## Chapter 4- Dynamic Programming\_

$$N \times (s_1 a) = E[R_{t+1} + \gamma \max_{a'} q_* (s_{t+1}) a') | s_t = s, A_t = a]$$

## [4.1] Policy Evaluation:

Computing the state value function Vx for any arbitrary function T is called policy evaluation

$$V_{\pi}(s) = E_{\pi}[G_{t} + S_{t} = s]$$

$$= E_{\pi}[R_{t+1} + Y G_{t+1} | S_{t} = s]$$

$$= E_{\pi}[R_{t+1} + Y G_{t+1} | S_{t} = s]$$

$$= E_{\pi}[R_{t+1} + Y G_{t+1} | S_{t} = s]$$

$$= \sum_{\alpha} \pi(\alpha|s) \sum_{s', x} p(s', x | s, \alpha) [x + Y V_{\pi}(s')]$$

$$= \sum_{\alpha} \pi(\alpha|s) \sum_{s', x} p(s', x | s, \alpha) [x + Y V_{\pi}(s')]$$

0

5

2

consider a siquence of approximate value functions Vo, Vi, --- each mapping St to R. The v. is chosen arbitrarily and each successive approximation is obtained by using the Billman equation for va as an update rule:

$$V_{k+1}(s) = E_{\pi}[R_{k+1} + \gamma V_{k}(s_{k+1}), |s_{k}=s]$$

$$= \sum_{\alpha} \pi(\alpha|s) \sum_{s', k'} \rho(s', k'|s, \alpha)[x + \gamma V_{k}(s')] \quad \forall s \in S$$

clearly  $v_k = v_\pi$  is a fixed point for this update rule because the Bellman equation for un assures us equality.

```
The sequence \{V_k\} can be shown in general to converge to V_{\pi} as k \to \infty.
 Herative Policy Evaluation: ((2) 77 2) 710 = (2) 71
  Enpet TT, the policy to be evaluated
 Algo parameter -> 0 [to determine accuracy]

Initialize V(s) 4 s e s+, arbitrarily except that V(terminal)=0
  = En [ ( Stat + ) En ( Stat 2 + ) ( Stat 2) | Stat ) At H = 1 900]
         Loop for each ses:
            ν τ ν (s)
ν(s) τ Σπ (a|s) Σρ (s', χ |s|a) [η+ γν (d')]
                 Δ - Max (Δ, (ν-ν(2)))
          man Pro
[4.2] Policy Improvements:
        Qπ (s,a) = E[R++++ Y Vπ (S++1) | S+= S, A+=a]
             = \sum_{s|n} p(s',n)[s,a)[n+3v_{\pi}(s')]
 ut π & π' be any pair of deteriminatic policies s. to US€ S
     Q_{\pi}(s)^{\pi'(s)} \geqslant V_{\pi}(s)
  Then IT' must be as good as, or better than, IT.
  mat is Vπ'(s) ≥ Vπ(s) Yses
        \pi'(s) = \pi(s) \ \forall s \in S \ \text{except} \ s' \text{ where}
(s', \pi'(s')) > V_{\pi}(s')
          then It is indeed better thom IT.
```

The

1

1

##

: Joans 9

 $V_{\pi}(s) \leq Q_{\pi}(s, \pi'(s))$   $= E[R_{th} + Y V_{\pi}(s_{t+1}) | s_{t} = s, A_{t} = \pi'(s)]$   $= E_{\pi'}[R_{th} + Y V_{\pi}(s_{t+1}) | s_{t} = s]$   $\leq E_{\pi'}[R_{th} + Y Q_{\pi}(s_{t+1}) | s_{t} = s]$   $\leq E_{\pi'}[R_{th} + Y Q_{\pi}(s_{t+1}) | s_{t} = s]$ 

 $= E_{\Pi^{1}} \left[ R_{\pm +1} + \gamma E_{\Pi^{1}} \left[ R_{\pm +2} + \gamma V_{\Pi^{1}} (S_{\pm +2}) \right] S_{\pm +1} + \lambda R_{\pm +2} + \gamma^{2} V_{\Pi^{1}} (S_{\pm +2}) \right] S_{\pm} = S$   $= E_{\Pi^{1}} \left[ R_{\pm +1} + \gamma R_{\pm +2} + \gamma^{2} V_{\Pi^{1}} (S_{\pm +2}) \right] S_{\pm} = S$   $\leq E_{\Pi^{1}} \left[ R_{\pm +1} + \gamma R_{\pm +2} + \gamma^{2} R_{\pm +3} + \gamma V_{\Pi^{1}} (S_{\pm +3}) \right] S_{\pm} = S$ 

the enqueries fire and be charmed in denoted

00 - 4 40 7V OF

Î

5

3

5

5

-

e

CI

RI

Q1

Q1

Q!

U

U

U

 $\leq E_{\Pi^1}[R_{\pm H} + \gamma R_{\pm + 2} + \gamma^2 R_{\pm + 2} - - | S_{\pm} = S]$ =  $V_{\Pi^1}(S)$ 

 $\pi'(s) \doteq \operatorname{argmax}_{A}(s,a)$   $= \operatorname{argmax}_{A} E[R_{\pm+1} + \gamma v_{\pi}(s_{\pm+1}) | s_{\pm} = s, A_{\pm} = a]$   $= \operatorname{argmax}_{A} \sum_{s,n} p(s',n)[x+\gamma v_{\pi}(s')]$ 

The greedy policy takes the actions that looks best in the short - town, so we know that it is as good as or better tran, the original policy. The process of making a new policy that improves the original policy, by making it greedy wint the value function is called policy improvement.

(2) TV < (('2)'77 (2)) > VIT (2)

them IT is indeed bottom terour n

Suppose,  $\pi'$  is as good as, but not better than  $\pi$ , Then  $V_{\pi} = V_{\pi}'$  and Vm(s) = max E[R++++ YVm(S++1) | St=S,At=a] = max \[ p(s',n|s,a)[x+ \( v\_{\pi}(s) ] - wash But this is same as Bellman Optimality have VII, must be V\* the go to 2. (4.3) Policy Iteration: Once a policy,  $\pi$ , has been improved using  $V_{\pi}$  to yield a better policy  $\pi'$ , we can then compute  $V_{\pi'}$  and improve it again to yield an even better  $\pi''$ .  $\pi_{0} \xrightarrow{E} \vee_{\pi_{0}} \xrightarrow{T} \pi_{1} \xrightarrow{E} \vee_{\pi_{1}} \xrightarrow{T} \pi_{2} \xrightarrow{E} \cdots \xrightarrow{T} \pi_{2} \xrightarrow{E} \vee_{*}$ Code pets al multiples pour je east louge tradrigum en O 1. Initialisation: V(s) ∈ R and π(s) ∈ A(s) authorabily v s∈ S. 2. Policy Evaluation: Loop for each ses:  $v \leftarrow V(s)$  $V(s) \leftarrow \sum_{s', x} p(s', x | s) \left[ \Re \pi(s) \right] \left[ \Re \pi(s') \right]$   $\Delta \leftarrow \max \left( \Delta, | V - V(s) | \right)$ 0 > 1 litm

4

3. Policy Improvement: policy-stable + bus

For each ses:

(2)  $\pi \rightarrow \text{naite} - \text{ab}$ If old-action ≠ π(s), then policy-stable = false of policy-stable, then stop and V & V\* and T & T\*; lle go to 2.

dud, so boop so as The seeggest os, but the

9

6

5

5

6

6

6

0

2

<u>e</u>

(Q"

**O**".

101

## [4.4] Value Horation:

Drawback of policy iteration is that each of its iterations impolies policy evaluation, which may stall be a protracted iterative computation requiring multiple sweeps through the state set if policy iteration is done iteratively, its convergence exactly to  $V_{\pi}$  occurs only in limit.

One important special case of policy evaluation is stopped after just one surep -> its called value theration

$$V_{k+1}(s) = \max_{\alpha} E[R_{k+1} + \gamma v_k(s_{k+1}) | s_k = s, A_k = \alpha]$$

$$= \max_{\alpha} \sum_{s',k} p(s',k|s_{l}\alpha) [k+\gamma v_k(s')]$$

$$= \sum_{\alpha} \sum_{s',k} p(s',k|s_{l}\alpha) [k+\gamma v_k(s')]$$

[((3) N. 8 + 16] ((3)) (3) (3) (3) (3) (3) (3) (3)

(1(2) V-VI (D) XOM -> A

0 > A liture

Pseudocode: What is all in I strait - I state to group make we heard Leveldary is great a great Loop: adam't outstand più a ver destre al see she and se Loop for each  $s \in S$ :  $v \leftarrow V(s)$ ,  $v \leftarrow$ Δ = max (Δ, | V - V (s\*) |) [E 2] Mayto (101/2 Productions) 0>4 lutum Outputs a deterministic policy,  $\pi \approx \pi_*$   $18.\pm$   $\pi(s) = \text{argmax}_{a \leq 1} p(s, \pi | s, a) [\pi + \chi V(s')]$ . 2 32 to publication 3 3 5 5 5 RECORDER SO A - WARRY OF BUSTON A SEE S cabing was safe I wanted day I A A 1 A 1 A 0 A TO & ARMEDIAL Schoolings in schools all late the wind and bringer