Planning in MDP

未知 Model-based

Recall: Value function $V^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid s_1 = s\right]$

Bellman equation: $V^\pi(s) = R(s,\pi) + \mathbb{E}_{s'\sim p(\cdot|s,\pi)}\left[V^\pi(s')\right]$

linear

$\hookrightarrow V^\pi = (I - \gamma \cdot P^\pi)^{-1} R^\pi$

Optimality ↓

$V^* = V^{\pi^*}$: $V^*(s) = \max_{a\in A}\Big(\underbrace{R(s,a) + \gamma \cdot \mathbb{E}_{s'\sim p(\cdot|s,a)}\left[V^*(s')\right]}_{Q^*(s,a)}\Big)$

Bellman Optimality Equation:

$$Q^*(s,a) = R(s,a) + \gamma \cdot \mathbb{E}_{s'\sim p(\cdot|s,a)}\left[\max_{a'} Q^*(s',a')\right]$$

non-linear

How to solve this non-linear system?

☆ def (Bellman Operator) $\mathcal{T}: \mathbb{R}^{S\times A} \to \mathbb{R}^{S\times A}$.

Optimality

Model / Environment

for any $f \in \mathbb{R}^{S\times A}$:

$$(\mathcal{T}f)(s,a) = R(s,a) + \gamma \cdot \mathbb{E}_{s'\sim p(\cdot|s,a)}\left[\max_{a'} f(s',a')\right]$$

$\forall s, a: \quad Q^*(s,a) = (\mathcal{T} Q^*)(s,a)$

fixed point of the Bellman Operator !

Optimal policy: $\pi^* = \pi_{Q^*}$

(π_f : greedy policy wrt. $f \in \mathbb{R}^{S \times A}$

$\pi_f(s) = \arg\max_{a \in A} f(s,a)$)

Planning: find an algorithm to compute π^*

for a given MDP $M = \langle S, A, \underline{P}, \underline{R}, \gamma \rangle$

Value Iteration 值迭代

LSVI-UCB.

initialize $f_0 = \vec{0}$

for $k = 1, 2, \cdots$: $\boxed{f_k \leftarrow \mathcal{T} f_{k-1}}$

Convergence: sequence $(f_k)_{k=1}^{\infty}$ converges to Q^*.

with $\| f_k - Q^* \|_\infty \le \gamma^k \cdot \frac{R_{max}}{1-\gamma}$ V_{max}

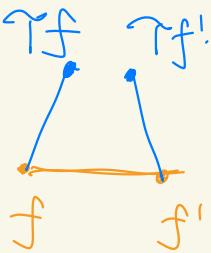
exponential convergence.

$$V = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t\right] \le R_{max} \cdot \sum_{t=1}^{\infty} \gamma^{t-1} = R_{max} / (1-\gamma)$$

Lemma $\mathcal{T}$ is γ-contraction under $\|\cdot\|_\infty$

i.e. $\forall f, f' \in \mathbb{R}^{S \times A}$:

$$\|\mathcal{T}f - \mathcal{T}f'\|_\infty \le \gamma \cdot \|f - f'\|_\infty$$



$$\|f\|_\infty = \max_{s,a} |f(s,a)|$$

$$\|f_k - Q^*\|_\infty$$
$$= \|\mathcal{T}f_{k-1} - \mathcal{T}Q^*\|_\infty$$
$$\le \gamma \cdot \|f_{k-1} - Q^*\|_\infty$$
$$\le \gamma^2 \cdot \|f_{k-2} - Q^*\|_\infty$$
$$\le \cdots \le \gamma^k \cdot \|f_0 - Q^*\|_\infty = \gamma^k \cdot \|Q^*\|_\infty = \gamma^k \cdot \frac{R_{max}}{1-\gamma}$$

① $Q^* = \mathcal{T}Q^*$

② γ-contraction

Proof: $(\mathcal{T}f)(s,a) = R(s,a) + \gamma \cdot \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\max_{a'} f(s',a')\right]$

$\forall (s,a)$: $(\mathcal{T}f)(s,a) - (\mathcal{T}f')(s,a)$

$$= \cancel{R(s,a)} + \gamma \cdot \mathbb{E}_{s'}\left[\max_{a'} f(s',a')\right]$$

$$- \cancel{R(s,a)} - \gamma \cdot \mathbb{E}_{s'}\left[\max_{a'} f'(s',a')\right]$$

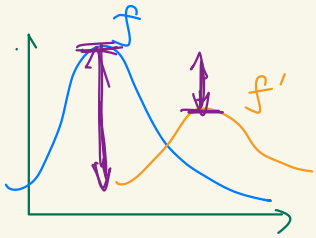
$$= \gamma \cdot \mathbb{E}_{s'}\left[\max_{a'} f(s',a') - \max_{a'} f'(s',a')\right]$$

$\mathbb{E}[X] \le \max X$

$$\le \gamma \cdot \max_{s'}\left(\max_{a'} f(s',a') - \max_{a'} f'(s',a')\right)$$

$$\le \gamma \cdot \max_{s'} \max_{a'} \left| f(s',a') - f'(s',a') \right|$$

$$= \gamma \cdot \|f - f'\|_\infty \qquad \square$$



Wlog $\max_{a'} f(s',a') > \max_{a'} f'(s',a')$.

$$\Rightarrow \max_{a'} f(s',a') - \max_{a'} f'(s',a')$$

$$\triangleq f(s',a_0) - \max_{a'} f'(s',a')$$

$$\le f(s',a_0) - f'(s',a_0) \le \max_{a'}\left(f(s',a') - f'(s',a')\right)$$

$$\Rightarrow \|f_k - Q^*\|_\infty \le \gamma^k \frac{R_{max}}{1-\gamma}$$

Approximate Q^*: f_k is ε-close to Q^*.

$$\gamma^k \cdot \frac{R_{max}}{1-\gamma} = \Big(\underbrace{1-(1-\gamma)}_{\to 0}\Big)^k \frac{R_{max}}{1-\gamma}$$

$\gamma \to 1$

$$= \Big(1-(1-\gamma)\Big)^{k(1-\gamma)\cdot\frac{1}{1-\gamma}} \cdot \frac{R_{max}}{1-\gamma}$$

$$e = \lim_{x\to 0}(1+x)^{\frac{1}{x}}$$

$$= \left(\frac{1}{e}\right)^{k\cdot(1-\gamma)} \cdot \frac{R_{max}}{1-\gamma} \le \varepsilon.$$

$$\Rightarrow e^{k(1-\gamma)} \geq \frac{R_{max}}{\varepsilon(1-\gamma)}.$$

$$\Rightarrow \boxed{k \geq \frac{1}{1-\gamma} \cdot \log \frac{R_{max}}{\varepsilon(1-\gamma)} = O\left(\frac{1}{1-\gamma}\right)}$$

effective horizon.

$$H \triangleq O\left(\frac{1}{1-\gamma}\right)$$

finite-horizon undiscounted

infinite-horizon discounted.

Finite-Horizon Interpretation of Value Iteration

Simplification: VI for V^*

initialize $f_0(s) = 0$ for all $s \in S$. $f \in \mathbb{R}^S$

for $k = 1, 2, \cdots$: $f_k(s) \leftarrow (\mathcal{T} f_{k-1})(s)$

Bellman Operator for state-value function:

$$V^* = \mathcal{T} V^*$$

$$(\mathcal{T} f)(s) = \max_a \left(R(s,a) + \gamma \cdot \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[f(s') \right] \right)$$

$f_0(s) = 0$

$$f_1(s) = \max_a \left(R(s,a) + \gamma \cdot 0 \right) = \max_a R(s,a)$$ 1-step reward

$$f_2(s) = \max_a \left(R(s,a) + \gamma \cdot \mathbb{E}_{s'} \left[f_1(s') \right] \right) = \max_a \left(R(s,a) + \gamma \cdot \mathbb{E}_{s'} \left[\max_{a'} R(s',a') \right] \right)$$

2-step reward

$\vdots$

Claim: f_k is optimal value function for cumulative

$$V^{\pi,k}(s) = \mathbb{E}_\pi \left[\sum_{t=1}^{k} \gamma^{t-1} r_t \,\middle|\, s_1 = s \right].$$

truncated version

$$f_k(s) = \max_\pi V^{\pi,k}(s)$$

(possibly) non-stationary

$$\boxed{f_k = V^{*,k}}$$

$P(\cdot \mid s, a, \underline{h})$

$\pi = \{\pi_h\}_{h=1}^{H}$ step

optimal for k-step truncated-V.

$$V^{*,k}(s) \geq V^{\pi^*,k}(s)$$

optimal for infinite-horizon!

$$\text{LHS} = V^{*,k}(s) \leq V^*(s)$$

$\exists$ policy $\begin{pmatrix} \text{stationary} \\ \text{non-stationary} \end{pmatrix}$

$$\text{RHS} = V^{\pi^*,k}(s)$$

$$= \mathbb{E}_{\pi^*}\left[\sum_{t=1}^{k} \gamma^{t-1} r_t \mid s_1 = s\right]$$

optimal at first k-steps

non-negative rewards afterward

$$\geq \mathbb{E}_{\pi^*}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid s_1 = s\right] - R_{max} \cdot \frac{\gamma^k}{1-\gamma}$$

$$= V^*(s) - R_{max} \cdot \frac{\gamma^k}{1-\gamma}$$

$\gamma^k + \gamma^{k+1} + \cdots$

$$V^*(s) \geq f_k(s) \geq V^*(s) - R_{max} \cdot \frac{\gamma^k}{1-\gamma}.$$

$$\Rightarrow \quad \| f_k - V^* \|_\infty \leq \gamma^k \cdot \frac{R_{max}}{1-\gamma}.$$

$$\| f_k - Q^* \|_\infty \leq \gamma^k \frac{R_{max}}{1-\gamma}.$$

Recall: Bellman Operator $\mathcal{T}$

$$f \in \mathbb{R}^{S \times A}: \quad (\mathcal{T} f)(s,a) = R(s,a) + \gamma \cdot \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a'} f(s',a') \right]$$

$\Rightarrow$ Value Iteration: $f_k \leftarrow \mathcal{T} f_{k-1}$

① $Q^* = \mathcal{T} Q^*$

② γ-contraction

$$\| \mathcal{T} f - \mathcal{T} f' \|_\infty \leq \gamma \cdot \| f - f' \|_\infty$$

$\downarrow$

guarantee $\| f_k - Q^* \|_\infty \leq \gamma^k \cdot \frac{R_{max}}{1-\gamma}.$

State-action function.

$\Leftrightarrow$ "Sample Complexity": $k \geq O\left(\frac{1}{1-\gamma} \log \frac{1}{\varepsilon}\right)$

effective horizon.

$\Rightarrow \| f_k - Q^* \|_\infty \leq \varepsilon.$

Greedy Policy: $\forall f \in \mathbb{R}^{S \times A}$, $\pi_f(s) = \arg\max_{a \in A} f(s,a).$

State function

Lemma $\| V^{\pi_f} - V^* \|_\infty \leq \frac{2}{1-\gamma} \cdot \| f - Q^* \|_\infty$

$$\mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \right]$$

Proof: $V^{\pi_f}(s) - V^*(s) \qquad \forall s \in S$

$$= Q^{\pi_f}(s, \pi_f(s)) - Q^*(s, \pi^*(s))$$

Add-and-Subtract.

$$= Q^{\pi_f}(s, \pi_f(s)) - Q^*(s, \pi_f(s)) \rightarrow (\mathrm{I})$$

$$+ Q^*(s, \pi_f(s)) - Q^*(s, \pi^*(s)) \rightarrow (\mathrm{II})$$

$$(\mathrm{I}) = R(s, \pi_f(s)) + \gamma \cdot \mathbb{E}_{s'}[V^{\pi_f}(s')]$$

$$- R(s, \pi_f(s)) - \gamma \cdot \mathbb{E}_{s'}[V^*(s')]$$

$$= \gamma \cdot \mathbb{E}_{s' \sim P(\cdot|s, \pi_f(s))}[V^{\pi_f}(s') - V^*(s')] \leq \gamma \cdot \|V^{\pi_f} - V^*\|_\infty$$

$$(\mathrm{II}) = Q^*(s, \pi_f(s)) - f(s, \pi^*(s)) + f(s, \pi^*(s)) - Q^*(s, \pi^*(s))$$

greedy wrt f.

$$\leq \|f - Q^*\|_\infty$$

$$\leq Q^*(s, \pi_f(s)) - f(s, \pi_f(s))$$

$$\leq \|f - Q^*\|_\infty$$

$f \cdot Q^{\pi_f}$

$$(\mathrm{II}) \leq 2 \cdot \|f - Q^*\|_\infty$$

$$\Rightarrow \|V^{\pi_f} - V^*\|_\infty \leq \gamma \cdot \|V^{\pi_f} - V^*\|_\infty + 2 \cdot \|f - Q^*\|_\infty \quad \square$$

Suboptimality gap $\|V^{\pi_f} - V^*\|_\infty \le \frac{2\gamma^H \cdot R_{max}}{(1-\gamma)^2}$

Note: non-stationary. (tighter bound)

$\pi_{Q^*,H}, \pi_{Q^*,H-1}, \cdots, \pi_{Q^*,1}, \pi_{arbitrary},$

$\Rightarrow \gamma^H \cdot \frac{R_{max}}{1-\gamma}$ (finite-horizon interpretation)

Value Iteration: 值迭代

policy optimization

$\boxed{Q^*: f_k \leftarrow \mathcal{T} f_{k-1}}$ $(\mathcal{T}f)(s,a) = R(s,a) + \gamma \cdot \mathbb{E}_{s'}[\max_{a'} f(s',a')]$

$V^*: f_k \leftarrow \mathcal{T} f_{k-1}$ $(\mathcal{T}f)(s) = \max_a \left(R(s,a) + \gamma \cdot \mathbb{E}_{s'}[f(s')]\right)$

policy evaluation

$Q^\pi: f_k \leftarrow \mathcal{T}^\pi f_{k-1}$ $(\mathcal{T}^\pi f)(s,a) = R(s,a) + \gamma \cdot \mathbb{E}_{s' \sim P(\cdot|s,a)}[f(s',\pi)]$

$V^\pi: f_k \leftarrow \mathcal{T}^\pi f_{k-1}$ $(\mathcal{T}^\pi f)(s) = R(s,\pi) + \gamma \cdot \mathbb{E}_{s'}[f(s')]$

Bellman Operator $\rightarrow$
- fixed point
- γ-contraction
- monotonicity.

Policy Iteration 策略迭代

Solving linear system

iterative method!

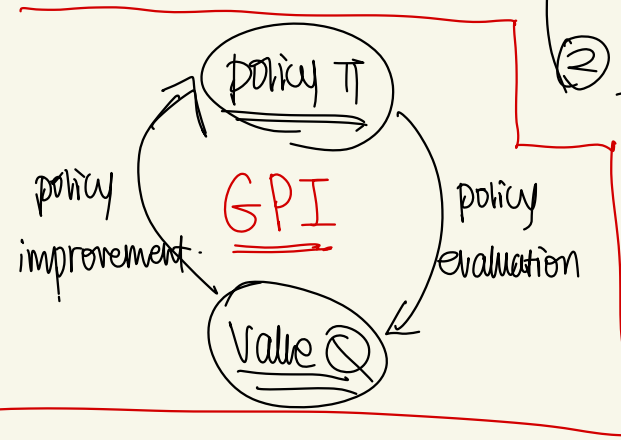
Algorithm: initialize π_0 any policy

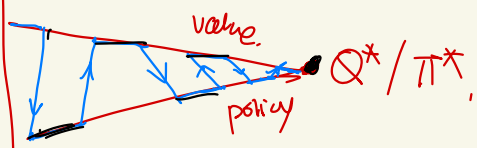
Bellman operator

for $k = 1, 2, \cdots$ { ① policy evaluation: find $Q^{\pi_{k-1}}$

② policy improvement: $\pi_k \leftarrow \pi_{Q^{\pi_{k-1}}}$

i.e. $\pi_k(s) = \arg\max_{a \in A} Q^{\pi_{k-1}}(s,a)$





1) **Sanity check**: if $\pi_k = \pi^*$ then $\pi_{k+1} = \pi^*$.

$\left(Q^{\pi_k} = Q^*, \quad \pi_{k+1} = \pi_{Q^*} = \pi^* \right)$

2) **Convergence** (monotone improvement)

Theorem (Policy Improvement Theorem)

① $\forall k$: $V^{\pi_{k+1}} \geq V^{\pi_k}$

(i.e. $V^{\pi_{k+1}}(s) \geq V^{\pi_k}(s) \ \forall s$)

② Unless $\pi_k = \pi^*$, then $\exists\, s \in S$:

such that $V^{\pi_{k+1}}(s) > V^{\pi_k}(s)$

PI → exact optimum: $|A|^{|S|}$

total # of deterministic policy

If $k \geq |A|^{|S|}$

then $\pi_k = \pi^*$

PD Lemma

Lemma (Performance Difference Lemma)

$\forall \pi, \pi'$, $\forall s \in S$:

$A^{\pi}(\tilde{s}, \pi')$: advantage of π' over π.

$$V^{\pi'}(s) - V^{\pi}(s) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{\tilde{s} \sim d^{\pi', s}}\left[Q^{\pi}(\tilde{s}, \pi'(\tilde{s})) - V^{\pi}(\tilde{s}) \right]$$

第一步走 π'

剩下的走 π.

(normalized) discounted occupancy

(for state value)

$$d^{\pi', s}(\cdot) = (1-\gamma) \cdot \sum_{t=1}^{\infty} \gamma^{t-1} d_t^{\pi', s}(\cdot)$$

$$d_t^{\pi', s}(\cdot) = \Pr[S_t = \cdot \mid S_1 = s]$$

Proof of $V^{\pi_{k+1}} \geq V^{\pi_k}$.

$\pi' = \pi_{k+1}$

$\pi = \pi_k$

By PD Lemma: $\forall s$

$$V^{\pi_{k+1}}(s) - V^{\pi_k}(s) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{\tilde{s}}\left[Q^{\pi_k}(\tilde{s}, \pi_{k+1}(\tilde{s})) - V^{\pi_k}(\tilde{s}) \right].$$

Greedy: $\pi_{k+1} = \pi_{Q^{\pi_k}} \rightarrow \pi_{k+1}(\tilde{s}) = \arg\max_a Q^{\pi_k}(\tilde{s}, a)$

$$\Rightarrow Q^{\pi_k}(\tilde{s}, \pi_{k+1}(\tilde{s})) = \max_a Q^{\pi_k}(\tilde{s}, a)$$

$$V^{\pi_k}(\tilde{s}) = \mathbb{E}_{a \sim \pi_k(\tilde{s})}\left[Q^{\pi_k}(\tilde{s}, a) \right].$$

$$A^{\pi_k}(s, \pi_{k+1}) \geq 0$$: non-negative advantage

$$\Rightarrow V^{\pi_{k+1}} \geq V^{\pi_k}. \quad \square.$$

If $\forall s \in S$: $V^{\pi_{k+1}}(s) = V^{\pi_k}(s)$.

$$A^{\pi_k}(s, \pi_{k+1}) = 0 \quad \forall s \in S$$

$\downarrow$

Bellman Optimality Equation

Recall: Policy Iteration
$$\begin{cases} ① \text{ evaluate } Q^{\pi_k} \\ ② \ \pi_{k+1} \leftarrow \pi_{Q^{\pi_k}} \end{cases}$$

Thm (Policy Improvement Theorem) $V^{\pi_{k+1}}(s) \geq V^{\pi_k}(s), \ \forall s.$

If $\pi_k \neq \pi^*$: $\exists\, s \in S$ s.t. $V^{\pi_{k+1}}(s) > V^{\pi_k}(s)$

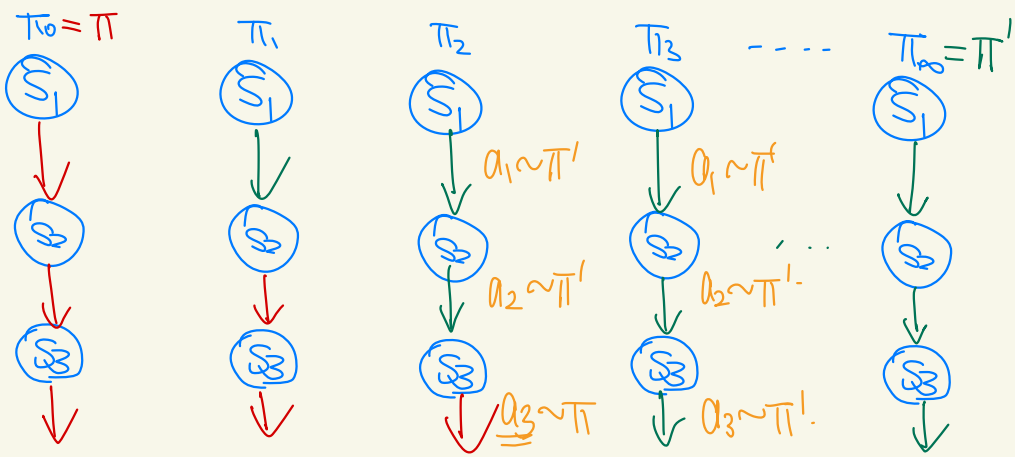
Performance-Difference lemma: $\forall \pi, \pi', \ \forall s \in S$

$$V^{\pi'}(s) - V^{\pi}(s) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{\tilde{s} \sim d^{\pi', s}(\cdot)} \left[Q^{\pi}(\tilde{s}, \pi'(\tilde{s})) - V^{\pi}(\tilde{s}) \right]$$

Advantage

one-step difference

Proof: Construct a sequence of (potentially non-stationary) policies $\{\pi_i\}_{i=0}^{\infty}$



π_i: "i steps of π', then π"

$$V^{\pi'}(s) - V^{\pi}(s) = V^{\pi_\infty}(s) - V^{\pi_0}(s)$$

$$= \sum_{i=0}^{\infty} \left(V^{\pi_{i+1}}(s) - V^{\pi_i}(s) \right)$$

One-step difference

e.g. $i=2$

MDP

Markovian

$$V^{\pi_3}(s) = \mathbb{E}\left[r_1 + \gamma \cdot r_2 \mid S_1 = s,\ a_{1:2} \sim \pi' \right]$$

$$+ \underbrace{\mathbb{E}\left[\sum_{t=3}^{\infty} \gamma^{t-1} r_t \mid S_1 = s,\ a_{1:2} \sim \pi',\ a_3 \sim \pi',\ a_{4:\infty} \sim \pi \right]}_{(\mathrm{I})}$$

$$V^{\pi_2}(s) = \mathbb{E}\left[r_1 + \gamma \cdot r_2 \mid S_1 = s,\ a_{1:2} \sim \pi' \right]$$

$$+ \underbrace{\mathbb{E}\left[\sum_{t=3}^{\infty} \gamma^{t-1} r_t \mid S_1 = s,\ a_{1:2} \sim \pi',\ a_3 \sim \pi,\ a_{4:\infty} \sim \pi \right]}_{(\mathrm{I})}$$

$$(\mathrm{I}) = \mathbb{E}\left[\sum_{t=3}^{\infty} \gamma^{t-1} r_t \mid S_1 = s,\ a_{1:2} \sim \pi',\ a_3 \sim \pi',\ a_{4:\infty} \sim \pi \right]$$

law of total expectation

$$= \mathbb{E}_{\tilde{s}}\left[\mathbb{E}\left[\sum_{t=3}^{\infty} \gamma^{t-1} r_t \mid \cancel{S_1 = s,\ a_{1:2} \sim \pi'},\ \boxed{S_3 = \tilde{s}},\ a_3 \sim \pi',\ a_{4:\infty} \sim \pi \right] \right]$$

by Markovian.

$$\tilde{s} \sim \Pr\left[S_3 = \cdot \mid S_1 = s \right]$$

$\mathbb{L}$ $d_3^{\pi',s}(\cdot)$: occupancy at time step 3.

$$= \mathbb{E}_{\tilde{s}}\left[\mathbb{E}\left[\sum_{t=3}^{\infty}\gamma^{t-1}r_t \,\middle|\, s_3=\tilde{s},\ a_3\sim\pi',\ a_{4:\infty}\sim\pi\right]\right]$$

$$= \gamma^2\cdot\mathbb{E}_{\tilde{s}}\left[\mathbb{E}\left[\sum_{t=1}^{\infty}\gamma^{t-1}r_t \,\middle|\, s_1=\tilde{s},\ a_1\sim\pi',\ a_{2:\infty}\sim\pi\right]\right]$$

shifting

$$= \gamma^2\cdot\mathbb{E}_{\tilde{s}}\left[Q^{\pi}(\tilde{s},\pi'(\tilde{s}))\right]$$

$$(\mathrm{II}) = \gamma^2\cdot\mathbb{E}_{\tilde{s}}\left[V^{\pi}(\tilde{s})\right].$$

$$\Rightarrow V^{\pi_3}(s)-V^{\pi_2}(s) = \gamma^2\cdot\mathbb{E}_{\tilde{s}\sim d_3^{\pi',s}(\cdot)}\left[Q^{\pi}(\tilde{s},\pi'(\tilde{s}))-V^{\pi}(\tilde{s})\right]$$

$$\Rightarrow V^{\pi_{i+1}}(s)-V^{\pi_i}(s) = \gamma^i\cdot\mathbb{E}_{\tilde{s}\sim d_{i+1}^{\pi',s}(\cdot)}\left[Q^{\pi}(\tilde{s},\pi'(\tilde{s}))-V^{\pi}(\tilde{s})\right]$$

$$V^{\pi'}(s)-V^{\pi}(s) = \sum_{i=0}^{\infty}\left(V^{\pi_{i+1}}(s)-V^{\pi_i}(s)\right)$$

$$= \sum_{i=0}^{\infty}\gamma^i\cdot\mathbb{E}_{\tilde{s}\sim d_{i+1}^{\pi',s}(\cdot)}\left[Q^{\pi}(\tilde{s},\pi'(\tilde{s}))-V^{\pi}(\tilde{s})\right]$$

$$\frac{1}{1-\gamma}\cdot d^{\pi',s}(\cdot) = \sum_{i=1}^{\infty}\gamma^{i-1}d_i^{\pi',s} = \sum_{i=0}^{\infty}\gamma^i\cdot d_{i+1}^{\pi',s}$$

normalized discounted occupancy

$$= \frac{1}{1-\gamma} \cdot \mathbb{E}_{\tilde{s} \sim d^{\pi', s}(\cdot)} \left[Q^{\pi}(\tilde{s}, \pi'(\tilde{s})) - V^{\pi}(\tilde{s}) \right]$$

$\square$

PD Lemma $\Rightarrow$ Policy Improvement Theorem.

Policy Iteration : exact guarantee, exp. time $|A|^{|S|}$

Value Iteration : approx. guarantee ; poly. time.

$$\|f_k - Q^*\|_\infty \le \gamma \cdot \|f_{k-1} - Q^*\|_\infty$$

$\downarrow$

Actually, Policy Iteration can have approx guarantee within polynomial time.

Theorem $\|Q^* - Q^{\pi_{k+1}}\|_\infty \le \gamma \cdot \|Q^* - Q^{\pi_k}\|_\infty$

Fact 1 $\mathcal{T}^{\pi_{k+1}} Q^{\pi_k} \ge \mathcal{T}^{\pi} Q^{\pi_k}$, $\forall \pi$. (pointwise)

greedy

Proof : $(\mathcal{T}^{\pi} Q^{\pi_k})(s,a) = R(s,a) + \gamma \cdot \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[Q^{\pi_k}(s', \pi) \right]$

$\pi_{k+1} = \pi_{Q^{\pi_k}} \Rightarrow Q^{\pi_k}(s, \pi_{k+1}) \ge Q^{\pi_k}(s, \pi)$, $\forall \pi$.

$\Rightarrow (\mathcal{T}^{\pi_{k+1}} Q^{\pi_k})(s,a) \ge (\mathcal{T}^{\pi} Q^{\pi_k})(s,a)$ $\square$

$\boxed{\text{Fact 2}}$ $\mathcal{T}^{\pi_{k+1}} Q^{\pi_k} \le Q^{\pi_{k+1}}$ (pointwise)

$\Longleftrightarrow$

Proof: $\mathcal{T}^{\pi_{k+1}} Q^{\pi_k} \le \mathcal{T}^{\pi_{k+1}} Q^{\pi_{k+1}}$

$$(\mathcal{T}^{\pi_{k+1}} f)(s,a) = R(s,a) + \gamma \cdot \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[f(s', \pi_{k+1}) \right]$$

(Q^{π} is the fixed point of $\mathcal{T}^{\pi}$)

It suffices to show: $Q^{\pi_k}(s, \pi_{k+1}) \le Q^{\pi_{k+1}}(s, \pi_{k+1})$

first step: π_{k+1}.

afterwards: different.

$$Q^{\pi_k}(s, \pi_{k+1}) = R(s, \pi_{k+1}) + \gamma \cdot \mathbb{E}_{s' \sim P(\cdot|s,\pi_{k+1})} \left[V^{\pi_k}(s) \right]$$

$$Q^{\pi_{k+1}}(s, \pi_{k+1}) = R(s, \pi_{k+1}) + \gamma \cdot \mathbb{E}_{s'} \left[V^{\pi_{k+1}}(s) \right]$$

If suffices to show: $V^{\pi_k}(s) \le V^{\pi_{k+1}}(s)$

(Policy Improvement Theorem) $\square$

1. $\mathcal{T}^{\pi_{k+1}} Q^{\pi_k} \ge \mathcal{T}^{\pi} Q^{\pi_k}$ $\forall \pi$.
2. $\mathcal{T}^{\pi_{k+1}} Q^{\pi_k} \le Q^{\pi_{k+1}}$

$\Longrightarrow$ $\mathcal{T}^{\pi} Q^{\pi_k} \le Q^{\pi_{k+1}}$.

Proof of the Theorem: $\|Q^* - Q^{\pi_{k+1}}\|_\infty \le \gamma \cdot \|Q^* - Q^{\pi_k}\|_\infty$

$$Q^* - Q^{\pi_{k+1}} \le Q^* - \mathcal{T}^{\pi} Q^{\pi_k}, \quad \forall \pi$$

$$= \mathcal{T} Q^* - \mathcal{T} Q^{\pi_k}$$

choose π as π^*.

$$\Rightarrow \|Q^* - Q^{\pi_{k+1}}\|_\infty \le \|\mathcal{T} Q^* - \mathcal{T} Q^{\pi_k}\|_\infty$$

$$\le \gamma \cdot \|Q^* - Q^{\pi_k}\|_\infty \quad \square$$

→ exponential convergence

Alternative Proof of the Policy Improvement Theorem:

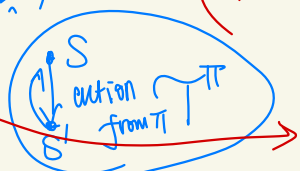
$$Q^{\pi_k} = \mathcal{T}^{\pi_k} Q^{\pi_k} \le \mathcal{T}^{\pi_{k+1}} Q^{\pi_k}.$$

(by Fact 1)

Apply $\mathcal{T}^{\pi_{k+1}}$ to Q^{π_k}:

$$Q^{\pi_k} \le \boxed{\mathcal{T}^{\pi_{k+1}}} Q^{\pi_k} \le \left(\mathcal{T}^{\pi_{k+1}}\right)^2 Q^{\pi_k} \le \cdots \le \left(\mathcal{T}^{\pi_{k+1}}\right)^\infty Q^{\pi_k}$$

Bellman Backup



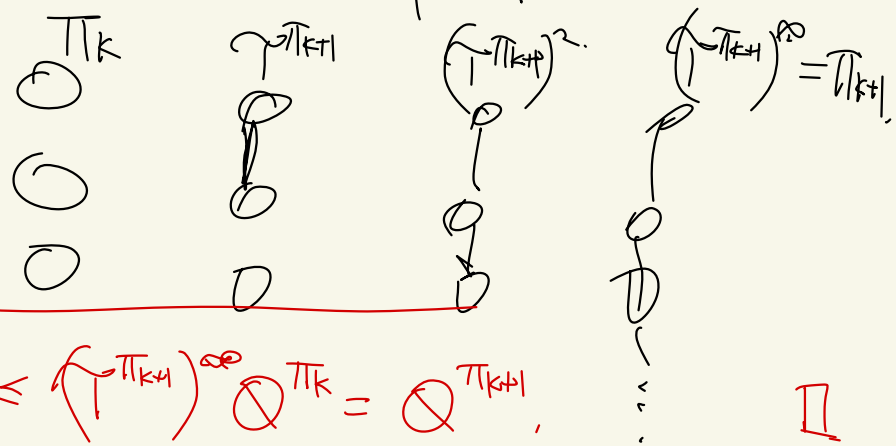
→ one-step difference.

Interp 1: value iteration wrt. $\mathcal{T}^{\pi_{k+1}}$

from arbitrary initialization (Q^{π_k})

$\rightarrow$ policy evaluation: $Q^{\pi_{k+1}}$

Interp 2: $\mathcal{T}^{\pi_{k+1}}$: one-step differeme.



$$\Rightarrow Q^{\pi_k} \le (\mathcal{T}^{\pi_{k+1}})^\infty Q^{\pi_k} = Q^{\pi_{k+1}}, \qquad \square$$

(Monotone Improvement)

Recall:

V^π, Q^π, V^*, Q^*

Value Iteration: $f_k \leftarrow \mathcal{T} f_{k-1}$

$$\|f_k - Q^*\|_\infty \le \gamma \cdot \|f_{k-1} - Q^*\|_\infty$$

Policy Iteration:
- ① evaluation : Q^{π_k}
- ② improvement : $\pi_{k+1} \leftarrow \pi_{Q^{\pi_k}}$

↳
- exact solution within exp. time.
- approx. solution within poly. time.

Linear Programming 线性规划

oracle

Primal Form. $\rightarrow p^*$

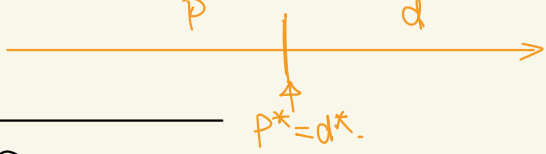
$$\boxed{\begin{aligned} \min_{x} \quad & c^T x \\ \text{s.t.} \quad & Ax \le b \\ & x \ge 0 \end{aligned}}$$

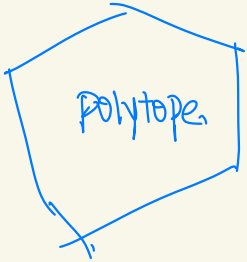
$\Longleftrightarrow$

Dual Form. $\rightarrow d^*$

$$\boxed{\begin{aligned} \max_{y} \quad & b^T y \\ \text{s.t.} \quad & A^T y \ge c \\ & y \ge 0 \end{aligned}}$$

- Weak Duality: $p^* \le d^*$. $\left(c^T x \le (A^T y)^T x = y^T A x \le y^T b \right)$
- Strong Duality: $p^* = d^*$. Minimax = Maximin.





Planning in MDP $\rightarrow$ LP oracle

Primal Form:

initial state distribution. (with $\forall s: d_0(s) > 0$)

$$\boxed{\begin{aligned} \min_{V \in \mathbb{R}^S} \quad & d_0^T V \\ \text{s.t.} \quad & V \ge \mathcal{T} V \end{aligned}}$$

NOT A VALUE FUNC.

1) Why linear ?
2) Why minimize ?
3) Why "$\ge$" rather than "$=$"?

$$\forall s \in S: \quad V(s) \ge \max_{a \in A} \left(R(s,a) + \gamma \cdot \mathbb{E}_{s' \sim P(\cdot|s,a)} [V(s')] \right)$$

$|S|$

1) $\Leftrightarrow$ linear constraints:

$$V(s) \geq R(s,a) + \gamma \cdot \mathbb{E}_{s' \sim P(\cdot|s,a)}[V(s')] \quad \forall a \in A$$

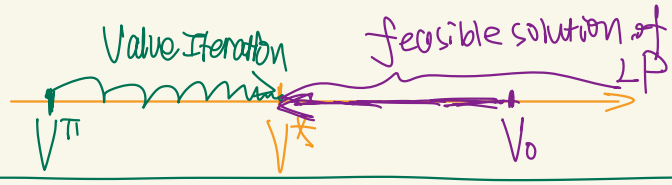
# of constraints

$|S| \times |A|$

Goal: $V = V^*$ is the optimal solution

Key observation: for any feasible point V

s.t. $V \geq \mathcal{T}V$: $V \geq V^*$ (pointwise)



Proof: Monotonicity 单调性. $V_1 \geq V_2 \Rightarrow \mathcal{T}V_1 \geq \mathcal{T}V_2$.
(for $\mathcal{T}$ and all $\mathcal{T}^\pi$)

$$V \geq \mathcal{T}V \geq \mathcal{T}^2 V \geq \cdots \geq \mathcal{T}^\infty V = V^* \quad \square$$

We "conceptually" run value iteration
(from an arbitrary V function)

Notice that V^* satisfies constraints

$\Rightarrow$ every feasible V is pointwise above V^*.

$\Rightarrow$ minimize $d_0^T V$ for a fully-supported V yields $V = V^*$.

Dual Form:

$$\max_{d \in \mathbb{R}^{S \times A},\, d \geq 0} \; d^T R$$

$R(s,a)_{(s,a) \in S \times A}$

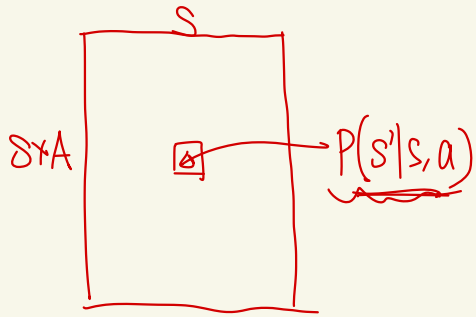
$$\text{s.t.} \; \Big(\sum_a d(s,a)\Big)_{s \in S} = \gamma P^T d + (1-\gamma) d_0$$

feasible d should be d^π for some π.

Marginalize out the actions

$P \in \mathbb{R}^{(S \times A) \times S}$

initial state distribution



Recall: discounted occupancy for state:

$$d^{\pi,s}(\cdot) = (1-\gamma) \cdot \sum_{t=1}^{\infty} \gamma^{t-1} d_t^{\pi,s}(\cdot)$$

$$d_t^{\pi,s}(\cdot) = \Pr_\pi[\, s_t = \cdot \mid s_1 = s \,]$$

$\Downarrow$

discounted occupancy for state-action: $\leftarrow d$

$$d^\pi(\cdot) = (1-\gamma) \cdot \sum_{t=1}^{\infty} \gamma^{t-1} d_t^\pi(\cdot) \qquad d_t^\pi(\cdot) = \Pr_\pi[\, (s_t, a_t) = \cdot \mid s_1 = s \,]$$

$\max_{d} d^T R$: maximize expected reward

optimal occupancy $d^* \Leftrightarrow$ optimal policy π^*

constraints for state-action occupancy

Bellman Flow Equation : $\left(\sum_a d(s,a)\right)_s = \gamma P^T d + (1-\gamma) d_0$

transition

$|S \times A| \times |S|$

$P(s' \mid s, a)$

marginalization

state-action occupancy $\rightarrow$ state occupancy

$$\sum_a d^\pi(s,a) = d^{\pi,s}(s)$$

$$\sum_a Pr_\pi[(s_t, a_t) = \cdot \mid s_1 = s] = Pr_\pi[s_t = \cdot \mid s_1 = s]$$

d_t^π : state-action occupancy at time t under policy π.

$d_{t+1}^{\pi,s} = P^T d_t^\pi$: state occupancy at time $t+1$ under policy π.

$$\gamma P^T d^\pi = (1-\gamma) \cdot \sum_{t=2}^{\infty} \gamma^{t-1} \cdot d^{\pi,s}$$

1	$d_1^\pi(s,a)$	P	$(1-\gamma) d_0$?	
γ	$d_2^\pi(s,a)$	P	$d_2^{\pi,s}(s)$	γ
γ^2	$d_3^\pi(s,a)$	P	$d_3^{\pi,s}(s)$	γ^2
$\vdots$	$\vdots$		$d_4^{\pi,s}(s)$	γ^3
			$\vdots$	

For any feasible d^π:

$$\pi(a \mid s) = \frac{d^\pi(s,a)}{\sum_a d^\pi(s,a)}$$

$\pi \Leftrightarrow d^\pi$