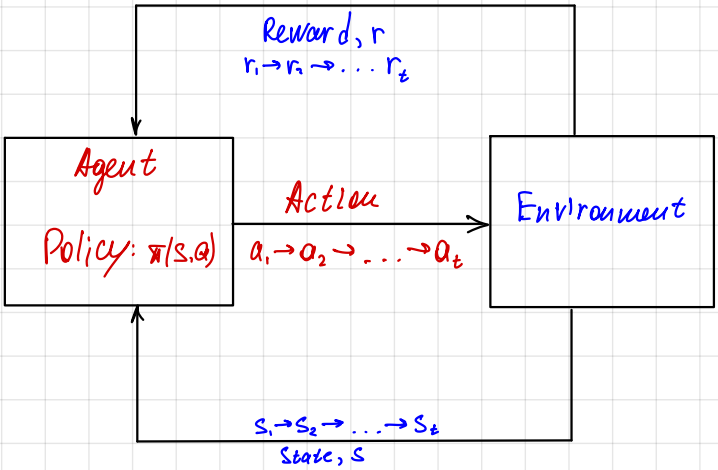
1. Overview

Policy : $\pi(s,a) = Pr(a=a \mid s=s)$



Value: $V_\pi(s) = E\left(\sum_t \gamma^t r_t \mid s_0 = s\right)$

2. Q-learning

$Q(s,a)$ = Quality of state / action pair

- combine Value $V_\pi(s)$ and Policy $\pi(s,a)$.

$$Q^{update}(s_t, a_t) = Q^{old}(s_t, a_t) + \alpha\left(r_t + \gamma \max_a Q(s_{t+1}, a) - Q^{old}(s_t, a_t)\right)$$