Value-based theory

Goal: Value-iteration converges to opt. policy

1. Bellman Optimal operator
2. Convergence.
3. It has a unique solution.

π^* opt.

$$q_*(s,a) = \underbrace{r} + \gamma \sum \sim\sim \max q \text{———}$$

$$= \max_{\pi} \mathbb{E}[G_t \mid s_t, a_t]$$

$S_t = s$

$A_t = a$

return starting at state s_t with a_t.

$$G_t = \sum_{k=t}^{\infty} \gamma^{k-t} r_k = r_t + \gamma \underbrace{\sum_{k=t+1}^{\infty} \gamma^{k-t} r_k}_{G_{t+1}}$$

$$\mathbb{P}(s', r \mid s, a)$$

$$q_*(s,a) = \cancel{\max_{\pi}}\, \mathbb{E}[R_t \mid S_t = s, A_t = a] + \max_{\pi} \mathbb{E}[\gamma G_{t+1} \mid \quad]$$

$$\sum_{r} r\, \mathbb{P}(r \mid s, a) = \sum_{r} r \sum_{s'} \mathbb{P}(s', r \mid s, a)$$

$$\underbrace{\gamma \max_{\pi}}_{\text{Ignore for now}} \mathbb{E}\left[G_{t+1} \mid A_t = a, S_t = s\right]$$

$$\left\{ \begin{array}{c} \text{Tower property} \\ \mathbb{E}_X\left[\mathbb{E}[Y|X]\right] = \mathbb{E}[Y] \\ s', a' \triangleq X \end{array} \right.$$

$$\mathbb{E}_{s',a'}\left[\underbrace{\mathbb{E}\left[G_{t+1} \mid A_{t+1} = a', S_{t+1} = s', A_t = a, S_t = s\right]}\right]$$

$$= \sum_{s',a'} p(s', a' | s, a)\, \mathbb{E}\left[G_{t+1} \cancel{\quad} \cancel{\quad}\right]$$

$$= \sum_{s',a'} p(s', a' | s, a) \overbrace{\mathbb{E}\left[G_{t+1} \mid S_{t+1} = s', A_{t+1} = a'\right]}^{q_\pi(s', a')}$$

$$P(s' \mid s,a)\, p(a' \mid s', s, a) = P(s' \mid s,a)\, \pi(a' \mid s')$$

$$\mathbb{E}[G_{t+1} \mid S_t = s, A_t = a] = \sum_{s',a'} P(s' \mid s,a)\, \pi(a' \mid s')\, q_\pi(s',a')$$

$$q_*(s,a) = \sum_r r \sum_{s'} \mathbb{P}(s', r \mid s,a) + \gamma \max_\pi \sum_{s',a'} P(s' \mid s,a)\, \overbrace{\pi(a' \mid s')\, q_\pi(s',a')}$$

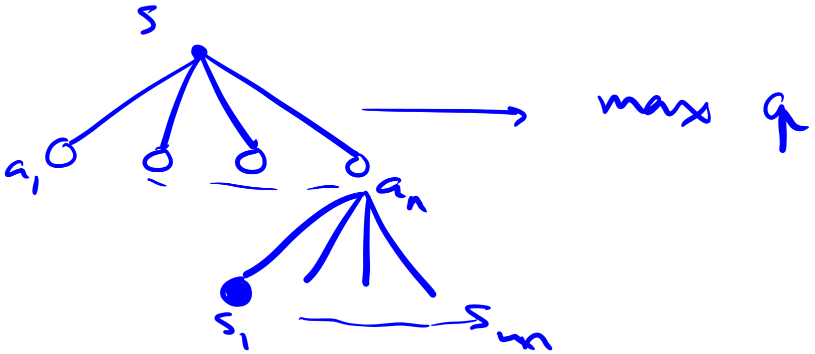
$$\hookrightarrow \gamma \max_\pi \sum_{s',a'} \mathbb{P}(s' \mid s,a) \max_{a'} q_\pi(s',a')$$

$$\sum \pi(a' \mid s') = 1$$

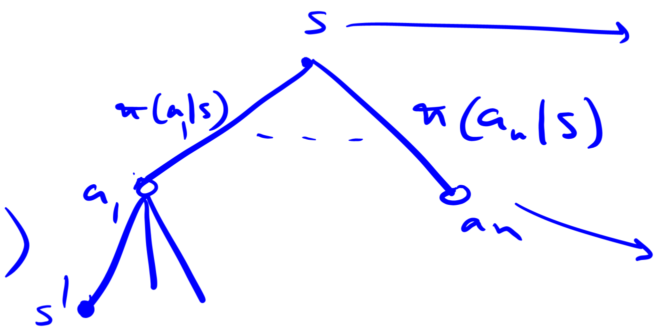
$$q_*(s,a) = \sum_{s',r} \mathbb{P}(s',r \mid s,a)\left[r + \gamma \max_{a'} q_*(s',a')\right]$$

Suppose π^* maximizes the right term.

$\hat{\pi}^*$: act greedy on first step then act as π^*.



π



$$V(s) = \sum_a \pi(a|s)\, q_\pi(s,a)$$

$q_\pi(s,a_1)$

$q_\pi(s,a)$

$$\max_{a_i} q_\pi(s,a_i)$$

If we choose the first action greedily the $V(s)$ increases.

$$\pi' \triangleq \arg\max_a q_\pi(s,a) \longrightarrow q_{\pi'} \geq q_\pi$$

$$\pi^* \longrightarrow \arg\max_a q_{\pi^*}(s,a)$$

$$q_*(s,a) = \sum_r r \sum_{s'} \mathbb{P}(s',r|s,a) + \gamma \max_\pi \sum_{s',a'} p(s'|s,a)\, \pi(a'|s')\, q_\pi(s',a')$$

$$\gamma \max_\pi \sum_{s'} \mathbb{P}(s'|s,a) \sum_{a'} \pi(a'|s')\, q_\pi(s',a')$$

$$\gamma \max_\pi \sum_{s'} \mathbb{P}(s'|s,a) \max_{a'} q_\pi(s',a')$$

$$\gamma \sum_{s'} \mathbb{P}(s'|s,a) \cancel{\max_{\pi,a'} q_\pi(s',a')}$$

$$\max_{a'} q_*(s',a')$$

$$\sum_{s',r} \mathbb{P}(s',r|s,a)$$

* π^* satisfies Bellman Opt. Eq.

$$f(\underbrace{q(s_1,a_1), \cdots\cdots, q(s_n,a_m)}_{Q}) = (\underbrace{q'(s_1,a_1), \cdots\cdots, q'(s_n,a_m)}_{Q'})$$

Value iteration.

$$f(Q_0) = Q_1$$
$$f(Q_1) = Q_2$$
$$\vdots$$
$f(Q_k) \approx$ converges to opt. Q

$$f(x) = x$$

$\Rightarrow$ T Bellman $\overset{\text{Opt.}}{\text{O}}$perator has a unique fixed point.

$T q_1 = q_2$

(Banach Fixed point theorem)

1. X is a non-empty complete metric space.
2. $T : X \to X$ is a contraction mapping on X.

$\Rightarrow$ T has a unique fixed point.

Cauchy Series: $\{x_i\}_{i=1}^{\infty}$ $\forall \varepsilon$ $\exists M$ s.t. for all $n, m \geq M$

$d(x_n, x_m) \leq \varepsilon$

distance metrice

Contraction Mapping. $T : X \to X$ is a contraction map. on X if $\alpha \in \mathbb{R}^+$, $\alpha < 1$ s.t. $\forall x, y \in X$

$$d(Tx, Ty) \leqslant \alpha d(x, y)$$

x_0 , $Tx_n = x_{n+1}$

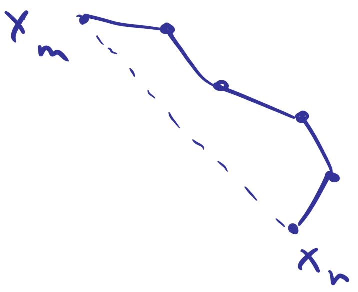
$$d(x_{k+1}, x_k) \underset{*}{\leqslant} \alpha d(x_k, x_{k-1})$$

$$d(x_m, x_n) \leqslant \sum_{i=0}^{n-m-1} d(x_{m+i}, x_{m+i+1})$$

$$\overset{*}{\leqslant} \alpha\, d(x_{m+i-1}, x_{m+i}) \leqslant \alpha^2 d(x_{m+i-2}, x_{m+i-1})$$

$$\vdots$$

$$\leqslant \alpha^{m+i} d(x_0, x_1)$$



$$d(x_m, x_n) \leq \overbrace{d(x_0, x_1)}^{C} \sum_{i=0}^{n-m-1} \alpha^{m+i} = C\alpha^m \frac{1-\alpha^{n-m}}{1-\alpha}$$

$$\leq C\alpha^m \cdot \frac{1}{1-\alpha} \leq \varepsilon$$

$n \geq m$

X_0

$$Tx = x$$

$$0 = d(Tx, x)$$

$$0 = d(Tx', x')$$

$$= d(\underbrace{Tx}_{x}, \underbrace{Tx'}_{x'}) \leq \alpha\, d(x, x')$$

$d \neq 0$

$\uparrow$

$x \neq x'$

$\Rightarrow \quad 1 \leq \alpha \quad \because$

Bellman Opt. Operator:

$$\hookrightarrow Tq(s,a) = \sum_{r,s'} \mathbb{P}(r,s' \mid s,a) \left[r + \gamma \max_{a'} q(s',a') \right]$$

$$T : \mathbb{R}^{|S| \times |A|} \to \mathbb{R}^{|S| \times |A|}$$

* T is a contraction mapping for a finite MDP.

$(\mathbb{R}^{|S| \times |A|}, d)$ is a complete metric space.

$d \downarrow$ L_p-norm

$$d \equiv L_\infty$$

$$\forall q_1, q_2 \in \mathbb{R}^{|S|\times|A|} : \text{Obj.: } \|Tq_1 - Tq_2\|_\infty \leq \alpha \underbrace{\|q_1 - q_2\|_\infty}$$

$\max|x-y|$

$\geq \max|x|$

$-\max|y|$

$$\|Tq_1 - Tq_2\|_\infty = \max_{s,a} |Tq_1(s,a) - Tq_2(s,a)|$$

$$= \gamma \max_{s,a} \sum_{r,s'} \mathbb{P}(s',r|s,a) \left| \max_{a'} q_1(s',a') - \max_{a'} q_2(s',a') \right|$$

$$\leq \gamma \max_{s,a} \underbrace{\sum_{r,s'} \mathbb{P}(s',r|s,a)}_{\sum_{s'} \mathbb{P}(s'|s,a)} \max_{a'} \left| q_1(s',a') - q_2(s',a') \right|$$

$$\leq \gamma \max_{s,a} \max_{s',a'} |q_1(s',a') - q_2(s',a')|$$

$$= \gamma \, \| q_1 - q_2 \|_\infty$$

$\Rightarrow$ Bellman Opt. Op. is a cont. map.

Theorem. Value Iteration converges to opt. policy in finite time in a finite MDP.

$|S|$ $\quad |A|$

$$q_{\pi_{k+1}} \geq q_{\pi_k}$$

If at some point $\pi_k = \pi_{k+1} \Rightarrow \pi_k = \pi^*$

$q_{\pi_{k+1}}$ can be strictly larger than q_{π_k} at most $|A|^{|S|}$ times.

the End.