MARKOV DECISION PROCESSES

dynamic setting

v and q-values

Tentative taxonom RL algorithms

MDP:

states : S set of states

S finite or S infinite

actions : A set of actions

set of reals

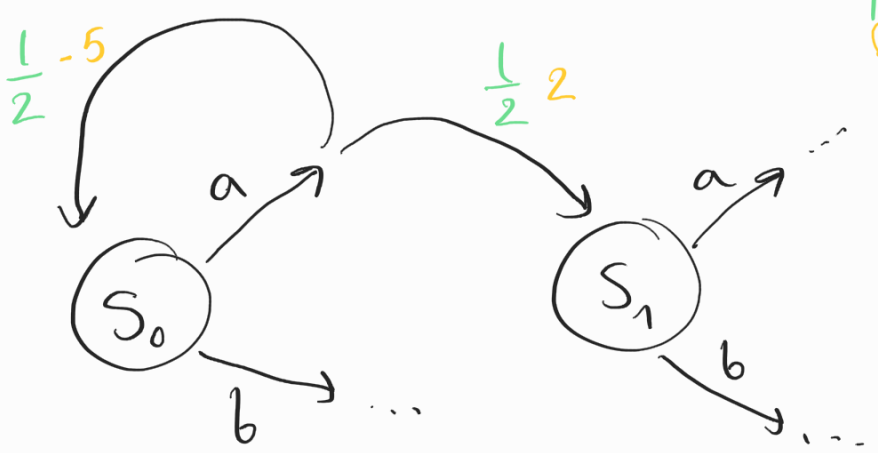
A finite or A infinite

probabilistic transition function: $\Delta : S \times A \to \mathcal{D}ist(S \times \mathbb{R})$

rewards

$\Delta(s, a)(s', r)$: probability that from state s playing action a, we go to state s' and get reward r

probability
reward



MDP : S , A , Δ

states actions transition function

Goal: Construct a strategy / policy

$\sigma : S \to A$

such that σ maximises

$\lambda \in (0,1)$

$$\mathbb{E}_{\sigma}\left[\sum_{t=0}^{\infty} \lambda^t r_t\right]$$

$r_0 + \lambda r_1 + \lambda^2 r_2 + \lambda^3 r_3 + \dots$

$s_0 \in S$ initial state

σ: $\sigma(s_0) = a_0$ $\quad \Delta(s_0, a_0)(s_1, r_0)$

$\sigma(s_1) = a_1$ $\quad \Delta(s_1, a_1)(s_2, r_1)$

$s_0, a_0, r_0, s_1, a_1, r_1, s_2, \dots$ →

play / trajectory / path

infinite or finite

Two cases:

- either eventually we reach a sink

$$\sum_{t=0}^{\infty} r_t$$ is actually finite

→ FINITE HORIZON

- or the trajectory may be infinite

$$\sum_{t=0}^{\infty} \lambda^t r_t$$

$\sum_{t=0}^{\infty} \gamma^t r_t$ may not be defined

→ DISCOUNTED

$\gamma \in (0,1)$: fixed constant

$$\sum_{t=0}^{\infty} \gamma^t r_t = r_0 + \gamma r_1 + \gamma^2 r_2 + \gamma^3 r_3 + \dots$$

$$\gamma^t \xrightarrow[t \to \infty]{} 0$$

Good news:

(1) well defined mathematically

(2) reasonable in practice