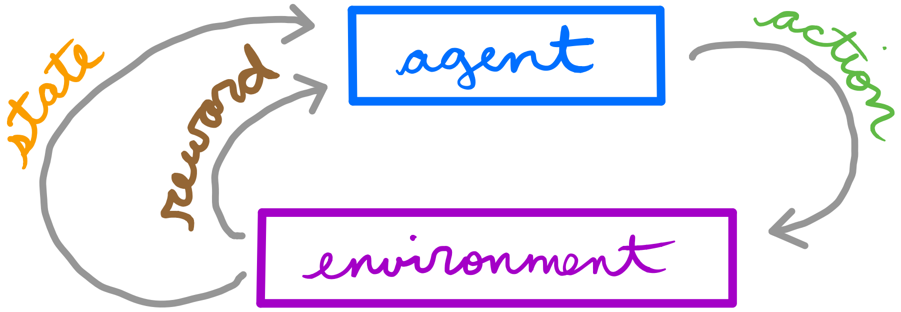


"Markov Decision Process"

S : all possible states

A : " " actions

R : reward distribution given (s, a)

$\mathbb{P}$: transition prob. to s_{t+1} given (s, a)

γ : discount factor

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

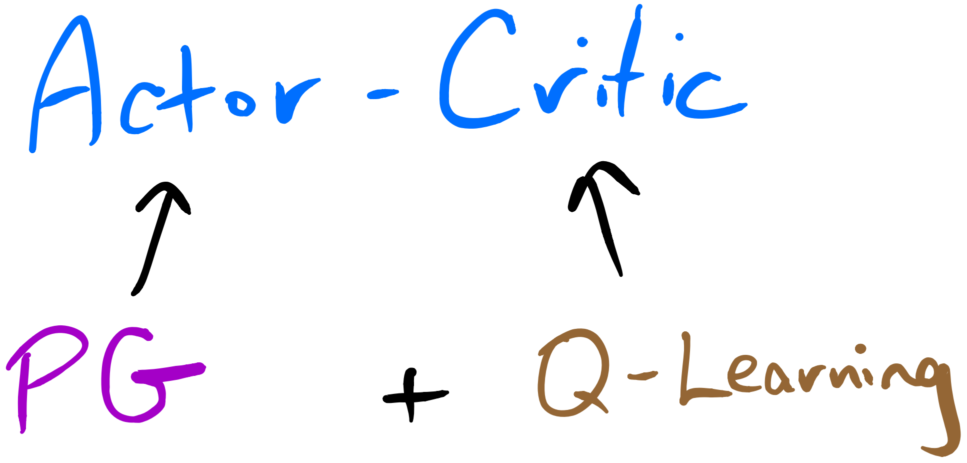
objective:

find $\pi^* = \max \left(\sum_{t>0} \gamma^t r^t \right)$

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