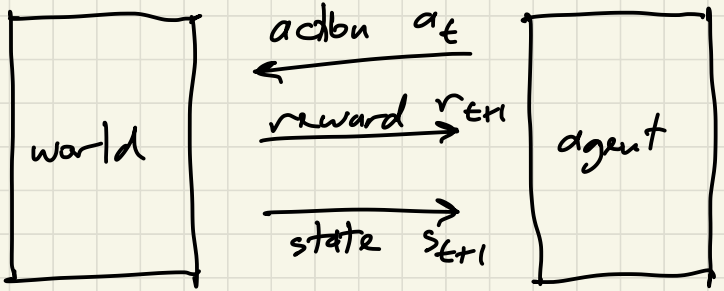
MDPs – Markov Decision Processes

– stochastic, discrete state, discrete action, state feedback (fully observable)

environment



t = time

3 system spaces: state $s_t \in S$, action $a_t \in A$, reward $r_t \in \mathbb{R}$

3 functions: model $p(s_{t+1} \mid s_t, a_t)$ = prob. of s_{t+1} given s_t, a_t

state transition function

reward $r_{t+1} = R(s_t, a_t)$ ← deterministic

we want to find → policy $\pi(a_t \mid s_t)$ = prob. of a_t given s_t

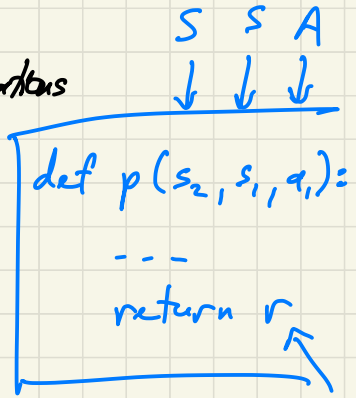
Notation

lowercase s, a, r, s_t, a_t ← values, samples, realizations

uppercase S_t, A_t, R_t ← random variables

$p(s_{t+1} \mid s_t, a_t) = \text{Prob}(S_{t+1} = s_{t+1} \mid S_t = s_t, A_t = a_t)$

$p : S \times S \times A \to \mathbb{R}$



$s_{t+1} \sim p(\cdot \mid s_t, a_t)$ ← s_{t+1} has (marginal) distribution given by p

s_{t+1} is sampled from p

Example $f(s)$ = expected (average) 1-step reward from state s

$$= \mathbb{E}[R(S_t, A_t) \mid S_t = s] \quad = \mathbb{E}[R(s,a) \mid s]$$

($R(S_t, A_t)$: R.V.)

$$= \mathbb{E}_{a \sim \pi(\cdot \mid s)}[R(s,a)]$$

$$= \sum_{a \in A} \pi(a \mid s) R(s,a)$$