(*). Contraction mapping Thm :-

$f : \underbrace{\mathbb{R}^n \to \mathbb{R}^m}$

1. 'f' which is $\underbrace{\text{contraction}}$:-

$x, y \in \mathbb{R}$.

$$\|f(x) - f(y)\| < \alpha \, \|x - y\|$$

$\boxed{0 \le \alpha < 1}$

↳ original.

$\Downarrow$

2. There is $\boxed{\text{a unique}}$ fixed point.

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

Method to compute fixed points :-

1. $\boxed{x_0 \in \mathbb{R}}$
2. Repeatedly apply 'f' on itself

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

$$x_0, x_1, x_2, \ldots \longrightarrow x^*$$

$\Downarrow$ satisfies

$$\boxed{f(x^*) = x^*}$$

Value Iteration :-

1. Identify a function.

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

(*) What is x & f ?

$$V = \begin{bmatrix} V(s_1) \\ V(s_2) \\ \vdots \\ V(s_n) \end{bmatrix}$$

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

$$\boxed{V = f(V).}$$

Exercise :- $V_1, V_2; \longrightarrow \|f(V_1) - f(V_2)\| \le \alpha \|V_1 - V_2\|. \rightarrow \boxed{\text{Exercise}}$

$\alpha \rightarrow \boxed{0 \le \alpha < 1}.$

1. Start with any V_0.
2. $V_{i+1}(s) \leftarrow f(V_i)(s) = \max_a \sum_{s'} P(s'|s,a)(r(s,a,s') + \gamma V_i(s'))$.
   $\hookrightarrow \boxed{\forall s \in S}$
3. $V_0, V_1, V_2, V_3, \cdots \rightarrow V^*$

$\hookrightarrow$ value iteration scheme.

chapter 5 :-

Model - free techniques.

$P(s'|s,a) \rightarrow \boxed{\text{not known.}}$

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

given : samples / trajectories.

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

M.C prediction :-

(assume determ. rewards)

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

$\hookrightarrow$ given a policy π

$\hookrightarrow$ Evaluate policy $\boxed{V^\pi}$

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

$\boxed{(s_1, s_2, s_3, \ldots)}$

$s_0, a_0 \sim$

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

$E[X]$; $x_1, x_2, x_3, \ldots$ i.i.d.

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

$H \to 0$

$T \to 1.$

$0, 0, 1, 0, 1, 0, \ldots$

$X = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t).$ $\leftarrow$ ~~t~~ T

$\nearrow$ Samples from this distribution

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

$v^{\pi}(s).$

$v^{\pi}(s_3)$

1. generate an episode following π:

s_2 s_3

Step 1

1. ~~$s_0^1 = s, a_0^1 \sim \pi, r_1^1$~~, $s_1, a_1^1 \sim \pi, r_2^1, s_2 \ldots r_t$

2. $s_0^2 = s_3, a_0 \sim \pi, r_1, s_2, a_1, \ldots$ r_t^2

$\vdots$ $r_t^{1000}.$

1000. $s_0^{1000} = s, \ldots$

Step 2.

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

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

$\vdots$

$R_{1000} = r_1^{1000} + \gamma r_2^{1000} + \ldots + \gamma^{t-1} r_t^{1000}$

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

Step 3

First visit:-

(*) $s_1, a_1, 5, s_3, a_3, 6, s_4, a_1, 3, s_2, a_1, 4, s_1, a_2, 8$.

$V(s_1) \rightarrow 5 + \gamma \cdot 6 + 3 \cdot \gamma^2 + 4\gamma^3 + 8\gamma^4 \rightarrow 8$

$V(s_2) \leftarrow 4 + \gamma 8$

$V(s_3) \leftarrow 6 + 3\gamma + 4\gamma^2 + 8\gamma^3$.

$V(s_4) \leftarrow 3 + 4\gamma + 8\gamma^2$.

Every-visit:-

(*). Prediction : π : V^{π}.

↓

Control :- computing optimal policy π^* using samples

Monte Carlo Exploring starts:-

$V(s) \rightarrow Q(s,a) \leftarrow$

$a \sim \pi$

s : $Q(s, a_1)$, $Q(s, a_2)$, ..., $Q(s, a_n)$

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

π_0

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

2. generate an episode from s_0, a_0 using π_0 ~random policy.

(*) for each s & action (s,a):-

1. $s, a, \cancel{r_1}, r_2, \ldots r_3 \ldots$ randomly
2. $s, a, - - .$ Step 1.
3. $s, a, - - .$

Evaluation

step2: $\begin{cases} R_1 \\ R_2 \\ \vdots \\ R_{1000} \end{cases}$

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

$*$ $\boxed{\pi_{i+1}(s) = \arg\max_a Q(s,a)}$. $\leftarrow$ improvement.

Q1:- Analogous to Policy Iteration

I. with same trajectory.

$Q(s,a) \leftarrow$ | R_1 | R_2 | $4+8\gamma+6\gamma^2$ | 6.. |

$s_1 a_2 \leftarrow$ | 0 | | | ... |

$s_3 a_1 \leftarrow$ | 7 | | |

| $6+3\gamma+7\gamma^2$ | |

trajectory 1:

s_3, a_2

$\rightarrow$ $s_1, a, 5, s_3, a_2, 6, s_1, a_1, 3, s_3, a_1, 7$

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

$R_1 = \boxed{5 + 6\gamma + 3\gamma^2 + 7\gamma^4}$; $R_2 = 3 + 7\gamma$.

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

s_3, a_1 :- 7

trajectory 2:- $s_1, a_1, 4, s_2, a_3, 8, s_1, a_1, 6.$

$4 + 8\gamma + 6\gamma^2$

6

$s, a_1 \leftarrow$ $\boxed{\text{low rewards.}}$

online Learning

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

on-policy first-visit MC control.

$\pi(s, a) \leftarrow$ $\boxed{\arg\max_a \tilde{Q}_t(s,a) \text{ w.p.}}$ $\leftarrow$

$\boxed{\text{random action } 1-\epsilon \text{ p.}}$