underbrace{\alpha(1-\alpha)^{n-i}}\, R_i$$

$$(1-\alpha)^n + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = 1$$

- Initialization

$Q(a) \leftarrow 0 \quad \forall a$

$N(a) \leftarrow 0 \quad \forall a$

Loop forever

$$A \leftarrow \begin{cases} \operatorname{argmax}_a Q(a) & \sim 1-\varepsilon \\ \text{'a' random}, & \sim \varepsilon \end{cases}$$

$R \leftarrow \text{Bandit}(A)$

$N(A) \leftarrow N(A) + 1$

$Q(A) \leftarrow Q(A) + \frac{1}{n}\left[R - Q(A)\right]$

$$\sum_{n=1}^{\infty} \alpha_n(a) = \infty \quad \text{and} \quad \sum_{n=1}^{\infty} \alpha_n^2(a) < \infty$$

Reading Assignment:

2.6 Opti Init Values

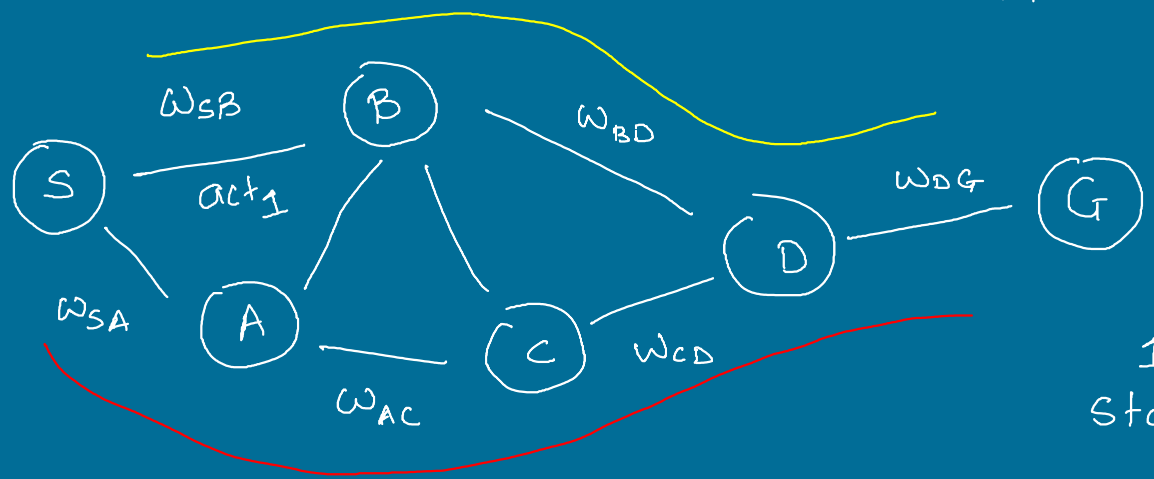
2.7 UCB Action Selection

2.8 Gradient Bandit Algorithm

Markov Decision Process:

(sequential)

Typical Search



action(state)

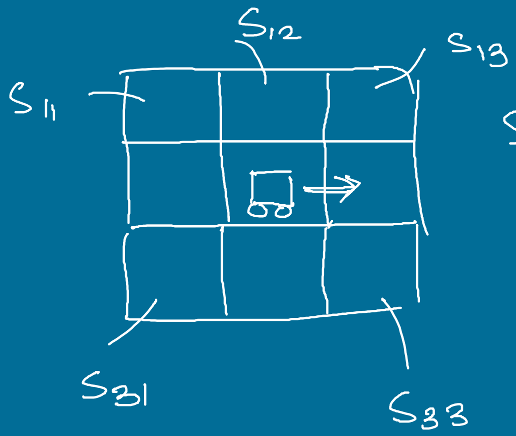
succ(state, action)

certain

1 certain 2

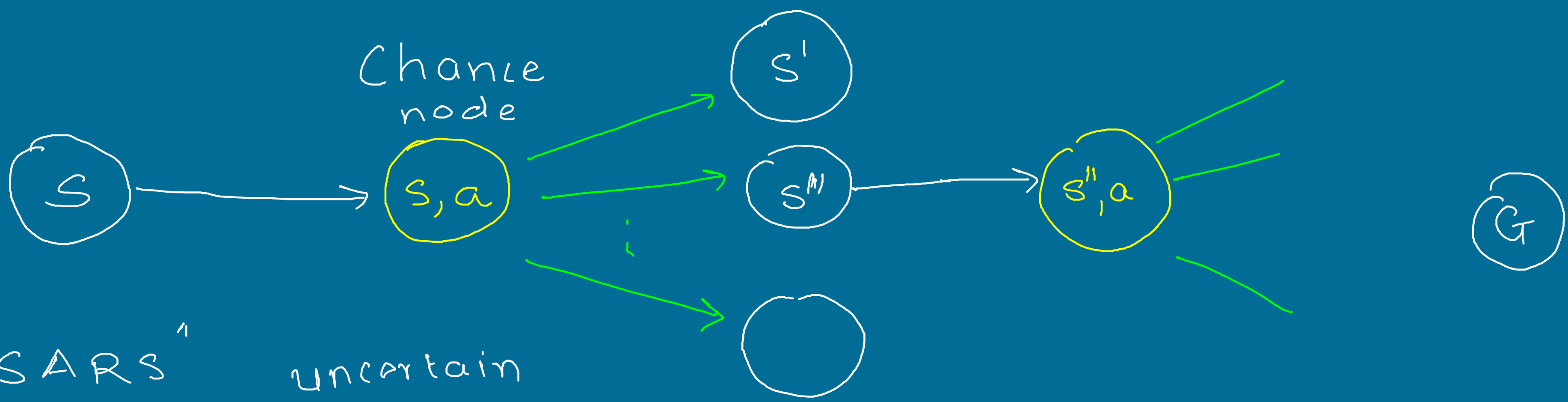
state, act, $\boxed{\text{weight}}$, state, act, $\boxed{\omega}$

$\boxed{\ }$ . . . , state = G



$\text{succ}(S_{22}, \rightarrow) \rightsquigarrow$

	Probability
S_{23}	0.5
S_{13}	0.05
S_{33}	0.05
S_{12}	0.2
S_{32}	0.2



"SARS"

uncertain

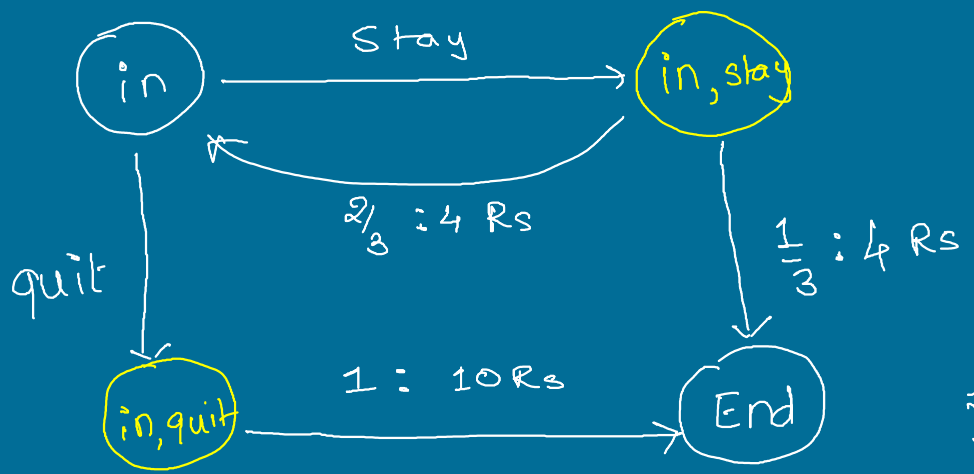
$$S_{t=0},\ A_{t=0},\ \underbrace{R_{t=1}}_{\text{non-deterministic}},\ S_{t=1},\ A_{t=1},\ R_{t=2},\ \ldots$$

uncertain

Example: Dice Game

- You choose to stay / quit
  (A) stay, (B) quit
- If you quit, you get 10 Rs and game ends
- if you stay, you get 4 Rs and then environment rolls a dice
  - + If the dice results in 2/4, game ends
  - + Otherwise, continue

Expected Total Reward



Policy

$\pi(\text{state}) = \text{action}$

$\pi(\text{in}) = \text{stay}$ | $\pi(\text{in}) = \text{quit}$

$\pi(\text{end}) \neq\ ?$ | 10 Rs.

Run 1: $S_0 = in$, $A_0 = stay$, $R_1 = 4Rs$, $S_1 = in$, $A_1 = stay$, $R_2 = 4Rs$, $S_2 = END$.

$\tilde{G}_0^1$ 8

Run 2: $S_0 = in$, $[A_0 = stay$, $R_1 = 4Rs$, $S_1 = in]$, $[A_1 = stay$, $R_2 = 4Rs$, $S_2 = in]$, $[A_2 = stay$, $R_3 = 4Rs$, $S_3 = in]$, $[A_3 = stay$, $R_4 = 4Rs$, $S_4 = END]$

0.44

$\tilde{G}_0^2$ 16

⋮

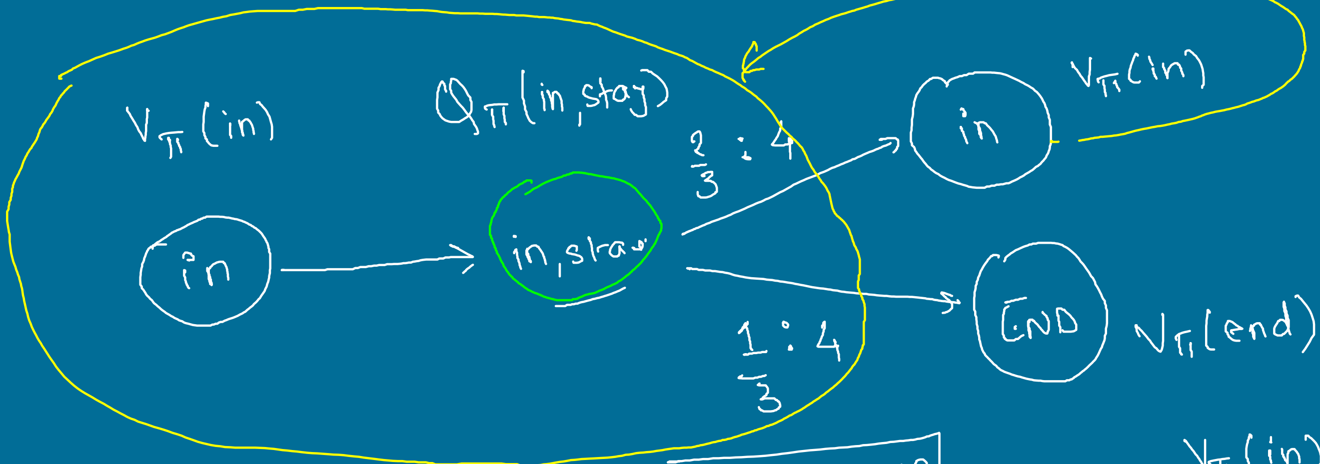
Run N

$$\frac{1000 \times 10}{1000 \times 12}$$

12

Avg Total Reward ~ $\boxed{\text{Expected Total Reward}}$ $\Rightarrow$ Expected Return $s = in$

$\pi(in) = stay$

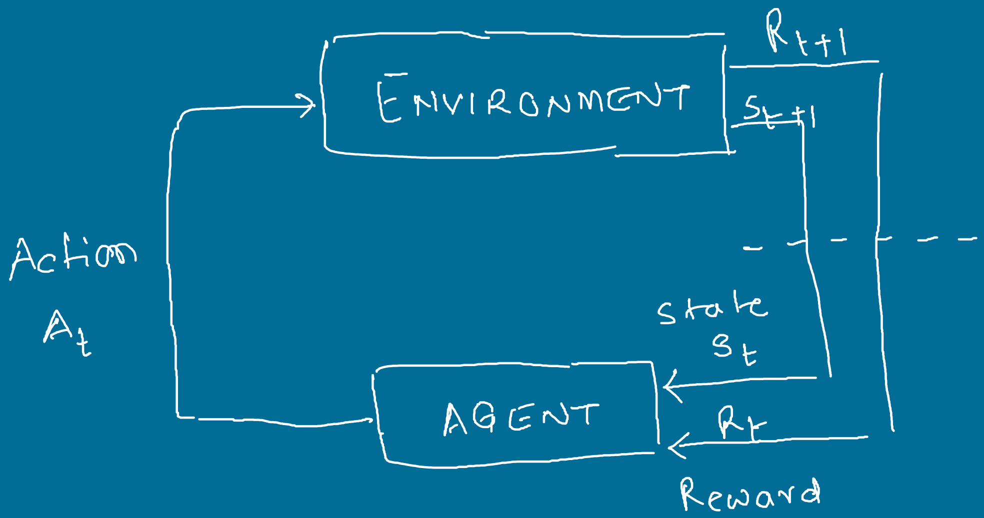


$\boxed{V_\pi(in) = 12}$

$$V_\pi(in) = \frac{2}{3}(4 + V_\pi(in)) + \frac{1}{3}(4 + \cancel{V_\pi(end)}^{\,0})$$

$$V_\pi(in) = 4 + \frac{2}{3} V_\pi(in)$$

Agent Environment Interface



$t = 0, 1, 2, \dots$

Trajectory

$$\overset{t}{S_0 \; A_0} \;\Big|\; \overset{t+1}{\underbrace{R_1 \; S_1 \; A_1}} \;\Big|\; \underbrace{R_2 \; S_2 \; A_2} \;\cdots$$

$S_t \in \mathcal{S}$, $A_t \in \mathcal{A}(S_t)$, $R_t \in \mathcal{R} \subseteq \mathbb{R}$

<u>Finite MDP</u> : $|\mathcal{S}| < \infty$, $|\mathcal{A}(s)| < \infty$, $|\mathcal{R}| < \infty$

$$p\left(S_{t+1} = s', R_{t+1} = r \;\middle|\; S_t = s, A_t = a\right)$$

$$p : \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} \longrightarrow [0, 1] \; ; \quad \sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r \mid s, a) = 1$$

$s \in \mathcal{S}$, $a \in \mathcal{A}(s)$.

$$\underbrace{p(s' \mid s, a)}_{\text{transition}} = \sum_{r} p(s', r \mid s, a)$$

$$r(s,a) = \sum_{r} r \sum_{s'} p(s', r \mid s, a)$$

$$\mathbb{E}\left[R_t \mid S_{t-1}=s, A_{t-1}=a\right]$$

$r(s,a)$

$r: S \times A \longrightarrow R$

final time step

$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \dots + R_T$$

(Return)

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+1+k}$$

$$\boxed{G_t = R_{t+1} + \gamma G_{t+1}}$$

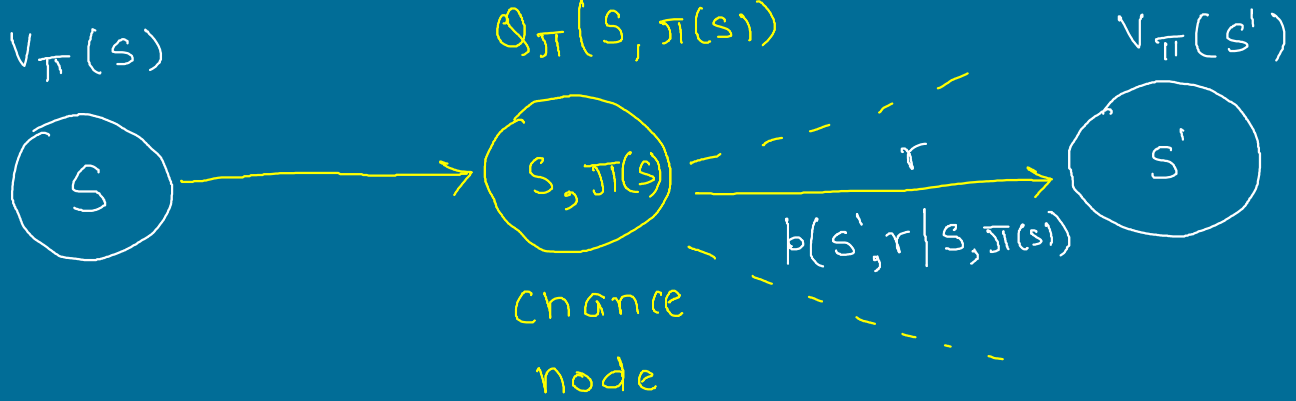
Value Function & Policy:

Value function
- function of states $V_\pi(s)$
- function of state-action $Q_\pi(s,a)$

$$V_\pi(s) = \mathbb{E}_\pi\left[G_t \mid S_t = s\right]$$

$$Q_\pi(s,a) = \mathbb{E}_\pi\left[G_t \mid S_t = s, A_t = a\right]$$

- $\pi : S \to A$ $\quad \pi(s) \in \mathcal{A}(s), \ s \in S$ Deterministic policy
- $\pi : S \times A \to [0,1]$ $\quad \pi(a \mid s)$ probability value



Deterministic Policy

$\pi(s)$

$$V_\pi(s) = \begin{cases} 0 & \text{if } s \text{ is the END / TERMINAL STATE} \\ Q_\pi(s,\pi(s)) & \end{cases}$$

$$Q_\pi(s,\pi(s)) = \sum_{r,s'} p(s',r|s,\pi(s))\left[r+\gamma V_\pi(s')\right]$$

$$\boxed{V_\pi(s) = \sum_{r,s'} p(s',r|s,\pi(s))\left[r+\gamma V_\pi(s')\right]}$$

$$V_{\pi}^{iter+1}(s) = \sum_{r,s'} p(s',r \mid s,\pi(s)) \left[r + \gamma V_{\pi}^{iter}(s') \right]$$

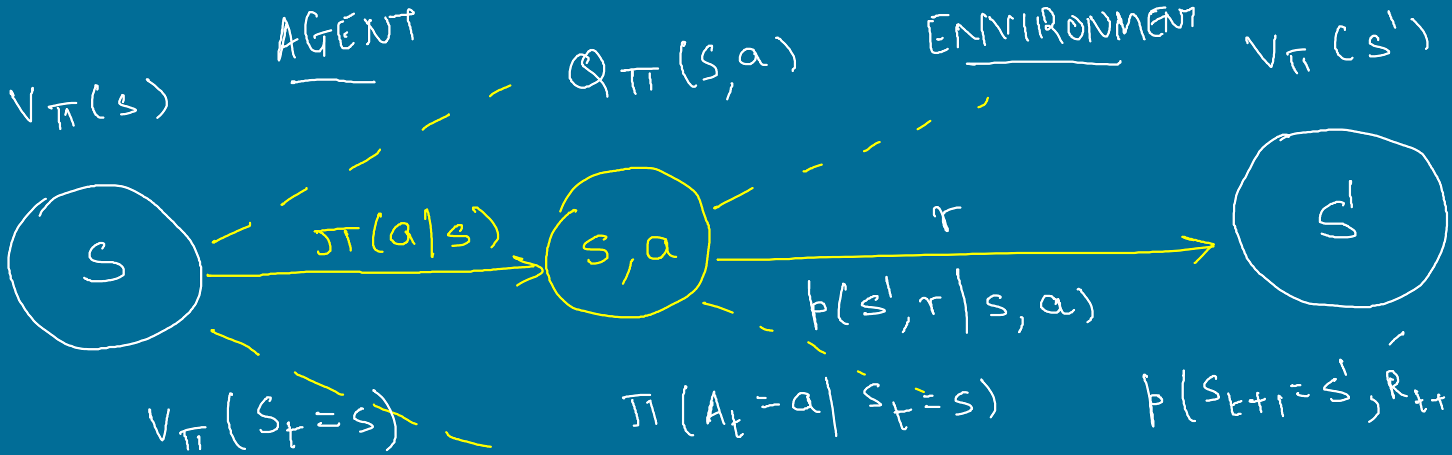
[Iterative Policy Evaluation]

$$\max_{s} \left| V_{\pi}^{iter+1}(s) - V_{\pi}^{iter}(s) \right| < \Delta$$

Non-deterministic Policy

$\pi(a \mid s)$

(Random)



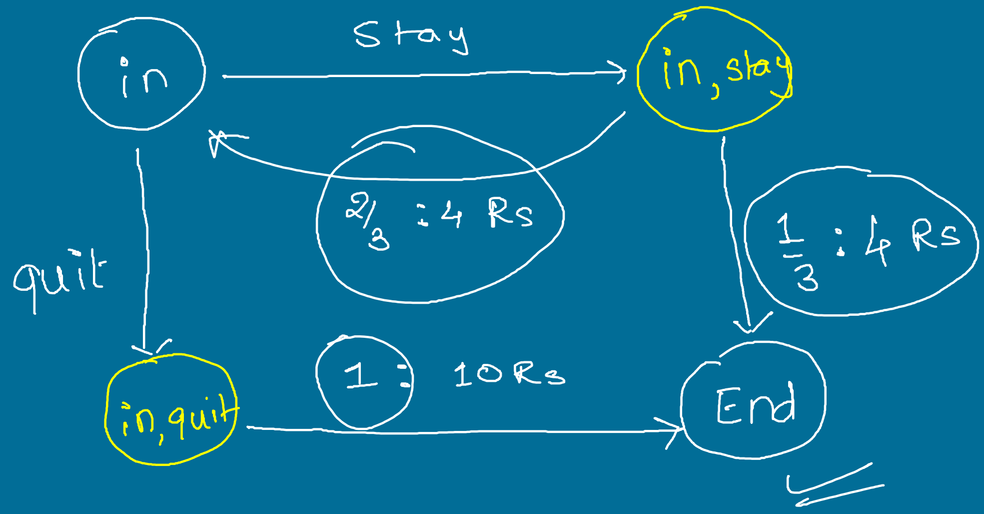
$\pi(A_t=a \mid S_t=s)$

$p(S_{t+1}=s', R_{t+1}=r \mid S_t=s, A_t=a)$

$$V_{\pi}(s) = \sum_{a} \pi(a \mid s) \sum_{s',r} p(s',r \mid s,a) \left[\underset{R_{t+1}}{r} + \gamma V_{\pi}(s') \right]$$

[s not an End/Terminal state]

$$V_\pi(s) = \sum_a \pi(a|s) \cdot Q_\pi(s,a)$$



$\gamma = 1$

deterministic

$\pi(\text{in}) = \text{stay}$

$\pi(\text{end}) = \{\ \}$

$V_\pi(\text{in}) = 12$

$V_\pi(\text{End}) = 0$

$\gamma = 1$

nondeterministic

$\pi(\text{stay}|\text{in}) = 0.5$

$\pi(\text{quit}|\text{in}) = 0.5$

$V_\pi(\text{in}) = \ ?$

$V_\pi(\text{end}) = 0$

$$V_\pi(\text{in}) = \left\{ \overset{1/2}{\pi(\text{stay}|\text{in})} \left(\overset{2/3}{p(\text{in},4|\text{in},\text{stay})} \cdot [4 + V_\pi(\text{in})] + \overset{1/3}{p(\text{end},4|\text{in},\text{stay})} \cdot [4 + \overset{0}{\cancel{V_\pi(\text{end})}}] \right\} +$$

$$\left\{ \overset{1/2}{\pi(\text{quit}|\text{in})} \left(\overset{1}{p(\text{end},10|\text{in},\text{quit})} \cdot [10 + \overset{0}{\cancel{V_\pi(\text{end})}}] \right\}$$

$$V_\pi(\text{in}) = \frac{1}{2}\left[\underset{A}{\frac{2}{3} \cdot (4 + V_\pi(\text{in}))} + \frac{1}{3} \cdot 4 \right] + \frac{1}{2}\left[\underset{B}{1 \cdot 10} \right] \quad \#$$

$$2 V_\pi(in) = \frac{2}{3} V_\pi(in) + 14$$

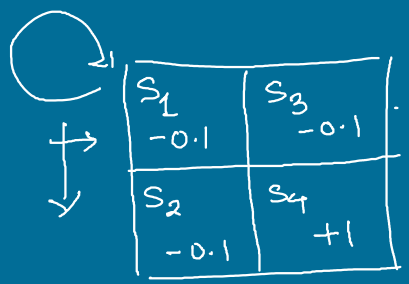
10 — 12

$$V_\pi(in) = \frac{3}{4} \times 14 = 10.5$$

< 12

> 12

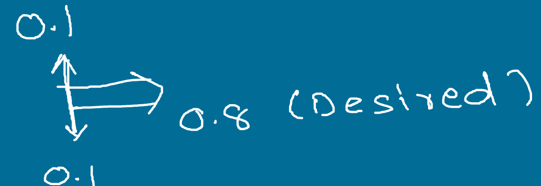
≈ 12

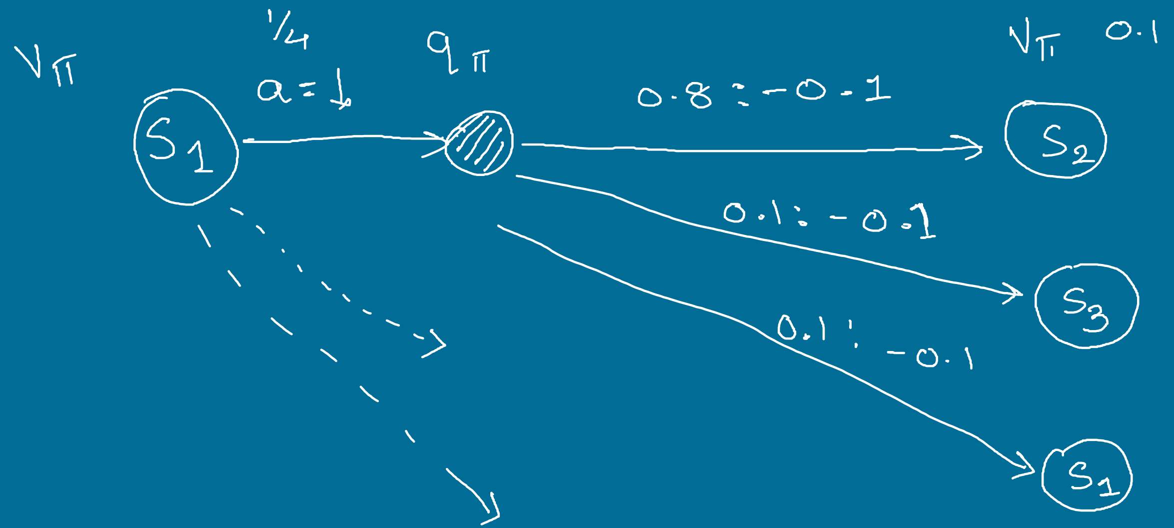


Grid World Example:

$\pi(a|s) = 1/4$ Policy $a \in \{\uparrow, \downarrow, \rightarrow, \leftarrow\}$

Environment is stochastic:



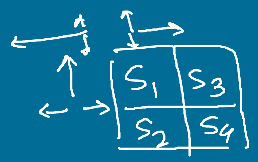


Value Function (state)

$$V_\pi(s) = \begin{cases} +1 & \text{if } s = S_4 \\ \sum_a \pi(a|s)\, q_\pi(s,a) & \end{cases}$$

(s is not a terminal state)

$$V_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\left[r + \gamma \cdot V_\pi(s')\right]$$



$$V_\pi(s_1) = \underbrace{\pi(\uparrow|s_1)}_{1/4} \cdot \left[0.8\left(-0.1 + \gamma V_\pi(s_1)\right) + 0.1\left(-0.1 + \gamma V_\pi(s_1)\right) \quad 0.1\left(-0.1 + \gamma V_\pi(s_3)\right) \right]$$

$$+ \underbrace{\pi(\downarrow|s_1)}_{1/4} \left[0.8\left(-0.1 + \gamma V_\pi(s_2)\right) + 0.1\left(-0.1 + \gamma V_\pi(s_3)\right) + 0.1\left(-0.1 + \gamma V_\pi(s_1)\right) \right]$$

$$+ \underbrace{\pi(\rightarrow|s_1)}_{1/4} \left[0.8\left(-0.1 + \gamma V_\pi(s_3)\right) + 0.1\left(-0.1 + \gamma V_\pi(s_1)\right) + 0.1\left(-0.1 + \gamma V_\pi(s_2)\right) \right]$$

$$+ \underbrace{\pi(\leftarrow|s_2)}_{1/4} \left[0.9\left(-0.1 + \gamma V_\pi(s_1)\right) + 0.1\left(-0.1 + \gamma V_\pi(s_2)\right) \right]$$

$$V_\pi(s_1) = \_\_ V_\pi(s_1) + \_\_ V_\pi(s_2) + \_\_ V_\pi(s_3) + \_\_ V_\pi(s_4) + C_1 \quad (1)$$

$$I \begin{bmatrix} V_\pi(s_1) \\ V_\pi(s_2) \\ V_\pi(s_3) \\ V_\pi(s_4) \end{bmatrix} - \begin{bmatrix} & & A & & \end{bmatrix} \begin{bmatrix} V_\pi(s_1) \\ V_\pi(s_2) \\ V_\pi(s_3) \\ V_\pi(s_4) \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} \quad \text{--- } (I)$$

$V_\pi(s_4) = +1$

$$(I - A)\, v = c \quad \Rightarrow \quad \boxed{v = (I - A)^{-1} c}$$

Iterative Policy Evaluation

$$V_\pi(s) = \sum_a \pi(a|s) \sum_{s', r} p(s', r | s, a)\left[r + \gamma \cdot V_\pi(s')\right]$$

$$V_\pi^{iter+1} = \sum_a \pi(a|s) \sum_{s', r} p(s', r | s, a)\left[r + \gamma\, V_\pi^{iter}(s')\right]$$

Input π, the policy to be evaluated
Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop:
    $\Delta \leftarrow 0$
    Loop for each $s \in \mathcal{S}$:
        $v \leftarrow V(s)$
        $\big(V(s) \leftarrow \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)\big[r + \gamma V(s')\big] \quad \big)$
        $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

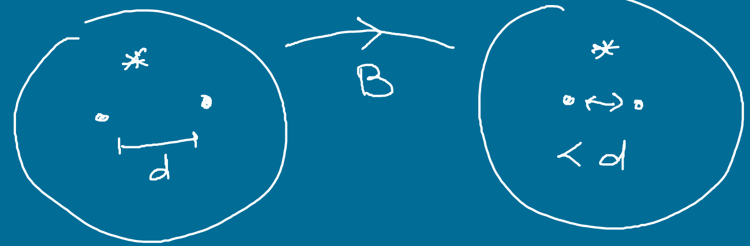
(I) Two arrays for maintaining V^{iter}, V^{iter+1}

(II) In-place

Q. Convergence? "Contraction" and "Fixed point thm"

$V^0 = V^{init}$

$$V^{i+1} = \underbrace{B}_{\text{contraction}} V^i = B^i V^{init} \qquad B : \mathbb{R}^4 \longrightarrow \mathbb{R}^4$$



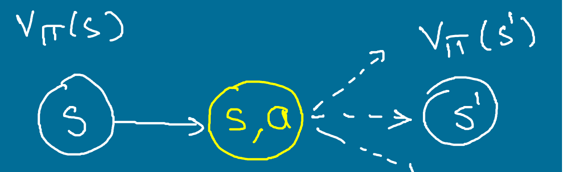
$B\,\underline{*} = \underline{*}$

08.04.2021 Summary

- Finite MDP
- Value functions < state value / state action

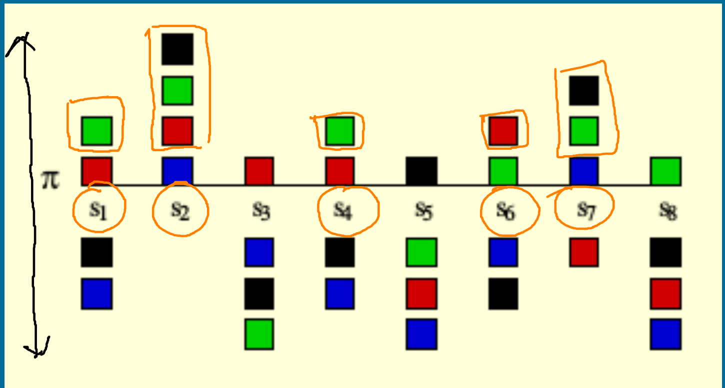
$V_\pi(s)$

$Q_\pi(s, a)$



[Bellman Eqn.]

Policy Improvement: Given a policy π



$$Q_\pi(s,a) = \sum_{s',r} p(s',r|s,a)[r + \gamma V_\pi(s')]$$

$$Q_\pi(s, GREEN) > Q_\pi(s, RED)$$

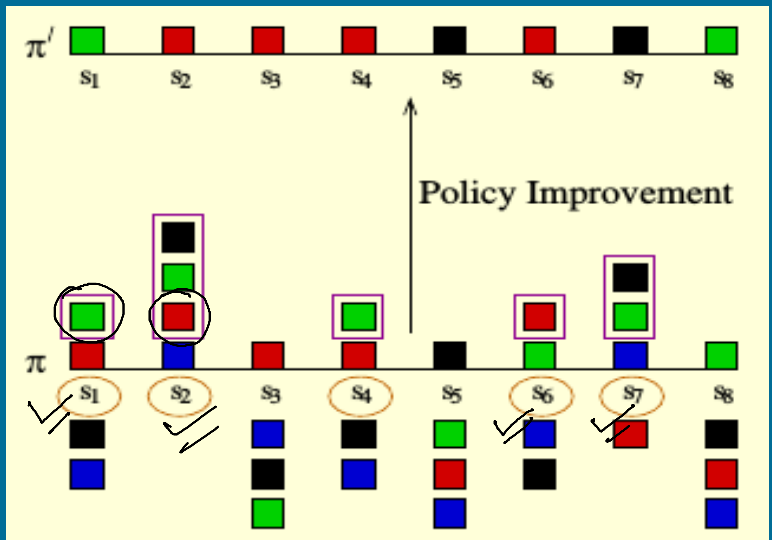
S_1 .... S_8

- Pick one or more improvable states and in them
- switch to an arbitrary improving action

Improvable states s_1, s_2, s_4, s_6, s_7

Improvable actions {[G], [B,G,R], [G], [R], [B,G]}

- Let the resulting policy be π'.



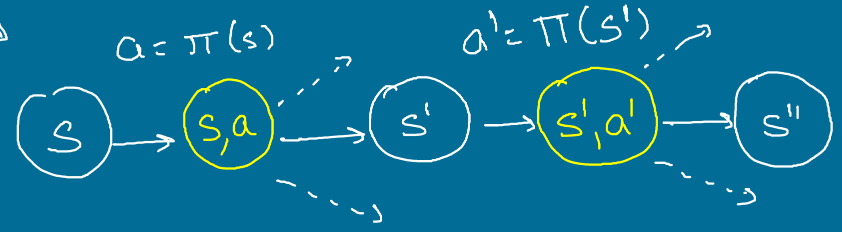
Policy Improvement Theorem:

(1) If π has no improvable states, then it is OPTIMAL else

(2) If π' is obtainable as above then

$\forall s \in S: V^{\pi'}(s) \geq V^{\pi}(s)$, $\exists s \in S: V^{\pi'}(s) > V^{\pi}(s)$

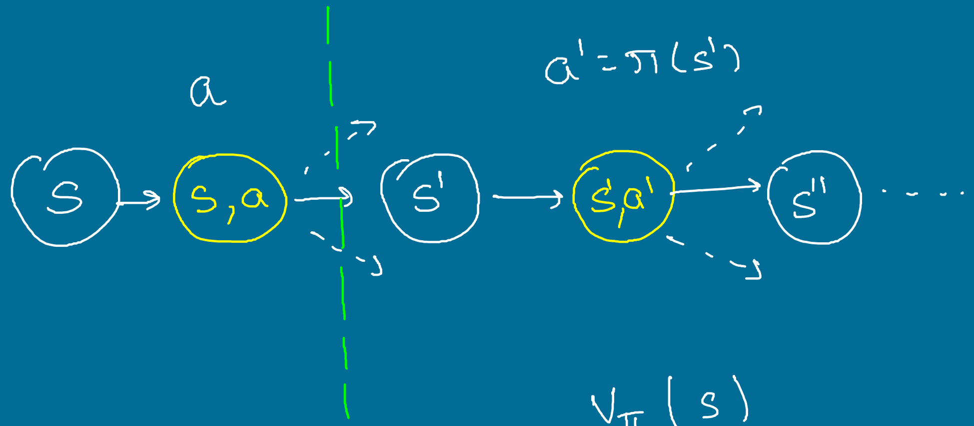
$\underbrace{Q_\pi(s, a)}$ $\quad\Big\vert\quad$ $\underbrace{Q_\pi(s, \pi(s)) = V_\pi(s)}$



$s \neq$ Terminal

$$V_\pi(s) = \mathbb{E}_\pi\left[G_t \mid S_t = s \right]$$

$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \dots$$



$V_\pi(s)$

$$Q_\pi(s, a) > Q_\pi(s, \pi(s)) \rightsquigarrow \boxed{Q_{\pi'}(s, \pi'(s)) > Q_\pi(s, \pi(s))}$$

π' is same as π everywhere except at s, $\pi'(s) = a$

$$a = \underset{act}{\operatorname{argmax}}\ Q_\pi(s, act)$$

- $X : S \to \mathbb{R}$ and $Y : S \to \mathbb{R}$, $\boxed{X \geq Y}$ if $\forall s \in S;\ X(s) \geq Y(s)$.
- $X > Y$ if $X \geq Y$ and $\exists s \in S;\ X(s) > Y(s)$.

- For policies $\pi_1, \pi_2 \in \Pi$, $\pi_1 \geq \pi_2$ if $V^{\pi_1} \geq V^{\pi_2}$

and $\pi_1 > \pi_2$ if $V^{\pi_1} > V^{\pi_2}$

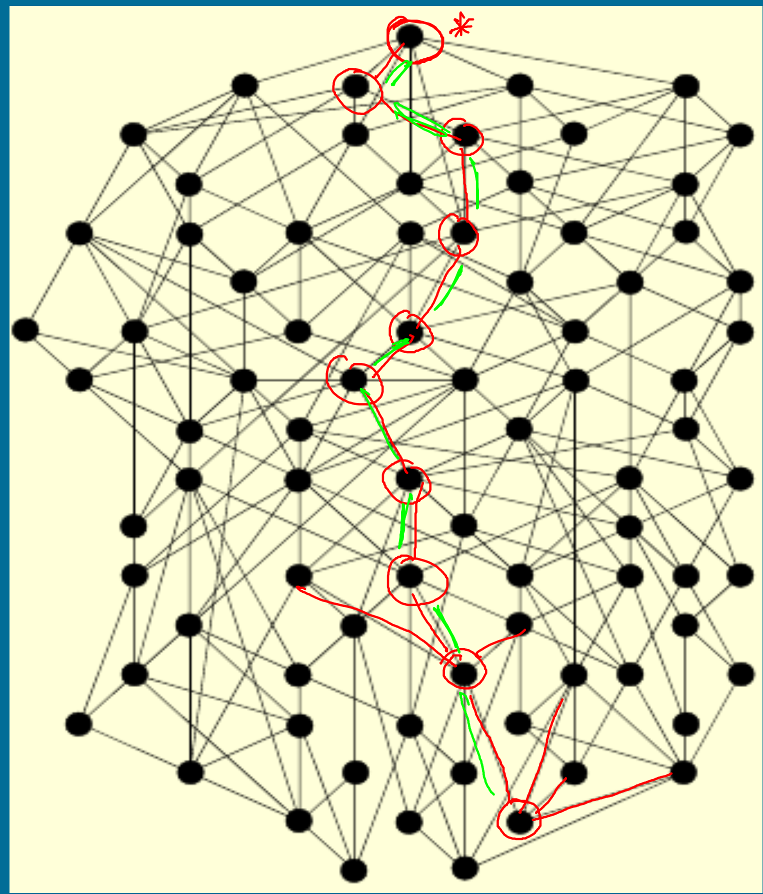
- Bellman Operator:

$$B^{\pi} : (S \to \mathbb{R}) \longrightarrow (S \to \mathbb{R})$$

$$V_\pi(s) = \sum_{s', r} p(s', r \mid s, \pi(s)) \left[r + \Gamma V_\pi(s') \right]$$

$a = \pi(s)$

$$(s) \longrightarrow (s,a) \longrightarrow (s')$$



"Policy POSET"

$$B^{\pi}(x(s)) = \sum_{s', r} p(s', r \mid s, \pi(s)) \left[r + \Gamma x(s') \right]$$

- For $\pi \in \Pi$, $X : S \to \mathbb{R}$ & $Y : S \to \mathbb{R}$

  if $X \geq Y$, then $B^{\pi}(X) \geq B^{\pi}(Y)$.

- For $\pi \in \Pi$, $X : S \to \mathbb{R}$

$$B^{\pi}(X) \longrightarrow B^{\pi}\left(B^{\pi}(X)\right) \longrightarrow B^{\pi}\left(B^{\pi}\left(B^{\pi}(X)\right)\right) \longrightarrow \cdots$$

$$(B^{\pi})^{2}(X) \qquad\qquad (B^{\pi})^{3}(X)$$

$$\boxed{\lim_{n\to\infty} (B^{\pi})^{n}(X) = V^{\pi}}$$

- $\pi, \pi' \in \Pi$ $\quad \boxed{(B^{\pi'})(V^{\pi})(s) = Q_{\pi}(s, \pi'(s))}$

$$a$$

$$B^{\pi'}(V^{\pi}(s)) = \sum_{s', r} p(s', r \mid s, \pi'(s)) \left[r + \gamma V_{\pi}(s')\right]$$

$$Q_\pi(s,a) = \sum_{s',r} p(s',r \mid s,a)\left[r + \gamma v_\pi(s')\right]$$

(I) π has no improvable states

$$\Rightarrow \forall \pi' \in \Pi, \quad V^\pi \geq B^{\pi'}(V^\pi) = Q_\pi(s, \pi'(s))$$

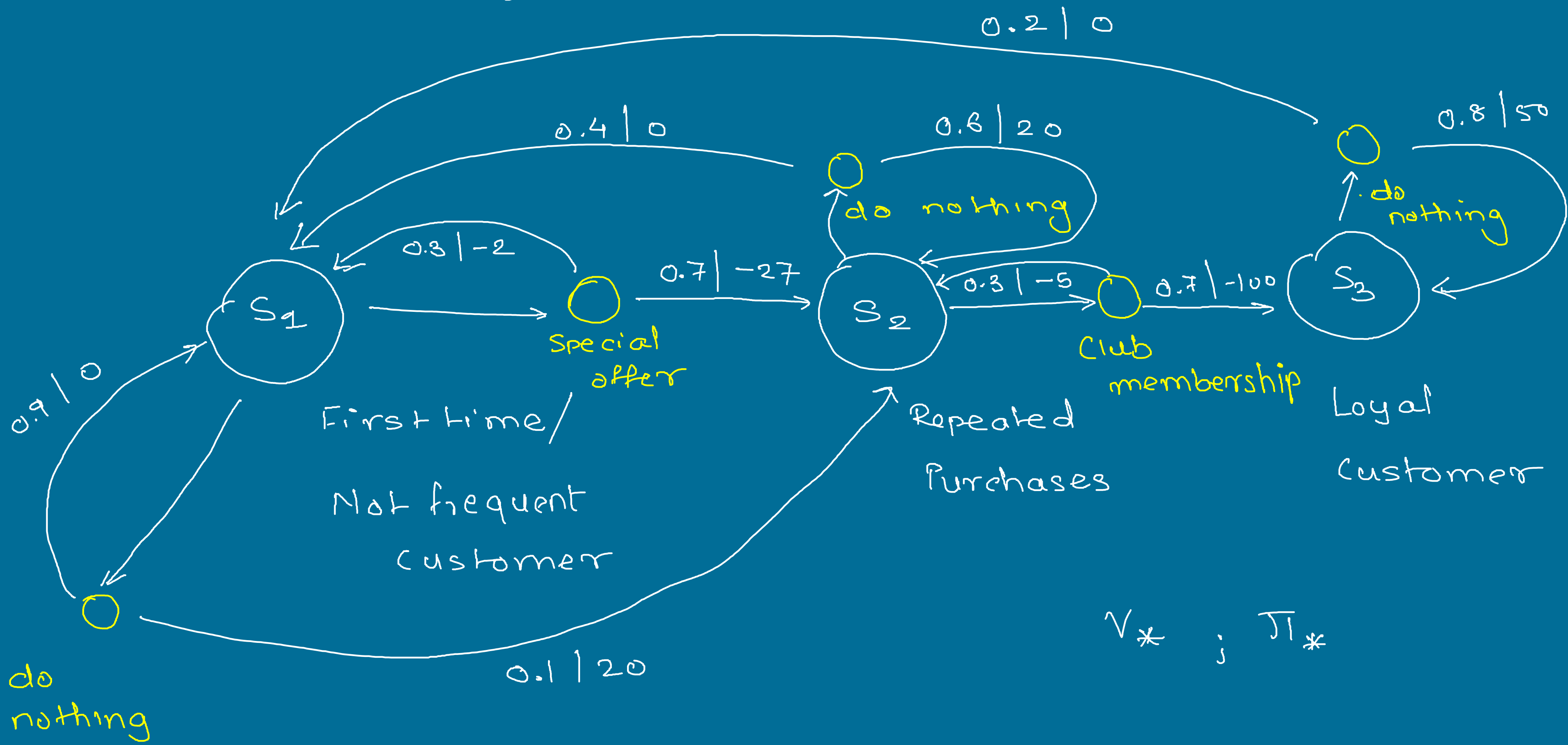
$$V^\pi \geq B^{\pi'}(V^\pi) \geq (B^{\pi'})^2 (V^\pi) \geq \dots \geq V^{\pi'}$$

(II) if π has improvable states and improvement yield π'

$$\Rightarrow \quad B^{\pi'}(V^\pi) > V^\pi$$

$$V^{\pi'} = (B^{\pi'})^n (V^\pi) > \quad \dots > (B^{\pi'})^2 (V^\pi) > B^{\pi'}(V^\pi) > V^\pi$$

Example: (Advertising Problem)



V_* ; π_*

1. Initialization
   $V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$

2. Policy Evaluation $\qquad \pi \longrightarrow V^{\pi}$
   Loop:
   $\quad \Delta \leftarrow 0$
   $\quad$ Loop for each $s \in \mathcal{S}$:
   $\qquad v \leftarrow V(s)$
   $\qquad V(s) \leftarrow \sum_{s',r} p(s', r \mid s, \pi(s))\big[r + \gamma V(s')\big]$
   $\qquad \Delta \leftarrow \max(\Delta, |v - V(s)|)$
   until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

3. Policy Improvement $\qquad \pi \rightarrow \pi'$
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
   $\quad$ *old-action* $\leftarrow \pi(s)$
   $\quad \pi(s) \leftarrow \operatorname{argmax}_a \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma V(s')\big]$
   $\quad$ If *old-action* $\neq \pi(s)$, then *policy-stable* $\leftarrow$ *false*
   If *policy-stable*, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2

Policy Iteration

π_1

$\downarrow$ Eval

V^{π_1}

$\downarrow$ Imp

π_2

$\downarrow$ Eval

V^{π_2}

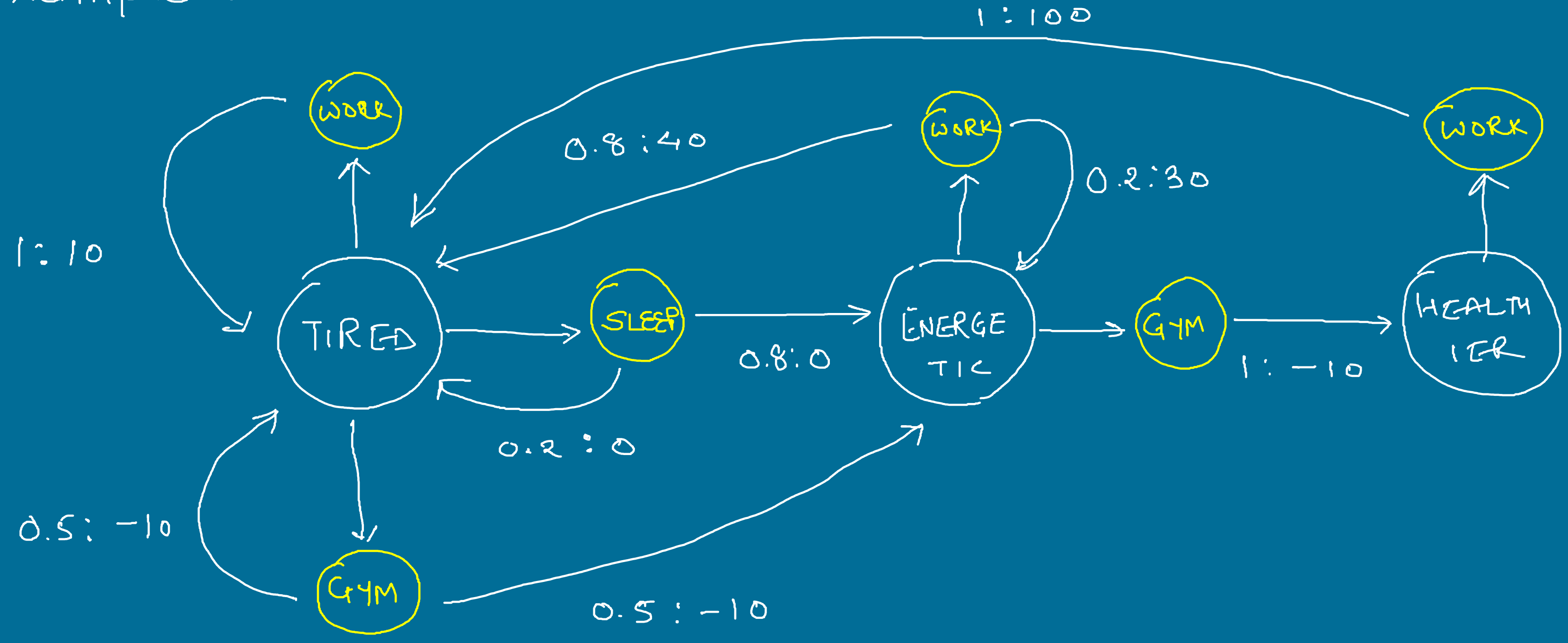
$\downarrow$ Imp

π_3

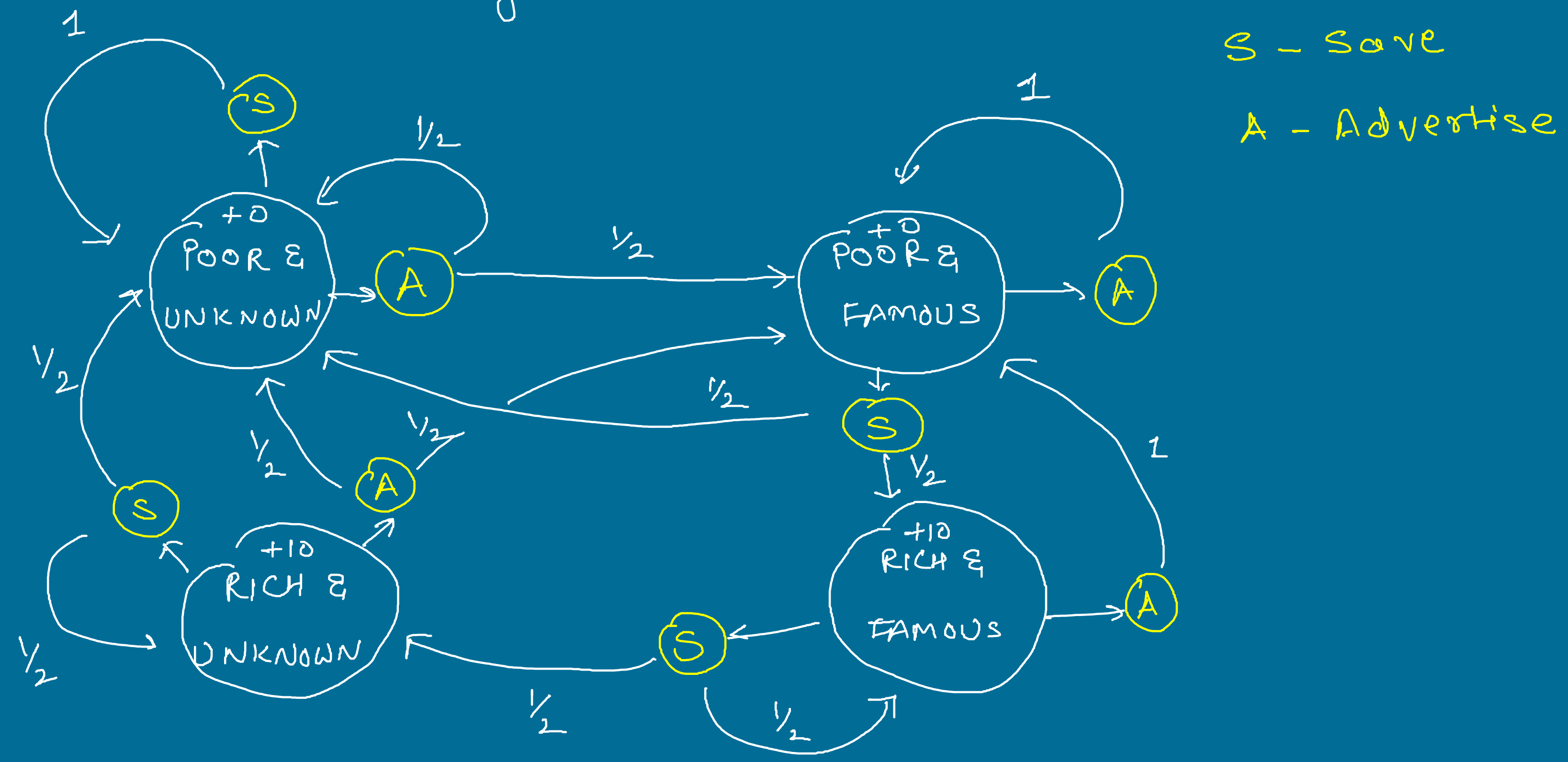
$\vdots$

(V_*, π_*)

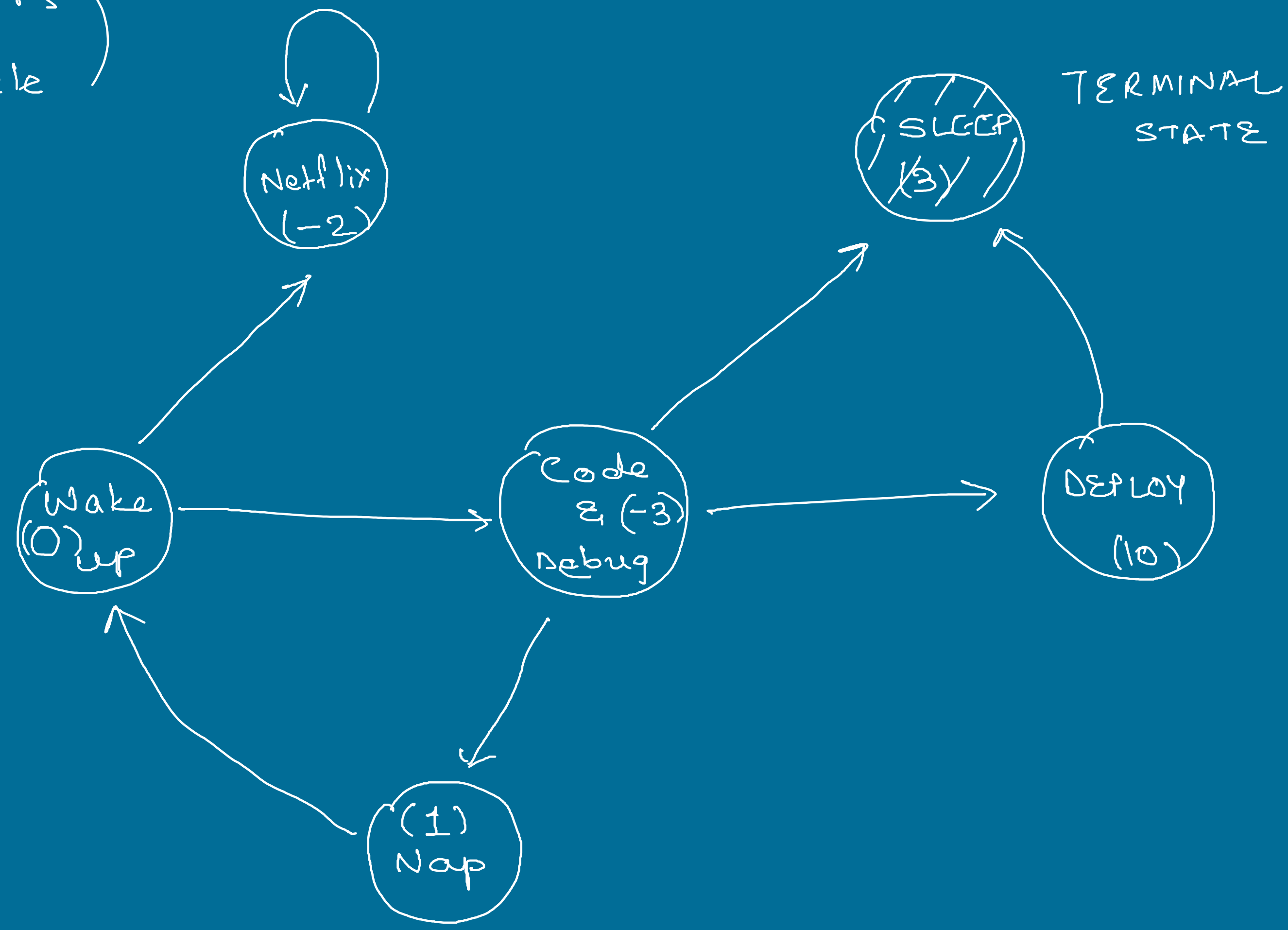
Example : (Actor's Life cycle)



Example: (Company Outreach)



Example : (Developer's Life Cycle)



Optimal Bellman Equations

$$V^*(s) = \max_a \left[\sum_{s',r} p(s',r|s,a) \left[r + \gamma V^*(s') \right] \right] \text{ (Value Iteration)}$$

$$Q^*(s,a) = \qquad \text{[Exercise]}$$

$$\pi^*(s) = \underset{a}{\text{argmax}} \; Q^*(s,a)$$

Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop:
| $\Delta \leftarrow 0$
| Loop for each $s \in \mathcal{S}$:
| $v \leftarrow V(s)$
| $V(s) \leftarrow \max_a \sum_{s',r} p(s',r|s,a)\left[r + \gamma V(s')\right]$
| $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

Output a deterministic policy, $\pi \approx \pi_*$, such that
  $\pi(s) = \text{argmax}_a \sum_{s',r} p(s',r|s,a)\left[r + \gamma V(s')\right]$

Mote Carlo Methods:

- Model is not known

Input: a policy π to be evaluated

Initialize:

$\quad V(s) \in \mathbb{R}$, arbitrarily, for all $s \in \mathcal{S}$

$\quad Returns(s) \leftarrow$ an empty list, for all $s \in \mathcal{S}$

Loop forever (for each episode):

$\quad$ Generate an episode following π: $\overset{s}{S_0}, A_0, R_1, S_1, A_1, R_2, \overset{s}{\ldots}, S_{T-1}, A_{T-1}, R_T$

$\quad G \leftarrow 0$

$\quad$ Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:

$\quad\quad G \leftarrow \gamma G + R_{t+1}$

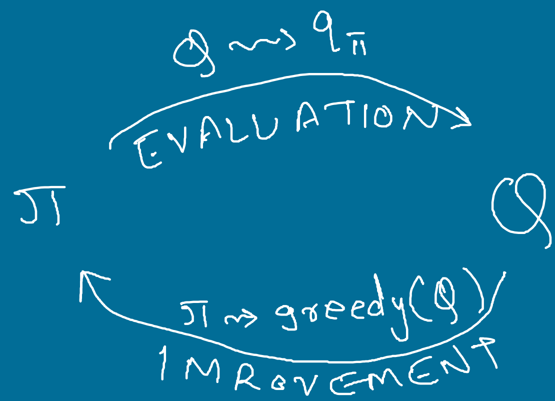
$\quad\quad$ Unless S_t appears in $S_0, S_1, \ldots, S_{t-1}$:

$\quad\quad\quad$ Append G to $Returns(S_t)$

$\quad\quad\quad V(S_t) \leftarrow \text{average}(Returns(S_t))$

Number of states $\uparrow\uparrow$

$p(s', r \mid s, a)$



First Visit MC / (Every Visit MC)

- On-policy methods

- Off-policy methods

On-policy first-visit MC control (for ε-soft policies), estimates $\pi \approx \pi_*$

Algorithm parameter: small $\varepsilon > 0$

Initialize:

$\quad \pi \leftarrow$ an arbitrary ε-soft policy

$\quad Q(s,a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

$\quad Returns(s,a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Repeat forever (for each episode):

$\quad$ Generate an episode following π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

$\quad G \leftarrow 0$

$\quad$ Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:

$\quad\quad G \leftarrow \gamma G + R_{t+1}$

$\quad\quad$ Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:

$\quad\quad\quad$ Append G to $Returns(S_t, A_t)$

$\quad\quad\quad Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$

$\quad\quad\quad A^* \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken arbitrarily)

$\quad\quad\quad$ For all $a \in \mathcal{A}(S_t)$:

$$\pi(a|S_t) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(S_t)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(S_t)| & \text{if } a \neq A^* \end{cases}$$

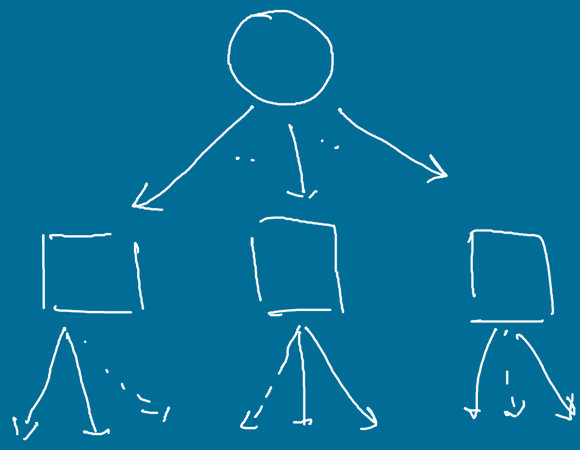
Game Theory

- pick a strategy for player that maximizes his/her utility given the strategies of other players

- Study of strategic decision making

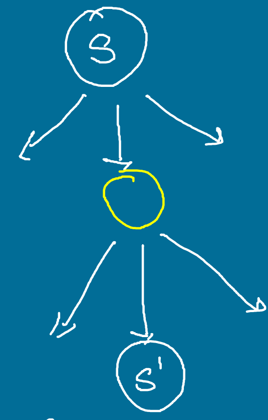
- John von Neumann, John Nash, stackleberg

Game Tree



UTILITIES

MDP



Probabilistic world outcomes

- Rewards
- Returns

Normal Form Game

		1 Actions		
		CRAM	DO-HW-TUT	PLAYGAME
	EASY	98	100	85
WORLD	HARD	97	90	65

Utility $\quad u_1(a_{1,k}, \text{world}) \qquad a_{1,k} \in A_1$

$$\frac{1}{2} \cdot 98 \qquad + \qquad \frac{1}{2} \cdot 97$$

$$\text{ExpUti} = p(\text{world} = \text{easy}) \cdot u_1(\text{cram}, \text{easy}) + p(\text{world} = \text{hard}) u_1(\text{cram}, \text{hard})$$

$u_1(S_1, \text{world}) \qquad S_1$ - strategy

$$S_1 = \left(p(a_{1,1}), p(a_{1,2}), \ldots, p(a_{1,n}) \right)$$

$$\rightarrow S_1 = \left(\tfrac{1}{2} \sim \text{cram}, \ \tfrac{1}{2} \sim \text{do-hw-tut}, \ 0 \sim \text{play game} \right)$$

$$\text{Exp Utility} = p(\text{cram}) \left[p(w = \text{easy}) u_1(\text{cram}, \text{easy}) + p(w = \text{hard}) u_1(\text{cram}, \text{hard}) \right]$$

$$+ \, p(\text{do-hw-tut}) \left[p(w = e) u_1(\text{d-h-t}, e) + p(w = h) u_1(\text{d-h-t}, h) \right]$$

$$= \frac{1}{2} \left[\frac{1}{2} \, 98 + \frac{1}{2} \, 97 \right] + \frac{1}{2} \left[\frac{1}{2} \, 100 + \frac{1}{2} \cdot 90 \right]$$

P2's ACTIONS

(u_1, u_2)

		PLAYER 2		
		ROCK	PAPER	SCISSORS
PLAYER 1	ROCK	0, 0	-1, 1	1, -1
	PAPER	1, -1	0, 0	-1, 1
	SCISSORS	-1, 1	1, -1	0, 0

JOINT UTILITIES

Goal: pick a strategy for player i that maximizes his utility given the strategies of other player.

$$\sum_{i=1}^{M} u_i(s) = 0$$

strategy profile

• Zero-Sum Game

Q. $P_2 \rightarrow$ Rock $\qquad P_1 \rightarrow S_1 = \underset{p(r)\;\;p(p)\;\;p(s)}{(0, 1, 0)}$

Q. $P_2 \rightarrow$ 50% Rock, 50% Paper $\qquad P_1 \rightarrow$ 50% scissor, 50% paper $\qquad$ paper

$E[u_1] = 1/4 \qquad E[u_1] = \frac{1}{2}$

Normal Form Game

Players : $1, \dots, M$

Pure strategies : $\mathcal{S}_i = \{ s_{i,1}, \dots, s_{i,n_i} \}$

Utility functions : $u_i(s_1, s_2, \dots, s_M)$

Non-deterministic strategy of ith player π_i – probability dist on $\mathcal{S}_i$.

- Dominant strategy

A strategy for player i, $s_{i,k}$ is "strictly dominant" if it is better than all other strategies of player i.

$$u_i(s_{i,k}, s_{-i}) > u_i(s_i, s_{-i}) .$$
$$\geq \qquad (s_i \neq s_{i,k}) \quad \text{(weakly dominant)}$$

- $u_i(s) = u_i(s_1, s_2, \ldots, s_M) = u_i(s_i, s_{-i})$

Payoff / Utility

Alphabet

$s_{-i} = (s_1, s_2, \ldots, s_{i-1}, s_{i+1}, \ldots, s_M)$

Roman

	A	B	C	D	E
i	2,10	4,7	4,6	5,2	3,8
ii	3,8	6,4	5,2	1,3	2,6
iii	5,3	3,1	2,2	4,1	3,0
iv	6,7	9,5	7,5	8,5	5,5

$\underline{\underline{(A, iv)}}$ ~ Interesting strategy profile

$\underline{u_i(s_i, s_{-i})}$

$\underset{7}{u_{alph}(A, iv)} \geq u_{Alph}(s_{Alp}, iv)$

$s_{Alp} \neq A$

$\underset{6}{u_{rom}(iv, A)} \geq u_{rom}(s_{rom}, A)$

$s_{rom} \neq iv$

- Is there always a dominant strategy?

Prisoner's Dilemma

		PRISONER 2	
		Cooperate	Defect
PRISONER 1	Cooperate	-1,-1	-6,0
	Defect	0,-6	-3,-3

$S^{NE} = (D, D)$

$u_1(s_1 = D, s_2 = D) \geq u_1(\underline{?}, s_2 = D)$

-3

$u_2(s_2 = D, s_1 = D) \geq u_2(\underline{?}, s_1 = D)$

-3

Social Welfare:

Sum of utilities of players

$SW(C, C) = -1 + (-1) = -2$

$SW(D, D) = -3 + (-3) = -6$

Nash Equillibrium: Nash equilibria are strategy profiles s, where none of the participant benefit from unilaterally changing their decision

$\{ s^{NE} = (s_i^{NE}, s_{-i}^{NE}) \}$ $\forall i$ $u_i(s_i^{NE}, s_{-i}^{NE}) \geq u_i(s_i, s_{-i}^{NE})$

$s_i \neq s_i^{NE}$

Professors Dilemma

		Student	
		Study	Games
Professor	Effort	1000,1000	0,-10
	Slack	-10,0	0,0

(Effort, Study) (strict)

(Slack, Games) (weak)

- Finding Pure Nash Equillibrium
  - Find dominating strategy, eliminate all other rows/colmn (recurse)
  - Remove a strictly dominated strategy (recurse)

Example:

2

1

	~~L~~	C	~~R~~
~~U~~	~~10,3~~	~~1,5~~	~~5,4~~
M	~~3,1~~	2,4	~~5,2~~
~~D~~	~~0,10~~	~~1,8~~	~~7,0~~

S^{NE}: (M, C)

Exercise:

	A	B	C	D	E
i	~~**2,4**~~	~~4,7~~	~~4,6~~	~~5,2~~	~~3,8~~
ii	~~3,8~~	~~6,4~~	~~5,2~~	~~1,3~~	~~2,6~~
iii	~~5,3~~	~~3,1~~	~~2,2~~	~~**9,1**~~	~~3,0~~
iv	6,7	~~9,5~~	~~**5,5**~~	~~8,5~~	~~4,5~~

Def| An action $s_i \in A_i$ weakly dominates action $b_i \in A_i$ for player i if

$$u_i(s_i, s_{-i}) \geq u_i(b_i, s_{-i}) \quad \forall\, s_{-i} \in A_{-i}$$

$[u_i(s_i, s_{-i}) < u_i(b_i, s_{-i})]$

and

$$u_i(s_i, s_{-i}) > u_i(b_i, s_{-i}) \quad \text{for some } s_{-i} \in A_{-i}.$$

It strictly dominates b_i if

$$u_i(s_i, s_{-i}) > u_i(b_i, s_{-i}) \quad \forall\, s_{-i} \in A_{-i}.$$

Def| An action $s_i \in A_i$ is weakly dominant if it weakly dominates every action in A_i. It is called strictly dominant if it strictly dominates every action in A_i.

Def| Weakly dominant strategy equillibrium of a game G,
strictly

$G := (N, A_i, u_i : \prod_i A_i \to \mathbb{R})$, in strategic form is defined as

the weakly dominant action profile, and is denoted by
strictly

$D^W(G)$.

$D^S(G)$

Algo| IESD: Iterated Elimination of strictly Dominated strategies
IEWD Weakly

	L	M	~~R~~
U	~~1,0~~	1,2	~~0,1~~
D	~~0,3~~	~~0,1~~	~~2,0~~

(U, M)

Def| A strategic form game is dominance solvable if IESD actions leads to a unique outcome.

	L	R
U	2,1	0,2
D	2,3	4,3

	L	R
U	3,1	2,0
M	4,0	1,1
D	4,4	2,4

(D, R)

(D, L)

Best Response Correspondence

$$B_i : A_{-i} \rightrightarrows A_i$$

$$B_i(s_{-i}) = \{ s_i \in A_i : u_i(s_i, s_{-i}) \geq u_i(b_i, s_{-i}) \ \forall b_i \in A_i \}$$

	L	M	R
U	1,0	1,2	0,2
D	0,3	1,1	2,0

$B_1(L) = \{U\}$ $\qquad B_1(R) = \{D\}$

$B_1(M) = \{U, D\}$ $\qquad B_2(U) = \{M, R\}$

$B_2(D) = \{L\}$

Proposition | For any 2-person game in strategic form G,

$(s_1^*, s_2^*) \in N(G)$ if and only if

$$s_1^* \in B_1(s_2^*), \ s_2^* \in B_2(s_1^*).$$

2

		H	T
	H	1, −1	−1, 1
1	T	−1, 1	1, −1

No Nash Equillibrium (pure strategies)

$$\left(\left(\tfrac{1}{2}, \tfrac{1}{2}\right), \left(\tfrac{1}{2}, \tfrac{1}{2}\right) \right)$$

$$E[u_1] = \overset{P(H)P(H)}{\underline{P(HH)}} \cdot u_1(H,H) + \overset{P(H)P(T)}{\underline{P(HT)}}\, u_1(HT) + \underline{P(TH)}\, u_1(TH) + P(\underline{TT})\, u_1(TT)$$

$$= 0$$

<u>Definition</u> A mixed strategy π_i for player i, is a probability distribution over his set of available actions A_i.

$$|A_i| = m \qquad \pi_i = (\pi_i^1, \ldots, \pi_i^m) \qquad \pi_i^k \geq 0, \quad \sum_{k=1}^{m} \pi_i^k = 1$$

- Let $\Delta(X)$ denote the set of all probability distributions on a set X.

$$\pi_i \in \Delta(A_i)$$

- $\pi = (\pi_1, \ldots, \pi_N)$ mixed strategy profile

- $(\pi_1, \ldots, \pi_N) = (\pi_i, \pi_{-i})$

BoS

$$\begin{array}{c|c|c|}
 & m & o \\
\hline
m & 2,1 & 0,0 \\
\hline
o & 0,0 & 1,2 \\
\hline
\end{array}$$

Example $\pi_1(m) = 1/3 \quad \pi_1(0) = 2/3 \qquad (1/3, 2/3)$

$$(\underbrace{(1/3, 2/3)}, (2/3, 1/3))$$ — mixed strategy profile

$p = \pi_1(m) \longrightarrow \pi_1(0) = 1 - \pi_1(m) = 1 - p$

$q = \pi_2(m) \longrightarrow \pi_2(0) = 1 - \pi_2(m) = 1 - q$

(p, q)

$$u_i(\pi) = \sum_{a \in A} P(a)\, u_i(a)$$

$$A = \prod_i A_i \qquad \pi \sim P(a) = \pi_1(a_1) \cdot \pi_2(a_2) \dots \pi_N(a_N)$$

	m	o
m	2,1	0,0
o	0,0	1,2

$$\pi = (p,q)$$

$$u_1(p,q) = 2pq + 0\,\cancel{p(1-q)} + 0\,\cancel{(1-p)q} + (1-p)(1-q)$$

$$u_1(p,q) = 1 + 3pq - p - q$$

$$u_2(p,q) = pq + 2(1-p)(1-q) = 2 - 2p - 2q + 3pq$$

$\#$ $(p,q) := \underline{(1, 1/3)}$ $\quad u_1(1, 1/3) = 2/3 \quad u_2(1, 1/3) = 1/3$

$$\mathrm{supp}(\pi_1) = \{m\} \qquad \mathrm{supp}(\pi_2) = \{m, o\}$$

Definition | The support of a mixed strategy π_i is the set of actions to which π_i assigns a positive probability.

$$\text{supp}(\pi_i) = \{ a \in \mathcal{A}_i \mid \pi_i(a) > 0 \}$$

Mixed Strategy Equillibrium:

Def | Best response correspondence of player i is the set of mixed strategies which are optimal given the other players' mixed strategies.

$$B_i(\pi_{-i}) = \underset{\pi_i \in \Delta(\mathcal{A}_i)}{\arg\max} \; u_i(\pi_i, \pi_{-i})$$

$$u_1(p,q) = 1+3pq-p-q$$

$$u_2(p,q) = pq + 2(1-p)(1-q) = 2-2p-2q+3pq$$

<u>Example</u>: $\pi_2 = (1/2, 1/2) \rightarrow q = 1/2$

$$B_1(\pi_2) = \underset{\pi_1 \in \Delta(A_1)}{\arg\max} \; 1+3p\tfrac{1}{2}-p-\tfrac{1}{2}$$

$$\pi_1 = (p, 1-p)$$

$$= \{(1,0)\} \qquad \boxed{u_1(p,q) = 1+3pq-p-q}$$

<u>Example</u>: $B_1(\pi_2 = (1,0)) = \{(1,0)\}$ $\quad B_1(q > \tfrac{1}{3}) = \{(1,0)\}$

$$B_1(q = \tfrac{1}{3}) = \{p \in [0,1]\} \quad B_1(q < \tfrac{1}{3}) = \{(0,1)\}$$

<u>Definition</u> | A mixed strategy equillibrium is a mixed strateg profile $(\pi_1^*, \ldots, \pi_n^*)$ such that, for all i

$$\pi_i^* \in \underset{\pi_i \in \Delta(\mathcal{A}_i)}{\arg\max}\ u_i(\pi_i, \pi_{-i}^*)$$

$$\boxed{\pi_i^* \in B_i(\pi_{-i}^*)}$$

<u>Example</u>: $\{((\overset{p}{1}, \overset{1-p}{0}), (\overset{q}{0}, \overset{1-q}{1})), ((0,1),(1,0)), ((2/3, 1/3), (1/3, 2/3))\}$

$$B_1((1/3, 2/3)) = (p, 1-p) \quad p \in [0,1]$$

$$B_2((2/3, 1/3)) = \text{exercise}$$

Remarks:

(1) $\pi_i \in B_i(\pi_{-i})$ iff every action in the support of π_i is itself a best response to π_{-i}.

(2) A mixed strategy profile π^* is MNE iff for each player i, each action in the support of π_i^* is a best response to π_{-i}^*.

- Each action in support of π_i^* yields the same expected payoff (utility) when played against π_{-i}^*, no other action yields a strictly higher payoff.

Proposition Every finite strategic form game has a mixed strategy equillibrium.

[Kakutani's FPT]

Definion | In a strategic form game, player i's mixed strategy π_i strictly dominates her action a_i' if

$$u_i(\pi_i, a_{-i}) > u_i(a_i', a_{-i}) \quad \forall a_{-i} \in A_{-i}$$

p-2

Example:

$$\# \quad \pi_1 = \begin{array}{c} \begin{matrix} T & M & B \end{matrix} \\ \left(0, \tfrac{1}{2}, \tfrac{1}{2}\right) \end{array}$$ dominates 'T' $\longrightarrow$

	L	R
T	1,1	1,0
M	3,0	0,3
B	0,1	4,1

P-1

$$u_1(\pi_1, L) = \frac{1}{2}\cdot 3 + \frac{1}{2}\cdot 0 = \frac{3}{2} > u_1(T, L)$$

$$u_1(\pi_1, R) = \frac{1}{2}\, 0 + \frac{1}{2}\cdot 4 = 2 > u_1(T, R)$$