ON-POLICY BELLMAN EQUATIONS

Fix $\sigma : S \to A$. How do we compute val_σ ?

$$\mathrm{val}_\sigma(s) = \mathbb{E}_\sigma\left[\sum_{i \geq 0} \gamma^i r_i\right]$$

$$\mathrm{val}_\sigma(s) = \sum_{s', r} \delta(s', r \mid s, \sigma(s)) \left[r + \gamma\, \mathrm{val}_\sigma(s') \right]$$

$\hookrightarrow$ this is a set of linear equations with unknowns $(\mathrm{val}_\sigma(s))_{s \in S}$

OPERATOR POINT OF VIEW

Define $F_S = \{ v : S \to \mathbb{R} \} \simeq \mathbb{R}^S$

$\Phi_\sigma : F_S \to F_S$

$$\Phi_\sigma(v)(s) = \sum_{s', r'} \delta(s', r \mid s, \sigma(s)) \left[r + \gamma v(s') \right]$$

Bellman's equations: $\Phi_\sigma(\mathrm{val}_\sigma) = \mathrm{val}_\sigma$

Theorem: Φ_σ is γ-contracting:

$$\| \Phi_\sigma(v) - \Phi_\sigma(v') \|_\infty < \gamma \| v - v' \|_\infty$$

So it has a unique fixed point: val_σ

Algorithm: $v_0(s) = 0$

$$v_{k+1} = \Phi_\sigma(v_k)$$

Stop when $\|v_{k+1} - v_k\| \leq \varepsilon$

Lemma: $v_k = \mathbb{E}_\sigma\left[\sum_{i=0}^{k-1} \gamma^i r_i\right]$

STRATEGY IMPROVEMENT ≡ ITERATION

$$\sigma \rightsquigarrow val_\sigma \rightsquigarrow \sigma'$$

$$\sigma'(s) = \underset{a \in A}{\operatorname{argmax}} \sum_{s', r} \delta(s', r \mid s, a)\left[r + \gamma\, val_\sigma(s')\right]$$

Theorem

if σ is not optimal then $val_\sigma < val_{\sigma'}$

Strategy improvement algorithm

$$\sigma \underset{\text{eval}}{\rightsquigarrow} val_\sigma \underset{\text{impr}}{\rightsquigarrow} \sigma_1 \underset{\text{eval}}{\rightsquigarrow} val_{\sigma_1} \underset{\text{impr}}{\rightsquigarrow} \sigma_2 \;\cdots\cdots\; \sigma_* \text{ optimal}$$

OFF-POLICY BELLMAN EQUATIONS

$$\mathrm{val}_*(s) = \max_{a \in A} \sum_{s', r} \delta(s', r \mid s, a) \left[r + \gamma \, \mathrm{val}_*(s') \right]$$

$$\Phi(v)(s) = \max_{a \in A} \sum_{s', r'} \delta(s', r \mid s, \sigma(s)) \left[r + \gamma v(s') \right]$$

Theorem : Φ is γ-contracting :

$$\| \Phi(v) - \Phi(v') \|_\infty \leq \gamma \| v - v' \|_\infty$$

So it has a unique fixed point : val_*

Value iteration algorithm :

$$v_0 \rightsquigarrow v_1 = \Phi(v_0) \rightsquigarrow v_2 = \Phi(v_1) \rightsquigarrow \dots$$

until $\| v_{k+1} - v_k \|_\infty \leq \varepsilon$