关于Par VI的若干问题

WangFangyikang

2023 年 3 月 28 日

目录

1 Computational Complexity **2**
1.1 Value Iteration 2
1.2 Value Iteration 3
1.3 Linear Programming Approach 4
1.3.1 原始问题 4
1.3.2 对偶问题 4

2 Computational Complexity **6**
2.1 Value Iteration 6
2.2 Value Iteration 7
2.3 Linear Programming Approach 8
2.3.1 原始问题 8
2.3.2 对偶问题 8

1 Computational Complexity

我们定义$L,(P,r,\gamma)$是确定一个MDP所需的bit-size，假定算术运算+，-，×，÷都使用单位时间(unit time)。我们希望找到一个算法，它能找到在$L,(P,r,\gamma)$、$|mathcalS|$、$|mathcalA|$的多项式时间内找到最优策略。如果一个算法能在$|S|$、$|A|$的多项式时间（与$L,(P,r,\gamma)$无关）内找到最优策略，则称该算法为strongly polynomial。

1.1 Value Iteration

引理 1.1 (收缩性). *对任意两个向量Q，$Q' \in \mathbb{R}^{|mathcalS||mathcalA|}$,$\|\tau Q - \tau Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty$。这说明了$\tau$是一个收缩映射。*

证明. 首先证明对于所有的$s \in \mathcal{S}, |V_Q(s) - V_{Q'(s)}| \leq \underset{a\in\mathcal{A}}{max}|Q(s,a) - Q'(s,a)|$。
不妨设$V_Q(s) > V_{Q'}(s)$(另一半是对称的)，记$a^* = \underset{a}{argmax} Q(s,a)$。
$|V_Q(s) - V_{Q'(s)}| = \underset{a\in\mathcal{A}}{max} Q(s,a) - \underset{a\in\mathcal{A}}{max} Q'(s,a) \leq Q(s,a^*) - Q'(s,a^*) \leq \underset{a\in\mathcal{A}}{max}|Q(s,a) - Q'(s,a)|$

$$
\begin{aligned}
\|\tau Q - \tau Q'\|_\infty &= \|r + \gamma P V_Q - r - \gamma P V_{Q'}\|_\infty \\
&= \gamma \|P(V_Q - V_{Q'})\|_\infty \\
&\leq \gamma \|V_Q - V_{Q'}\|_\infty \\
&= \gamma \max_s |V_Q - V_{Q'}| \\
&\leq \max_s \max_a |Q(s,a) - Q'(s,a)| \\
&= \gamma \|Q - Q'\|_\infty
\end{aligned}
\tag{1}
$$

□

由该引理可见，经历了k次迭代后，$\|Q_k - Q^*\|_\infty \leq \gamma^k \|Q_0 - Q^*\|_\infty$，因此$\lim_{k\to\infty} Q_k - Q^* = 0$,算法收敛。

引理 1.2 (Q-Error Amplification)). *对任意向量$Q \in \mathbb{R}^{|mathcalS||mathcalA|}, V^{\pi_Q} \geq V^* - \frac{2\|Q-Q^*\|_\infty}{1-\gamma}\mathbb{1}$*
其中$\mathbb{1}$为全1向量。

证明. 对于固定的s以及$a = \pi_Q(s), (\pi_Q(s) = \underset{a\in\mathcal{A}}{argmax} Q(s,a))$

$$\begin{aligned}
V^*(s) - V^{\pi_Q}(s) &= Q^*(s,\pi^*(s)) - Q^{\pi_Q}(s,a) \\
&= Q^*(s,\pi^*(s)) - Q^*(s,a) + Q^*(s,a) - Q^(\pi_Q)(s,a) \\
&= Q^*(s,\pi^*(s)) - Q^*(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V^*(s') - V^{\pi_Q}(s')] \\
&\leq Q^*(s,\pi^*(s)) - Q(s,\pi^*(s)) + Q(s,a) - Q^*(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V^*(s') - V^{\pi_Q}(s')] \\
&\leq 2\|Q - Q^*\|_\infty + \gamma \|V^* - V^{\pi_Q}\|_\infty
\end{aligned} \tag{2}$$

□

定理 1.1 (Q-value iteration convergence)**.** 设$Q^{(0)} = 0, Q^{(k+1)} = \tau Q^{(k)}, k = 0, 1, \ldots \pi^{(k)=pi_{Q^{(k)}}}$,当$k \geq \frac{\log \frac{2}{(1-\gamma)^2\epsilon}}{1-\gamma}$,
$V^{\pi^{(k)}} \geq V^* - \epsilon\mathbb{1}$，即$k$次迭代后$V^{\pi^{(k)}}$和$V^*$非常接近。

证明. $\|Q^{(k)} - Q^*\|_\infty = \|\tau^k Q^{(0)} - \tau^k Q^*\|_\infty \leq \gamma^k \|Q^{(0)} - Q^*\|_\infty = (1 - ()1 - \gamma)^k \leq \frac{exp(-(1-\gamma)k)}{1-\gamma}$

□

1.2 Value Iteration

策略迭代算法从任意一个策略π_0出发，并对k=0,1,2,...重复接下来的两步： 1.策略评估。计算Q^{π_k}
2.策略提升。更新策略：$\pi_{k+1} = \pi_{Q^{\pi_k}}$,即$\pi_{k+1}(s) = \underset{a\in\mathcal{A}}{argmax} Q^{\pi_k}(s,a)$

引理 1.3. *1.*$Q^{\pi_{k+1}} \geq \tau Q^{\pi_k} \geq Q^{\pi_k}$
2.$\|Q^{\pi_{k+1}} - Q^*\|_\infty \leq \gamma \|Q^{\pi_k} - Q^*\|_\infty$

证明. 首先证明$\tau Q^{\pi_k} \geq Q^{\pi_k}$。注意到策略迭代中的策略都是确定策略，颖对于所有$k, s$，$V^{\pi_k}(s) = Q^{\pi_k}(s, \pi_k(s))$。

$$\begin{aligned}
\tau Q^{\pi_k}(s,a) &= r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[\max_{a'} Q^{\pi_k}(s',a')] \\
&\geq r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[Q^{\pi_k}(s',\pi_k(s'))] = Q^{\pi_k}(s,a)
\end{aligned} \tag{3}$$

再证明$Q^{\pi_{k+1}} \geq \tau Q^{\pi_k}$，这需要先证明$Q^{\pi_{k+1}} \geq Q^{\pi_k}$：
$Q^{\pi_k} = r + \gamma P^{\pi_k} Q^{\pi_k} \leq r + \gamma P^{\pi_{k+1}} Q^{\pi_k} \leq \sum_{t=0}^{\infty} \gamma^t (P^{\pi_{k+1}})^t r = Q^{\pi_{k+1}}$
第一个不等式是因为π_{k+1}是greedy policy，所以一定比π_k更好，第二个不等式由递归得到。因此

$$\begin{aligned}
Q^{\pi_{k+1}}(s,a) &= r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[Q^{\pi_{k+1}}(s',\pi_{k+1}(s'))] \\
&\geq r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[Q^{\pi_k}(s',\pi_{k+1}(s'))] \\
&= r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[\max_{a'} Q^{\pi_k}(s',a')] = \tau Q^{\pi_k}(s,a)
\end{aligned} \tag{4}$$

（1）式证明完成。下面证明（2）式：
$\|Q^* - Q^{\pi_{k+1}}\|_\infty \le \|Q^* - \tau Q^{\pi_k}\|_\infty = \|\tau Q^* - \tau Q^{\pi_k}\|_\infty \le \gamma \|Q^* - Q^{\pi_k}\|_\infty$
证明完成。 □

定理 1.2 (Policy iteration convergence)**.** 设$Q^{\pi_0} = 0, \pi_0$为任意初始策略。当$k \ge \frac{\log \frac{1}{(1-\gamma)\epsilon}}{1-\gamma}$，第$k$个策略有这样的*performance bound:*
$Q^{\pi_k} \ge Q^* - \epsilon$

证明.

$$\|Q^* - Q^{\pi_k}\|_\infty \le \gamma \|Q^* - Q^{\pi_{k-1}}\|_\infty \quad \le \gamma^k \|Q^* - Q^{\pi_0}\|_\infty = \gamma^k \|Q\|_\infty^{\|} \quad = (1-(1-\gamma))^k \|Q\|_\infty^{\|} \le \frac{\exp(-(1-\gamma)k)}{1-\gamma} \tag{5}$$

□

1.3 Linear Programming Approach

线性规划的方法可以在严格多项式时间内解决问题。

1.3.1 原始问题

最初的想法是求解

$\min_{V \in \mathbb{R}^{|\mathcal{S}|}} \quad \sum_s \mu(s) V(s)$
subject to $V(s) \ge \max_{a \in \mathcal{A}} [r(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s')] \qquad \forall s \in \mathcal{S}$
但这不是LP问题。将其转化为LP问题，得到：

$\min_{V \in \mathbb{R}^{|\mathcal{S}|}} \quad \sum_s \mu(s) V(s)$
subject to $V(s) \ge r(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s') \qquad \forall a \in \mathcal{A}, s \in \mathcal{S}$
其中，$\mu(s)$是初始状态分布。如果μ has full support,那么该问题的唯一解就是$V^*(s)$。

证明. 证明利用了τ的单调性，即$V \ge V'$时，有$\tau V \ge \tau V'$。
令$V' = \tau V$，则$\tau V \ge \tau V' = \tau^2 V$。迭代得到：$V \ge \tau^\infty = V^*$。
因此该约束条件下得到的解都是$V \ge V^*$的情况，由于目标函数是求 $\min_{V \in \mathbb{R}^{|\mathcal{S}|}} \quad \sum_s \mu(s) V(s)$，
最终得到的解即$V = V^*$。 □

1.3.2 对偶问题

对每一个LP都存在一个对偶问题，原问题的决策变量对应对偶问题的约束条件，原问题的约束条件对应对偶问题的决策变量。

对于固定的策略π，定义关于状态和动作的visitation measure：

$$d_{s_0}^{\pi}(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t Pr^{\pi}(s_t = s, a_t = a|s_0) \tag{6}$$

其中$Pr^{\pi}(s_t = s, a_t = a|s_0)$是从状态$s_0$出发，经过策略$\pi$，到达$s_t = s, a_t = a$的概率。并记$d_{\mu}^{\pi}(s,a) = \mathbb{E}_{s_0\sim\mu}[d_{s_0}^{\pi}(s,a)]$
对于任意的状态s有：

$$\sum_a d_{\mu}^{\pi}(s,a) = (1-\gamma)\mu(s) + \gamma\sum_{s',a'}P(s|s',a')d_{\mu}^{\pi}(s',a') \tag{7}$$

证明. 左边：

$$\begin{aligned}\sum_a d_{\mu}^{\pi}(s,a) &= d_{\mu}^{\pi}(s)\\ &= \mathop{\mathbb{E}}_{s_0\sim\mu}[d_{s_0}^{\pi}(s)]\\ &= \mathop{\mathbb{E}}_{s_0\sim\mu}[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t Pr^{\pi}(s_t = s|s_0)]\end{aligned} \tag{8}$$

右边：

$$\begin{aligned}&\gamma\sum_{s',a'}P(s|s',a')d_{\mu}^{\pi}(s',a') + (1-\gamma)\mu(s)\\ =&\gamma\sum_{s',a'}P(s|s',a')\mathop{\mathbb{E}}_{s_0\sim\mu}[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t Pr^{\pi}(s_t = s', a_t = a'|s_0)] + (1-\gamma)\mu(s)\\ =&\mathop{\mathbb{E}}_{s_0\sim\mu}[(1-\gamma)\sum_{t=0}^{\infty}\sum_{s',a'}\gamma^{t+1}P(s|s',a')Pr^{\pi}(s_t = s', a_t = a'|s_0)] + (1-\gamma)\mu(s)\\ =&\mathop{\mathbb{E}}_{s_0\sim\mu}[(1-\gamma)\sum_{t=0}^{\infty}\sum_{s',a'}\gamma^{t+1}Pr^{\pi}(s|s_0)] + (1-\gamma)\mu(s)\\ =&\mathop{\mathbb{E}}_{s_0\sim\mu}[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t Pr^{\pi}(s_t = s|s_0)]\end{aligned} \tag{9}$$

所以左边=右边。 □

定义一个状态-动作多面体：
$\mathcal{K}_{\mu} := \{d|d \geq 0 \quad and \quad \sum_a d(s,a) = (1-\gamma)\mu(s) + \gamma\sum_{s',a'}P(s|s',a')d(s',a')\}$
可以看到，$\mathcal{K}_{\mu}$是所有可以的状态-动作分别的集合，即$d \in \mathcal{K}_{\mu}$当且仅当存在一个平稳的策略π使得$d_{\mu}^{\pi} = d$。
有了这些定义，对偶问题可以写成：
max $\frac{1}{1-\gamma}\sum_{s,a} d_{\mu}(s,a)r(s,a)$

subject to $d \in \mathcal{K}_\mu$
目标函数可以看作是每个状态-动作对的密度乘上奖励，加权求和来求总奖励，求这个总奖励的最大值。解这个对偶LP求得一个解d^*,那么就可以求得最优策略：
$\pi^*(s,a) = \frac{d^*(s,a)}{\sum_{a'} d^*(s,a')}$
另一种求最优策略的方法是求$\underset{a}{argmax} \quad d^*(s,a')$。

2 Computational Complexity

我们定义$L,(P,r,\gamma)$是确定一个MDP所需的bit-size，假定算术运算+，-，×，÷都使用单位时间(unit time)。我们希望找到一个算法，它能找到在$L,(P,r,\gamma)$、$|mathcalS|$、$|mathcalA|$的多项式时间内找到最优策略。如果一个算法能在$|S|$、$|A|$的多项式时间（与$L,(P,r,\gamma)$无关）内找到最优策略，则称该算法为strongly polynomial。

2.1 Value Iteration

引理 2.1 (收缩性). 对任意两个向量Q，$Q' \in \mathbb{R}^{|mathcalS||mathcalA|}$,$\|\tau Q - \tau Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty$。这说明了$\tau$是一个收缩映射。

证明. 首先证明对于所有的$s \in \mathcal{S}, |V_Q(s) - V_{Q'(s)}| \leq \underset{a\in\mathcal{A}}{max}|Q(s,a) - Q'(s,a)|$。
不妨设$V_Q(s) > V_{Q'}(s)$(另一半是对称的)，记$a^* = \underset{a}{argmax} Q(s,a)$。
$|V_Q(s) - V_{Q'(s)}| = \underset{a\in\mathcal{A}}{max} Q(s,a) - \underset{a\in\mathcal{A}}{max} Q'(s,a) \leq Q(s,a^*) - Q'(s,a^*) \leq \underset{a\in\mathcal{A}}{max}|Q(s,a) - Q'(s,a)|$

$$
\begin{aligned}
\|\tau Q - \tau Q'\|_\infty &= \|r + \gamma P V_Q - r - \gamma P V_{Q'}\|_\infty \\
&= \gamma \|P(V_Q - V_{Q'})\|_\infty \\
&\leq \gamma \|V_Q - V_{Q'}\|_\infty \\
&= \gamma \max_s |V_Q - V_{Q'}| \\
&\leq \max_s \max_a |Q(s,a) - Q'(s,a)| \\
&= \gamma \|Q - Q'\|_\infty
\end{aligned}
\tag{10}
$$

□

由该引理可见，经历了k次迭代后，$\|Q_k - Q^*\|_\infty \leq \gamma^k \|Q_0 - Q^*\|_\infty$，因此$\lim_{k\to\infty} Q_k - Q^* = 0$,算法收敛。

引理 2.2 (Q-Error Amplification)). 对任意向量$Q \in \mathbb{R}^{|mathcalS||mathcalA|}, V^{\pi_Q} \geq V^* - \frac{2\|Q-Q^*\|_\infty}{1-\gamma}\mathbb{1}$
其中$\mathbb{1}$为全*1*向量。

证明. 对于固定的s以及$a = \pi_Q(s), (\pi_Q(s) = \underset{a\in\mathcal{A}}{argmax} Q(s,a))$

$$\begin{aligned} V^*(s) - V^{\pi_Q}(s) &= Q^*(s,\pi^*(s)) - Q^{\pi_Q}(s,a) \\ &= Q^*(s,\pi^*(s)) - Q^*(s,a) + Q^*(s,a) - Q^(\pi_Q)(s,a) \\ &= Q^*(s,\pi^*(s)) - Q^*(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[V^*(s') - V^{\pi_Q}(s')] \\ &\le Q^*(s,\pi^*(s)) - Q(s,\pi^*(s)) + Q(s,a) - Q^*(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[V^*(s') - V^{\pi_Q}(s')] \\ &\le 2\|Q - Q^*\|_\infty + \gamma\|V^* - V^{\pi_Q}\|_\infty \end{aligned} \tag{11}$$

□

定理 2.1 (Q-value iteration convergence)**.** 设$Q^{(0)} = 0, Q^{(k+1)} = \tau Q^{(k)}, k = 0, 1, \ldots \pi^{(k)=pi_{Q^{(k)}}}$,当$k \ge \frac{\log\frac{2}{(1-\gamma)^2\epsilon}}{1-\gamma}$,
$V^{\pi^{(k)}} \ge V^* - \epsilon\mathbb{1}$，即$k$次迭代后$V^{\pi^{(k)}}$和$V^*$非常接近。

证明. $\|Q^{(k)} - Q^*\|_\infty = \|\tau^k Q^{(0)} - \tau^k Q^*\|_\infty \le \gamma^k\|Q^{(0)} - Q^*\|_\infty = (1-()1-\gamma)^k \le \frac{exp(-(1-\gamma)k)}{1-\gamma}$

□

2.2 Value Iteration

策略迭代算法从任意一个策略π_0出发，并对k=0,1,2,...重复接下来的两步： 1.策略评估。计算Q^{π_k}
2.策略提升。更新策略：$\pi_{k+1} = \pi_{Q^{\pi_k}}$,即$\pi_{k+1}(s) = \underset{a\in\mathcal{A}}{argmax} Q^{\pi_k}(s,a)$

引理 2.3. *1.*$Q^{\pi_{k+1}} \ge \tau Q^{\pi_k} \ge Q^{\pi_k}$
2.$\|Q^{\pi_{k+1}} - Q^*\|_\infty \le \gamma\|Q^{\pi_k} - Q^*\|_\infty$

证明. 首先证明$\tau Q^{\pi_k} \ge Q^{\pi_k}$。注意到策略迭代中的策略都是确定策略，颖对于所有$k, s$，$V^{\pi_k}(s) = Q^{\pi_k}(s, \pi_k(s))$。

$$\begin{aligned} \tau Q^{\pi_k}(s,a) &= r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[\max_{a'} Q^{\pi_k}(s',a')] \\ &\ge r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[Q^{\pi_k}(s',\pi_k(s'))] = Q^{\pi_k}(s,a) \end{aligned} \tag{12}$$

再证明$Q^{\pi_{k+1}} \ge \tau Q^{\pi_k}$，这需要先证明$Q^{\pi_{k+1}} \ge Q^{\pi_k}$：
$Q^{\pi_k} = r + \gamma P^{\pi_k} Q^{\pi_k} \le r + \gamma P^{\pi_{k+1}} Q^{\pi_k} \le \sum_{t=0}^{\infty} \gamma^t (P^{\pi_{k+1}})^t r = Q^{\pi_{k+1}}$
第一个不等式是因为π_{k+1}是greedy policy，所以一定比π_k更好，第二个不等式由递归得到。因此

$$\begin{aligned} Q^{\pi_{k+1}}(s,a) &= r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[Q^{\pi_{k+1}}(s',\pi_{k+1}(s'))] \\ &\ge r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[Q^{\pi_k}(s',\pi_{k+1}(s'))] \\ &= r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[\max_{a'} Q^{\pi_k}(s',a')] = \tau Q^{\pi_k}(s,a) \end{aligned} \tag{13}$$

（1）式证明完成。下面证明（2）式：
$\|Q^* - Q^{\pi_{k+1}}\|_\infty \le \|Q^* - \tau Q^{\pi_k}\|_\infty = \|\tau Q^* - \tau Q^{\pi_k}\|_\infty \le \gamma \|Q^* - Q^{\pi_k}\|_\infty$
证明完成。 □

定理 2.2 (Policy iteration convergence)**.** 设$Q^{\pi_0} = 0, \pi_0$为任意初始策略。当$k \ge \frac{\log \frac{1}{(1-\gamma)\epsilon}}{1-\gamma}$，第$k$个策略有这样的*performance bound:*
$Q^{\pi_k} \ge Q^* - \epsilon$

证明.

$$\|Q^* - Q^{\pi_k}\|_\infty \le \gamma \|Q^* - Q^{\pi_{k-1}}\|_\infty \quad \le \gamma^k \|Q^* - Q^{\pi_0}\|_\infty = \gamma^k \|Q\|_\infty^{\|} \quad = (1-(1-\gamma))^k \|Q\|_\infty^{\|} \le \frac{\exp(-(1-\gamma)k)}{1-\gamma} \tag{14}$$

□

2.3 Linear Programming Approach

线性规划的方法可以在严格多项式时间内解决问题。

2.3.1 原始问题

最初的想法是求解

$\min_{V \in \mathbb{R}^{|\mathcal{S}|}} \quad \sum_s \mu(s) V(s)$
subject to $V(s) \ge \max_{a \in \mathcal{A}} [r(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s')] \qquad \forall s \in \mathcal{S}$
但这不是LP问题。将其转化为LP问题，得到：

$\min_{V \in \mathbb{R}^{|\mathcal{S}|}} \quad \sum_s \mu(s) V(s)$
subject to $V(s) \ge r(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s') \qquad \forall a \in \mathcal{A}, s \in \mathcal{S}$
其中，$\mu(s)$是初始状态分布。如果μ has full support,那么该问题的唯一解就是$V^*(s)$。

证明. 证明利用了τ的单调性，即$V \ge V'$时，有$\tau V \ge \tau V'$。
令$V' = \tau V$，则$\tau V \ge \tau V' = \tau^2 V$。迭代得到：$V \ge \tau^\infty = V^*$。
因此该约束条件下得到的解都是$V \ge V^*$的情况，由于目标函数是求 $\min_{V \in \mathbb{R}^{|\mathcal{S}|}} \quad \sum_s \mu(s) V(s)$，
最终得到的解即$V = V^*$。 □

2.3.2 对偶问题

对每一个LP都存在一个对偶问题，原问题的决策变量对应对偶问题的约束条件，原问题的约束条件对应对偶问题的决策变量。

对于固定的策略π，定义关于状态和动作的visitation measure：

$$d_{s_0}^{\pi}(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t Pr^{\pi}(s_t = s, a_t = a|s_0) \tag{15}$$

其中$Pr^{\pi}(s_t = s, a_t = a|s_0)$是从状态$s_0$出发，经过策略$\pi$，到达$s_t = s, a_t = a$的概率。并记$d_{\mu}^{\pi}(s,a) = \mathbb{E}_{s_0\sim\mu}[d_{s_0}^{\pi}(s,a)]$
对于任意的状态s有：

$$\sum_a d_{\mu}^{\pi}(s,a) = (1-\gamma)\mu(s) + \gamma\sum_{s',a'}P(s|s',a')d_{\mu}^{\pi}(s',a') \tag{16}$$

证明. 左边：

$$\begin{aligned}\sum_a d_{\mu}^{\pi}(s,a) &= d_{\mu}^{\pi}(s)\\ &= \mathop{\mathbb{E}}_{s_0\sim\mu}[d_{s_0}^{\pi}(s)]\\ &= \mathop{\mathbb{E}}_{s_0\sim\mu}[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t Pr^{\pi}(s_t = s|s_0)]\end{aligned} \tag{17}$$

右边：

$$\begin{aligned}&\gamma\sum_{s',a'}P(s|s',a')d_{\mu}^{\pi}(s',a') + (1-\gamma)\mu(s)\\ =&\gamma\sum_{s',a'}P(s|s',a')\mathop{\mathbb{E}}_{s_0\sim\mu}[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t Pr^{\pi}(s_t = s', a_t = a'|s_0)] + (1-\gamma)\mu(s)\\ =&\mathop{\mathbb{E}}_{s_0\sim\mu}[(1-\gamma)\sum_{t=0}^{\infty}\sum_{s',a'}\gamma^{t+1}P(s|s',a')Pr^{\pi}(s_t = s', a_t = a'|s_0)] + (1-\gamma)\mu(s)\\ =&\mathop{\mathbb{E}}_{s_0\sim\mu}[(1-\gamma)\sum_{t=0}^{\infty}\sum_{s',a'}\gamma^{t+1}Pr^{\pi}(s|s_0)] + (1-\gamma)\mu(s)\\ =&\mathop{\mathbb{E}}_{s_0\sim\mu}[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t Pr^{\pi}(s_t = s|s_0)]\end{aligned} \tag{18}$$

所以左边=右边。 □

定义一个状态-动作多面体：
$\mathcal{K}_{\mu} := \{d|d \geq 0 \quad and \quad \sum_a d(s,a) = (1-\gamma)\mu(s) + \gamma\sum_{s',a'}P(s|s',a')d(s',a')\}$
可以看到，$\mathcal{K}_{\mu}$是所有可以的状态-动作分别的集合，即$d \in \mathcal{K}_{\mu}$当且仅当存在一个平稳的策略π使得$d_{\mu}^{\pi} = d$。
有了这些定义，对偶问题可以写成：
max $\frac{1}{1-\gamma}\sum_{s,a} d_{\mu}(s,a)r(s,a)$

subject to $d \in \mathcal{K}_\mu$

目标函数可以看作是每个状态-动作对的密度乘上奖励，加权求和来求总奖励，求这个总奖励的最大值。解这个对偶LP求得一个解d^*,那么就可以求得最优策略：

$\pi^*(s,a) = \frac{d^*(s,a)}{\sum_{a'} d^*(s,a')}$

另一种求最优策略的方法是求$\underset{a}{argmax} \quad d^*(s,a')$。