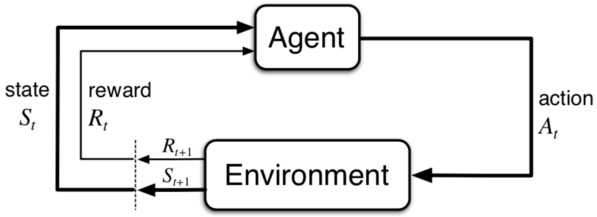
Reinforcement Learning Cheat Sheet

Recap

$\mathbb{E}[X] \doteq \sum_{x_i} x_i \cdot Pr\{X = x_i\}$
$\mathbb{E}[X|Y = y_j] = \sum_{x_i} x_i \cdot Pr\{X = x_i|Y = y_j\}$
$\mathbb{E}[X|Y = y_j] = \sum_{z_k} Pr\{Z = z_k|Y = y_j\} \cdot \mathbb{E}[X|Y = y_j, Z = z_k]$

Agent-Environment Interface



The Agent at each step t receives a representation of the environment's *state* $S_t \in \mathcal{S}$ and it selects an action $A_t \in \mathcal{A}(s)$. One time step later, as a consequence of its action, the agent receives a *reward*, $R_{t+1} \in \mathcal{R} \subseteq \mathbb{R}$ and goes to the new state S_{t+1}.
The MDP and agent together thereby give rise to a sequence or trajectory that begins like this:
$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3 ...$

Return

The *return* is a some specific function of reward sequence. When there is a natural notion of final time step (T), the agent-environment interaction breaks naturally into sub-sequences (*episodes*). Each episodes ends in a special state called *terminal state*. $\mathcal{S}^+$ is the set of all states plus the terminal state.
The *total discounted return* is expressed as the sum of rewards (opportunely discounted with γ):

$$\begin{aligned} G_t &\doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... \\ &= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \quad [3.8] && (1) \\ &= R_{t+1} + \gamma G_{t+1} \quad [3.9] && (2) \end{aligned}$$

Where γ is the *discount factor* and T is the final time step. It can be infinite. When there is a natural notion of final time step, we have the *episodes*.

https://github.com/linker81/Reinforcement-Learning-CheatSheet

Policy

A *policy* is a mapping from a state to probabilities of selecting each possible action

$$\pi(a|s) \qquad (3)$$

That is the probability of select an action $A_t = a$ if $S_t = s$.

Markov Decision Process

A finite **Markov Decision Process**, MDP, is defined by:
finite set of states: $s \in \mathcal{S}$, finite set of actions: $a \in \mathcal{A}$
dynamics:
$p(s', r|s, a) \doteq Pr\{S_t = s', R_t = r|S_{t-1} = s, A_{t-1} = a\}$ [3.2]
state transition probabilities:
$p(s'|s, a) \doteq Pr\{S_t = s'|S_{t-1} = s, A_{t-1} = a\}$ [3.4]
expected reward for state-action:

$$\begin{aligned} r(s, a) &\doteq \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a] \\ &= \sum_{r \in \mathcal{R}} r \cdot \sum_{s' \in \mathcal{S}} p(s', r|s, a) \quad [3.5] \end{aligned}$$

expected reward for state-action-nextstate:

$$\begin{aligned} r(s', s, a) &\doteq \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a, S_t = s'] \\ &= \sum_{r \in \mathcal{R}} r \cdot \frac{p(s', r|s, a)}{p(s'|s, a)} \quad [3.6] \end{aligned}$$

Value Functions

State-Value function describes *how good* is to be in a specific state s under a certain policy π. Informally, is the expected return (expected cumulative discounted reward) when starting from s and following π

$$\begin{aligned} v_\pi(s) &\doteq \mathbb{E}_\pi[G_t|S_t = s] \quad [3.12] && (4) \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s] \quad [3.12] && (5) \\ &= \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a) \Big[r + \gamma v_\pi(s')\Big] \quad [3.14] && (6) \end{aligned}$$

Action-Value function (Q-Function) describes *how good* is to perform a given action a in a given state s under a certain policy π. Informally, is the expected return when starting from s, taking action a and following π

$$\begin{aligned} q_\pi(s, a) &\doteq \mathbb{E}_\pi[G_t|S_t = s, A_t = a] \quad [3.13] && (7) \\ &= \sum_{s',r} p(s', r|s, a) \Big[r + \gamma \sum_{a'} \pi(a'|s') q_\pi(a', s')\Big] \quad [Ex\ 3.17] && (8) \end{aligned}$$

Relation between Value Functions

$$\begin{aligned} v_\pi(s) &= \sum_a \pi(a|s) \cdot q_\pi(s, a) \quad [Ex\ 3.12] && (9) \\ &= \mathbb{E}_\pi[q_\pi(s, a)|S_t = s] \quad [Ex\ 3.18] && (10) \end{aligned}$$

$$\begin{aligned} q_\pi(s, a) &= \sum_{s',r} p(s', r|s, a) \Big[r + \gamma v_\pi(s')\Big] \quad [Ex\ 3.13] && (11) \\ &= \mathbb{E}\Big[R_{t+1} + \gamma v_\pi(s')|S_t = s, A_t = a\Big] \quad [Ex\ 3.19] && (12) \end{aligned}$$

Optimal Value Functions

$$\begin{aligned} v_*(s) &\doteq \max_\pi v_\pi(s) \quad [3.15] && (13) \\ &= \max_a \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1})|S_t = s, A_t = a] \quad [3.18] \\ &= \max_a \sum_{s',r} p(s', r|s, a) \Big[r + \gamma v_*(s')\Big] \quad [3.19] \end{aligned}$$

$$\begin{aligned} q_*(s, a) &\doteq \max_\pi q_\pi(s, a) \quad [3.16] && (14) \\ &= \mathbb{E}[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a')|S_t = s, A_t = a] \\ &= \sum_{s',r} p(s', r|s, a) \Big[r + \gamma \max_{a'} q_*(s', a')\Big] \quad [3.20] \end{aligned}$$

$$v_*(s) = \max_{a \in A(s)} q_{\pi_*}(s, a) \qquad (15)$$

Intuitively, the above equation express the fact that the value of a state under the optimal policy **must be equal** to the expected return from the best action from that state.

Relation between Optimal Value Functions

$$v_*(s) = \max_a \sum_{s',r} p(s', r|s, a) \Big[r + \gamma \sum_{a'} \pi(a'|s') q_*(s', a')\Big] \quad [Ex\ 3.25] \qquad (16)$$

$$q_*(s, a) = \sum_{s',r} p(s', r|s, a) \Big[r + \gamma v_*(s')\Big] \qquad (17)$$