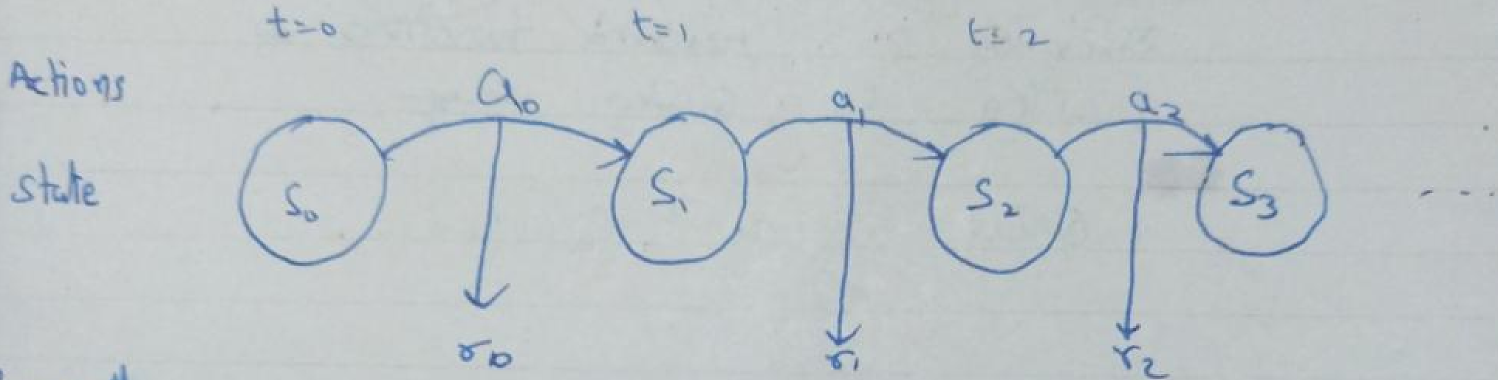
Bellman equation

$$V_\pi(s) = r(s,a) + \gamma V_\pi(s')$$

Bellman equation writes value of a decision Problem for a given state in terms of immediate reward from the action taken.

At $t=0$
solution / Proof



Reward

Present value Reward	r_0	γr_1	$\gamma^2 r_2$

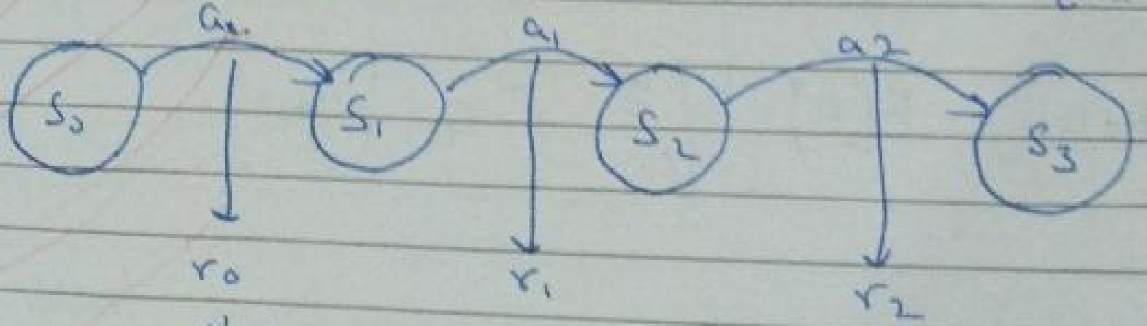
$$R(s) = r_0 + \gamma r_1 + \gamma^2 r_2$$

Date ....

So after performing action a_0 our agent moves to new state S' here S_1

$t=0$ $\quad$ $t=1$ $\quad$ $t=2$ $\quad$ $t=3$

$t'=0$ $\quad$ $t'=1$ $\quad$ $t'=2$



$$R'(S) = r_1 + \gamma r_2 + \gamma^2 r_3 \ldots$$

now our new Value of cumulative reward will be.

$$R(S) = r_0 + \gamma(r_1 + \gamma r_2 + \gamma^2 r_3 \cdots$$

$$R(S) = r_0 + \gamma R(S_1)$$

$$\therefore E_\pi[R(S_t) | a_t] = E_\pi[v_\pi(S_t) | a_t]$$

$$V(S) = r_0 + \gamma V(S')$$

Similarly $\because$ reward function is written in similar way

$$Q(S) = r_0 + Q(S')$$

Spiral

Expected Value.

Expectation values of R_{t+1} given that we know that the current state is s. The formula for this is

$$E_\pi [R_{t+1} / S_t = s] = \sum_{r \in R} r \, p(r/s)$$

the probability of the appearance of reward r is conditioned on the state s.

Relation with bellman equation is that for stochastic decisions

Bellman equation for $V_\pi(s) = E_\pi [R_t | S_t = s]$, $V_\pi(s) = r_0 + \gamma V_\pi(s')$

OPTIMALITY

It is best path/policy the agent follows to accumulate maximum rewards.

$$V_*(s) = \max_\pi V_\pi(s)$$

$$q_*(s,a) = \max_\pi q_\pi(s,a)$$

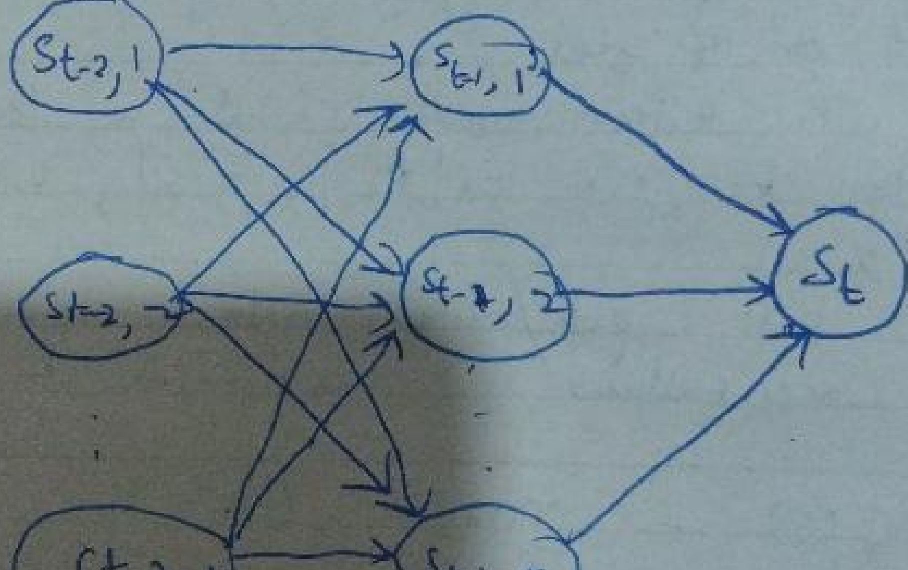
Date ....

Bellman equation for optional policy

$$v_\pi(s_{t-1}) = r(s_{t-1}, a) + \gamma v_\pi(s_t)$$

If our actions are according to optimal policy then $v^*(s_{t-1}) = \max_a \{ r(s_{t-1}, a) + \gamma v^*(s_t) \}$

$v^*(s)$ = optimal state value for stable state s corresponding to optimal policy.



$$v^*(s_{t-1}) = \max_a \{ r(s_{t-1}, a) + \gamma v^*(s_t) \}$$

Date ....

Generalized policy iteration

Policy iteration

$\pi_0 \xrightarrow{E} V_{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} V_{\pi_1} \xrightarrow{I} \pi_2 \cdots \quad \pi_*$

E : Evaluation of V_π using policy π
I : Improvement of policy π using present value function.

1. Initialize π_0 randomly
2. Iterate following steps until π_i converges to π^*

Policy Evaluation.

action given by random policy

$$V(s) = \sum_{s' \in S} P(s, \pi(s), s')[r(s, \pi(s), s') + \gamma V^{\pi}(s')]$$

for all states s

Policy Improvement

$$\pi_{i+1}(s) = \arg\max_a \sum_{s' \in S} P(s, a, s')[r(s, a, s') + \gamma V^{\pi}(s')]$$

Date ....

Value iteration

Bellman equation gives us a recursive definition of the optimal value:

$$V^*(s) = \max_{a \in A} \sum P(s'|s, a)[R(s, a, s') + \gamma V^*(s')$$

Initializ ~~e~~ $V_0(s) = 0$ for all states s

Iterate the Bellman update till convergence

$$V_{i+1}(s) \leftarrow \max_{a \in A} \sum_{s' \in S} P(s'|s, a)[R(s,a,s') + \gamma V_i(s')$$

We run this till values of all states starts converging then ~~we~~ when we find the optimal value function we find the optimal policy using the optimal value function

Date ....

Convergence of PI

Let π_{k+1} be greedy policy w.r.t V^{π_k}, then

$$V^{\pi_k}(s) \leq V^{\pi_{k+1}}(s), \quad \forall s$$

This is guaranteed by the Policy Improvement theorem which states if π' is greedy w.r.t to V^{π}, then $V^{\pi'} \geq V^{\pi}$

and since the no. of Deterministic policies is finite : $|A|^{|S|}$

Thus, the algorithm can make only finitely many improvements before it must converge to some policy.

Value iteration Convergence.

Proof using Contraction Mapping and Banach Fixed point theorem.

let u, v two value functions then

$$\|Tu - Tv\|_\infty \leq \gamma \|u - v\|$$

So T is a γ-contraction

Time complexity (Policy Evaluation : $O(|S|^3)$)
Policy improvement : $O(|S||A|)$

memory : $O(|S|)$

S... zl

From banach fixed point theorem

- v^* is a fixed point of T
- iteration $v_{k+1} = Tv_k$

Converges linearly to v^*, at rate γ

$$\|v_k - v^*\|_\infty \leq \gamma^k \|v_0 - v^*\|_\infty$$

Once $v_k = v^*$

then greedy policy.

$$\pi_k(s) = \arg\max_a \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma v_k(s')]$$

- Per iteration time : $O(|S||A|)$
  iteration needed : $O\left(\frac{\log 1/\epsilon}{1-\gamma}\right)$

best for large state spaces approximation.

Spiral