# 折扣模型最优策略的结构

董泽清　刘　克
（中国科学院应用数学研究所）

**摘　　要**

本文研究了折扣马尔可夫决策规划（以下简记为MDP）最优策略的结构．证明了：任给一策略 $\pi^* = (\pi_0^*, \pi_1^*, \cdots, \pi_n^*, \pi_{n+1}^*, \cdots)$，若它是$\beta$折扣最优的，则随机平稳策略$\pi_0^{*\infty} = (\pi_0^*, \pi_0^*, \cdots)$也是$\beta$折扣最优的；对任何$n(\geq 1)$，我们也给出了随机平稳策略$\pi_n^{*\infty} = (\pi_n^*, \pi_n^*, \cdots)$也是$\beta$折扣最优的充分条件．还证明了：任给一随机平稳策略$\pi_0^{\infty} = (\pi_0, \pi_0, \cdots)$，若它是$\beta$折扣最优的，则必可分解为若干个$\beta$折扣最优的（决定性）平稳策略，而且这些平稳策略的凸组合就得到原来的$\pi_0^{\infty}$．因此，较完满地解决了折扣模型最优策略的结构．

## §1　引　　言

我们沿用〔1〕的概念和记号，不重述．本文所研究的MDP：$\{S, (A(i), i\in S), b, r, V_\beta\}$，其中状态空间$S$、每个状态可用的行动（决策）集$A(i)$均为可列集，状态转移律簇$q$是时齐的，报酬函数$r$是有界的（界记作$M$），以及折扣目标

$$V_\beta(\pi; i) = \sum_{t=0}^{\infty} \beta^t \sum_{\substack{a\in A(j)\\ j\in S}} P_\pi\{Y_t = j,\ \Delta_t = a \mid Y_0 = i\},$$
$$\pi\in\Pi,\ i\in S, \qquad (1)$$

其中$\beta\in(0,1)$是给定的．

一个策略$\pi^*\in\Pi$，若对任给的$i\in S$、$\pi\in\Pi$均有

$$V_\beta(\pi^*;\ i)\geq V_\beta(\pi; i) \qquad (2)$$

则称$\pi^*$是$\beta$折扣最优的或$\pi^*$是$\beta$折扣最优策略，简称为最优的或最优策略．

若一个（决定性）马氏策略$\pi^* = (f_0^*, f_1^*, f_2^*, \cdots)$是最优的，则对同一折扣因子$\beta$，$f_0^{*\infty}$也是最优的，这乃一熟知的事实．设一个策略$\pi^* = (\pi_0^*, \pi_1^*, \pi_2^*, \cdots, \pi_n^*, \pi_{n+1}^*, \cdots)$是最优的，那么对任何非负整数$n$，时刻$n$的（随机）决策规则$\pi_n^*$构成的平稳策略$\pi_n^{*\infty}$对同一$\beta$是否也是最优的？除$n=0$之外，回答是否定的；我们也给出了$\pi_n^{*\infty}$也是最优策略的条件；以及给出了修改$\pi_n^{*\infty}$的办法，使修改后的$\pi_n^*$所构成的平稳策略，对同一$\beta$也是最优的．

Chitgopekar〔3〕已证明：若$f^\infty$与$g^\infty$是两个（决定性）$\beta$折扣最优策略，则它们的随机化（即凸组合）构成的随机平稳策略也是$\beta$折扣最优的．反问题：若随机平稳策略$\pi_0^\infty = (\pi_0^\infty, \pi_0^\infty, \pi_0^\infty, \cdots)$是$\beta$折扣最优的，令

$$A'(i) = \{a: \pi_0(a\mid i) > 0\},\qquad i\in S, \qquad (3)$$

则任取一个$f(i)\in A'(i)$，$i\in S$，问这样定义的$f$所构成的平稳策略$f^\infty$对同一$\beta$是否也是最

*1984年2月20日收到．

优的？回答是肯定的．即最优随机平稳策略，可分解若干个决定性平稳策略，它们对同一$\beta$都是最优的．

这样，我们就完满地解决了折扣模型最优策略的结构．粗略地说，一个策略是最优的，当而且仅当，在每个时刻，对每个可能到达的状态都必须选用最优的行动．

## §2．最优策略的性质

令Б为定义在$S$上的可列维列向量空间，对任给的$V$，$V'\in$Б．若对每个$i\in S$，均有$V(i)\geq V'(i)$，则记作$V\geq V'$；若$V\geq V'$且存在$i\in S$使$V(i)>V(i)$，则记作$V>V'$．令

$$B\equiv\{V:-(1-\beta)^{-1}Me\leq V\leq(1-\beta)^{-1}Me, V\in Б\},\qquad(4)$$

其中$M$是报酬函数$r$的绝对值之界，$e\in$Б，且每个分量均为1的列向量．F为全体决策函数所成之集．

对任给的$f\in F$、$V\in B$，定义映射$T_f$为

$$T_fV=r(f)+\beta Q(f)V,\qquad(5)$$

其中$r(f)$为一列向量，其第$i$个分量为$r(i,f(i))$，$Q(f)$为一（次）随机矩阵，其第$(i,j)$个元素为$q(j\mid i,f(i))$，$i,j\in S$，$\beta\in(0,1)$．在下面的讨论中，我们总认为$\beta$是事先给定的．

我们不加证明地引用如下结果（证明可见〔1〕的第5章），把它们归并成一个引理．

**引理1** 对任给的$f\in F$，$V\in B$，我们有：

1）$T_f$把B映射到B；

2）$T_f$为一单调映射，即对任给的$V$、$V'\in B$．若$V\geq V'$，则$T_fV\geq T_fV'$；

3）$T_f$有唯一不动点$V_\beta(f^\infty)\in B$，即

$$T_fV_\beta(f^\infty)=V_\beta(f^\infty),$$

其中$V_\beta(f^\infty)$的第$i$个分量$V_\beta(f^\infty;i)$由（1）式给出，只要取$\pi=f^\infty$：

4）若$T_fV\geq V$，则$V_\beta(f^\infty)\geq V$，$V\in B$；

5）设$\pi'=(g_0,g_1,g_2,\cdots)\in\prod_m^d$，$\pi\in\prod$，则

$$T_{g_0}T_{g_1}\cdots T_{g_n}V_\beta(\pi)=V_\beta(g_0,g_1,\cdots,g_n,\pi)\to V_\beta(\pi'),\ \text{当}\ n\to\infty.$$

**定理1** 设$\pi^*=(\pi_1^*,\pi_2^*,\pi_3^*,\cdots)\in\prod$是$\beta$折扣最优的，且$\pi_1^*=f\in F$，则$f^\infty$对同一$\beta$也是最优的．

**证** 对任给的$i\in S$，我们均有

$$V_\beta(\pi^*;i)=r(i,f(i))+\beta\sum_{j\in S}q(j\mid i,f(i))V_\beta({}^1\pi^*(i,f(i));j).$$

其中

$${}^1\pi^*(i,f(i))=(\pi_0',\pi_1',\pi_2',\cdots)\in\prod$$

具有

$$\pi_t'(\cdot\mid s_0,a_0,s_1,a_1,\cdots,s_t)\equiv\pi_{t+1}^*(\cdot\mid i,f(i),s_1,a_0,s_1,a_1,\cdots,s_t).$$
$$a_n\in A(s_n),\ s_n\in S,\ n=0,1,2,\cdots,t.$$

即${}^1\pi^{(i,f(i))}$是由$\pi$往后推移一个时段所构成的策略．而去掉的那个时段，系统必处于状态$i$、

选用行动$f(i)$．由于$\pi^*$是最优的．则

$$V_\beta(\pi^*;i)\leq r(i,f(i))+\beta\sum_{j\in S}q(j\,|i,f(i))V_\beta(\pi^*;j)$$

$$=[T_fV_\beta(\pi^*)](i),\qquad i\in S$$

即

$$V_\beta(\pi^*)\leq T_fV_\beta(\pi^*),$$

由引理（1.4）有 $V_\beta(\pi^*)\leq V_\beta(f^\infty)$，由$\pi^*$是最优的，故$V_\beta(\pi^*)\equiv V_\beta(f^\infty)$． ∎

**系** （Blackwell〔2〕的定理2）若$\pi=(f_0,f_1,f_2,\cdots)\in\Pi_m^d$是最优的，则$f_0^\infty$对同一$\beta$也是最优的．

设$\pi=(f_0,f_1,\cdots,f_n,\pi_{n+1},\pi_{n+2},\cdots)$是$\beta$折扣最优的，那么，$f_n^\infty$对同一$\beta$是否也是最优的？回答是否定的．反例如下：

例 取$S=\{0,1\}$，$A(0)=\{1,2\}$，$A(1)=\{1\}$，$q(1\,|0,1)=q(1\,|0,2)=q(1\,|1,1)=1$，$r(0,1)=1$，$r(0,2)=r(1,1)=0$．

显然$F=\{f,g\}$，其中$f(i)=1$，$i=0,1$；$g(0)=2$，$g(1)=1$．

由于F是有限集，故存在平稳策略是最优的（如见〔1〕的第2章）．易见$f^\infty$是最优的（对任给的$\beta\in(0,1)$均成立），其最优报酬向量为

$$V_\beta(f^\infty;\ 0)=\sum_{t=0}^{\infty}\beta^t\sum_{j\in S}P_\pi\{Y_t=j\,|Y_0=0\}r(j,f(j))=1,$$

$$V_\beta(f^\infty;\ 1)=\sum_{t=0}^{\infty}\beta^t\sum_{j\in S}P_\pi\{Y_t=j\,|Y_0=1\}r(j,f(j))=0;$$

以及$V_\beta(g^\infty;i)=0$，$i\in S$，即$g^\infty$不是最优的（对任给的$\beta\in(0,1)$均成立）．若取$\pi=(f,g,g,\cdots)$，我们有

$$V_\beta(\pi;\ 1)=0$$

$$V_\beta(\pi;\ 0)=r(0,f(0))+\beta\sum_{j\in S}g(j\,|0,f(0))V_\beta(g^\infty)=1,$$

故

$$V_\beta(\pi)\equiv V_\beta(f^\infty),$$

即$\pi$是最优的，但$g^\infty$不是最优的．因此，一般来讲，$\pi=(f_0,f_1,\cdots,f_n,\pi_{n+1},\pi_{n+2},\cdots)$是最优的，对同一$\beta$，$f_n^\infty$不一定是最优的．下面我们给出一个充分条件，它保证$f_n^\infty$仍是最优的．为此，先引入在一个策略下“可实现的历史”这一概念．

**定义** 设$\pi=(\pi_0,\pi_1,\pi_2,\cdots)\in\Pi$，对任给的$t(\geq 1)$，若存在历史$h_{t-1}(\pi)\equiv(i_0,a_0,i_1,a_1,\cdots,i_{t-1},a_{t-1})$使

$$\pi_0(a_0\,|i_0)q(i_1\,|i_0,a_0)\pi_1(a_1\,|h_0(\pi),i_1)q(i_2\,|i_1,a_1)\cdots\cdot$$
$$q(i_{t-1}\,|i_{t-2},a_{t-2})\pi_{t-1}(a_{t-1}\,|h_{t-2}(\pi),i_{t-1})>0$$

则称$h_{t-1}(\pi)$为使用策略$\pi$，于$t=0$时从状态$i_0$出发的条件下，直到$t-1$时刻可实现的历史，其中$a_n\in A(i_n)$，$i_n\in S$，$n=0,1,\cdots,t-1$．对给定的$\pi$、$t$，这种可实现的历史之全体$\{h_{t-1}(\pi)\}$记做$H_{t-1}(\pi)$．对给定的$j\in S$、$\pi\in\Pi$，若存在某个可实现的历史$h_{t-1}(\pi)\equiv(i_0,a_0,i_1,a_1,\cdots,i_{t-1},a_{t-1})\in H_{t-1}(\pi)$使$g(j\,|i_{t-1},a_{t-1})>0$，则称使用策略$\pi$，于$t=0$时从$i_0$

出发的条件下，通过可实现的历史$h_{t-1}(\pi)$在时刻$t$可达状态$j$，简称为用策略$\pi$、$t=0$从$i_0$出发，于时刻$t$可达$j$。

**定理2** 若$\pi=(f_0, f_1, \cdots, f_n, \pi_{n+1}, \pi_{n+2}, \cdots)\in\Pi$是$\beta$折扣最优的，且对任给的$j\in S$必存在$i_0\in S$使得，在使用策略$\pi$、$t=0$从状态$i_0$出发，于时刻$n$可达状态$j$，则$f_n^\infty$对同一$\beta$也是最优的。

**证** 对任给的$r(r=1,2,\cdots, n)$，$i_k\in S$ $(k=0,1,\cdots, r-1)$，令

$$r\pi(i_0, f_0(i_0), i_1, f_1(i_1), \cdots, i_{r-1}, f_{r-1}(i_{r-1}))\equiv r_\pi(h_{r-1}(\pi))$$
$$=(f_r, f_{r+1}, \cdots, f_n, r\pi'_{n+2}, \cdots)\in\Pi, \tag{6}$$

其中

$$r_{\pi'_{n+t}}(\cdot\,|s_0, f_r(s_0), \cdots, s_{n-r}, f_n(s_{n-r}), s_{n-r+1}, a_{n-r+1}\cdots, s_{n-r+t})$$
$$=\pi_{n+t}(\cdot\,|h_{r-1}(\pi), s_0, f_r(s_0), \cdots, s_{n-r}, f_n(s_{n-r}), s_{n-r+1}, a_{n-r+1}, \cdots, s_{n-r+t})$$
$$a_k\in A(s_k),\quad k=n-r+1,\ n+r+2,\ \cdots,\ n-r+t,$$
$$s_j\in S,\quad j=0,1,\cdots,\ n-r+t,\ t=r+1, r+2,\ \cdots,$$

具有$h_{r-1}(\pi)=(i_0, f_0(i_0), i_1, f_1(i_1), \cdots, i_{r-1}, f_{r-1}(i_{r-1}))$。

由于$\pi$是$\beta$折扣最优的，及定理1知

$$V_\beta(\pi)\equiv V_\beta(f_0^\infty)\geq V_\beta(\pi'),\quad \pi'\in\Pi. \tag{7}$$

由(6)、(1)及(7)，对任给的$h_{n-1}(\pi)=(i_0, f_0(i_0), \cdots, i_{n-1}, f_{n-1}(i_{n-1}))$我们有

$$V_\beta(\pi; i_0)=r(i_0, f_0(i_0))+\beta\sum_{i_1}q(i_1|i_0, f_0(i_0))V_\beta({}^1\pi(i_0, f_0(i_0)); i_1)$$
$$=r(i_0, f_0(i_0))+\beta\sum_{i_1}q(i_1|i_0, f_0(i_0))r(i_1, f_1(i_1))$$
$$+\beta^2\sum_{i_1}q(i_1|i_0, f_0(i_0))\sum_{i_2}q(i_2|i_1, f_1(i_1))V_\beta({}^2\pi^{(i_0, f_0(i_0), i_1, f_1(i_1))}; i_2)$$
$$=\cdots=r(i_0, f_0(i_0))+\beta\sum_{i_1}q(i_1|i_0, f_0(i_0))r(i_1, f_1(i_1))$$
$$+\beta^2\sum_{i_1}q(i_1|i_0, f_0(i_0))\sum_{i_2}q(i_2|i_1, f_1(i_1))r(i_2, f_2(i_2))+\cdots$$
$$+\beta^{n-1}\sum_{i_1}q(i_1|i_0, f_0(i_0))\sum_{i_2}q(i_2|i_1, f_1(i_1))\cdots\sum_{i_{n-1}}q(i_{n-1}|i_{n-2}, f_{n-2}(i_{n-2}))$$
$$\cdot r(i_{n-1}, f_{n-1}(i_{n-1}))$$
$$+\beta^n\sum_{i_1}q(i_1|i_0, f_0(i_0))\cdots\sum_{i_n}q(i_n|i_{n-1}, f_{n-1}(i_{n-1}))V_\beta({}^{n\pi_{n-1}(h_{n-1}(\pi))}; i_n)$$
$$\leq r(i_0, f_0(i_0))+\beta\sum_{i_1}q(i_1|i_0, f(i_0))r(i_1, f_1(i_1))$$
$$+\beta^2\sum_{i_1}q(i_1|i_0, f_0(i_0))\sum_{i_2}q(i_2|i_1, f_1(i_1))r(i_2, f_2(i_2))+\cdots$$
$$+\beta^{n-1}\sum_{i_1}q(i_1|i_0, f_0(i_0))\cdots\sum_{i_{n-1}}q(i_{n-1}|i_{n-2}, f_{n-2}(i_{n-2}))r(i_{n-1}, f_{n-1}(i_{n-1}))$$
$$+\beta^n\sum_{i_1}q(i_1|i_0, f_0(i_0))\cdots\sum_{i_n}q(i_n|i_{n-1}, f_{n-1}(i_{n-1}))V_\beta(\pi; i_n), \tag{8}$$

上式中未标明的求和范围，均是在$S$上求和（下同），若用向量、矩阵来写，再用引理1.5）有

$$V_\beta(\pi)\leq V_\beta(f_0, f_1, \cdots, f_{n-1}, \pi)\leq V_\beta(\pi),$$

因此，（8）式全成立等式，并得到

$$\beta^n\sum_{i_1} q(i_1|i_0, f_0(i_0))\cdots\sum_{i_n} q(i_n|i_{n-1}, f_{n-1}(i_{n-1}))V_\beta({}^n\pi^{(h_{n-1}(\pi))}; i_n)$$
$$=\beta^n\sum_{i_1} q(i_1|i_0, f_0(i_0))\cdots\sum_{i_n} q(i_n|i_{n-1}, f_{n-1}(i_{n-1}))V_\beta(\pi; i_n),$$

由于$\pi$是最优的，因此

$$V_\beta({}^n\pi^{(h_{n-1}(\pi))}; i_n)\leq V_\beta(\pi; i_n),\quad i_n\in S.$$

故，对任一可实现历史$h_{n-1}(\pi)$，于时刻$n$可达的那些$i_n$必有

$$V_\beta({}^n\pi^{(h_{n-1}(\pi))}; i_n)=V_\beta(\pi; i_n).\qquad(9)$$

由所设，对每个$j\in S$，必存在$i_0\in S$，使得在用策略$\pi$，$t=0$从状态$i_0$出发的条件下，于时刻$n$可达$j$．因此，对每个$i_n$，均必存在可实现历史$h_{n-1}(\pi)$，使

$$V_\beta({}^n\pi^{h_{n-1}(\pi)}; i_n)=V_\beta(\pi: i_n).$$

对每个$i_n$，用这样的历史，我们就确定了一个$\beta$折扣最优策略

$$(f_n, {}^n\pi'_{n+1}, {}^n\pi'_{n+2}, \cdots)\equiv{}^n\pi'\quad\Pi,\qquad(10)$$

再由定理1，就得到$f_n^\infty$对同$-\beta$也是最优的．∎

从（9）式我们看到，若不要求对所有状态出发同时达最优，而$\pi=(f_0, f_1, \cdots, f_n, \pi_{n+1}, \pi_{n+2}, \cdots)$是最优的，凡是使用策略$\pi$，从某个初始（$t=0$时）状态出发，于时刻$n$可达的那些状态，然后在$t=0$改为从这些可达状态出发，用策略${}^n\pi'$（由（10）给出），对同一$\beta$它也是最优的．但对于使用策略$\pi$，从任何初始状态出发，于时刻$n$均不可达的那些状态，在$t=0$时，改为从这些不可达状态出发、用策略${}^n\pi'$，就无法保证它还是最优的．这实质上就是Bellman的“最优化原理”对我们所研究的模型成立．

如何修改$f_n$，使修改后的$f_n$能构成一个平稳最优策略？我们有下结果．

**定理3** 若$\pi=(f_0, f_1, \cdots, f_n, \pi_{n+1}, \pi_{n+2}, \cdots)$是$\beta$折扣最优的，令

$$A^*(i)=\{a: r(i,a)+\beta\sum_j q(j|i,a)V_\beta(\pi; j)=V_\beta(\pi; i),\quad a\in A(i)\}.$$

$$\overline{B}=\{i:\ \text{存在}i_0\in S,\ \text{使用策略}\pi,\ t=0\ \text{从}i_0\ \text{出发，于时刻}n\ \text{可达状态}\ i,\ i\in S\}$$

且定义$f_n^*$为

$$f_n^*(i)=\begin{cases} f_n(i), & \text{当}i\in\overline{B},\\ a, & \text{当}i\notin\overline{B},\ a\in A^*(i)\end{cases}\qquad i\in S$$

则$f_n^{*\infty}$对同一$\beta$也是最优的．

这个定理的证明，从下面的定理7立即可得．为了简单起见，当$i\notin\overline{B}$时可取$f_n^*(i)=f_0(i)$．

一个问题：定理1中之$\pi^*$的初如决策规则若是随机的$\pi_0^*$，会有什么结果？我们先叙述一引理．

对任给的$i\in S$，$\pi_0(\cdot|i)$是$A(i)$上的概率分布．类似于（5）我们定义一个映射$\mathcal{T}_{\pi 0}$为：

$$[\mathcal{T}_{\pi_0}V](i)=\sum_{a\in A(i)}\pi_0(a|i)[r(i,a)+\beta\sum_j q(j|i,a)V(j)],$$
$$V\in B, i\in S.\qquad(11)$$

若用向量、矩阵符号来写，(11) 可写作

$$\mathcal{T}_{\pi_0}V=\pi_0 r+\beta\pi_0 qV,\qquad V\in\mathbf{B},\tag{12}$$

其中$\mathcal{T}_{\pi_0}V$，$\pi_0 r$，$\beta\pi_0 qV$ 均为列向量，其第 $i$ 个分量分别为$[\mathcal{T}_{\pi_0}V](i)$，$\sum_{a\in\mathbf{A}(i)}\pi_0(a|i)r(i,a)$，以及$\beta\sum_{a\in\mathbf{A}(i)}\pi_0(a|i)\sum_j q(j|i,a)V(j)$.

**引理 2** 对任给的随机规则$\pi_0$、$V\in\mathbf{B}$，按 (11)(或 (12)) 定义的$\mathcal{T}_{\pi_0}$，我们有

1) $\mathcal{T}_{\pi_0}$是把 B 映射到 B 的单调映射；2) 若$\mathcal{T}_{\pi_0}V>V$，则$V_\beta(\pi_0^\infty)>V$；

3) 若$\mathcal{T}_{\pi_0}V\geq V$，则$V_\beta(\pi_0^\infty)\geq V$；4) 若$\mathcal{T}_{\pi}V<V$，则$V_\beta(\pi_0^\infty)<V$；

5) 若 $\mathcal{T}_{\pi_0}V\leq V$，则$V_\beta(\pi_0^\infty)\leq V$.

**证** 1) $\forall\, i\in\mathbf{S}$，由 (11)

$$[\mathcal{T}_{\pi_0}V](i)\leq\sum_{a\in\mathbf{A}(i)}\pi_0(a|i)[|r(i,a)|+\beta\sum_j q(j|i,a)|V(j)|]$$

$$\leq M+\beta M(1-\beta)^{-1}=M(1-\beta)^{-1},$$

故 $\mathcal{T}_{\pi_0}$把 B 映射到 B．设$V$、$V'\in\mathbf{B}$且$V\geq V'$，由 (12)，

$$\mathcal{T}_{\pi_0}V-\mathcal{T}_{\pi_0}V'=\pi_0 r+\beta\pi_0 qV-(\pi_0 r+\beta\pi_0 qV')=\beta\pi_0 q(V-V')\geq 0,$$

即$\mathcal{T}_{\pi_0}$是单调映射．

2) 由 1)，$\mathcal{T}_{\pi_0}\mathcal{T}_{\pi_0}V\equiv\mathcal{T}_{\pi_0}^2V\geq\mathcal{T}_{\pi_0}V>V$，即$\pi_0 r+\beta\pi_0 q\pi_0 r+\beta^2\pi_0 qV>V$，用归纳法，对任何正整数 $n$ 有$\mathcal{T}_{\pi_0}^n V\geq\mathcal{T}_{\pi_0}V>V$，即

$$\pi_0 r+\beta\pi_0 q\pi_0 r+\beta^2\pi_0 q\pi_0 q\pi_0 r+\cdots+\beta^{n-1}\pi_0 q\pi_0 q\cdots\pi_0 q\pi_0 r+$$

$$\overbrace{\beta^n\pi_0 q\pi_0 q\cdots\pi_0 q\pi_0}^{n\text{重}\pi_0 q}qV\geq\pi_0 r+\beta\pi_0 qV>V,\tag{13}$$

最左端最后一项的绝对值$\leq\beta^n(1-\beta)^{-1}M\cdot e\to 0$，当 $n\to\infty$时 (按分量分别取极限)，其中 $e$ 是每个元素均为 1 的列向量．而

$$V_\beta(\pi_0^\infty)=\pi_0 r+\beta\pi_0 q\pi_0 r+\beta^2\pi_0 q\pi_0 q\pi_0 r+\cdots+\beta^{n-1}\pi_0 q\pi_0 q\cdots\pi_0 q\pi_0 r$$

$$+\overbrace{\beta^n\pi_0 q\pi_0 q\cdots\pi_0 q\pi_0}^{n\text{重}\pi_0 q}q\pi_0 r+\cdots$$

上式右端前 $n$ 项与 (13) 式最左端的前 $n$ 项完全相同，在 (13) 式两端令 $n\to\infty$，得

$$V_\beta(\pi_0^\infty)\geq\mathcal{T}_{\pi_0}V>V.$$

3)、4) 及 5) 的证明完全类似于 2) 的证明 ∎

**定理 4** 设策略$\pi^*=(\pi_0^*,\pi_1^*,\pi_2^*,\cdots)\in\Pi$是$\beta$折扣最优的，则$\pi_0^{*\infty}=(\pi_0^*,\pi_0^*,\pi_0^*,\cdots)$对同一$\beta$也是最优的．

**证** 由于$V_\beta(\pi^*)=\pi_0^* r+\beta\pi_0^* q\pi_1^* r+\beta^2\pi_0^* q\pi_1^* q\pi_2^* r+\cdots+\beta^n\pi_0^* q\pi_1^* q\cdots\pi_{n-1}^* q\pi_n^* r+\cdots$

$$=\pi_0^* r+\beta\pi_0^* q[\pi_1^* r+\beta\pi_1^* q\pi_2^* r+\cdots+\beta^{n-1}\pi_1^* q\pi_2^* q\cdots\pi_{n-1}^* q\pi_n^* r+\cdots]$$

$$=\pi_0^* r+\beta\pi_0^* qV_\beta({}^1\pi^*(\cdot,\cdot)),$$

按分量来写，$\forall\, i\in\mathbf{S}$

$$V_\beta(\pi^*;i)=\sum_{a\in\mathbf{A}(i)}\pi_0^*(a|i)r(i,a)+\sum_{a\in\mathbf{A}(i)}\pi_0(a|i)\sum_j q(j|i,a)V_\beta({}^1\pi^*(i,a);j),$$

其中${}^1\pi^*(i,a)=(\pi_0^{*\prime},\pi_1^{*\prime},\pi_2^{*\prime},\cdots)$ 具有$\pi_t^{*\prime}(\cdot|s_0,a_0,s_1,a_1,\cdots,s_t)=\pi_{t-1}^*(\cdot|i,a,s_0,a_0,s_1,a_1,\cdots,s_t)$ $a_k\in\mathbf{A}(s_k)$，$s_k\in\mathbf{S}$，$k=0,1,\cdots,t$，$t=0,1,2,\cdots$，

显然${}^1\pi^*(\cdot,\cdot)\in\Pi$ (但依赖于初始状态与初始行动)，由于$\pi^*$是最优的，因此

$$V_\beta(\pi^*) \le \pi_0^* r + \beta \pi_0^* q V_\beta(\pi^*) = \mathcal{T}_{\pi_0} V_\beta(\pi^*)$$

由引理2.3有$V_\beta(\pi_0^{*\infty}) \ge V_\beta(\pi^*)$，由于$\pi^*$的最优性，必有$V_\beta(\pi_0^{*\infty}) = V_\beta(\pi^*)$，即$\pi_0^{*\infty}$对同一$\beta$也是最优的．∎

下面的定理5与6的证明，分别类似于定理2与3的证明，现仅叙述如下．

**定理5** 若$\pi = (\pi_0, \pi_1, \cdots, \pi_n, \pi_{n+1}, \cdots) \in \Pi$是$\beta$折扣最优的，且对任给的$j \in \mathbf{S}$，必存在$i_0$ ($i_0 \in \mathbf{S}$)，使得在用策略$\pi$时，$t = 0$从$i_0$出发，于时刻$n$可达状态$j$，则$\pi_n^\infty$对同一$\beta$也是最优的，其中$\pi_n$，对每个$j \in \mathbf{S}$，只要任取一通过$h_{n-1}(\pi)$于时刻$n$可达$j$的历史，$\pi_n(\cdot \mid h_{n-1}(\pi), j)$就完全确定了．

**定理6** 若$\pi = (\pi_0, \pi_1, \cdots, \pi_n, \pi_{n+1}, \cdots) \in \Pi$是$\beta$折扣最优的，任给$i \in \mathbf{S}$，若$i \in \bar{\mathbf{B}}$，则任取一个元$(\cdot \mid i)$为$\mathbf{A}^*(i)$上的概率分布，则定义$\pi_n^*$为

$$\pi_n^*(\cdot \mid i) = \begin{cases} \pi_n(\cdot \mid h_{n-1}(\pi), i), & \text{当} i \in \bar{\mathbf{B}}, \\ \bar{\pi}(\cdot \mid i), & \text{当当} i \in \bar{\mathbf{B}}, \end{cases}$$

则$\pi_n^{*\infty}$对同一$\beta$也是最优的，其中$\bar{\mathbf{B}}$，$\mathbf{A}^*(i)$的定义同定理3，$h_{n-1}(\pi)$为使用策略$\pi$的一个可实现历史、并通过它于时刻$n$可达$i$．

上面诸定理的证明本质上未用到报酬函数$r$的有界性．

## §3 最优随机平稳策略的分解

现在来回§1中提出的第二个问题．令

$$V_\beta^*(i) = \sup_{\pi \in \Pi} V_\beta(\pi; i), \qquad i \in \mathbf{S},$$

即$V_\beta^*$为最优报酬向量，我们有（见〔1〕的第五章）

$$V_\beta^*(i) = \sup_{a \in \mathbf{A}(i)} [r(i,a) + \beta \sum_j q(j \mid i, a) V_\beta^*(j)], \quad i \in \mathbf{S}, \tag{14}$$

$\mathbf{A}^*(i)$由定理3中所给出，注意到那里的$\pi$是最优的，所以$V_\beta(\pi) \equiv V_\beta^*$．$\mathbf{A}^*(i)$实为状态$i$可用的最优行动集（若$\mathbf{A}^*(i)$非空时）．

**引理3** 若存在$\beta$折扣最优策略$\pi^* = (\pi_0^*, \pi_1^*, \cdots)$，则对一切$i \in \mathbf{S}$，$\mathbf{A}^*(i)$均是非空的．

**证** 由于$\pi^*$是最优的，对一切$i \in \mathbf{S}$，由（14）式有

$$\begin{aligned} V_\beta^*(i) = V_\beta(\pi^*; i) &= \sum_{a \in \mathbf{A}(i)} \pi_0^*(a \mid i)[r(i,a) + \beta \sum_j q(j \mid i, a) V_\beta(\pi^{*(i,a)}; j)] \\ &\le \sum_{a \in \mathbf{A}(i)} \pi_0^*(a \mid i)[r(i,a) + \beta \sum_j q(j \mid i, a) V_\beta^*(j)] \\ &\le \sup_{a \in \mathbf{A}(i)} [r(i,a) + \beta \sum_j q(j \mid i, a) V_\beta^*(j)] = V_\beta^*(i), \end{aligned} \tag{15}$$

因此，上式全成立等式，其中

$$V_\beta(\pi^{*(i,a)}; j) = \sum_{t=1}^{\infty} \beta^{t-1} \sum_{\substack{b \in \mathbf{A}(k) \\ k \in \mathbf{S}}} P_\pi\{\mathbf{Y}_t = k, \Delta_t = b \mid \mathbf{Y}_0 = i, \Delta_0 = a, \mathbf{Y}_1 = j\} r(k,b),$$

由（14），我们总有

$$r(i,a) + \beta \sum_j q(j \mid i, a) V_\beta^*(j) \le V_\beta^*(i), \ a \in \mathbf{A}(i), \ i \in \mathbf{S},$$

由（15）成立等式，因此，对一切使$\pi_0^*(a_0 \mid i) > 0$的$a_0$均应有

$$r(i, a_0) + \beta \sum_j q(j \mid i, a_0) V_\beta^*(j) = V_\beta^*(i),$$

故，$\mathbf{A}^*(i)$ 非空，$i\in\mathbf{S}$．

**定理 7** 一个策略$\pi=(\pi_0,\pi_1,\cdots,\pi_n,\pi_{n+1},\cdots)$是$\beta$折扣最优的充要条件是对每个$n\geq 0$，时刻$n$的决策规则$\pi_n$局限在最优行动集上也构成一个概率分布，即在使用策略$\pi$下，通过任一可实现历史$h_{n-1}(\pi)$，于时刻$n$可达$j$的条件下，选取行动$a\in\mathbf{A}^*(j)$的概率为$\pi_n(a\,|h_{n-1}(\pi),\ j)$，且有$\sum_{a\in\mathbf{A}^*(j)}\pi_n(a\,|h_{n-1}(\pi),\ j)=1$．

**证** 必要性：设$\pi^*=(\pi_0^*,\pi_1^*,\cdots,\pi_n^*,\cdots)$是最优的，由引理3，知$\pi_0^*$对任何$i\in\mathbf{S}$，$\pi_0^*(\cdot\,|i)$均为$\mathbf{A}^*(i)$上的概率分布，$\mathbf{A}^*(i)$由定理3给出，而且（15）全成立等式，即

$$\begin{aligned}V_\beta^*(i)&=V_\beta(\pi^*;\,i)=\sum_{a\in\mathbf{A}^*(i)}\pi_0^*(a\,|i)\big[r\ (i,a)+\beta\sum_j q(j\,|i,a)V_\beta(\pi^{*(i,a)};\,j)\big]\\&=\sum_{a\in\mathbf{A}^*(i)}\pi_0^*(a\,|i)\big[r(i,a)+\beta\sum_j q(j\,|i,a)V_\beta^*(j)\big]\\&=E_{\pi^*}\big[r(\mathbf{Y}_0,\Delta_0)\,|\mathbf{Y}_0=i\big]+\sum_{a\in\mathbf{A}^*(i)}\pi_0^*(a\,|i)\,\beta\sum_j q(j\,|i,a)V_\beta(\pi^{*(i,a)};\,j)\\&=E_{\pi^*}\big[r(\mathbf{Y}_0,\Delta_0)\,|\mathbf{Y}_0=i\big]+\sum_{a\in\mathbf{A}^*(i)}\pi_0^*(a\,|i)\beta\sum_j q(j\,|i,a)V_\beta^*(j)．\end{aligned}$$

由于总有

$$V_\beta(\pi^{i,a);}\ j)\leq V_\beta^*(j),\quad j\in\mathbf{S},$$

因此对一切使$P_{\pi^*}\{\Delta_0=a,\mathbf{Y}_1=j\,|\mathbf{Y}_0=i\}=\pi_0^*(a\,|i)q(j\,|i,a)>0$的$(i,a,j)$均有

$$V_\beta(\pi^{*(i,a)};\,j)=V_\beta^*(j),\quad j\in\mathbf{S},\ a\in\mathbf{A}^*(i),\ i\in\mathbf{S}．$$

同理可证明

$$\begin{aligned}V_\beta^*(i)&=V_\beta(\pi^*;\,i)=E_{\pi^*}\big[r(\mathbf{Y}_0,\Delta_0)\,|\mathbf{Y}_0=i\big]+\beta E_{\pi^*}\big[r(\mathbf{Y}_1,\Delta_1)\,|\mathbf{Y}_0=i\big]\\&+\beta^2\sum_{a\in\mathbf{A}^*(i)}\pi_0^*(a\,|i)\ \sum_j q(j\,|i,a)\sum_{b\in\mathbf{A}^*(j)}\pi_1(b\,|i,a,j)\sum_k q(k\,|j,b)\cdot V_\beta(\pi^{*(i,a,j,b)};\,k)\\&=E_{\pi^*}\big[r(\mathbf{Y}_0,\Delta_0)\,|\mathbf{Y}_0=i\big]+\beta E_{\pi^*}\big[r(\mathbf{Y}_1,\Delta_1)\,|\mathbf{Y}_0=i\big]\\&+\beta^2\sum_{a\in\mathbf{A}^*(i)}\pi_0^*(a\,|i)\ \sum_j q(j\,|i,a)\sum_{b\in\mathbf{A}^*(j)}\pi_1^*(b\,|i,a,j)\sum_k q(k\,|j,b)V_\beta^*(k),\end{aligned}$$

且当$P_{\pi^*}\{\Delta_0=a,\mathbf{Y}_1=j\,|\mathbf{Y}_0=i\}>0$时，$\pi_1^*(\cdot\,|i,a,j)$仅是$\mathbf{A}^*(j)$上的概率分布，以及当

$$P_{\pi^*}\{\Delta_0=a,\mathbf{Y}_1=j,\Delta_1=b,\mathbf{Y}_2=k\,|\mathbf{Y}_0=i\}>0$$

时有

$$V_\beta(\pi^{*(i,a,j,b)};\,k)=\mathbf{V}_\beta^*(k),\ a\in\mathbf{A}^*(i),\ b\in\mathbf{A}^*(j),\ i,\ j,\ k\in\mathbf{S},$$

其中

$$V_\beta(\pi^{*(i,a,j,b)};\,k)=\sum_{t=2}^{\infty}\beta^{t-2}E_{\pi^*}\big[r(\mathbf{Y}_t,\Delta_t)\,|h_1(\pi),\mathbf{Y}_2=k\big],$$

而$h_1(\pi)=(i,a,j,b)$．

用归纳法我们可以证明：当$P_{\pi^*}\{\Delta_0=a_0,\mathbf{Y}_1=i_1,\Delta_1=a_1,\cdots,\mathbf{Y}_n=i_n\,|\mathbf{Y}_0=i\}>0$时，$\pi_n^*(\cdot\,|h_{n-1}(\pi^*),\ i_n)$仅是$\mathbf{A}^*(i_n)$上的概率分布，当

$$P_{\pi^*}\{\Delta_0=a_0,\mathbf{Y}_1=i_1,\Delta_1=a_1,\cdots,\mathbf{Y}_n=i_n,\Delta_n=a_n,\mathbf{Y}_{n+1}=i_{n+1}\,|\mathbf{Y}_0=i\}>0$$

时有

$$V_\beta(\pi^{*(h_n(\pi^*))};\,i_{n+1})=V_\beta^*(i_{n+1}),\quad i_{n+1}\in\mathbf{S},$$

其中$h_n(\pi^*)=(i,a_0,i_1,a_1,\cdots,i_n,a_n)$，$a_t\in\mathbf{A}^*(i_t)$，$i_t\in\mathbf{S}$，$t=1,2,\cdots,n$．

这就证明了，$\pi^*$在各个时刻选取行动的规则局限在最优行动集上，必为一概率分布．

充分性：设$\pi^* = (\pi_0^*, \pi_1^*, \pi_2^*, \cdots, \pi_n^*, \pi_{n+1}^*, \cdots)$ 中的$\pi_n^*$ $(n \geq 0)$ 均是局限于最优行动集上的概率分布，它也为一个策略，对任给的$i \in \mathbf{S}$,

$$V_\beta(\pi^*; i) = E_\pi^*\left[\sum_{t=0}^{\infty} \beta^t r(\mathbf{Y}_t, \Delta_t) \,\middle|\, \mathbf{Y}_0 = i\right]$$

$$= E_\pi^*\left\{\left[\sum_{t=0}^{N-1} \beta^t r(\mathbf{Y}_t, \Delta_t) + \beta^N V_\beta(\pi^{*(\mathbf{Y}_0, \Delta_0, \cdots, \mathbf{Y}_{N-1}, \Delta_{N-1})}; \mathbf{Y}_N)\right] \,\middle|\, \mathbf{Y}_0 = i\right\}$$

其中$\pi^{*(\mathbf{Y}_0, \Delta_0, \cdots, \mathbf{Y}_{N-1}, \Delta_{N-1})}$的第 $n+1$ 个决策规则为

$$\pi_n^{*(i, \Delta_0, \cdots, \mathbf{Y}_{N-1}, \Delta_{N-1})}(\cdot \,|\, \mathbf{Y}_N, \Delta_N, \cdots, \mathbf{Y}_{N+n})$$
$$= \pi_{N+n}^*(\cdot \,|\, i, \Delta_0, \cdots, \mathbf{Y}_{N-1}, \Delta_{N-1}, \mathbf{Y}_N, \Delta_N, \cdots, \mathbf{Y}_{N+n}),\ n \geq 0,\ N \geq 1.$$

而
$$E_\pi^*\left\{\left[\sum_{t=0}^{N-1} \beta^t r(\mathbf{Y}_t, \Delta_t) + \beta^N V_\beta^*(\mathbf{Y}_N)\right] \,\middle|\, \mathbf{Y}_0 = i\right\}$$

$$\leq E_\pi^*\left\{\left[\sum_{t=0}^{N-1} \beta^t r(\mathbf{Y}_t, \Delta_t) + \beta^N \left(V_\beta(\pi^{*(\mathbf{Y}_0, \Delta_0, \cdots, \mathbf{Y}_{N-1}, \Delta_{N-1})}; \mathbf{Y}_N) + \frac{2M}{1-\beta}\right)\right] \,\middle|\, \mathbf{Y}_0 = i\right\}$$

$$= E_\pi^*\left\{\left[\sum_{t=0}^{N-1} \beta^t r(\mathbf{Y}_t, \Delta_t) + \beta^N V_\beta(\pi^{*(\mathbf{Y}_0, \Delta_0, \cdots, \mathbf{Y}_{N-1}, \Delta_{N-1})}; \mathbf{Y}_N) \,\middle|\, \mathbf{Y}_0 = i\right\}\right.$$

$$\left. + \beta^N \frac{2M}{1-\beta} = V_\beta(\pi^*; i) + \beta^N \frac{2M}{1-\beta},\right.$$

即

$$V_\beta(\pi^*; i) + \beta^N \frac{2M}{1-\beta} \geq E_\pi^*\left\{\left[\sum_{t=0}^{N-1} \beta^t r(\mathbf{Y}_t, \Delta_t) + \beta^N V_\beta^*(\mathbf{Y}_N)\right] \,\middle|\, \mathbf{Y}_0 = i\right\} \quad i \in \mathbf{S},\ N \geq 1, \tag{16}$$

由$\mathbf{A}^*(j)$ 的定义，我们有

$$E_\pi^*\left\{\left[\beta^{N-1} r(\mathbf{Y}_{N-1}, \Delta_{N-1}) + \beta^N V_\beta^*(\mathbf{Y}_N)\right] \,\middle|\, \mathbf{Y}_0 = i\right\}$$
$$= \beta^{N-1} \sum_{\substack{a \in \mathbf{A}^*(j) \\ j \in \mathbf{S}}} E_\pi^*\left\{\left[r(j, a) + \beta V_\beta^*(\mathbf{Y}_N)\right] \,\middle|\, \mathbf{Y}_0 = i,\ \mathbf{Y}_{N-1} = j,\ \Delta_{N-1} = a\right\}$$
$$\cdot P_\pi^*\{\mathbf{Y}_{N-1} = j,\ \Delta_{N-1} = a \,|\, \mathbf{Y}_0 = i\}$$
$$= \beta^{N-1} \sum_{\substack{a \in \mathbf{A}^*(j) \\ j \in \mathbf{S}}} \left\{r(j, a) + \beta \sum_k q(k \,|\, j, a) V_\beta^*(k)\right\} P_\pi^*\{\mathbf{Y}_{N-1} = j,\ \Delta_{N-1} = a \,|\, \mathbf{Y}_0 = i\}$$
$$= \beta^{N-1} \sum_{a, j} V_\beta^*(j) P_\pi^*\{\mathbf{Y}_{N-1} = j,\ \Delta_{N-1} = a \,|\, \mathbf{Y}_0 = i\} = \beta^{N-1} E_\pi^*\left[V_\beta^*(\mathbf{Y}_{N-1}) \,\middle|\, \mathbf{Y}_0 = i\right],$$

将上式代入 (16)，则有

$$V_\beta(\pi^*; i) + \beta^N \frac{2M}{1-\beta} \geq E_\pi^*\left\{\left[\sum_{t=0}^{N-2} \beta^t r(\mathbf{Y}_t, \Delta_t) + \beta^{N-1} V_\beta^*(\mathbf{Y}_{N-1})\right] \,\middle|\, \mathbf{Y}_0 = i\right\}$$
$$= \cdots = E_\pi^*\left\{\left[r(\mathbf{Y}_0, \Delta_0) + \beta V_\beta^*(\mathbf{Y}_1)\right] \,\middle|\, \mathbf{Y}_0 = i\right\} = V_\beta^*(i),\ i \in \mathbf{S},\ N \geq 1,$$

然后，令$N \to \infty$得$V_\beta(\pi^*; i) \geq V_\beta^*(i)$, $i \in \mathbf{S}$,

由$V_\beta^*(i)$ 是最优报酬 ($t = 0$，从 $i$ 出发)，知$\pi^*$是最优的． ∎

这个定理说明：一个策略是最优的，当而且仅当在每个时刻的决策规则，对可达的那些状态都必须选用最优行动．因此，若

$$\pi_0^\infty = (\pi_0, \pi_0, \cdots) \in \Pi_s,$$

是最优的随机平稳策略，则$\pi_0(a|i)>0$的$a$，必属于$\mathbf{A}^*(i)$，$i\in\mathbf{S}$，即由（3）式定义的$\mathbf{A}'(i)\subset\mathbf{A}^*(i)$，$i\in\mathbf{S}$．因此，对每个$i\in\mathbf{S}$，任取一$f(i)\in\mathbf{A}'(i)$，则$f^\infty$对同一$\beta$也必须是最优的．即$\pi_0^*$必是若干个（可能是无限个）决策函数的凸组合而成，而且这些决策函数之每一个，都单独构成一个平稳策略是最优的（对同一$\beta$）．

$\Pi^*$，$\Pi_m^{d*}$、$\Pi_m^*$、$\Pi_s^{d*}$，$\Pi_s^*$分别表示$\Pi$，$\Pi_m^d$，$\Pi_m$，$\Pi_s^d$、$\Pi_s$中的最优策略集．$C(A)$表示集A的凸包，则我们已证明下结果．

**定理8**　若MDP：$\{S,(A(i),i\in S),q,r,V_\beta\}$存在最优策略，必有$C(\Pi_s^{d*})=\Pi_s^*\subset\Pi^*$；类似地有$C(\Pi_m^{d*})=\Pi_m^*\subset\Pi^*$，这几$C(\Pi_m^{d*})$是诸最优（决定性）马氏策略，各时刻决策规则的凸组合，所成的凸包．

从而我们较完满地搞清了折扣MDP最优策略的结构．

## 参　考　文　献


〔1〕董泽清，“马尔可夫决策规划”讲义，科学院应用数学所油印，1981年．

〔2〕Blackwell, D., Discrete dynamic programming, Ann. Math. Statist. Vol. 33 (1962), 719—726.

〔3〕Chitgopekar, S. S., Denumerable state Markovian sequential control processes: On randomizations of optimal policies, Noval Res. Logist. Quart. Vol. 22 (1975), 567—573.

〔4〕郭世贞，折扣目标最优策略中的最小方差策略问题，昆明工学院油印，1983年．