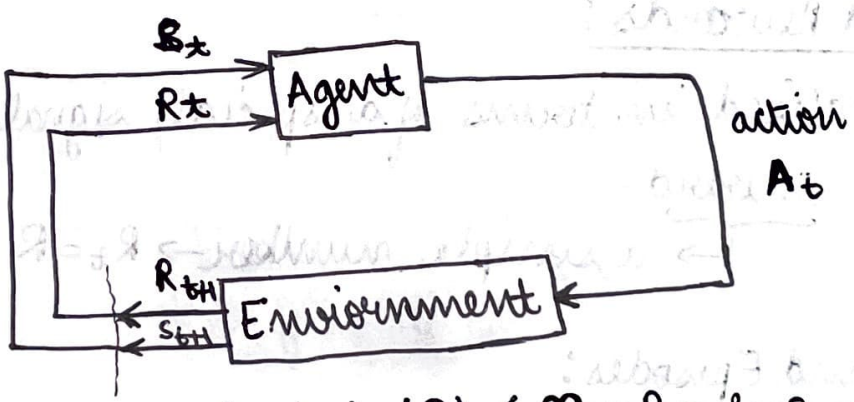
Chapter 3 - Finite Markov Decision Processes

MDPs are a classical formalisation of sequential decision making where actions influence immidiate rewards & ~~subs~~ subsequent situations.

[3.1] Agent - Environment Interface:

Agent → learner and decision maker

Environment → comprising everything outside the data.



In finite MDP $|S|, |A|, |R| < \infty$, R_t & S_t have well defined probability distributions dependant only on preceding state and action.

$$p(s', r | s, a) = Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\}$$

$$\forall s', r, s, a$$

$$p : S \times R \times S \times A \to [0, 1]$$

$$\sum_{s' \in S} \sum_{r \in R} p(s', r | s, a) = 1$$

State transition property:

$$p(s'|s,a) = Pr\{S_t = s' | S_{t-1} = s, A_{t-1} = a\} = \sum_{r} p(s', r | s, a)$$

Expected reward for state action pair

$$r(s,a) \doteq E[R_t | S_{t-1} = s, A_{t-1} = a] = \sum_{r \in R} r \sum_{s' \in S} p(s', r | s, a)$$

Expected reward for state-action-next state

$$r(s,a,s') \doteq E[R_t | S_{t-1} = s, A_{t-1} = a, S_t = s'] = \sum_{r \in R} r \frac{p(s', r | s, a)}{p(s'|s,a)}$$

[3.2] Goals and Rewards:

↳ formalised in terms of a special signal called reward.

↳ a simple number → $R_t \in R$

[3.3] Returns and Episodes:

$$G_t \doteq R_{t+1} + R_{t+2} ---- + R_T$$

↳ final time step

↓ for episodic task

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} ---- = \sum_{0}^{\infty} \gamma^k R_{t+k+1}$$

↓ for continuous tasks.

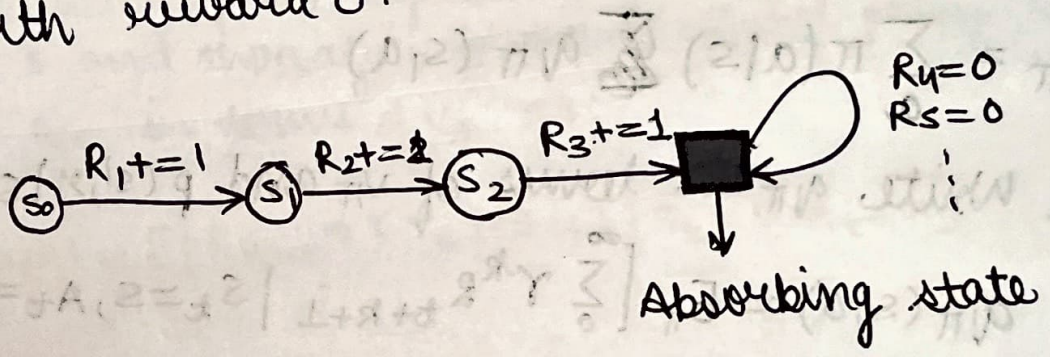
$0 \leq \gamma \leq 1$

↳ discount rate

$$\boxed{G_t = R_{t+1} + \gamma G_{t+1}}$$

[3.4] Unified notation for episodic and continuing tasks:

We have defined return as a sum over finite numbers as well as infinite nos. These 2 can be unified by considering episode termination to be entering of special absorbing state, that transitions only to itself with reward 0.



[3.5] Policies & Value functions:

Value functions → functions of states that estimate how good it is for agent to be in a given state.

Policy → a mapping from states to probabilities of selecting each possible action.

$\pi(a|s)$ → probability that action $A_t = a$ is selected at $S_t = s$.

If current state is S_t what is the expectation of R_{t+1}

$$E[R_{t+1}] = \sum_a \pi(a|s_t) \sum_{s',r} r \, p(s', r | s_t, a)$$

$$v_\pi(s) \doteq E_\pi[G_t | S_t] = E_\pi\left[\sum_0^\infty \gamma^k R_{t+k+1} | S_t = s\right] \quad \forall s \in S$$

$$q_\pi(s,a) \doteq E_\pi[G_t | S_t = s, A_t = a] = E_\pi\left[\sum_0^\infty \gamma^k R_{t+k+1} | S_t = s, A_t = a\right]$$

Ex: 3.12 Write v_π in terms of q_π

$$v_\pi = E_\pi\left[\sum_0^\infty \gamma^k R_{t+k+1} \mid S_t = s\right]$$

$$q_\pi(s,a) = E_\pi\left[\sum_0^\infty \gamma^k R_{t+k+1} \mid S_t = s, A_t = a\right]$$

$$v_\pi = \sum_a \pi(a|s)\, q_\pi(s,a)$$

Ex: 3.13 Write q_π in terms of v_π and $p(s', r|s,a)$

$$q_\pi(s,a) = E_\pi\left[\sum_0^\infty \gamma^k R_{t+k+1} \mid S_t = s, A_t = a\right]$$

$$v_\pi(s) = E_\pi\left[\sum_0^\infty \gamma^k R_{t+k+1} \mid S_t = s\right]$$

$$p(s', r|s,a) = P_r\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\}$$

$$q_\pi(s,a) = \sum_{s',r} p(s', r|s,a)\left[r + \gamma v_\pi(s')\right]$$

$$v_\pi(s) \doteq E_\pi[G_t \mid S_t = s]$$

$$= E_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s]$$

$$= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s,a)\left[r + \gamma E_\pi[G_{t+1} \mid S_{t+1} = s']\right]$$

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s,a)\left[r + \gamma v_\pi(s')\right] \quad \forall s \in S$$

↳ Bellman equation for v_π

[3.6] Optimal Policies and optimal value functions:

$$v_*(s) \doteq \max_\pi v_\pi(s) \quad \forall s \in S$$

$$q_*(s,a) \doteq \max_\pi q_\pi(s,a) \quad \forall s \in S, a \in A(s)$$

For state action pair, expected return for taking action a. in state s and thereafter following optimal policy, we can write q_* in terms of v_* as:

$$q_*(s,a) = E\left[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a\right]$$

Exercise 3.25 → v_* in terms of q_*

$$v_* = \max_a q_*(s,a)$$

Exercise 3.26 → Write q_* in v_* & $p(s', r|s,a)$

$$q_*(s,a) = \sum_{s',r} p(s', r|s,a)\left[r + \gamma v_*(s')\right]$$

Exercise 3.27 → π_* in terms of q_*

$$a_* = \arg\max_a \pi_*(a|s)$$

$$= \arg\max_a q_*(s,a)$$

Exercise 3.28 → π_* in terms of v_* & $p(s', r|s,a)$

$$\pi_* = \arg\max_a \sum p(s', r|s,a) E\left[r + \gamma v_*(s')\right]$$