

(\*) Contraction mapping Thm:-

(1) 'f' which is contraction:-

$$\underbrace{f: \mathbb{R}^n \rightarrow \mathbb{R}^m}_{x, y \in \mathbb{R}^n}$$

$$\underbrace{|f(x) - f(y)|}_{\substack{\uparrow \\ 0 \leq \alpha < 1}} < \underbrace{\alpha}_{\substack{\uparrow \\ \text{original}}} \underbrace{|x - y|}_{\substack{\uparrow \\ \text{original}}}$$

$\Downarrow$

2. There is a unique fixed point.

$$\exists x : f(x) = x.$$

Method to compute fixed points:-

1.  $x_0 \in \mathbb{R}$

2. Repeatedly apply 'f' on itself

$$x_0, \underbrace{f(x_0)}_{x_1}, \underbrace{f(f(x_0))}_{x_2}, \dots, \underbrace{f(x_2)}_{x_3} = f(f(f(x_0)))$$

$$x_0, x_1, x_2, \dots \rightarrow x^*$$

$\Downarrow$  satisfies

$$f(x^*) = x^*$$

Value Iteration 1-

i. Identify a function.

$$V(s) = \max_{a \in A} \sum_{s' \in S} \underbrace{P(s'|s,a)}_{\text{known}} \left[ \underbrace{r(s,a,s')}_{\text{known}} + \gamma \underbrace{V(s')}_{\text{known}} \right]$$

(\*) What is  $\gamma$  &  $\underline{f}$ ?

$$\gamma = \left[ \frac{V(s_1)}{V(s_2)} \right]$$

$$f(V) = \begin{bmatrix} \max_a \sum_{s' \in S} P(s'|s_1, a) (r(s_1, a, s') + \gamma V(s')) \\ \max_a \sum_{s' \in S} P(s'|s_2, a) (r(s_2, a, s') + \gamma V(s')) \\ \vdots \end{bmatrix}$$

$$V = f(V).$$

Exercise :-  $V_1, V_2 \rightarrow \|f(V_1) - f(V_2)\| \leq \alpha \|V_1 - V_2\| \rightarrow$  Exercise  
 $\downarrow$   
 $0 \leq \alpha < 1$

1. Start with any  $V_0$ .

2.  $V_{i+1}(s) \leftarrow f(V_i)(s) = \max_a \sum_{s'} P(s'|s, a) (r(s, a, s') + \gamma V_i(s'))$   
 $\rightarrow \forall s \in S$

3.  $V_0, V_1, V_2, V_3, \dots \rightarrow V^*$

$\rightarrow$  Value iteration scheme.

Chapter 5 :-

Model-free techniques.

$P(s'|s, a)$   $\rightarrow$  not known.

Not known :- transition prob., structure of reward,

given: samples / trajectories.

$\rightarrow [s_0, a_0, r_1, s_1, a_1, r_2, \dots]$

M.C prediction :-

$\rightarrow$  given a policy  $\pi$

$\rightarrow$  evaluate policy  $V^\pi$

(assume determ. rewards)

$s, a \rightarrow r(s, a)$  is fixed.

$$V^\pi(s) = E \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_t \sim \pi \right]$$

$\downarrow$   
 $(s_1, s_2, s_3, \dots)$

Q: Sample average, Law of Large no's:-

$$E[X]; \quad x_1, x_2, x_3, \dots \text{ i.i.d.}$$

$$\frac{\sum x_n}{n} \xrightarrow{\text{a.s.}} E[X].$$

$$X = \sum_{t=0}^{\infty} \gamma^t v(s_t, a_t)$$

→ Samples from this distribution

$$E[X] = v^\pi(s).$$

$$\pi, v(s)$$

1. generate an episode following  $\pi$ :

Step 1:  $s_0 = s, a_0 \sim \pi, r_0, s_1, a_1 \sim \pi, r_1, s_2, \dots, r_t$

Step 2:  $s_0 = s, a_0 \sim \pi, r_0, s_1, a_1 \sim \pi, r_1, s_2, \dots, r_t$

Step 3:  $s_0 = s, a_0 \sim \pi, r_0, s_1, a_1 \sim \pi, r_1, s_2, \dots, r_t$

Step 2:

$$R_1 = \frac{r_1 + \gamma r_2 + \gamma^2 r_3 + \dots + \gamma^{t-1} r_t}{1 + \gamma + \gamma^2 + \dots + \gamma^{t-1}}$$

$$R_2 = \frac{r_1 + \gamma r_2 + \gamma^2 r_3 + \dots + \gamma^{t-1} r_t}{1 + \gamma + \gamma^2 + \dots + \gamma^{t-1}}$$

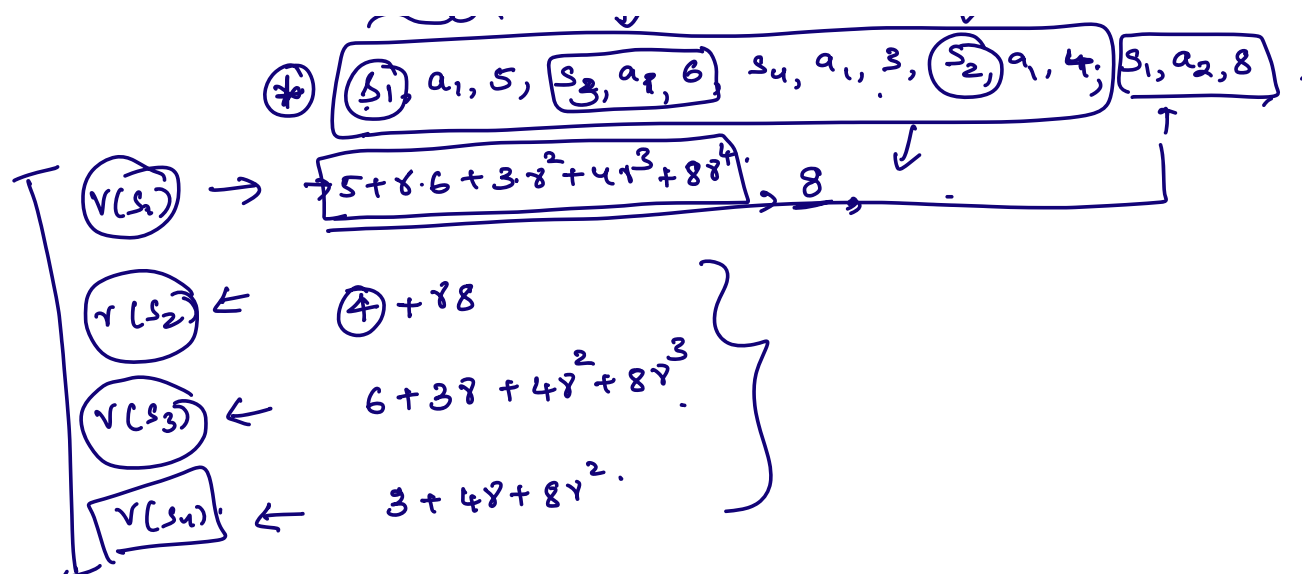
$$\vdots$$

$$R_{1000} = \frac{r_1 + \gamma r_2 + \gamma^2 r_3 + \dots + \gamma^{t-1} r_t}{1 + \gamma + \gamma^2 + \dots + \gamma^{t-1}}$$

Step 3:

$$v^\pi(s) \approx \frac{\sum R_i}{i}$$

First visit:-



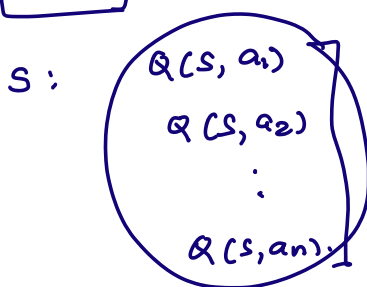
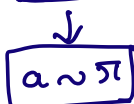
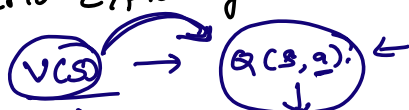
Every-visit:-

(\*) Prediction:  $\pi : \underline{V}^\pi$

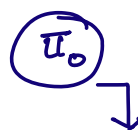
↓

Control :- computing optimal policy  $\underline{\pi}^*$  using samples

Monte Carlo Exploring starts:-

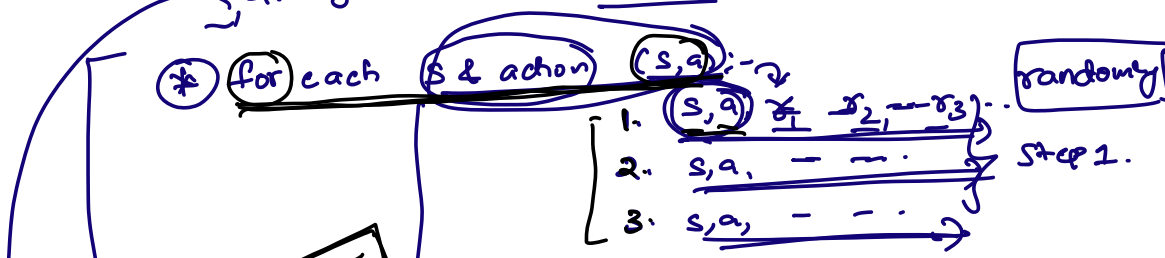


$$Q^*(s) = \arg \max_i Q(s, a_i)$$



(\*) 1. choose  $s_0, a_0$  randomly.

2. generate an episode from  $\underline{s_0, a_0}$  using  $\pi_0 \sim$  random policy



Evaluation

step 2:  $\begin{cases} R_1 \\ R_2 \\ \vdots \\ R_{1000} \end{cases}$

step 3:-  $Q(s, a) \leftarrow \frac{R_1 + R_2 + \dots + R_n}{n}$

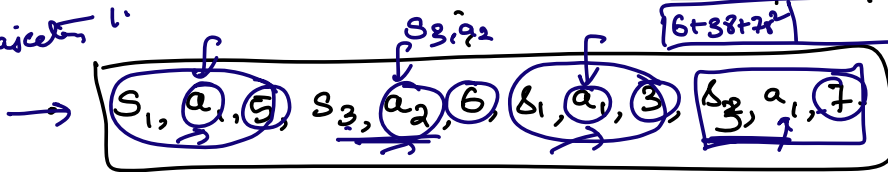
\*  $\pi_{\pi}(s) = \arg \max_a Q(s, a)$   $\leftarrow$  improvement

Q:- Analogous to Policy Iteration

I. with same trajectory.

$Q(s_1, a_1)$	$\leftarrow$	<table> <tr> <td><math>R_1</math></td> <td><math>R_2</math></td> <td><math>4 + 8\gamma + 6\gamma^2</math></td> <td><math>6..</math></td> </tr> </table>	$R_1$	$R_2$	$4 + 8\gamma + 6\gamma^2$	$6..$
$R_1$	$R_2$	$4 + 8\gamma + 6\gamma^2$	$6..$			
$s_1, a_2$	$\leftarrow$	<table> <tr> <td>0</td> <td></td> <td></td> <td>...</td> </tr> </table>	0			...
0			...			
$s_3, a_1$	$\leftarrow$	<table> <tr> <td>7</td> <td></td> <td></td> <td></td> </tr> </table>	7			
7						
$s_3, a_2$	$\leftarrow$	<table> <tr> <td><math>6 + 3\gamma + 7\gamma^2</math></td> <td></td> <td></td> <td></td> </tr> </table>	$6 + 3\gamma + 7\gamma^2$			
$6 + 3\gamma + 7\gamma^2$						

trajectory 1



$$\sum_{t=0}^{\infty} \gamma^t r_t$$

$$R_1 = 5 + 6\gamma + 3\gamma^2 + 7\gamma^3; \quad R_2 = 3 + 7\gamma.$$

$$s_3, a_2 :- 6 + 3\gamma + 7\gamma^2$$

$$s_3, a_1 :- 7$$

trajectory 2:-  $(s_1, a_1) 4, (s_2, a_2) 8, (s_1, a_1) 6.$

$$\frac{4 + 8\gamma + 6\gamma^2}{6}$$

$$s_1(a_1) \leftarrow \text{low rewards.}$$

online learning

Estimate of  $\tilde{Q}_t(s, a)$   $\forall s, a$ .

on-policy first-visit MC control.

$$\pi(s, a) \leftarrow \begin{cases} \arg \max_a \tilde{Q}_t(s, a) \text{ w.p. } \leftarrow \\ \text{random action } 1 - \epsilon \end{cases}$$