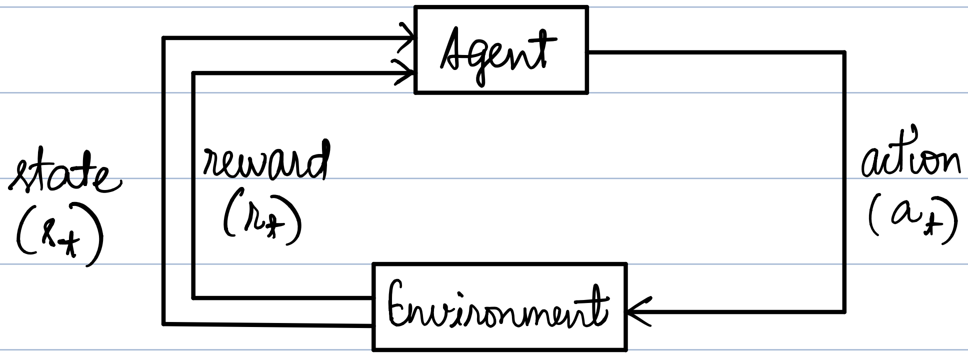
Chapter 3: Finite MDP

30/07/19

3.1. Agent-Environment Interface

- agent ← learner, decision maker
- environment ← everything outside agent



Thus, resulting trajectory: $S_0, A_0, R_0, S_1, A_1, R_1 \ldots$

- S, A, R are finite in a finite MDP

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

where $s, s' \in S$, $r \in R$, $a \in A(s)$

defines the dynamics of the MDP

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

$$\sum_{s' \in S} \sum_{r \in R} p(s', r \mid s, a) = 1 \quad \forall\, s \in S,\ a \in A(s)$$

In a MDP, p characterizes the environment's dynamics

↳ Probability of each possible value for S_t and R_t depends only on the immediate preceding state and action, S_{t-1} and A_{t-1}

state transition probability :

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

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

expected reward :

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

$$r : S \times A \rightarrow \mathbb{R}$$

3.2. Goals and Rewards

- reward hypothesis
  - all goals can be expressed as expected reward maximisation of a specific scalar signal

3.3. Returns and Episodes

- expected return (for episodic tasks)

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

$T \leftarrow$ final time step (terminal state)

- discounted return (for continuing tasks)

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

$0 \leq \gamma \leq 1 \leftarrow$ discount rate

- $G_t \doteq R_{t+1} + \gamma G_{t+1}$

3.5. Policy and Value functions

- policy $\longrightarrow \pi(a|s)$
- value function (state value function)
  - $\hookrightarrow v_\pi(s) \doteq E_\pi[G_t \mid S_t = s]$
- action value function
  - $\hookrightarrow q_\pi(a, s) \doteq E_\pi[G_t \mid S_t = s, A_t = a]$

- value function is a measure of the "goodness" of a state

3.12 : $v_\pi(s) = \sum_{a \in A} \pi(a|s) \cdot q_\pi(a, s)$

3.13 : $q_\pi(s) = E[G_t \mid s, a]$

$$= E[R_{t+1} + \gamma G_{t+1} \mid s, a]$$

$$= \sum_{a \in A} p(s'|s,a) \cdot [R_{t+1} + \gamma V(s')]$$

Bellman Equation :

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

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

$$= \sum_{a \in A} \pi(a|s) \, E[q_\pi(a,s) \mid S_t = s]$$

$$= \sum_{a \in A} \pi(a|s) \sum_{s'} \sum_{r} p(s', r|s,a) \cdot [r + \gamma v_\pi(s')]$$

3.6. Optimal policy and Value functions

$$\pi^* \leftarrow v_{\pi^*}(s) \geq v_\pi(s) \quad \forall s \in S$$

$$q_*(s,a) = \max_\pi q_\pi(s,a)$$

26/08/19

$$\begin{aligned}
v_\pi(s) &= E\left[G_t \mid S_t = s\right] \\
&= E\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s\right] \\
&= E\left[R_{t+1} \mid S_t = s\right] + \gamma\, E\left[G_{t+1} \mid S_t = s\right] \\
&= \sum_a \pi(a|s) \sum_{s'} R^a_{ss'} P^a_{ss'} + \gamma \sum_a \pi(a|s) \sum_{s'} P^a_{ss'} v_\pi(s') \\
&= \sum_a \pi(a|s) \sum_{s'} P^a_{ss'} \cdot \left[R^a_{ss'} + \gamma v_\pi(s') \mid S_{t+1} = s'\right]
\end{aligned}$$

here, $P^a_{ss'} = p(s' \mid s, a)$ and $R^a_{ss'} = E\left[R_{t+1} \mid S_t = s, A_t = a, S_{t+1} = s'\right]$