Definitions

Action (A): All the possible moves agent can make.

State (S): Current situation returned by environment

Reward (R): An immediate return sent back from the environment to evaluate last action

Policy (π): Strategy that agent employs to determine next action based on current state.

Value (V): Expected long-term return with discount, as opposed to short term reward.

$V^{\pi}(s) \rightarrow$ long term return of current state under policy π

Q-value: $Q^{\pi}(s,a) \rightarrow$ long term return of current state s, taking action a under policy π.

Q-Learning

Reward $= Q(s,a) \rightarrow$ probability of a reward in action-state pair } Q-value

$$Q(s,a) = r + \gamma \max_{a'} Q(s',a')$$

- r: reward by making action a
- next reward r'
- next state s'
- action in next state a'

update

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a') - Q(s,a))$$