1) Value Iteration

~~from Q-functions to policies~~

remember $\pi^*(s) = \underset{a \in \mathcal{A}}{\operatorname{argmax}} Q^*(s,a)$

$Q^t \approx Q^*$

is $\underline{\pi^t(s)} = \underset{a \in \mathcal{A}}{\operatorname{argmax}} Q^t(s,a)$

a good policy?

Theorem:

$$\boxed{V^{\pi^t}(s) \geq V^*(s) - \underbrace{\frac{2\gamma^t}{1-\gamma} \|Q^0 - Q^*\|_\infty}_{} \quad \forall s \in \mathcal{S}}$$

Proof:

assume:

$$\underline{V^{\pi^t}(s) - V^*(s) \geq \gamma} \underset{s' \sim P(s, \pi^t(s))}{\mathbb{E}} \left[V^{\pi^t}(s') - V^*(s') \right]$$

$$- \underbrace{2 \|Q^t - Q^*\|_\infty}$$

$$\boxed{\|Q^t - Q^*\| \leq \gamma^t \|Q^0 - Q^*\|_\infty}$$

$$\underbrace{V^{\pi^t}(s) - V^*(s)} \ge \gamma \underbrace{\mathbb{E}_{s'}[\underbrace{V^{\pi^t}(s') - V^*(s')}]} - 2\gamma^t \|Q^0 - Q^t\|_\infty$$

$$\ge \gamma \mathbb{E}_{s'}\left[\gamma \mathbb{E}[V^{\pi^t}(s'') - V^*(s'')] - 2\gamma^t \|Q^0 - Q^t\|\right]$$

$$-2\gamma^t \|Q^0 - Q^t\|_\infty$$

$$\ge \cancel{\gamma^k \mathbb{E}_{s^k}[V^{\pi^t}(s^k) - V^*(s^k)]} \to 0$$

$\gamma < 1$

$$-2\underbrace{\sum_{\ell=0}^{k} \gamma^{t+\ell}} \|Q^0 - Q^t\|_\infty$$

let $k \to \infty$

$$\sum_{\ell=0}^{\infty} \gamma^{t+\ell} = \gamma^t \sum_{\ell=0}^{\infty} \gamma^\ell = \frac{1}{1-\gamma} \cdot \gamma^t$$

proof of assumption

1) write V^{π^t} & V^* in terms of Q^{π^t} & Q^*

$V^\pi(s) = Q^\pi(s, \underline{\pi(s)})$

2) add & subtract Q-fns.

3) use definition of π^t

2) Policy Iteration

Alg (PI)

initialize $\pi^0 : \mathcal{S} \to \mathcal{A}$

for $t = 0, 1, \ldots$

Policy Evaluation: $Q^{\pi^t}(s,a)\ \forall s,a)$

Policy Improvement:

$$\forall s \quad \pi^{t+1}(s) = \arg\max_{a \in \mathcal{A}} Q^{\pi^t}(s,a)$$

aside: PE-exact $\quad V = R + \gamma P V$

$$\mathbb{R}^S \to V^{\pi^t} = (I - \gamma P)^{-1} R$$

$$Q^{\pi^t}(s,a) = r(s,a) + \mathbb{E}_{s' \sim P(s,a)}[V^{\pi^t}(s')]$$

Two key properties

1) Monotonic Improvement

$$Q^{\pi^{t+1}}(s,a) \geq Q^{\pi^t}(s,a) \quad \forall s,a$$

2) convergence

$$\|V^* - V^{\pi^t}\|_\infty \leq \gamma^t \|V^* - V^{\pi^0}\|_\infty$$

$$\boxed{\begin{array}{l}\underline{\text{Lemma}} \text{ (Monotonic Improvement)} \\ \qquad Q^{\pi^{t+1}}(s,a) \geq Q^{\pi^t}(s,a)\end{array}}$$

<u>Proof</u>:

$$Q^{\pi^{t+1}}(s,a) - Q^{\pi^t}(s,a)$$

$$= \gamma \mathbb{E}_{s' \sim P(s,a)}\left[\underbrace{V^{\pi^{t+1}}(s') - V^{\pi^t}(s')} \right]$$

$$\star \; \underline{Q^{\pi^{t+1}}(s', \underbrace{\pi^{t+1}(s')}) - Q^{\pi^t}(s', \underbrace{\pi^t(s')})}$$

$$\geq Q^{\pi^t}(s', \pi^{t+1}(s')) \cancel{- Q^{\pi^t}(s', \pi^t(s'))}$$

$$+ Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) \cancel{= Q^{\pi^t}(s', \pi^t(s'))}$$

$$Q^{\pi^{t+1}}(s,a) - Q^{\pi^t}(s,a).$$

$$\geq \gamma \mathop{\mathbb{E}}_{s'} \left[Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^{t+1}(s')) \right]$$

iterate K times

$$\geq 0 \qquad \square$$

Q: does lemma imply that

~~$V^{\pi^{t+1}}(s) \geq V^{\pi^t}(s) \quad \forall s$?~~

$$\underbrace{Q^{\pi^{t+1}}(s, \pi^{t+1}(s))}_{V^{\pi^{t+1}}(s)} \geq \underline{Q^{\pi^t}(s, \pi^{t+1}(s))}$$

because $\pi^{t+1} = \arg\max Q^{\pi^t}(s,a)$

$$\geq \underbrace{Q^{\pi^t}(s, \pi^t(s))}_{V^{\pi^t}(s)}$$

<u>Theorem</u>: Convergence

$$\| V^{\pi^t} - V^* \|_\infty \leq \gamma^t \| V^{\pi_0} - V^* \|_\infty$$

Proof sketch:

1) use Q fn; Bellman opt.
2) monotic improvement
3) $\pi^{t+1} = \arg\max Q^{\pi^t}(s,a)$
4) Basic Inequalities

Value & Policy Iteration

- converge exponentially fast
  γ^t

- exact solution?
  $\gamma^t \to 0$ as $t \to \infty$
  But $\gamma^T > 0 \quad \forall T$

- only policy iteration is guaranteed to find exact optimal policy in finite # steps (HW1)

3) Finite Horizon MDP

$$\mathcal{M} = \{ \underline{S}, \underline{A}, \underline{P}, \underline{r}, H, \mu_0 \}$$

$$H \in \mathbb{N}^+ = \{1, 2, 3, \ldots\}$$

length of time

$$\mu_0 \in \Delta(S)$$ initial state distribution

time-varying policies

$$\pi = (\pi_0, \pi_1, \pi_2, \ldots \pi_{H-1})$$

$$\pi_t : S \to A$$

Value & Q fn.

$$V_t^{\pi}(s) = \mathbb{E}\left[\sum_{k=t}^{H-1} r(s_k, a_k) \middle| \begin{array}{l} s_t = s \\ s_{k+1} \sim P(s_k, a_k) \\ a_k = \pi_k(s_k) \end{array} \right]$$

$$Q_t^{\pi}(s, a) = \mathbb{E}\left[\sum_{k=t}^{H-1} r(s_k, a_k) \middle| \begin{array}{l} s_t = s, a_t = a \\ s_{k+1} \sim P \quad a_k = \pi_k^{(t)} \end{array} \right]$$

Bellman Equation:

$$Q_t^{\pi}(s,a) = r(s,a) + \mathbb{E}_{s' \sim P(s,a)}\left[V_{t+1}^{\pi}(s')\right]$$

4) Dynamic Programming

to find $\pi^* = (\pi_0^*, \dots \pi_{H-1}^*)$

start at $H-1$: (note: $V_H^*(s) = 0$)

$$Q_{H-1}^*(s,a) = r(s,a)$$

$$\pi_{H-1}^* = \arg\max_{a \in \mathcal{A}} Q_{H-1}^*(s,a)$$

$$V_{H-1}^* = Q_{H-1}^*(s, \pi_{H-1}^*(s))$$

$$\to Q_t^*(s,a) = r(s,a) + \mathbb{E}_{s' \sim P(s,a)}\left[V_{t+1}^*(s')\right]$$

$$\to \pi_t^*(s) = \arg\max Q_t^*(s,a)$$

Dynamic Programming terminates in π steps.

<u>exact</u> π^*