运筹学小丛书

马氏决策浅说

董泽清　刘克著

辽宁教育出版社

·运筹学小丛书·

马氏决策浅说

董泽清
刘　克　著

辽宁教育出版社
一九八六年·沈阳

马氏决策浅说

董泽清　刘　克　著

辽宁教育出版社出版　辽宁省新华书店发行
（沈阳市南京街6段1里2号）　沈阳新华印刷厂印刷

字数：57,000　　开本：787×1092 1/32 印张：3
印数：1—1,800
1986年5月第1版　　1986年5月第1次印刷

责任编辑：杨　力　　责任校对：李晓晶
封面设计：周晓风　　插　　图：韩　梅

统一书号：7371·255　　定价：0.43 元

出版说明

运筹学是二十世纪四十年代开始形成的一门学科，是现代数学的一个重要组成部分。在科学技术迅速发展的今天，运筹学有着广泛的应用。

为了向广大读者普及运筹学知识，在中国运筹学会的关心和支持下，尤其是在徐利治教授的积极倡导和组织下，编辑出版了这套《运筹学小丛书》。

这套丛书用通俗的语言系统地介绍了运筹学中各个分支的基础知识和应用方法，其中包括规划论、对策论、排队论等方面的二十多个专题。在编写内容上，注重科学性、知识性和趣味性相结合，论述一般先从实例谈起，由浅入深，引出完整的数学理论。并且，大部分内容的引出方法都是初等的，因此，凡是具有高中以上文化程度的读者，都可以阅读。

目　录

第一章 概 述

§1.1 引 言

决策，俗称拍板、定案．人们在社会活动中会遇到各种各样的决策（选取行动、措施、方案等）问题．小至个人的日常生活问题，大至一个国家、甚至全球性的问题都需要作决策．由于问题本身性质的差别、研究的角度不同，所以研究进行科学地决策的学科已有多个．如研究确定性多阶段（或连续）最优决策的“动态规划”；研究具有无后效结构的动态随机系统最优序贯决策的“马尔可夫决策规划”；基于主观概率与效用函数有机结合之上的“决策分析”；研究两（多）人具有竞争性决策的“对策论”；研究多指标决策问题的“多目标决策”；作为管理手段的“决策支持系统”等．有人主张管理就是决策．总之，“决策”(decision) 已成为现代应用数学研究的重要对象．其中有一类决策问题，不是作一次决策就完结，而是要在一系列的（或连续的）时刻点上都要作决策，而且系统状态的转移又是随机的——事先无法确切预知．在每个观察时刻，根据观察得到的系统状态，从它可用的行动 (action，措施、方案等）集中选用其一（即作决策），系统下次可能出现的状态是不肯定的（具有某种随机规律），决策者应观察下次实际出现的状态（即

收集新信息），然后据此再作新的决策，如此一步一步地进行下去，这种决策称为“序贯决策”．有一类序贯决策，其系统状态的转移规律具有无后效性，即已知现时系统所处的状态，采取选用的行动之后，下次系统转移到哪个状态，虽然事先不能确切预知，但下次转移到的状态所服从的概率规律是已知的，且与系统以前的发展历史无关．用句数学语言来讲，就是系统状态的转移规律具有马尔可夫性．系统状态的这种转移规律与选用的行动两者，交互作用决定系统的发展进程．同时，根据观察到的状态与选用的行动，将获得一定的经济效益（报酬、费用，在工程技术上有着广泛的含义）．在各个时刻选取行动的目的是使系统运行的全过程，在某种意义下，达到最优运行效果，即选取控制系统发展的最优策略．粗略地说，策略将告诉决策者在各个时刻选取行动的规则．

研究这种状态转移规律具有无后效结构的动态随机系统的最优序贯决策的学科就是“马尔可夫决策规划”（以后简记为MDP）．由于状态转移规律的这种特殊结构，勿需每进行一步就观察系统实际出现的状态、再选取行动，这样一步一步地进行，而可事先经过全盘的理论分析，就可确定控制系统发展的最优策略．

不难想象，在任一时刻选取行动，最好依赖于在那个时刻以前的全部历史信息．因为这样会使决策者对自然界的适应能力达到最大．但人们往往是幸运的，在许多实际的序贯决策问题中进行决策时，总能不顾已收集的部分历史信息，甚至可以不顾已收集的全部历史信息，而并不损失在全过程

上的运行效果．这使问题变得大为简单．如用现代化火炮系统打敌人的高速飞行器．若用太多的历史数据来描述敌飞行器在未来短时间的运行轨迹（一般近似地看作直线），决定射击提前量，并不一定好．事实上，当敌飞行器已转弯，再用拐弯前的观察数据去计算未来短时间的轨迹（直线），显然不好．再如研究某种疾病的遗传规律，以便采取有关措施，减少发病率，也不用去查太老祖宗的患病史．这为MDP的研究提供了实际背景．

例1.1 机器最优维修策略问题．

现以简化的机器维修问题作为例子．设等周期（如一天）地考察一台运行的机器，在每周期的初始时刻观察它的运行情况．每次观察时，机器可处于两个状态之一：正常运行（记作1）；出了故障（记作2）．在任一周期，若正常运行可得收益10元，到下一周期初，仍处于正常运行的可能性为0.7，处于出故障状态的可能性为0.3，即系统状态转移是随机的．处于正常运行状态可用的行动只有一个——继续生产（记作a_1）．若处于故障状态，则有两个行动可供选用：快修（记作 a_2），需付费用5元（即收益为负5元），而该时段能修复的可能性为0.6；或常规修理（记作a_3），需付费用2元，且在该时段能修复的可能性为0.4．其状态转移图，如图1、图2所示．问题是在各个周期初，根据观察到的状

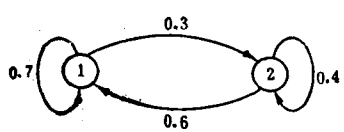

图1 快修转移图

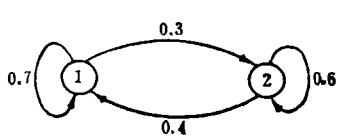

图2 常规修理转移图

态，如何选取行动，使整个考察期内的某种期望收益达最大．这就是该例的最优化问题．图中箭头↱表示转移方向，上面的数字表示相应的转移概率，i 表示编号为 i 的状态，$i=1$，2。

不难看出，每阶段的状态转移规律与报酬依赖于选用的行动a．以$q(j\mid i,\ a)$表示在时刻t观察到的系统状态为i，选用的行动为a，于$t+1$时刻转移到状态j的概率；$r(i,a)$ 表示在时刻t观察到的状态为i，选用的行动为 a 所获得的报酬．从$q(j\mid i,a)$与$r(i,\ a)$的写法已表明它们与t以前系统的历史无关，将它们的值列入表1.1—1．

表1.1—1

状态 i	行动 a	$q(j\mid i,a)$ $j=1$	$q(j\mid i,a)$ $j=2$	$r(i,a)$ （元）
1	a_1	0.7	0.3	10
2	a_2	0.6	0.4	−5
	a_3	0.4	0.6	−2

令$f(1)=a_1$，$f(2)=a_2$；$g(1)=a_1$，$g(2)=a_3$．f 与 g称为（决定性）决策规则．f 表示当观察到的系统状态为1时，选用行动a_1；当观察到的系统状态为2时，选用行动a_2．对g有类似地解释．

若在 $t=0$时所用的决策规则是从f 与g中选用一个，记作f_0，且从状态i出发，则获得报酬 $r(i,\ f_0(i))$，$t=1$ 时机

器转移到状态 j 的概率为 $q(j \mid i, f_0(i))$，$i, j=1, 2$．设 $t=1$时选用的决策规则也是 f 与g中之一，记作 f_1，由于状态转移是随机的，则获得的报酬也是随机的，其期望报酬为 $\sum_{j=1}^{2} q(j \mid i, f_0(i))r(j, f_1(j))$．$t=2$ 时转移到状态k的概率为 $q(k \mid j, f_1(j))$，$k, j=1, 2$．依此类推，得一决定性决策规则叙列 $(f_0, f_1, f_2, \cdots) \triangleq \pi$，称为一策略（严格定义在1.3节）．由于报酬是从 $t=0$ 时开始计算的，基于经济上利率$\alpha(>0)$的考虑，未来t时段的单位报酬，折合成 $t=0$ 时值 $\beta^t\left(\beta=\frac{1}{1+\alpha}\right)$．因此$t=0$时从 i 出发，长期的期望折扣总报酬为

$$r(i, f_0(i)) + \beta\sum_{j=1}^{2} q(j \mid i, f_0(i))r(j, f_1(j))$$
$$+ \beta^2\sum_{k=1}^{2}\sum_{j=1}^{2} q(j \mid i, f_0(i))q(k \mid j, f_1(j))r(k, f_2(k))$$
$$+ \cdots \triangleq V_\beta(\pi; i), \quad i=1, 2. \qquad (1.1\text{—}1)$$

这组值（$i=1$，2)是衡量该问题修理策略优劣的准则，它显然是初始状态 i 和使用的策略π的函数（认为β是给定的）．由于这种策略π 有无限多个，因此，最优策略是否存在？若存在，如何把它找出来？对不同的 π，把（1.1—1）计算出来，再找一个最优的，显然是不可行的．能否在小范围内就可找到最优策略？均是本门学科所应研究的问题．

这个例子，自然可推广到多个状态（如按磨损程度分级）的情形，修理措施（行动）也可分为许多级．

下面我们仅就周期观察的情形来叙述 MDP，即离散时间MDP。

§1.2 离散时间MDP的基本组成部分

一般说，一个离散时间 MDP，是由如下意义的五部分所组成 $\{S, (A(i), i\in S), q, r, V\}$.

（1）S为状态空间，它是被考察系统所有可能状态之（非空）集. 例1.1中 $S=\{1, 2\}$. 为简单起见，本书假定S为一有限集.

（2）$A(i)$为状态i可用的行动集，$i\in S$，例1.1中$A(1)=\{a_1\}$，$A(2)=\{a_2, a_3\}$. 为简单起见，假定对所有$i\in S$，$A(i)$均为有限集.

（3）q 是系统状态的转移律族，族的参数是可用的行动. 假定q是时齐的马尔可夫转移律族（可放宽）. $q(j|i, a)$表示在任一时刻 $t(t=0, 1, 2, \cdots)$系统处于状态 i，选用行动 $a\in A(i)$ 的条件下，于 $t+1$ 时刻转移到状态 j的概率，它与系统在t以前的历史（$2t$重状态与行动组）无关(即马尔可夫转移律)；也与时刻 t 无关（即时齐的转移律）. 具有

$$q(j|i, a)\geqslant 0,\ a\in A(i),\ i, j\in S \text{ 且}$$

$$\sum_{j\in S} q(j|i, a)=1,\ a\in A(i),\ i\in S.$$

（对折扣模型（见第三章），若改为$\sum_{j\in S} q(j|i, a)\leqslant 1$，

$a\in A(i)$，$i\in S$,理论发展是完全平行的；这表示，系统状态未找全.）

（4）令 $\Gamma=\{(i,\ a)\colon a\in A(i),\ i\in S\}$，$r$是定义在$\Gamma$上的单值实函数，称为报酬 (reward) 函数. $r(i,a)$表示在任一时刻t，系统处于状态i，选用行动 a 时，所获得的报酬（当r是费用或别的含义时，其理论发展是平行的）.

例1.1中之q与r由表1.1—1给出.

（5）准则(Criteria,指标、目标) V是定义在$\Pi\times S$上的单值实函数，其中Π是全体（允许）策略集（严格定义见1.3节)，S是状态空间. 对任意给定的策略$\pi\in\Pi$,状态$i\in S$，$V(\pi;\ i)$表示$t=0$时从状态i出发的条件下，用策略 π 的准则值，它是衡量诸策略优劣的标准. 在 1.4 节将给出几个常用的准则.

当 S、$A(i)$、$i\in S$、q、r、V均给定时，我们就认为给定了一个具体的 MDP，即给定了一个具体的模型. 在本章的余下部分，我们认为S、$A(i)$、$i\in S$、q、r 均已给定.

§1.3 策略(Policy)类

定义1 定义在状态空间S上的映象 f，使得对每个$i\in S$ 有$f(i)\in A(i)$，则称 f 为决定性决策规则，或称为决策函数. 全体决策函数所成之集，记作F.

令

$\mathcal{N}_0=\{0,\ 1,\ 2,\ \cdots\}$，$\mathcal{N}=\{1,\ 2,\ 3,\ \cdots\}$.

定义2 任一决策函数叙列

$(f_0, f_1, f_2, \cdots) = \pi$, $f_t \in F$, $t \in \mathcal{N}_0$,

就称为（决定性）一马氏策略，其中 f_t 是时刻t选用行动的规则，显然它不依赖于时刻 t 以前系统的历史．即只要知道时刻t，及t时刻系统所处的状态 i，按 π 选用的行动 $f_t(i)$ 就唯一决定了．全体马氏策略所成之集记作$\Pi^d{}_m$，称为马氏策略类．

若在时刻 $t(t \in \mathcal{N}_0)$选用行动的规则（记作 π_t）虽然不依赖于t以前系统的历史，但是随机地选用行动，即若t时刻系统处于状态i_t，选用行动a的概率为$\pi_t(a \mid i_t)$，具有$\pi_t(a \mid i_t) \geqslant 0$且 $\sum_{a \in A(i)} \pi_t(a \mid i_t) = 1$，则称这种 $\pi = (\pi_0, \pi_1, \pi_2, \cdots)$ 为随机马氏策略．全体如此策略所成之集记作Π_m，称它为随机马氏策略类．

定义 3 若在时刻t选用行动的规则π_t，不仅是随机的，而且依赖于系统在t以前的历史，则这样的 $\pi = (\pi_0, \pi_1, \pi_2, \cdots)$ 是最一般的策略，简称为策略．全体策略所成之集记作 Π，称它为策略空间．

一般策略，作为控制系统运行的规则，使用时是不方便的．因为各个时刻选取行动的规则，随时间是改变的，而且既是随机的，又要时时记住系统以前的历史．所以，对特殊的策略我们更感兴趣．

定义 4 设 $\pi = (f_0, f_1, f_2, \cdots) \in \Pi^d{}_m$，若对每个$t \in \mathcal{N}$均有 $f_t \equiv f_0$，则称它为（决定性）平稳策略，记作 f_0^∞．全体平稳策略所成之集记作$\Pi^d{}_s$，称它为平稳策略类．

设$\pi = (\pi_0, \pi_1, \pi_2, \cdots) \in \Pi_m$，若对任何 $t \in \mathcal{N}_0$ 均有

$\pi_t \equiv \pi_0$，则称π为随机平稳策略，记作π_0^∞，全体如此策略所成之集记作Π_s，称它为随机平稳策略类.

从上面的定义，不难看出各策略类之间满足如下关系：

$\Pi^d{}_s \subset \Pi_s \subset \Pi_m \subset \Pi$，$\Pi^d{}_s \subset \Pi^d{}_m \subset \Pi_m \subset \Pi$.

也不难想象F包含的元素数目与$\Pi^d{}_s$包含的元素数目一样多.

在例1.1中所用的策略为一马氏策略 $\pi=(f_0, f_1, f_2, \cdots)$。若 $f_t=g$，则在时刻 t，当系统处于状态1（即正常生产）就只有继续生产，这是由于 $g(1)=a_1$；当系统处于状态2（即出了故障），则采取快修，这是由于$g(2)=a_3$.

对该例 $\Pi^d{}_m$ 包含的元素数目为无限个，但平稳策略类$\Pi^d{}_s$仅包含两个元素f^∞与g^∞.

§1.4 准 则

令 Y_t、Δ_t 分别表示在时刻 t，所观察到的系统状态、选用的行动. 由于状态转移律是随机的，在时刻 t 选用行动的规则也可能是随机的，所以Y_t、Δ_t一般是随机变量(Y_0除外)，故$(Y_0, \Delta_0, Y_1, \Delta_1, \cdots)$为一离散时间参数的随机过程. 当然这种随机过程依赖于所用的策略. 用策略 $\pi\in\Pi$ 时，所对应的这种随机过程 $(Y_0, \Delta_0, Y_1, \Delta_1, \cdots)$，为了符号简单起见，记作$\mathscr{L}(\pi)$. 注意$\mathscr{L}(\pi)$当MDP的前三元组$\{S, (A(i), i\in S), q\}$给定之后，仅依赖所用的策略$\pi\in\Pi$，与MDP的后两元组$\{r, V\}$无关. 这也是我们不把该学科称为马尔可夫决策过程的理由.

定义 5 对任给的$\pi\in\Pi$，所对应的随机过程$\mathscr{L}(\pi)$，称为由π产生的马尔可夫决策过程.

对任给的 $\pi\in\Pi$，设由 π 产生的马尔可夫决策过程为 $\mathscr{L}(\pi)=(Y_0, \Delta_0, Y_1, \Delta_1, \cdots)$，令

$$R_t(\pi)=r(Y_t,\Delta_t),\ t\in\mathcal{N}_0, \qquad (1.4-1)$$

其中r为报酬函数. 由于$\triangle_t$、Δ_t是随机变量，所以$R_t(\pi)$也是随机变量.

常用的准则（它是衡量策略优劣的标准）有以下三种.

（1）有限阶段准则（目标）

由于$R_t(\pi)$是随机变量，要给出衡量策略优劣的标准，一个方便的办法是采取数学期望的意义. 对给定的自然数N、选用的策略$\pi\in\Pi$以及$i\in S$，令

$$V_N(\pi;\ i)=\sum_{t=0}^{N}E_\pi[R_t\ (\pi)\mid Y_0=i]$$

$$=\sum_{t=0}^{N}\sum_{\substack{a\in A(j)\\ j\in S}}P_\pi\{Y_t=j,\Delta_t=a\mid Y_0=i\}r(j,a), \qquad (1.4-2)$$

其中E_π表示对P_π求数学期望，$P_\pi\{Y_t=j,\ \Delta_t=a\mid Y_0=i\}$表示在用策略$\pi\in\Pi$，$t=0$从$i$出发的条件下，于时刻$t$转移到状态$j$、选用行动$a$的条件概率；$V_N(\pi;\ i)$意即用策略$\pi$，在$t=0$时从$i$出发的条件下，直到时刻$N$获得的期望总报酬.

设$V_N(\pi)$表示一个列向量，其第i个分量为$V_N(\pi;i)$，$i\in S$. $V_N(\pi)$就称为“N阶段准则”，或称为“N阶段目标函数”. 有时为了简便起见，我们不指明阶段数，统称为“有