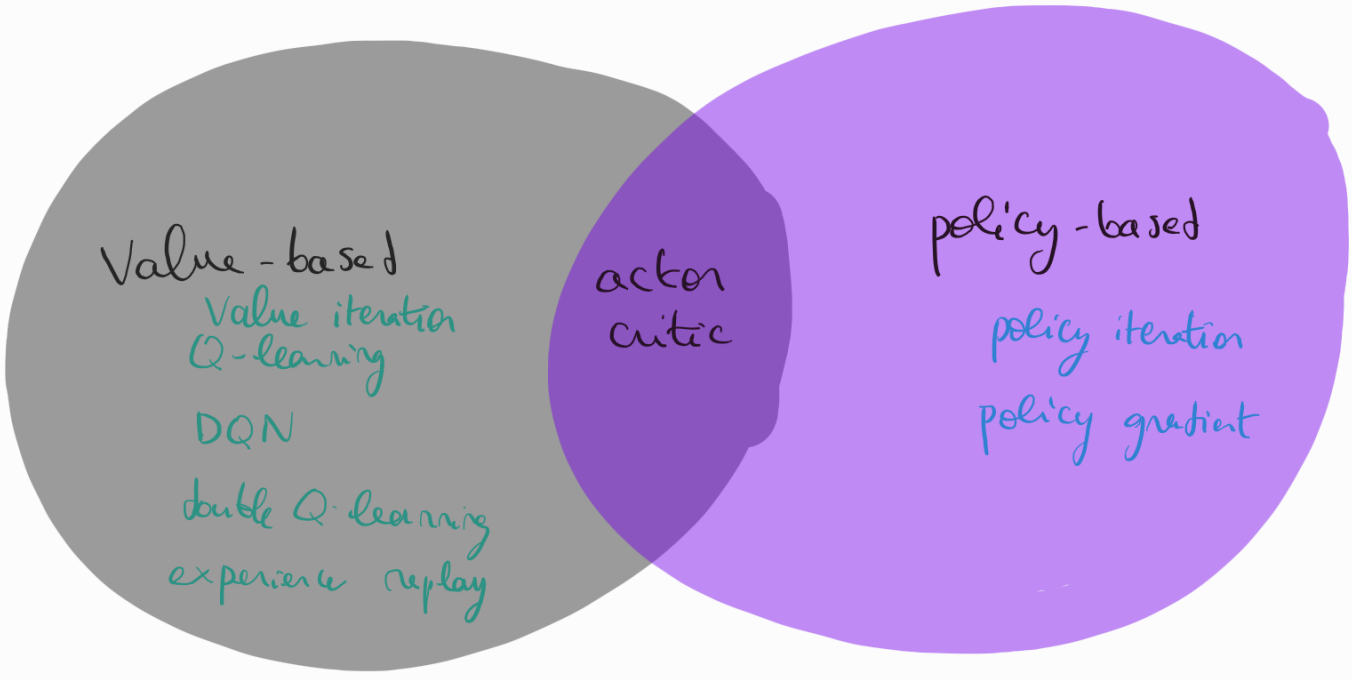
ACTOR CRITIC METHODS



Value-based :

- Monte Carlo
- Temporal difference

Policy-based :

policy gradient

TD :

$v_w : S \longrightarrow \mathbb{R}$ represented by neural network

w : parameters

current state s

choose a using ε-greedy from $v_w(s)$

sample (s, a, r, s')

$$\delta \leftarrow r + \gamma v_w(s') - v_w(s)$$

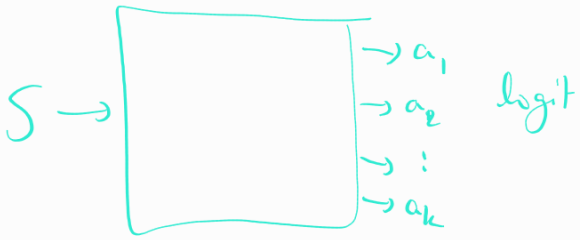
$$w \leftarrow w + \alpha \delta \nabla_w v_w(s)$$

PG

$\pi_\theta : S \rightarrow \mathcal{D}(A)$ represented by neural network

θ: parameters

current state s



Choose a using $\pi_\theta(s)$

Sample (s, a, r, s')

$$\delta \leftarrow r + \gamma \max_{a'} \pi_\theta(s', a') - \pi_\theta(s, a)$$

$$\theta \leftarrow \theta + \beta \, \delta \, \nabla_\theta \log \pi_\theta(a|s)$$

Terminology:

Critic : value function

Actor : policy

One step Actor Critic

$v_w : S \longrightarrow \mathbb{R}$ represented by neural network

w: parameters

$\pi_\theta : S \longrightarrow \mathcal{D}(A)$ represented by neural network

θ: parameters

current state s

choose a using $\pi_\theta(s)$

sample (s, a, r, s')

$$\delta \leftarrow r + \gamma\, v_w(s') - v_w(s)$$

$$w \leftarrow w + \alpha\, \delta\, \nabla_w v_w(s) \qquad \text{TD}$$

$$\theta \leftarrow \theta + \beta\, \delta\, \nabla_\theta \log \pi_\theta(a|s) \qquad \text{PG}$$

INCREASING ROBUSTNESS WITH TRUST REGIONS

Regularisation

Kullback-Leibler divergence:

$$KL(\pi_{old} \| \pi_\theta) = \mathop{\mathbb{E}}_{s}\left[\sum_a \pi_{old}(a|s) \log \frac{\pi_\theta(a|s)}{\pi_{old}(a|s)}\right]$$

new objective:

$$\mathop{\mathbb{E}}_{\pi}[R(\pi\theta)] - \eta\, KL(\pi_{old} \| \pi_\theta)$$

TRPO / PPO