$v : S \to \mathbb{R}$

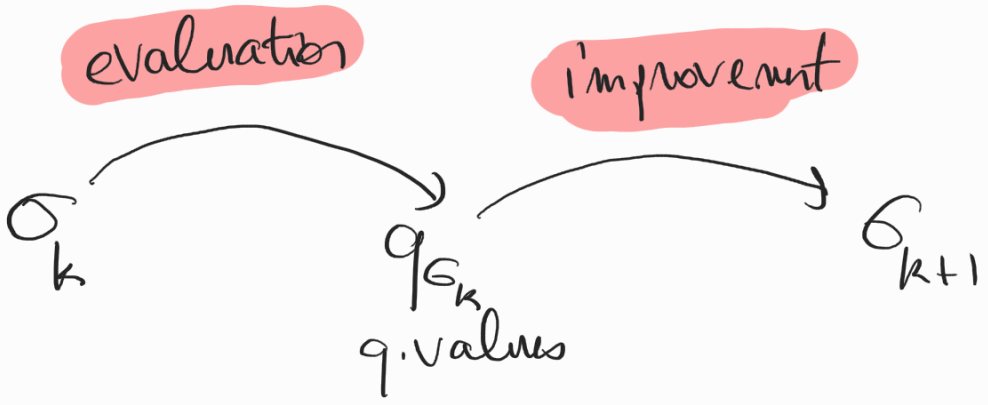
Value iteration $v_0 \to v_1 \to v_2 \to v_3 \dots$

policy iteration $\sigma_0 \to \sigma_1 \to \sigma_2 \to \sigma_3 \dots$

$\sigma : S \to A$

=

strategy improvement



improvement:

$$\sigma_{k+1}(s) = \operatorname*{argmax}_{a \in A} q_{\sigma_k}(s,a)$$

evaluation:

$$\sigma \longrightarrow q_\sigma$$

On-policy Bellman equations for q-values:

$$q_\sigma(s,a) = \sum_{s',r} \Delta_{(s,a)}(s',r)\left(r + \gamma\, q_\sigma(s', \sigma(s'))\right)$$

$s' \in S$
$r \in \mathbb{R}$

$val_\sigma(s')$

"policy evaluation" : (variant of VI) where σ is fixed

$$q_\sigma^0(s,a) = 0 \quad \text{for all } s, a$$

For $k = 0, \dots$:

$$q_\sigma^{k+1}(s,a) = \sum_{\substack{s' \in S \\ r \in \mathbb{R}}} \Delta(s,a)(s', r)\left(r + \gamma\, q_\sigma^k(s', \sigma(s'))\right)$$

until: $\|q_\sigma^{k+1} - q_\sigma^k\| < \varepsilon$

PI:

σ_0 arbitrary policy

iterate:

- evaluation $(\sigma_k) \longrightarrow q_{\sigma_k}$

- $\sigma_{k+1} = \text{improve}(\sigma_k, q_{\sigma_k})$

- $\theta_{k+1} = \text{Improve}(\theta_k, q_{\theta_k})$

MDPs

Bellman equations

↳ Value iteration

↳ Policy iteration