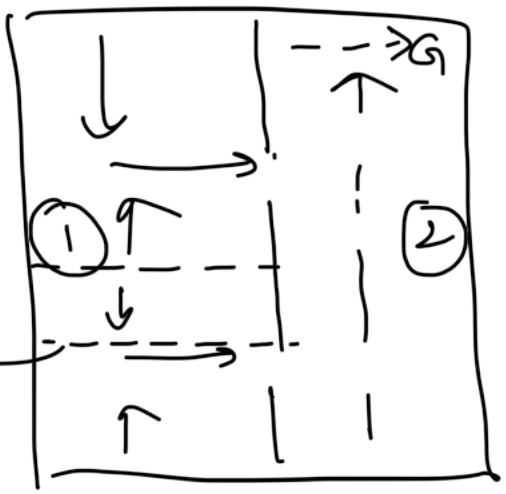
Week - 11 RL

L1 Hierarchial RL

- provides temporal abstrac^n
- transfer / reusability of the func^{ns}
- more powerful / "meaningful" state abstrac^{ns}

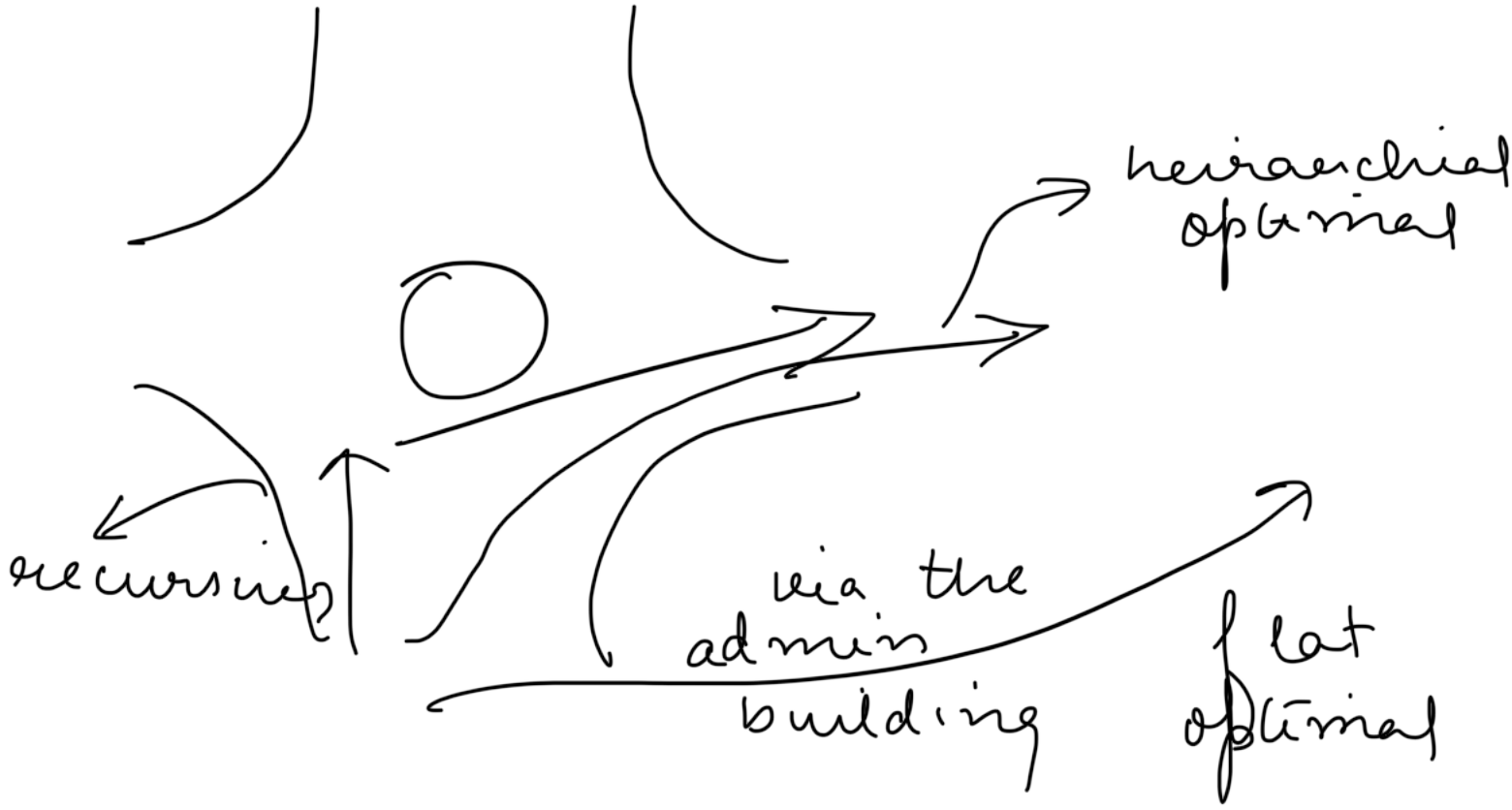
L2 Types of Optimality



then the policy will change

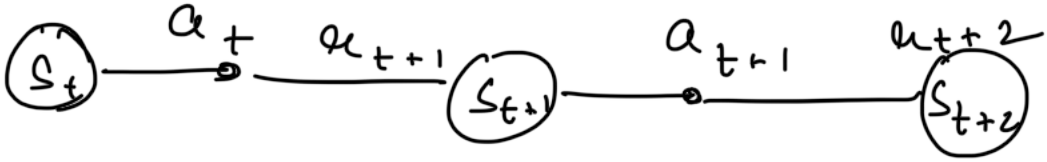
- there are many types of optimality -
  1. hierarchial optimal
  2. recursively optimal (at least each subproblem is optimal)
  3. flat optimal

Eg. going to hostel zone via OC.



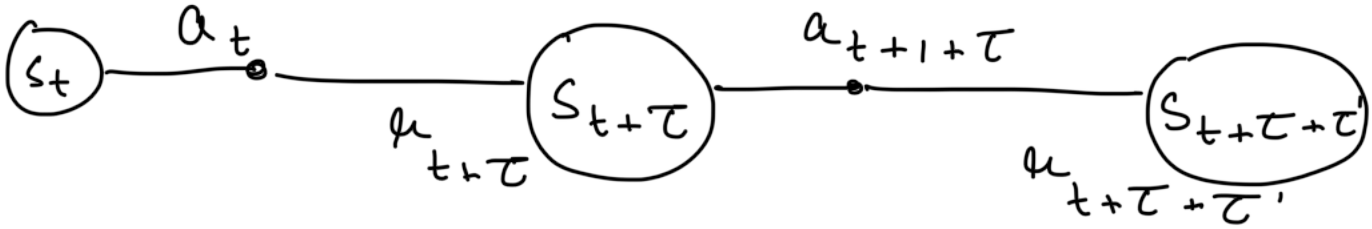
L3 Semi-Markov Decision Process

- SMDP (actions with durations)



(earlier)

now,



here, τ is referred to as the holding time

$\langle S, A, P, R \rangle$ $\quad p(s', \tau \mid s, a)$

$$R = E[r \mid s, a, s', \tau]$$

- SMDP Q Learning

$$Q(s,a) = Q(s,a) + \alpha[r_{t+\tau} + \gamma^{t+\tau-1} \max_{a'} Q(s', a') - Q(s,a)]$$

$$R_{t+\tau} = R_{t+1} + \gamma R_{t+1} + \ldots + \gamma^{\tau-1} R_{t+\tau}$$

L4 Options

- It helps to choose macro actions. Kind of sub-problem to the main problem.
- We will be using the encapsulated policy concept. In this, basically we being the policy for each sub-problem under one umbrella.
- Initiation Set, $I_0 \subseteq S$
- Policy, Π_0, $S \rightarrow A$
- Termination β, $S \rightarrow [0, 1]$

$\therefore$, define the options as

$$O = \langle I_0, \Pi_0, \beta \rangle$$

- Markov options – π_o depends only on the current state.
- Semi-markov options – π_o depends on the history, exactly, since when the option had started.

5 Learning With Options

- SMDP Q Learning

$Q(s, a) \Rightarrow$ that we defined above

- Now, we come to Intra-Option Q learning

$Q(s, o)$

$s_1 \dots s_n$

$\pi_o \rightarrow a_1 \dots a_n$

$$Q(s, a_1) = Q(s_1, a_1) + \alpha \left[R_1 + \gamma \max_a Q(s_1, a) - Q(s_1, a_1) \right]$$

similarly for,

$Q(s_1, a_2)$

$\vdots$

$$Q(s_1, o) = Q(s_1, a) + \alpha \left[r_1 + \gamma Q(s_2, o) - Q(s_1, o) \right]$$

↳ if not ending at s_2.

if it terminates at s_2, then

$\gamma Q(s_2, 0) \Rightarrow \gamma \max_a Q(s_2, a)$