Lec-6, IT567, 24-25

If at $t = T$, you land at the terminal state, ideally, what should be G_T? $G_t = R_{t+1} + R_{t+2} + \dots + R_T$

0
1   The time of termination, T, is a RV that normally varies from episode to episode

→ continuing task :- $(T = \infty)$ but then $G_t \to \infty$ or become unbounded $(+\infty / -\infty)$ easily.

Now, we need the concept of discounting.

$$G_t \triangleq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots +$$

$$= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

$0 \le \gamma \le 1$ or $\in [0,1]$
↓
discount rate.

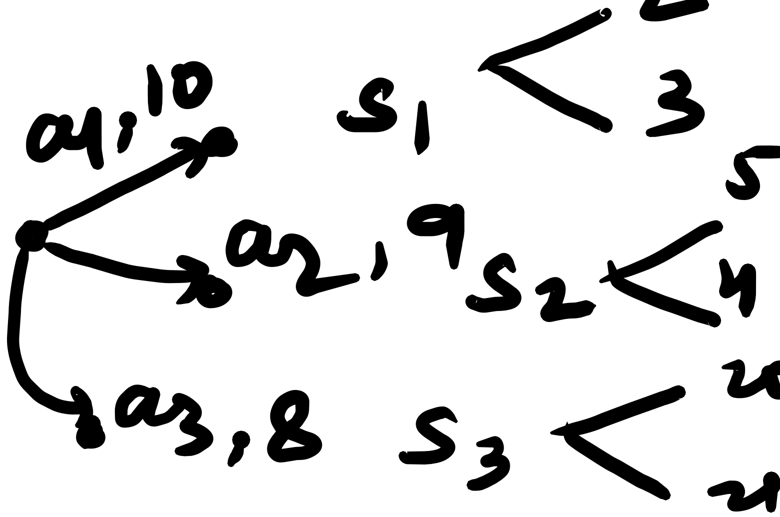
Discount rate determines the present value of future rewards :- a reward received k time steps in the future is worth only γ^{k-1} times what it would be worth if it were received immediately.

As long as $\{R_k\}$ is bounded & $\gamma < 1$

G_t is finite valued — H.W. ?

$\gamma = 0$:- agent is myopic (max. immediate reward) → but may not lead to max. of cum. reward

$\gamma \simeq 1$: agent is farsighted.



$$G_t = R_{t+1} + \gamma\left[R_{t+2} + \gamma R_{t+3} + \cdots\right]$$
$$= R_{t+1} + \gamma G_{t+1}$$

The recursive relationship works for $\forall t < T$, even if termination occurs at $t+1$, if we define $G_T = 0$

$t \to 0$ to $T-1$ (episodic)

$T-1 \to S_T, R_T$

↳ terminal state

ex- Suppose reward is non-zero & constant (2) & $\gamma < 1$

Continuing task.

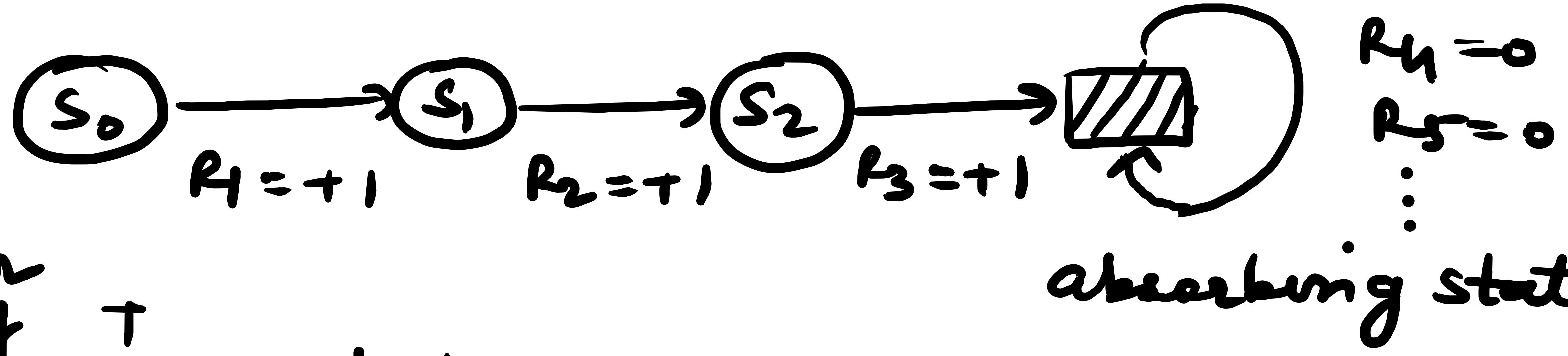
find $G_t \to \frac{2}{1-\gamma}$,

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \quad (\because \gamma < 1, R_t = 2 \ \forall t)$$

$$= 2\left[1 + \gamma + \gamma^2 + \gamma^3 + \dots\right] = \frac{2}{1-\gamma}$$

To keep notation intact b/w episodic & continuing tasks
(special absorbing state)

ex-



another way

$$G_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$$

(includes the possibility that $T \to \infty$ or $\gamma = 1$ (but not both))

bounded seq:- we say that a seq. $\{a_n\}$ is bounded if it is both bounded from below & from above. i.e., $\exists\, k$ & K s.t. $\forall n$ $k \le a_n \le K$

Policies & value functions :-

Value function of a state s under a policy π, denoted as $V_\pi(s)$, is the expected return when starting in s & "following π thereafter". For MDPs,

$$V_\pi(s) \triangleq E\left[G_t \mid S_t = s\right] = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s\right]$$

$$\forall s \in \mathcal{S}$$