SARSA vs Q-learning

on vs off policy learning

ϵ strategy: has two competing goals:

(1) exploration:

(2) exploitation

off-policy learning: use two strategies:

one for exploring and one for exploiting

We design a policy improvement algorithm in the statistical framework

Objective: learn $q_* : S \times A \to \mathbb{R}$

initialise $q : S \times A \to \mathbb{R}$ (0 or random)

REPEAT:

$S \leftarrow S_0$

while (trajectory not over):

choose action a using ε-greedy strategy from q

sample (s, a) : r, s'

CASE ①: SARSA (on-policy):

(s, a, r, s', a')

evaluate using ε-greedy

$$q(s,a) = q(s,a) + \alpha\left[r + \gamma\, q(s', a(s')) - q(s,a)\right]$$

CASE ②: Q-learning (off-policy)

evaluating using greedy

$$q(s,a) = q(s,a) + \alpha\left[r + \gamma \max_{a' \in A} q(s', a') - q(s,a)\right]$$