

"Markov Decision Process"

S : all possible states

A : " " actions

R : reward distribution
given (S, a)

P : transition prob.
to S_{t+1}
given (S, a)

γ : discount factor

$$\boxed{S \xrightarrow{\pi} A}$$

objective:

$$\text{find } \pi^* = \max_{\pi} \left(\sum_{t=0}^{\infty} \gamma^t r^t \right)$$

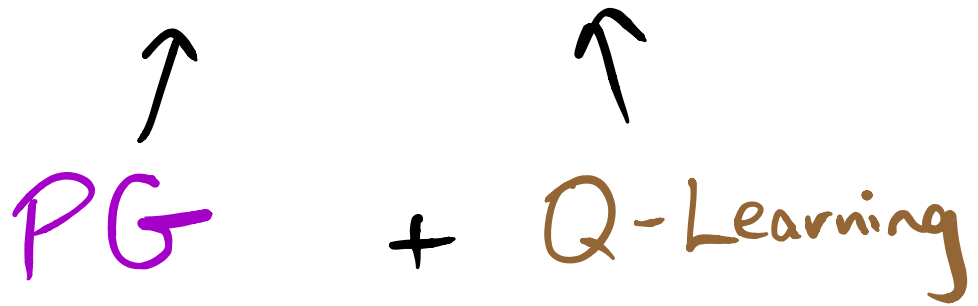
value function: $V^{\pi}(s)$

Q-value function: $Q^{\pi}(s, a)$

$$Q^*(s, a) \approx Q(s, a, \theta)$$

Actor - Critic

PG + Q-Learning



The diagram illustrates the components of the Actor-Critic reinforcement learning architecture. At the top, the text 'Actor - Critic' is written in blue. Below it, 'PG' (Policy Gradient) is written in purple and 'Q-Learning' is written in brown, separated by a plus sign. Two black arrows point upwards from 'PG' and 'Q-Learning' towards 'Actor - Critic', indicating that both methods contribute to the overall architecture.