第2卷 第3期 自 动 化 学 报 Vol. 2, No. 3
1964 年 7 月 ACTA AUTOMATICA SINICA July, 1964

可控馬尔可夫鏈的一种最优决策

吳 滄 浦

摘 要

本文研究了一种最优馬尔可夫控制系統，这种控制系統以統計規律依賴于决定序列的馬尔可夫鏈描述。我們称决定序列为决策。存在一具有下述性質的目标状态，一旦系統到达此状态，状态就不再改变。我們的目的是要选取一决策，使所有从每一初始状态出发最終到达此目标状态的概率都达到最大。我們先提出在平稳决策集合中求最优决策的决策迭代法。然后証明，此决策在包含平稳及不平稳决策的决策集合上也是最优的。

一、問題的提法

某些最优控制过程可以归結为一类最优可控馬尔可夫鏈的模型。Bellman 首先研究了这类模型的漸近性質[1,2]。Howard 研究了两种这类模型的最优决策的求解問題[3]。在一种模型中，他假定任何决策下的馬尔可夫鏈都是正則的，指标是依賴于极限轉移概率的某种平均期望效益。在另一种模型中，未来阶段的效益要"打折扣"，这时，指标是依賴于轉移概率的总期望效益。Eaton 和 Zadeh 研究了一种具有吸收型目标状态的最优可控馬尔可夫鏈[4]，指标是依賴于轉移概率的总期望代价。在他們的模型中，决策集合只包含平稳决策。Blackwell 和 Derman 还进一步研究了决策集合含有不平稳决策及随机决策时最优决策的性質[5,6,7,8]。

以最小化总期望代价为指标，可以描写实际問題中某些平均值的指标，如平均过渡历程时間、平均誤差等，但不能描写某些概率指标。例如，在 Понтрягин 所研究的一类最优追踪問題中[9]，被追点在相空間中的运动是一个馬尔可夫过程，追点的运动是可控的，最优控制的指标是追点命中被追点的概率。在这問題中，如果把点在相空間中的运动用格子点的离散形式表示，則追踪过程可归結为具有吸收型目标状态的可控馬尔可夫鏈。这时，最优指标不能用上述工作中的期望值指标表示，而应該以到达目标状态的概率（下称命中概率）作为指标。

本文将以最大化命中概率作为指标，研究在包含不平稳决策的决策集合中，具有吸收型目标状态的可控馬尔可夫鏈的最优决策。

具有吸收型目标状态的可控馬尔可夫鏈可描述如下。在时間的进程中，系統只在給定的离散时刻 $t_0, t_1, t_2, \cdots$ 发生状态的改变。系統所有可能的状态数是有限的，我們以 $i = 1, 2, \cdots, n, n+1$ 表示之。在状态改变后，系統的状态变成某一新状态的概率，只取决于改变前它所处的状态以及所取的控制。在任一时刻 $t_k (k = 0, 1, 2, \cdots)$ 之前，如系統处于状态 $i (i = 1, 2, \cdots, n)$，則可供选择的控制 d_i 有 l_i 个。d_i 选定后，相应的各轉移概率 $p_{ij}(d_i) (j = 1, 2, \cdots, n, n+1)$ 也就确定。$i = n+1$ 表示目标状态。从目标状态出发的轉移概率 $p_{n+1,j}$ 則是固定的：$p_{n+1,1} = p_{n+1,2} = \cdots = p_{n+1,n} = 0$，$p_{n+1,n+1} = 1$，即

目标状态是一个吸收状态.

由此可見，在时刻 t_k，鏈的轉移概率矩陣 P 取决于所选取的一組控制 $d_1^k, d_2^k, \cdots, d_n^k$. 我們以向量 $\mathbf{d}^{(k)} = (d_1^k, d_2^k, \cdots, d_n^k)$ 表示之，称之为第 k 阶段的决定[1]. 矩陣 P 为 $\mathbf{d}^{(k)}$ 的函数

$$P(\mathbf{d}^{(k)}) = \begin{pmatrix} p_{11}(d_1^{(k)}) & \cdots & p_{1n}(d_1^{(k)}) & p_{1,n+1}(d_1^{(k)}) \\ \cdots\cdots & \cdots & \cdots\cdots & \cdots\cdots \\ p_{n1}(d_n^{(k)}) & \cdots & p_{nn}(d_n^{(k)}) & p_{n,n+1}(d_n^{(k)}) \\ 0 & \cdots & 0 & 1 \end{pmatrix}.$$

如果系統在时刻 t_0 之前所处的状态給定，各时刻 $t_0, t_1, \cdots$ 前的决定 $\mathbf{d}^{(0)}, \mathbf{d}^{(1)}, \cdots$ 都給定，則表征系統变化的随机过程——馬尔可夫鏈也就确定. 以 $\boldsymbol{\delta}$ 表示序列 $\{\mathbf{d}^{(0)}, \mathbf{d}^{(1)}, \cdots\}$，并称之为决策. 于是，系統状态变化过程的統計規律取决于所选取的决策.

以 x_i 表示由 t_0 前的初始状态 $i(i = 1, 2, \cdots, n)$ 最終到达目标状态的概率，則 x_i 为决策 $\boldsymbol{\delta}$ 的函数 $x_i(\boldsymbol{\delta})$；以 $\mathbf{x}(\boldsymbol{\delta})$ 表示 n 維列向量 $\langle x_1(\boldsymbol{\delta}), x_2(\boldsymbol{\delta}), \cdots, x_n(\boldsymbol{\delta})\rangle$. 我們的問題是：

1. 是否存在这样的 $\boldsymbol{\delta}^*$，使得对一切其他的 $\boldsymbol{\delta}$ 都有：$\mathbf{x}(\boldsymbol{\delta}^*) \geqq \mathbf{x}(\boldsymbol{\delta})$[2]. $\boldsymbol{\delta}^*$ 称为最优决策.

2. 如果最优决策存在，怎样求出 $\boldsymbol{\delta}^*$ 及相应的命中概率 $\mathbf{x}(\boldsymbol{\delta}^*)$.

当各时刻 $t_0, t_1, \cdots$ 前所取的决定全都一样时，决策 $\{\mathbf{d}, \mathbf{d}, \mathbf{d}, \cdots\}$ 称为平稳决策. 对于平稳决策，我們就以其决定元 $\mathbf{d}$ 表示之. 我們先証明，利用第三节的决策迭代法，从任意选取的平稳决策出发，經过有限次迭代，可以得到这样的平稳决策 $\mathbf{d}^*$，使得对一切平稳决策 $\mathbf{d}$ 都有 $\mathbf{x}(\mathbf{d}^*) \geqq \mathbf{x}(\mathbf{d})$. 然后証明，对于这一 $\mathbf{d}^*$ 以及任何不平稳决策 $\boldsymbol{\delta}$ 都有 $\mathbf{x}(\mathbf{d}^*) \geqq \mathbf{x}(\boldsymbol{\delta})$. 于是，最优决策存在，可以在平稳决策中得到，而且可以用决策迭代法求出.

二、命中概率的性质

現在研究平稳决策下命中概率的性质. 首先，在平稳决策下，所研究的过程形成一均匀馬尔可夫鏈[3]，这鏈的統計規律由轉移概率矩陣 $P(\mathbf{d})$ 完全确定. 下面以矩陣 $P(\mathbf{d})$ 表示这一馬尔可夫鏈，并称之为鏈 $P(\mathbf{d})$.

根据馬尔可夫鏈的理論[10,11]，給定决策 $\mathbf{d}$ 后，鏈 $P(\mathbf{d})$ 的全部状态可分成两类：恆返状态和非恆返状态，而恆返状态則組成若干恆返子鏈. 显然，目标状态 $n+1$ 是恆返状态，而且自身形成一恆返子鏈. 如果鏈 $P(\mathbf{d})$ 除了状态 $n+1$ 所形成的恆返子鏈外，沒有其他的恆返子鏈，我們就称决策 $\mathbf{d}$ 为一致决策.

由命中概率的定义，$\mathbf{x}(\mathbf{d})$ 应滿足等式

$$\mathbf{x}(\mathbf{d}) = Q(\mathbf{d})\mathbf{x}(\mathbf{d}) + \mathbf{q}(\mathbf{d}), \tag{1}$$

---

1) d_i^k 表示控制的方式，不是数. 但如果以号碼 $1, 2, \cdots, l_i$ 去标志各控制方式，則可以把 d_i^k 看作是在这 l_i 个整数上取值的变数. 本文将如此理解.

2) 本文以黑体拉丁字母 $\mathbf{a}$ 表示分量为 $a_1, a_2, \cdots, a_n$ 的 n 維列向量，記作 $\mathbf{a} = \langle a_1, a_2, \cdots, a_n\rangle$. 記号 $\mathbf{a} \geqq \mathbf{b}$ 表示对所有 $i = 1, 2, \cdots, n$，$a_i \geqq b_i$. 記号 $\mathbf{a} \geqslant \mathbf{b}$ 表示对所有 $i = 1, 2, \cdots, n$，$a_i \geqq b_i$，同时至少有一 i 使得 $a_i > b_i$.

3) 即轉移概率不随时間变化而保持固定的馬尔可夫鏈.

其中 $Q(\mathbf{d})$ 是 $n \times n$ 矩陣

$$Q(\mathbf{d}) = \begin{pmatrix} p_{11}(\mathbf{d}) & p_{12}(\mathbf{d}) & \cdots & p_{1n}(\mathbf{d}) \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ p_{n1}(\mathbf{d}) & p_{n2}(\mathbf{d}) & \cdots & p_{nn}(\mathbf{d}) \end{pmatrix}, \tag{2}$$

$\mathbf{q}(\mathbf{d})$ 是 n 維列向量

$$\mathbf{q}(\mathbf{d}) = \langle p_{1,n+1}(\mathbf{d}),\ p_{2,n+1}(\mathbf{d}),\ \cdots,\ p_{n,n+1}(\mathbf{d}) \rangle. \tag{3}$$

考虑未知量 $\mathbf{x}$ 的綫性方程組

$$\mathbf{x} = Q(\mathbf{d})\mathbf{x} + \mathbf{q}(\mathbf{d}), \tag{4}$$

不难証明，这方程組有唯一解的充分必要条件是决策 $\mathbf{d}$ 为一致决策.

实际上，如果 $\mathbf{d}$ 不是一致决策，則在状态 $1, 2, \cdots, n$ 中存在有恆返子鏈. 这时，矩陣 $I - Q(\mathbf{d})$ 是降秩的，因而(4)沒有唯一解.

反之，設 $\mathbf{d}$ 是一致决策，則状态 $1, 2, \cdots, n$ 全是非恆返状态，因之

对一切 $1 \leqslant i \leqslant n$，

$$\lim_{N\to\infty} p_{i,n+1}^{(N)}(\mathbf{d}) = 1^{1)}.$$

故存在这样的 N，使得

对一切 $1 \leqslant i \leqslant n$，

$$\sum_{j=1}^{n} p_{ij}^{(N)}(\mathbf{d}) < 1,$$

因而矩陣 $Q(\mathbf{d})$ 的最大特征数小于 1. 由此可知，从任一初始向量 $\mathbf{x}^{(0)}$ 开始，利用递推公式

$$\mathbf{x}^{(k+1)} = Q(\mathbf{d})\mathbf{x}^{(k)} + \mathbf{q}(\mathbf{d}) \tag{5}$$

进行迭代，序列 $\{\mathbf{x}^{(k)}\}$ 必收斂于方程(4)的唯一解 $\mathbf{x} = \mathbf{1}^{2)}$.

当 $\mathbf{d}$ 不是一致决策时，状态 $1, 2, \cdots, n$ 中含有恆返子鏈，这些子鏈中任何状态的命中概率均为零，即向量 $\mathbf{x}(\mathbf{d})$ 中必含有零分量. 但这时 $\mathbf{x} = \mathbf{1}$ 仍为方程組(4)的解，可見它有无穷多个解. 在这种情况下，从任意初始向量 $\mathbf{x}^{(0)}$ 出发，利用递推公式(5)进行迭代，序列 $\{\mathbf{x}^{(k)}\}$ 不一定收斂于 $\mathbf{x}(\mathbf{d})$. 但是如果选择适当的初始向量 $\mathbf{x}^{(0)}$，則我們仍可得到这种收斂. 为此，引入下述概念. 把除了目标状态而外的全部状态分成两类：凡其命中概率大于零的統称优状态，凡其命中概率等于零的統称劣状态. 当然，談到状态的优劣，总是对給定的决策 $\mathbf{d}$ 而言的.

如果我們这样选择初始向量 $\mathbf{x}^{(0)}$，使得对所有在决策 $\mathbf{d}$ 下属于劣状态的状态 i，分量 $x_i^{(0)} = 0$，对属于优状态的状态 i，分量 $x_i^{(0)}$ 可以任意选取 [下称此为条件 (6)]，則由公式(5)所确定的序列 $\{\mathbf{x}^{(k)}\}$ 必收斂于 $\mathbf{x}(\mathbf{d})$：

$$\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x}(\mathbf{d}). \tag{6}$$

这一論断的証明見附录引理一.

---

1) $p_{i,n+1}^{(N)}$ 表示矩陣 P 的 N 次乘冪 P^N 的第 i 行第 $n+1$ 列元素.

2) $\mathbf{1}$ 表示分量全是 1 的 n 維列向量.

三、决策迭代法

以 $\mathbf{d}^*$ 表示平稳决策集合中的最优决策，应用动态规划的决策迭代法求 $\mathbf{d}^*$，这方法由下列两个基本步驟組成.

步驟 I: 对前一次迭代所得的 $\mathbf{d}^{(r)}$（在第一步迭代时，则对任意选定的初始决策 $\mathbf{d}^{(0)}$），計算相应的命中概率 $\mathbf{x}(\mathbf{d}^{(r)})$. 为此，只須列出矩陣 $P(\mathbf{d}^{(r)})$，令所有恆返子鏈中的状态所对应的分量 $x_i(\mathbf{d}^{(r)}) = 0$，其余非零分量則由方程組(4)中相应的等式联立解出.

步驟 II: 根据前一步驟算出的 $\mathbf{x}(\mathbf{d}^{(r)})$，由下列公式

$$z_i^{(r+1)} = \max_{d_i}\left[\sum_{j=1}^{n} p_{ij}(d_i)x_j(\mathbf{d}^{(r)}) + p_{i,n+1}(d_i)\right], \quad 1 \leqslant i \leqslant n \tag{7}$$

計算 $z_i^{(r+1)}$. 当 $z_i^{(r+1)} = x_i(\mathbf{d}^{(r)})$ 时，保留 $d_i^{(r)}$ 不变，否則就用使上式方括号取最大值的 d_i 替代 $d_i^{(r)}$. 从 $i = 1$ 到 $i = n$ 进行这样的計算和替代，我們或者得到一个取代旧决策 $\mathbf{d}^{(r)}$ 的新决策 $\mathbf{d}^{(r+1)}$，或者使决策 $\mathbf{d}^{(r)}$ 不換.

如果将 $\mathbf{d}^{(r)}$ 換成 $\mathbf{d}^{(r+1)}$，我們就継續对 $\mathbf{d}^{(r+1)}$ 施行步驟 I、II，如此循环迭代，直到决策不換为止. 現在証明，对任意选定的初始决策 $\mathbf{d}^{(0)}$，这种迭代程序能在有限步内收敛于最优平稳决策 $\mathbf{d}^*$.

首先注意，不可能存在这样的状态：在迭代前决策 $\mathbf{d}^{(r)}$ 下是优状态，而在迭代后决策 $\mathbf{d}^{(r+1)}$ 下却成为劣状态. 換句話說，如果状态 i 在决策 $\mathbf{d}^{(r+1)}$ 下是劣状态，則有 $x_i(\mathbf{d}^{(r)}) = 0$. 这事实从决策迭代法看来是直观的，但严格証明却并非簡易的，証明見附录引理二.

考虑决策 $\mathbf{d}^{(r+1)}$ 下的递推公式(5)，并取 $\mathbf{x}(\mathbf{d}^{(r)})$ 作为初始向量 $\mathbf{x}^{(0)}$. 由上述事实，对于决策 $\mathbf{d}^{(r+1)}$，$\mathbf{x}^{(0)}$ 满足条件(6)，故由此确定的序列 $\{\mathbf{x}^{(k)}\}$ 收敛：

$$\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x}(\mathbf{d}^{(r+1)}). \tag{8}$$

另一方面，因

$$\begin{aligned}
\mathbf{x}^{(1)} &= Q(\mathbf{d}^{(r+1)})\mathbf{x}^{(0)} + \mathbf{q}(\mathbf{d}^{(r+1)}) = Q(\mathbf{d}^{(r+1)})\mathbf{x}(\mathbf{d}^{(r)}) + \\
&\quad + \mathbf{q}(\mathbf{d}^{(r+1)}) = \mathbf{z}^{(r+1)} \geqq \mathbf{x}(\mathbf{d}^{(r)})^{1)}, \\
\mathbf{x}^{(2)} &= Q(\mathbf{d}^{(r+1)})\mathbf{x}^{(1)} + \mathbf{q}(\mathbf{d}^{(r+1)}) \geqq Q(\mathbf{d}^{(r+1)})\mathbf{x}(\mathbf{d}^{(r)}) + \\
&\quad + \mathbf{q}(\mathbf{d}^{(r+1)}) = \mathbf{z}^{(r+1)} \geqq \mathbf{x}(\mathbf{d}^{(r)}), \\
&\cdots\cdots\cdots\cdots\cdots
\end{aligned}$$

[依此类推，对一切 $k = 1, 2, \cdots$，都有 $\mathbf{x}^{(k)} \geqq \mathbf{x}(\mathbf{d}^{r})$].

由(8)可得 $\mathbf{x}(\mathbf{d}^{(r+1)}) \geqq \mathbf{x}(\mathbf{d}^{(r)})$.

不仅如此，如果 $\mathbf{d}^{(r+1)} \neq \mathbf{d}^{(r)}$，則

$$\mathbf{x}(\mathbf{d}^{(r+1)}) \geqq \mathbf{x}(\mathbf{d}^{(r)}).$$

事实上，如果 $\mathbf{x}(\mathbf{d}^{(r+1)}) = \mathbf{x}(\mathbf{d}^{(r)})$，則这时

$$\begin{aligned}
z_i^{(r+1)} &= \sum_{j=1}^{n} p_{ij}(d_i^{(r+1)})x_j(\mathbf{d}^{(r)}) + p_{i,n+1}(d_i^{(r+1)}) = \\
&= \sum_{j=1}^{n} p_{ij}(d_i^{(r+1)})x_j(\mathbf{d}^{(r+1)}) + p_{i,n+1}(d_i^{(r+1)}) = x_i(\mathbf{d}^{(r+1)}) = x_i(\mathbf{d}^{(r)})
\end{aligned}$$

1) 这里 $\mathbf{z}^{(r+1)}$ 为 n 維列向量，其分量 $z_i^{(r+1)}$，$(1 \leqslant i \leqslant n)$ 由(7)确定.

对 $i = 1, 2, \cdots, n$ 都成立，因而 $\mathbf{d}^{(r+1)} = \mathbf{d}^{(r)}$ 与假設矛盾.

由此可見，每換一次决策，指标必严格上升（即至少有一分量加大），所以不可能有同一决策循环出現的情况. 由于决策总数是有限的，故在有限步內，必出現决策不再替換的情况. 假設这时决策是 $\mathbf{d}^{(r)}$，則它必为最优平稳决策.

事实上，这时对任一决策 $\mathbf{d}$ 必有

$$\sum_{j=1}^{n} p_{ij}(\mathbf{d})x_j(\mathbf{d}^{(r)}) + p_{i,n+1}(\mathbf{d}) \leqq x_i(\mathbf{d}^{(r)}), \quad 1 \leqslant i \leqslant n$$

即

$$Q(\mathbf{d})\mathbf{x}(\mathbf{d}^{(r)}) + q(\mathbf{d}) \leqq \mathbf{x}(\mathbf{d}^{(r)}).$$

考虑决策 $\mathbf{d}$ 下的递推公式(5)，并取 $\mathbf{x}^{(0)} = \mathbf{0}$[1]，則有

$$\mathbf{x}^{(1)} = \mathbf{q}(\mathbf{d}) \leqq Q(\mathbf{d})\mathbf{x}(\mathbf{d}^{(r)}) + \mathbf{q}(\mathbf{d}) \leqq \mathbf{x}(\mathbf{d}^{(r)}),$$

$$\mathbf{x}^{(2)} = Q(\mathbf{d})\mathbf{x}^{(1)} + \mathbf{q}(\mathbf{d}) \leqq Q(\mathbf{d})\mathbf{x}(\mathbf{d}^{(r)}) + \mathbf{q}(\mathbf{d}) \leqq \mathbf{x}(\mathbf{d}^{(r)}),$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

依此类推，对一切 $k = 1, 2, \cdots$，都有

$$\mathbf{x}^{(k)} \leqq \mathbf{x}(\mathbf{d}^{(r)}).$$

但

$$\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x}(\mathbf{d}),$$

故

$$\mathbf{x}(\mathbf{d}) \leqq \mathbf{x}(\mathbf{d}^{(r)}). \quad (\text{对一切 } \mathbf{d})$$

因此 $\mathbf{d}^{(r)}$ 是最优平稳决策 $\mathbf{d}^*$. 与此同时，我們也就証明了在平稳决策集合中最优决策是存在的[2].

四、考虑不平稳决策的情况

現在要証明，上面所求得的最优平稳决策 $\mathbf{d}^*$ 在包含不平稳决策的决策集合中也是最优的. 为此，先推导最优指标函数 $\mathbf{x}(\mathbf{d}^*)$ 的一个重要性質.

以 $\mathbf{x}^*$ 記 $\mathbf{x}(\mathbf{d}^*)$，根据动态規划的最优化原理[3]，$\mathbf{x}^*$ 应滿足下列方程

$$\mathbf{x}^* = \max_{\mathbf{d}} [Q(\mathbf{d})\mathbf{x}^* + \mathbf{q}(\mathbf{d})]. \tag{9}$$

下面証明，这方程的解 $\mathbf{x}^*$ 可以由动态規划的函数迭代法求出. 就是說，如果取 $\mathbf{x}^{(0)} = \mathbf{0}$，并依下列递推公式定义序列 $\{\mathbf{x}^{(k)}\}$：

$$\mathbf{x}^{(k+1)} = \max_{\mathbf{d}} [Q(\mathbf{d})\mathbf{x}^{(k)} + \mathbf{q}(\mathbf{d})], \quad k = 0, 1, 2, \cdots \tag{10}$$

則有

$$\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x}^*. \tag{11}$$

为了簡化起見，我們把式(10)写成

$$\mathbf{x}^{(k+1)} = T\mathbf{x}^{(k)},$$

考虑决策 $\mathbf{d}^*$ 下的递推公式(5)

$$\mathbf{x}^{(k+1)} = Q(\mathbf{d}^*)\mathbf{x}^{(k)} + \mathbf{q}(\mathbf{d}^*),$$

1) $\mathbf{0}$ 表示分量全是零的 n 維列向量.

2) 对于最优平稳决策的存在性，可以給出另一完全独立的証明.

3) 也可以由第三节 d^* 的求法得出.

或者写成

$$\mathbf{x}^{(k+1)} = T^*\mathbf{x}^{(k)}.$$

因为对任意的 $\mathbf{x}\geqq\mathbf{0}$，

$$\max_{\mathbf{d}}[Q(\mathbf{d})\mathbf{x}+\mathbf{q}(\mathbf{d})]\geqq Q(\mathbf{d}^*)\mathbf{x}+\mathbf{q}(\mathbf{d}^*),$$

即

$$T\mathbf{x}\geqq T^*\mathbf{x};$$

当 $\mathbf{x}=\mathbf{x}^{(0)}=\mathbf{0}$ 时，得

$$T\mathbf{x}^{(0)}\geqq T^*\mathbf{x}^{(0)}.$$

由这两个不等式推得

$$T^{(2)}\mathbf{x}^{(0)}=T(T\mathbf{x}^{(0)})\geqq T^*(T^*\mathbf{x}^{(0)})=T^{*(2)}\mathbf{x}^{(0)}.$$

依此，对一切 $k=1,2,\cdots$，可推得

$$T^{(k)}\mathbf{x}^{(0)}\geqq T^{*(k)}\mathbf{x}^{(0)}. \tag{12}$$

因 $\mathbf{x}^{(0)}=\mathbf{0}$，不論 $\mathbf{d}^*$ 如何，条件(6)恆滿足，故

$$\lim_{k\to\infty}T^{*(k)}\mathbf{x}^{(0)}=\mathbf{x}^*.$$

于是，由式(12)得

$$\liminf_{k\to\infty}T^{(k)}\mathbf{x}^{(0)}\geqq\mathbf{x}^*. \tag{13}$$

另一方面，注意到 $\mathbf{x}^*\geqq\mathbf{0}$，而 $\mathbf{x}^*$ 滿足式(9)，故对任意平稳决策 $\mathbf{d}$，我們有

$$T\mathbf{x}^{(0)}=\max_{\mathbf{d}}[\mathbf{q}(\mathbf{d})]\leqq\max_{\mathbf{d}}[Q(\mathbf{d})\mathbf{x}^*+\mathbf{q}(\mathbf{d})]=\mathbf{x}^*.$$

依此，对一切 $k=1,2,\cdots$ 可推得

$$T^{(k)}\mathbf{x}^{(0)}\leqq\mathbf{x}^*.$$

于是，

$$\limsup_{k\to\infty}T^{(k)}\mathbf{x}^{(0)}\leqq\mathbf{x}^*. \tag{14}$$

合并式(13)和(14)，就得到式(11).

現在考虑任一不平稳决策 $\boldsymbol{\delta}$，而 $\boldsymbol{\delta}=\{\mathbf{d}^{(0)},\mathbf{d}^{(1)},\mathbf{d}^{(2)},\cdots\}$. 令

$$\begin{aligned}
&\mathbf{y}^{(1)}=\mathbf{q}(\mathbf{d}^{(0)}),\\
&\mathbf{y}^{(2)}=\mathbf{y}^{(1)}+Q(\mathbf{d}^{(0)})\mathbf{q}(\mathbf{d}^{(1)}),\\
&\mathbf{y}^{(3)}=\mathbf{y}^{(2)}+Q(\mathbf{d}^{(0)})Q(\mathbf{d}^{(1)})\mathbf{q}(\mathbf{d}^{(2)}),\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
&\mathbf{y}^{(k)}=\mathbf{y}^{(k-1)}+Q(\mathbf{d}^{(0)})Q(\mathbf{d}^{(1)})\cdots Q(\mathbf{d}^{(k-2)})\mathbf{q}(\mathbf{d}^{(k-1)}),\\
&\cdots\cdots\cdots\cdots\cdots\cdots.
\end{aligned}$$

由命中概率的定义可得，在决策 $\boldsymbol{\delta}$ 下的命中概率 $\mathbf{x}(\boldsymbol{\delta})$ 为

$$\mathbf{x}(\boldsymbol{\delta})=\lim_{k\to\infty}\mathbf{y}^{(k)}. \tag{15}$$

把序列 $\{\mathbf{y}^{(k)}\}$ 与由式(10)所确定的且初始向量 $\mathbf{x}^{(0)}=\mathbf{0}$ 的序列 $\{\mathbf{x}^{(k)}\}$ 进行比較：

$$\begin{aligned}
&\mathbf{x}^{(1)}=\max_{\mathbf{d}}[\mathbf{q}(\mathbf{d})]\geqq\mathbf{q}(\mathbf{d}^{(0)})=\mathbf{y}^{(1)},\\
&\mathbf{x}^{(2)}=\max_{\mathbf{d}}[\mathbf{q}(\mathbf{d})+Q(\mathbf{d})\mathbf{x}^{(1)}]\geqq\mathbf{q}(\mathbf{d}^{(0)})+Q(\mathbf{d}^{(0)})\mathbf{x}^{(1)}\geqq\\
&\qquad\geqq\mathbf{y}^{(1)}+Q(\mathbf{d}^{(0)})\mathbf{q}(\mathbf{d}^{(1)})=\mathbf{y}^{(2)},\\
&\qquad\cdots\cdots\cdots\cdots\cdots\cdots.
\end{aligned}$$

依此类推，对任何 $k=1,2,\cdots$，都有

$$\mathbf{x}^{(k)} \geqq \mathbf{y}^{(k)}.$$

于是，由式(11)及(15)可得

$$\mathbf{x}(\mathbf{d}^*) \geqq \mathbf{x}(\delta).$$

这就是說，最优平稳决策 $\mathbf{d}^*$ 在包含不平稳决策的决策集合中也是最优的. 这样，我們就証明了第一节中最优决策的存在，而且可以在平稳决策集合中得到. 因此，当我們考虑这种类型的可控馬尔可夫鏈的最优控制时，便可以只在平稳决策的集合中，应用上述的决策迭代法求出最优决策.

如果我們知道平稳决策集合中存在着一致决策，則显然，它們同时就是最优决策. 不过在这种情况下，我們往往仍然不知究竟决策集合中哪些决策是一致的. 应用上述决策迭代法，可以有效地将它們求出. 如果我們事先不知道决策集合中是否存在一致决策，那么上述决策迭代法也可以当作一个检驗一致决策是否存在的方法.

不难証明，本文定义的一致决策和 Eaton & Zadeh 定义的适当决策[4]是等价的. 本文的决策迭代法也可作为检驗 Eaton 問題中适当决策是否存在的方法. Eaton 問題中的最优决策是在适当决策集合內得到的，因此如果一致决策存在，我們便可找到这样的决策，使它在本文意义下和 Eaton 意义下都是最优的. 也就是說，在这种情况下，存在着双指标最优决策. 如果不存在一致决策，則 Eaton 意义下的最优决策不存在，但本文意义下的最优决策仍然存在，而且仍可由上述决策迭代法求出.

附 录

以 $G(\mathbf{d})$ 和 $B(\mathbf{d})$ 分别表示在决策 $\mathbf{d}$ 下由所有优状态和劣状态的附标所組成的集合. 利用通用的表示法，可以写成

$$G(\mathbf{d})=\{i:\ x_i(\mathbf{d})>0,\quad 1\leqslant i\leqslant n\},$$
$$B(\mathbf{d})=\{i:\ x_i(\mathbf{d})=0,\quad 1\leqslant i\leqslant n\}.$$

記号 $i\in G(\mathbf{d})$ 表示 i 属于集合 $G(\mathbf{d})$，$A\cup B$ 表示 AB 两集合之和，$A\cap B$ 表示 AB 两集合之交.

引理一 設对所有 $i\in B(\mathbf{d})$，$x_i^{(0)}=0$，对所有 $i\in G(\mathbf{d})$，$x_i^{(0)}$ 任意，$\mathbf{x}^{(1)},\mathbf{x}^{(2)},\cdots$ 由公式(5)确定，則

$$\lim_{k\to\infty}\mathbf{x}^{(k)}=\mathbf{x}(\mathbf{d}). \tag{A.1}$$

証明 对給定的决策 $\mathbf{d}$，我們把状态的編号重排，使得 $1,2,\cdots,s$ 是劣状态，$s+1,s+2,\cdots,n$ 是优状态. 这时，对所有 $1\leqslant i\leqslant s$，$x_i^{(0)}=0$，$p_{i,s+1}(d_i)=p_{i,s+2}(d_i)=\cdots=p_{i,n}(d_i)=p_{i,n+1}(d_i)=0$. 由式(5)得，对所有 $k=1,2,\cdots$，$x_1^{(k)}=x_2^{(k)}=\cdots=x_s^{(k)}=0$，于是

$$\lim_{k\to\infty}x_i^{(k)}=0=x_i(\mathbf{d}).\quad 1\leqslant i\leqslant s. \tag{A.2}$$

对后面 $n-s$ 个分量，由式(5)則有

$$\left.\begin{aligned} x_{s+1}^{(k)}&=\sum_{j=s+1}^{n}p_{s+1,j}(d_{s+1})x_j^{(k)}+p_{s+1,n+1}(d_{s+1}),\\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\ x_n^{(k)}&=\sum_{j=s+1}^{n}p_{nj}(d_n)x_j^{(k)}+p_{n,n+1}(d_n).\end{aligned}\right\} \tag{A.3}$$

因为对 $s+1\leqslant i\leqslant n$，$\lim_{N\to\infty}p_{i,n+1}^{(N)}(d_i)=x_i(\mathbf{d})>0$，故存在这样的 N，使得 $\sum_{j=1}^{n}p_{ij}^{(N)}(d_i)<1$. 这样，方程組

(A.3)右边 x_j 的系数矩陣的最大特征数小于 1．由此可知，从任何初始的 $x_{s+1}^{(0)}, x_{s+2}^{(0)}, \cdots, x_n^{(0)}$ 出发，利用公式(5)进行迭代，所得 $x_{s+1}^{(k)}, x_{s+2}^{(k)}, \cdots, x_n^{(k)}$ 必收敛于方程組 (A.3) 的唯一解，但 $x_{s+1}(\mathbf{d}), x_{s+2}(\mathbf{d}), \cdots, x_n(\mathbf{d})$ 是它的一个解，故得

$$\lim_{k\to\infty} x_i^{(k)} = x_i(\mathbf{d}). \quad s+1 \leqslant i \leqslant n \tag{A.4}$$

合并式(A.2)和式(A.4)，就得式(A.1)．

引理二　設 $\mathbf{d}^{(r)}$ 是迭代前决策，$\mathbf{d}^{(r+1)}$ 是迭代后决策，則 $G(\mathbf{d}^{(r)}) \cap B(\mathbf{d}^{(r+1)})$ 是空集．

証明　为簡化起見，引入下列記号：$GG = G(\mathbf{d}^{(r)}) \cap G(\mathbf{d}^{(r+1)})$，$GB = G(\mathbf{d}^{(r)}) \cap B(\mathbf{d}^{(r+1)})$，$BG = B(\mathbf{d}^{(r)}) \cap G(\mathbf{d}^{(r+1)})$，$GG = G(\mathbf{d}^{(r)}) \cap G(\mathbf{d}^{(r+1)})$．

采用反証法．假定 GB 不是空集．因对 $i \in GB$，$j \in GG \cup BG$，$p_{ij}(d_i^{(r+1)}) = 0$，而对 $j \in BB$，$x_j(\mathbf{d}^{(r)}) = 0$，故 $z_i^{(r+1)} = \sum_{j\in GB} p_{ij}(d_i^{(r+1)}) x_j(\mathbf{d}^{(r)}) \geqq x_i(\mathbf{d}^{(r)})$．显然，这时至少存在 $i \in GB$，使得

$$z_i^{(r+1)} = \sum_{j\in GB} p_{ij}(d_i^{(r+1)}) x_j(\mathbf{d}^{(r)}) = x_i(\mathbf{d}^{(r)}).$$

以 $\overline{GB}$ 記 GB 中使上述等式成立之子集，$\widehat{GB}$ 記 $\overline{GB}$ 对 GB 之余集，則 $\overline{GB}$ 不是空集．这时可能出現两种情况．其一，$\widehat{GB}$ 是空集，其二，$\widehat{GB}$ 不空．

先考虑第一种情况．这时，$GB = \overline{GB}$，故由决策迭代規則，对所有 $i \in GB$，$d_i^{(r+1)} = d_i^{(r)}$，于是 $p_{ij}(d_i^{(r+1)}) = p_{ij}(d_i^{(r)})$．因而对所有 $i \in GB$，$j \in GG \cup BG$ 及 $j = n+1$，$p_{ij}(d_i^{(r)}) = p_{ij}(d_i^{(r+1)}) = 0$．考虑决策 $\mathbf{d}^{(r)}$ 下的递推公式(5)，当 $i \in GB$ 时，

$$x_i^{(k+1)} = \sum_{j\in GB\cup BB} p_{ij}(d_i^{(r)}) x_j^{(k)}. \tag{A.5}$$

因为对 $i \in BB \cup BG$，$j = GB \cup GG$ 及 $j = n+1$，$p_{ij}(d_i^{(r)}) = 0$，故对 $i \in BB \cup BG$，

$$x_i^{(k+1)} = \sum_{j\in BB\cup BG} p_{ij}(d_i^{(r)}) x_j^{(k)}. \tag{A.6}$$

因此，如果取 $\mathbf{x}^{(0)}$ 使得对所有 $i \in GB \cup BB \cup BG$，$x_i^{(0)} = 0$，則由式(A.5)，(A.6)可見，对这些 i，

$$x_i^{(k)} = 0. \quad k = 1, 2, \cdots$$

另一方面，由引理一，$\lim_{k\to\infty} x_i^{(k)} = x_i(\mathbf{d}^{(r)})$．于是对 $i \in GB$ 得 $x_i(\mathbf{d}^{(r)}) = 0$，这与 GB 的定义矛盾．

其次考虑第二种情况．如果 $\max_{i\in\overline{GB}} x_j(\mathbf{d}^{(r)}) > \max_{i\in\widehat{GB}} x_j(\mathbf{d}^{(r)})$，則因对 $i \in \overline{GB}$，$d_i^{(r+1)} = d_i^{(r)}$，故

$$x_i(\mathbf{d}^{(r)}) = \sum_{j\in\overline{GB}} p_{ij}(d_i^{(r)}) x_j(\mathbf{d}^{(r)}) + \sum_{j\in\widehat{GB}} p_{ij}(d_i^{(r)}) x_j(\mathbf{d}^{(r)}).$$

可見在决策 $\mathbf{d}^{(r)}$ 下，$\overline{GB}$ 中使 $x_i(\mathbf{d}^{(r)})$ 达到 $\max_{i\in\overline{GB}} x_j(\mathbf{d}^{(r)})$ 的那些状态組成一恆返子鏈，因而对这些状态，$x_i(\mathbf{d}^{(r)}) = 0$，这和 $\overline{GB}$ 的定义相矛盾．反之，如果 $\max_{i\in\overline{GB}} x_j(\mathbf{d}^{(r)}) \leqq \max_{i\in\widehat{GB}} x_j(\mathbf{d}^{(r)})$，則对 $i \in \widehat{GB}$，

$$z_i^{(r+1)} = \sum_{j\in\overline{GB}} p_{ij}(d_i^{(r+1)}) x_j(\mathbf{d}^{(r)}) + \sum_{j\in\widehat{GB}} p_{ij}(d_i^{(r+1)})$$

$$x_j(\mathbf{d}^{(r)}) \leqq \max_{i\in\widehat{GB}} x_j(\mathbf{d}^{(r)}) \sum_{j\in GB} p_{ij}(d_i^{(r+1)}) \leqq \max_{i\in\widehat{GB}} x_j(\mathbf{d}^{(r)}).$$

这又与 $\widehat{GB}$ 的定义相矛盾．因此，GB 不是空集是不可能的．

参　考　文　献

[1] Bellman, R., Dynamic Programming, Princeton University Press, Princeton, New Jersey, 1957.

[2] Bellman, R., A Markovian Decision Process, Journal of Mathematics and Mechanics, **6** (1957), No. 5, 679—684.

[3] Howard, R. A., Dynamic Programming and Markov Processes, John Wiley & Sons Inc., New York, 1960 (动态規划与馬尔柯夫过程，李为政等譯，上海科学技术出版社，1963).

[4] Eaton, J. H. & Zadeh, L. A., Optimal Pursuit Strategies in Descrete-state Probabilistic Systems, *Trans. of ASME, Journal of Basic Engineering*, **84** (1962), No. 1, 23—29.
[5] Blackwell, D. V., On the Functional Equation of Dynamic Programming, *Journal of Mathematical Analysis and Applications*, **2** (1961), No. 2, 273—276.
[6] Blackwell, D. V., Discrete dynamic Programming, *The Annals of Mathematical Statistics*, **33** (1962), No. 2, 719—726.
[7] Derman, C., On Sequential Decisions and Markov Chains, *Management Science*, **9** (1962), No. 1, 16—24.
[8] Derman, C., Stable Sequential Control Rules and Markov Chains, *Journal of Mathematical Analysis and Applications*, **6** (1963), No. 2, 257—265.
[9] Понтрягин, Л. С., Болтянский, В. Г., Гамкрелидзе, Р. В., Мищенко, Е. Ф., Математическая теория оптимальных процессов, ГИФМЛ, Москва, 1961.
[10] Feller, W., An Introduction to Probability Theory and its Applications, John Wiley & Sons Inc., 1957.
[11] Романовский, В. И., Дискретные цепи Маркова, ГИТТЛ, Москва, 1949 (疏散的馬尔可夫鏈，梁友棋譯，科学出版社，北京，1958).

AN OPTIMAL POLICY FOR CONTROLLING THE CONTROLLABLE MARKOV CHAINS

WU CHANG-PU

This paper is concerned with one type of the optimal Markov controlled systems. The controlled system is described by a Markov chain whose statistical property depends on the sequence of decisions that we call a policy. There exists an objective state with the property that once the system reaches this state, it remains unchanged forever. Our purpose is to choose a policy which maximizes all the probabilities that the system ever reaches this objective state from every initial state. First we give a policy-iteration method for obtaining an optimal policy over the set of stable policies. We then prove such a policy is also optimal over the set containing both stable and unstable policies.