1.3b Computational Complexity

3. Value Iteration for Finite Horizon MDPs

① Set $Q_{H-1}(s,a) = r_{H-1}(s,a)$

② For $h = H-2, \cdots, 0$:

$$Q_h(s,a) = r_h(s,a) + \gamma E_{s' \sim P(s,a)}\left[\max_{a \in A} Q_{h+1}(s', a')\right]$$

By Theorem 1.9, $Q_h = Q_h^*$

and $\pi(s,h) = \arg\max_{a \in A} Q_h^*(s,a)$ is an optimal policy.

---

4. The Linear Programming Approach, LP

Recap: Value iter & policy iter depend polynomially on $\frac{1}{1-\gamma}$

not on $|S|, |A|$.

Goal: use LP to provide a polynomial time algorithm.

I: The primal LP

Let $V \in \mathbb{R}^{|S|}$ be variables:

$$\min \sum_s \mu(s) V(s)$$

$$\text{s.t. } V(s) \geq r(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s') \quad \forall a \in A, \forall s \in S.$$

if μ has full support, the optimal value function

V^* is the unique solution to this LP.

Pf: Let π^* be π_{V^*} (Theorem 1.8)

$\forall s, a$, Let π be the policy that takes action a at state s first, then follows π^*.

$$V^*(s) \geq V^{\pi}(s) = r(s,a) + \gamma \sum_{s'} P(s'|s,a) V^{\pi^*}(s')$$
$$= r(s,a) + \gamma \sum_{s'} P(s'|s,a) V^*(s')$$

$\Rightarrow V^*$ satisfies the constraints.

For any other V satisfied the constraints, we have

$$V(s) \geq r(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s')$$

Let a be $\pi^*(s)$, then

$$V(s) \geq r(s,\pi^*(s)) + \gamma E_{s' \sim P(s,\pi^*(s))}[V(s')]$$
$$\geq r(s,\pi^*(s)) + \gamma E_{s' \sim P(s,\pi^*(s))}[r(s',\pi^*(s')) + \gamma E_{s''}[V(s'')]]$$
$$\geq \cdots$$
$$\geq E\left[\sum_{t=0}^{\infty} \gamma^t r^t(s_t, a_t) \mid \pi^*, s_0 = s\right]$$
$$= V^*(s)$$

Therefore, when μ has full support, $\mu V \geq \mu V^*$ $(\mu \geq 0)$

II: The dual LP

For a fixed policy π, let's introduce some notations:

① visitation measure:

$$d_{s_0}^{\pi}(s,a) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t Pr^{\pi}(s_t = s, a_t = a | s_0)$$

where $Pr^{\pi}(s_t = s, a_t = a | s_0)$ is the probability that $s_t = s$, $a_t = a$ after starting at state s_0 and following π.

② for a distribution μ over S,

$$d_\mu^\pi(s,a) = E_{s_0 \sim \mu}[d_{s_0}^\pi(s,a)]$$

We now show that for all $s \in S$

$$\sum_a d_\mu^\pi(s,a) = (1-\gamma)\mu(s) + \gamma \sum_{s',a'} P(s|s',a') d_\mu^\pi(s',a')$$

Pf: $\sum_a d_\mu^\pi(s,a) = \sum_a \sum_{s^v} d_{s^v}^\pi(s,a)\mu(s^v)$

$$= \sum_a \sum_{s^v} \mu(s^v)(1-\gamma)\sum_{t=0}^{\infty}\gamma^t P_r^\pi(s_t=s, a_t=a | s_0=s^v)$$

$$= (1-\gamma)\mu(s) + \sum_a \sum_{s^v} \mu(s^v)(1-\gamma)\sum_{t=1}^{\infty}\gamma^t P_r^\pi(s_t=s, a_t=a | s_0=s^v)$$

$$:= (1-\gamma)\mu(s) + L$$

$$L = \sum_{s^v}\mu(s^v)\,\gamma\left[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t \sum_a P_r^\pi(s_{t+1}=s, a_{t+1}=a | s_0=s^v)\right]$$

$$= \sum_{s^v}\mu(s^v)\,\gamma\left[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t \sum_{s',a'} P_r^\pi(s_t=s', a_t=a' | s_0=s^v)\, P(s|s',a')\right]$$

$$= \gamma\sum_{s^v}\mu(s^v)\sum_{s',a'}P(s|s',a')\left[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t P_r^\pi(s_t=s', a_t=a' | s_0=s^v)\right]$$

$$= \gamma\sum_{s^v}\mu(s^v)\sum_{s',a'}P(s|s',a')\, d_{s^v}^\pi(s',a')$$

$$= \gamma\sum_{s',a'}P(s|s',a')\, d_\mu^\pi(s',a') \qquad \square$$

Note: the sum operations can be exchange due to Dominanted convergence Th.

[Def] state-action polytope:

$$K_\mu := \{ d \mid d \geq 0,\ \sum_a d(s,a) = (1-\gamma)\mu(s) + \gamma\sum_{s',a'}P(s|s',a')\, d(s',a') \}$$

Proposition 1.15 $\quad d \in K_\mu \iff \exists$ stationary policy π s.t. $d = d_\mu^\pi$

Pf: $(\Leftarrow)$ Obviously through proof above.

$(\Rightarrow)$. Assume $d \in K_\mu$, we have $d \geq 0$,

$$\sum_a d(s,a) = (1-\gamma)\mu(s) + \gamma \sum_{s',a'} P(s|s',a') d(s',a')$$

Let $\pi(a|s) = \dfrac{d(s,a)}{\sum_{a'} d(s,a')}$

$$d(s,a) = (1-\gamma)\mu(s)\pi(a|s) + \gamma \sum_{s',a'} P^\pi_{(s',a'),(s,a)} d(s',a')$$

$$\Rightarrow d = (1-\gamma) P_0^\pi + \gamma d P^\pi$$

$$\Rightarrow d = P_0^\pi \left[(1-\gamma)(I-\gamma P^\pi)^{-1}\right]$$

With lemma 1.6

$$\left[(1-\gamma)(I-\gamma P^\pi)^{-1}\right]_{\substack{(s,a)\\(s',a')}} = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}^\pi(s_t=s', a_t=a' | s_0=s, a_0=a)$$

we have

$$\begin{aligned}
d(s,a) &= \sum_{s',a'} \mu(s')\pi(a'|s')(1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}^\pi(s_t=s, a_t=a | s_0=s', a_0=a') \\
&= \sum_{s'} \mu(s')(1-\gamma)\sum_{t=0}^{\infty}\gamma^t \sum_{a'} \mathbb{P}^\pi(s_t=s, a_t=a | s_0=s', a_0=a')\pi(a'|s') \\
&= \sum_{s'} \mu(s')(1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr{}^\pi(s_t=s, a_t=a | s_0=s') \\
&= \sum_{s'} \mu(s') d_{s'}^\pi(s,a) \\
&= d_\mu^\pi(s,a) \qquad \square
\end{aligned}$$

Dual LP

W.R.T. variable $d \in \mathbb{R}^{|S|\cdot|A|}$,

$$\max \quad \frac{1}{1-\gamma} \sum_{s,a} d_\mu(s,a) r(s,a)$$

$$\text{s.t.} \quad d \in K_\mu$$

If d^* is the solution, provided μ has full support, we have $\pi^*(a|s) = \dfrac{d^*(s,a)}{\sum_{a'} d^*(s,a')}$ is optimal.

Pf: By Lemma 1.6. we can rewrite (π is stationary)

$$\frac{1}{1-\gamma} d_\mu^\pi \cdot r = Q^\pi \cdot \mathbb{1}$$

Since $d_\mu^\pi(s,a) = \sum_{s'} \mu(s')(1-\gamma) \sum_{t=0}^{\infty} \gamma^t P_r^\pi(s_t=s, a_t=a | s_0=s')$

$$= \sum_{s'} \mu(s')(1-\gamma) \sum_{t=0}^{\infty} \gamma^t \sum_{a'} \mathbb{P}^\pi(s_t=s, a_t=a | s_0=s', a_0=a') \pi(a'|s')$$

$$= \sum_{s',a'} \mu(s') \pi(a'|s') \left[(1-\gamma)(I-\gamma P^\pi)^{-1}\right]_{(s',a'),(s,a)}$$

then $\frac{1}{1-\gamma} \sum_{s,a} d_\mu^\pi(s,a) r(s,a)$

$$= \sum_{s,a} \sum_{s',a'} \mu(s') \pi(a'|s') (I-\gamma P^\pi)^{-1}_{(s',a')(s,a)} r(s,a)$$

$$= \sum_{s',a'} \mu(s') \pi(a'|s') \sum_{s,a} (I-\gamma P^\pi)^{-1}_{(s',a')(s,a)} r(s,a)$$

$$= \sum_{s',a'} Q^\pi(s',a')$$

Use proposition 1.15. the dual LP try to maximize Q^π, where $\pi(a|s) = \dfrac{d(s,a)}{\sum_{a'} d(s,a')}$, then it is obvious for the rest proof $\square$