马尔科夫决策规划

董泽清

中国科学院应用数学研究所
1981年7月

目 录

马尔科夫决策规划

(Markovian Decision Programming)

第一章 绪论

§1. 引言

在人类的社会实践中，会遇到各种各样的决策（措施、作用、行动等）问题。其中有一类决策问题，当人们选定决策之后，其结果是不确定的。选取决策的目的，是使得在某种意义上达到最优，从而改进作决策的进程。研究这种不确定性决策的理论称为决策分析（Decision Analysis）。如要决定是否开采某油井？可供选取的决策是"开采"或"不开采"。其结果是"有油"或"无油"均是可能的，我们根据各种资料判断得：有油的可能性为 0.3，无油的可能性为 0.7；如开采，且有油将获利 48 元，如开采，且无油则亏本 2 元；如不开采，则既不亏本也不获利。我们应该采取什么决策？如开采，其平均获利为

$$48\times 0.3-2\times 0.7=13\text{（元）}。$$

因此在平均目标的意义下，开采是有利的。如果决策者有足够大的资本，在此目标下，当有油的可能性大于 0.04 时，决定开采都是有利的。

决策分析中，其结果的不确定性是用主观概率（如上例中，有油的可能性为 0.3 等）来刻划的；各个可能结果的优劣是以"效用"（如上例中，开采有油将获利 48 元等）函数来刻划的。这里应着重指出，决策分析中使用的是"主观概率"而不是概率论中的概率，为了区别起见，后者称为"客观概率"。

客观概率是随机事件的一种客观属性，是唯一决定的。而主观概率是决策者对随机事件是否发生的一种信任程度，它是根据各种已有资料对随机事件的客观属性的一种合理判断，随着人们对随机事件的深入了解，会修改这种主观信任程度，但决不是主观臆断。如我随机地掷一枚铜板，若连续10次均掷出了正面，问下一次仍掷出正面的概率为多少？也许有人会回答仍是 $\frac{1}{2}$，为何有这样的回答呢？他是基于铜板的质量完全为均匀的假设之上。若确实如此，那么连续10次均掷出正面的客观概率 $=(\frac{1}{2})^{10}=\frac{1}{1024}\doteq 0.001$，这是一小概率事件，居然在一回（连续掷10次当一回）试验中发生了，这只能促使我们怀疑该铜板是不均匀的。因此若连续掷10次都出现了正面，那么掷第11次仍出正面的主观概率就应该大于 $\frac{1}{2}$。其具体数值可借助贝叶斯定理来确定（如见 Parff 等[1965]）。事实上这是根据样本信息对随机事件的客观属性更接近真实的反映，但带有主观的成份，当然，若客观概率已知，我们就以客观概率作为主观概率。但事实上，客观概率往往是不知道的，只好用主观概率。

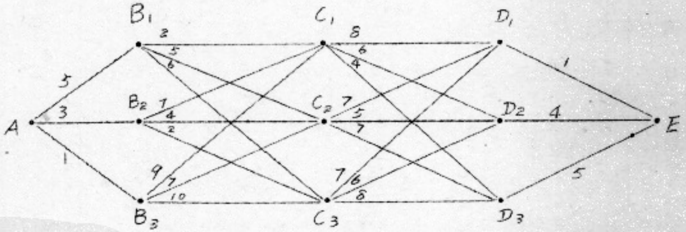
决策分析已发展了一套套较科学的方法来评定结果的主观概率，以及评定结果优劣的效用理论。但仍带有艺术品质，未严格数学化。（如见 Keeney [1978]）。

虽然决策分析也讨论多周期决策问题，用决策树的方法进行分析。但实际上仅对小数目的周期才可行。

决策分析首先成功地应用于石油和天然气工业的重大决策问题。Matheson [1969] 研究了新产品的引进，美国对火星的无人探险，以及核动力引入墨西哥国家动力系统的可能性的决策问题。总之应用范围极为广泛。

还有一类决策问题，虽然所作决策的结果是完全确定的，但必须在一系列彼此相互联系的阶段（一般理解为时间）上都要作决策，每个阶段，在各个可能状态上的诸决策均有一定的经济效益，而且所有阶段的总的经济效益是各个阶段经济效益的某种形式的和。这使得问题变得较复杂。研究这种带动态性的决策理论是“动态规划”。特别应该指出：在各个阶段作决策时，不能只顾眼前的利益，必须与长远利益结合起来考虑。只顾眼前利益是一种近视眼策略，经常不会是最优的。问题是在各个阶段应该选取什么决策，使总的经济效益达最优。

如从A地要修一条路（输油管、输气管、水渠、线路等）到E地，中间可供选择的途经地有若干个，如下图。两点连线上的数字表示距离（造价、获利等）。应如何选取一条从A地到E地的最短路线（最低造价、最大获利）。



我们用向后归纳法来求解这个问题。

令 $f_n(i)$ 表示从 i 地出发，按最短路线走，到E地还有 n 段要走的最短距离，$n=1,2,3,4$，i 取 A，B_j，C_j，D_j，$j=1,2,3$；

$r(i,j)$ 表示从 i 地到 j 地的距离。

1) $n=1$.

$f_1(D_1)=r(D_1,E)=1$，$f_1(D_2)=r(D_2,E)=4$，

$$f_1(D_3)=r(D_3,E)=5.$$

2) $n=2$

$$f_2(C_1)=\min_i\{r(C_1,D_i)+f_1(D_i)\}$$
$$=\min\{8+1,6+4,4+5\}=9.$$

故从 C_1 到 E 的最短距离为 9，在 C_1 的最优决策是到 D_1 或 D_3.

$$f_2(C_2)=\min_i\{r(C_2,D_i)+f_1(D_i)\}$$
$$=\min\{7+1,5+4,7+5\}=8.$$

故从 C_2 到 E 的最短距离为 8，最优决策是到 D_1.

$$f_2(C_3)=\min_i\{r(C_3,D_i)+f_1(D_i)\}$$
$$=\min\{7+1,6+4,8+5\}=8.$$

故从 C_3 到 E 的最短距离为 8，最优决策为到 D_1.

3) $n=3$

$$f_3(B_1)=\min_i\{r(B_1,C_i)+f_2(C_i)\}=11$$

故从 B_1 到 E 的最短距离为 11，最优决策是到 C_1.

$$f_3(B_2)=\min_i\{r(B_2,C_i)+f_2(C_i)\}=10,$$

故从 B_2 到 E 的最短距离为 10，最优决策是到 C_3。

$$f_3(B_3)=\min_i\{r(B_3,C_i)+f_2(C_i)\}=15$$

故从 B_3 到 E 的最短距离为 15，最优决策是到 C_2。

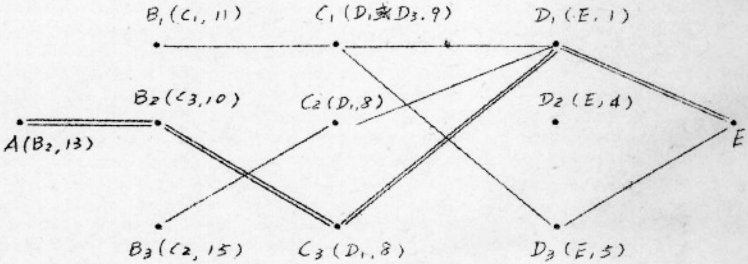
4) $n=4$

$$f_4(A)=\min_i\{r(A,B_i)+f_3(B_i)\}=13,$$

故从 A 到 E 的最短距离为 13，最优决策是到 B_2，而 B_2 的

最优决策是到 C_3，C_3 的最优决策是到 D_1，D_1 到 E．故最优路线为 $A_1 \longrightarrow B_2 \longrightarrow C_3 \longrightarrow D_1 \longrightarrow E$．

在各点标上它的最优决策，及它到E的最短距离，仅保留最短路（子路），得下图（从A到E的最短路线用双线标出）。标出这个图是便于作数值计算的。



我们为何能用如上的向后归纳法求解呢：这是因为有贝尔曼的最优化原则所保证。所谓最优化原则，对于上述例子，要为："最短路的子路（即部分路）也是最短路"，用最优化原则导出 $f_n(i)$ 满足的泛函方程，然后从泛函方程求最优解．

我们应着重指出：向后归纳法，丰富了求解内容，我们不仅求出了从A到E的最短路线，而且同时求出了任一地点到E的最短路线；同时它比穷举计法（即把所有从A到E的可能的路线的距离均求出来，然后选一个最短的路线）优越多。它把指数型（乘积型）的计算量（上例中为 $3 \times 3 \times 3 = 27$ 条路）化为线性（加法型）的计算量（上例中 $3+3+3=9$）．问题越大，向后归纳法的优越性越明显．

动态规划已成功地解决了许许多多的实际问题，如见 Bellman 与 Dreyfus〔1962〕，但动态规划的数学基础还不够严格，"最优化原则"成立的范围有多大，至今未搞清楚。

在上例中A，B_j，C_j，D_j及E称为状态。A有三个可供选择的决策：下阶段到B_1或B_2或B_3，D_1，D_2均只有一个决策（即无决策可供选择）：下阶段到E，等等。

还有一类决策问题，既要在一系列彼此联系的时刻上均要作决策，而且每次决策的结果也是不确定的。这使问题变得更加错综复杂。必须每经历一个时段，就观察实际发生的结果，即收集新的信息，然后再选取新的决策。也就是要作序贯决策。马尔科夫决策规划就是研究一类特殊的（状态转移规律具有无后效性）序贯决策问题。

在序贯决策中，如果存放和提取信息的费用可忽略不计，则显然在给定时刻选取决策，最好依赖于在那个时刻以前收集的全部历史信息，即人类应该作序贯决策，因为，只有这样做才能使人类对自然界的适应性达最大。但人们往往是幸运的，对许多实际问题，人类总能不顾已收集的部分历史信息，甚至不顾已收集的全部历史信息，因此作决策时，只依赖余下那部份历史信息已足够了，这使问题又变得简单一些。如用现代大炮打高速敌机，若用太多的信息描述敌机的运行轨迹，并不一定好。比如敌机已转弯，再用转弯前的观察数据去拟合轨迹，反而不好。在研究某种疾病的遗传规律时，也不用去查太老的祖宗的患病史。

马尔科夫决策规划，从一开始就置于严格的数学基础之上，成为随机运筹学的一个主要分枝，也可以说是应用数学的一个分枝。它是使用客观概率。

§2　马尔科夫决策规划研究内容及例

马尔科夫决策规划（以后简记为MDP）是研究如下动态

概率系统的最优化问题。该系统可连续地或周期地被观察，在观察时刻，决策者根据观察到的状态，从可用的决策集中选取其一，并予以实施，则有两个结果：1)系统状态转移的概率规律就确定了，且与该时刻以前的历史无关，即具有马尔科夫(无后效)性；2)将获得一定的经济效益（可以是钱，也可以是物在工程技术领域，有着广泛的解释），也与历史无关。系统发展的不同路径将获得不同的经济效益。问题是在各个时刻应如何选取决策，使系统处于最优运行。即选取最优策略，所谓策略，它将告诉决策在各个时刻如何选取决策，当然"最优"与衡量策略优劣的指标有关。

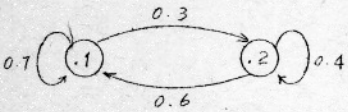
例 1. 存贮最优化问题

周期地（如以月为周期）检查某物品的存贮量（即系统的状态），每次检查之后，可供选取的决策是物品允许的定货量，把新进之货加到已有的存贮中，系统的转移律由相继检查时刻之间对该物品的（随机）需求模型所决定。比如，各时期的需求量是独立、同分布的随机变量，其分布是已知的。所涉及的经济效益有：新增物品的定货费，库存物品的存贮费，不足物品的损失费等。衡量目标（指标）一般是每单位时间的平均费用。今后我们将看到：在相当一般的条件下，存在一对正整数 s^*, S^* ($s^* < S^*$)，当检查时发现存货量 $\leq s^*$，则进货到 S^*；当检查时的存货量 $> s^*$，则不进货，乃一最优策略。

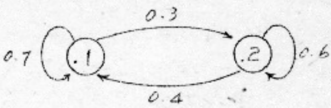
例 2. 机器维修的最优化问题

我们周期地（如一小时）观察该机器，有两个可能的状态：正常生产（以 1 表示），出了故障（以 2 表示）。发生的故障可修理复原。在任何周期，如机器生产，可获报酬 10 元，下一周期仍处于状态 1 的概率为 0.7，而转移到状态 2 的概率为 0.3。如机器发生了故障，我们有两种措施可供采取：快修记

作1），需要费用5元（即报酬为-5元），一周期能修好的概率为0.6；另一个是常规修理（记作2），需要费用2.5元，一周期能修好的概率为0.4，转移图如下：



快修转移图



常规修理转移图

其中箭头"⌒"表示转移方向，上面的数值表示相应的转移概率。(i) 表示状态 i，$i=1,2$。

问题是在各个时刻应如何决策，使某种总的平均报酬达最大。

这个问题可推广为：机器可分为 $L+1$ 个状态 $\{0,1,\cdots,L\}$，状态 i 表示机器处于 i 级磨损，$i=1,2,\cdots,L-1$，0 表示新机器，L 表示不能用的坏机器，在状态 i 可供选择的措施是：照常生产，或更换成一个新机器，其费用为 C_i（$C_i<C_{i+1}$，$i=1,2,\cdots,L-1$），更换成新机器的费用为 C_L。如机器磨损服从马尔科夫规律，今后我们将看到，在相当弱的条件下，存在一个 i^*，当检查发现机器的状态 $i<i^*$，则照常生产；当 $i\geq i^*$ 时，则更换为新机器，将是一个最优策略。

§3. 基本假设和定义

我们首先研究周期地观察系统的情形，为简便计，我们假定在时刻 $t=0,1,2,\cdots$ 处观察系统（t 可以只取有限值）

如无特殊声明，我们总假定系统状态的转移律族是时间上齐次的（次）随机律。

（离散时间）MDP是由如下定义的五重组$\{S, (A(i), i\in S), q, r, V\}$所构成：

1. S是一非空集，称为系统的"状态集"或"状态空间"，S的元素称为"状态"，今后将用有足标或无足标的小写英文字母$s, i, j, \cdots$来表示。如无特殊声明，我们总假设S为一可列集（即有限集或可数无限集），关于S的结构不作任何假定。

2. $A(i)$ $(i\in S)$也为一非空集，称为状态i可用的"决策"（作用、行动、措施等）集。$A(i)$的元素称为"决策"。今后将用有足标或无足标的小写英文字母$a, b, \cdots$表示。如无特殊声明，我们总假定$A(i)$为一可列集。

3. 对每个$t\geq 0$，令

$$h_t=(i_0, a_0, i_1, a_1, \cdots, s_t, a_t),\quad a_n\in A(i_n),$$
$$i_n\in S,$$
$$n=0, 1, \cdots, t.$$

称为系统直到时刻t的一个"历史"。全体如此历史所成之集记H_t，称为"系统直到时刻t的历史集"。

q是一族时间上齐次的马尔科夫转移律，即对任何$h_{t-1}\in H_{t-1}$，$i_t\in S$，$a_t\in A(i_t)$，$q(\cdot|h_{t-1}, i_t, a_t)$与$q(\cdot|i_t, a_t)$均有定义且

$$q(j|h_{t-1}, i_t, a_t)=q(j|s_t, a_t),\quad j\in S,\quad t=0, 1, 2, \cdots$$

其中h_{-1}表示无历史，以及

$$q(j|i_t, a_t)\geq 0,\quad \sum_{j\in S} q(j|i_t, a_t)\leq 1$$

即对任给的 $i_t \in S$，$\{q(j|i_t, a_t): j\in S\}$ 为一族次随机向量，其参数为在 i_t 处可用的决策 $a_t \in A(i_t)$。

用话来说，即每逢系统处于状态 i，选取决策 $a\in A(i)$，则不管系统的历史如何，下次转移到状态 j 的概率为 $q(j|i,a)$。这个概率有时写作 $q_{ij}(a)$。

4. 令 $\Gamma=\{(i,a): a\in A(i), i\in S\}$，$r$ 是定义在 Γ 上的单值实函数，称为系统的"报酬函数"（注意，负的报酬就是费用）。每逢系统处于状态 i，选取决策 a，则我们将获得（一周期期望）报酬 $r(i,a)$。它与系统的历史无关。如无特殊声明，我们总假设 r 是有界的。即存在一个正数 M，使得 r 的绝对值 $|r(i,a)|\le M$，$a\in A(i)$，$i\in S$，（以后我们将放弃这个假设）。

5. V 是定义在 $S\times\Pi$*) 上单值实函数，称为系统的"目标函数"或"最优性准则"。其中 Π 是全体策略所成之集，严格定义将在下面给出。常用的目标函数将在§5中给出。

研究的问题有：在各个时刻如何选取决策（即选取什么策略）使 $V(i)$，对所有 $i\in S$ 同时达到极值，这种策略是否存在？或在什么意义下存在？能否在较小的策略类中找到？怎样把一个复杂的问题化为一个简单的问题？以及获得最优策略的算法，特别是对特殊模型的特殊有效算法等等。

注意我们只研究目标函数达最大值的问题。如要求最小值，只要用 $-r$ 代替 r，仍求最大值即可，或把 r 解释为费用，理论发展完全是平行的。

注1. S，$A(i)$（$i\in S$）均可为有限集，至少一个为可数无穷集，甚至各种各样的非可数无限集。转移律族 q 的时间

---

*) $S\times\Pi$ 表示集 S 与集 Π 的笛卡儿乘积集。

参数可以是离散的，连续的；有限时段，无限时段；β可以是时齐的、非时齐的，半马尔科夫的；β也可以是未知的。报酬函数r可以是有界，各种各样无界的，甚至r与系统的历史有关。状态信息可以是部份可观察的，状态信息延迟的。目标函数也可以是各种各样的。几种成份的任何一个组合均构成一个可研究的模型。

注2. 我们之所以假定S为一可列集，一方面是尽可能照顾到应用的广泛性；另一方面也是为了避免一般状态空间带来的测度论上的纷扰。因为S非可数时，许多叙述的合理性应当予以证明。

注3. 在实际应用时，对于状态i可用的决策a，与对于状态$j(\neq i)$可用的同一决策a，所代表的实际内容可以完全不同。a仅表示在可用的决策集中编号为a的决策。

一个"策略"（规则，计划）（policy, Strategy, Plan, rule）π，是一个序列$\pi=\{\pi_0, \pi_1, \pi_2, \cdots\}$，其中对每个$t\geq 0$，$h_{t-1}\in H_{t-1}$，$i_t\in S$，$\pi_t(\cdot|h_{t-1}, i_t)$是$A(i_t)$上的一个概率分布。其概率意义为：当用策略$\pi$时，系统历史$h_{t-1}$已发生，在时刻$t$系统到达状态$i_t$的条件下，选取决策$a\in A(i_t)$的概率为$\pi_t(a|h_{t-1}, i_t)$。即$\pi_t$是时刻$t$选取决策的规则。我们总有$0\leq\pi_t(a|h_{t-1}, i_t)\leq 1$，且有$\sum_{a\in A(i_t)}\pi_t(a|h_{t-1}, i_t)=1$。注意$h_{-1}$理解为无历史。全体策略之集记作$\Pi$。一个决定性策略，意即$\pi=\{\pi_0, \pi_1, \pi_2, \cdots\}$的每个$\pi_t$均为退化分布，即对每个$h_{t-1}\in H_{t-1}$，$i_t\in S$总存在一个$a_t\in A(i_t)$使得$\pi_t(a_t|h_{t-1}, i_t)=1$，$t\geq 0$。全体决定性策略所成之集记作$\Pi^d$。

我们对特殊策略特别感兴趣，一个策略$\pi=\{\pi_0, \pi_1, \pi_2\cdots\}$，

如果对每个 $t \geq 0$，它的 π_t 只依赖于时刻 t 所处的状态 i_t，即 $\pi_t(\cdot|h_{t-1}, i_t) \equiv \pi_t(\cdot|i_t)$，则称为"随机马氏策略"，有时也称它为"无记忆策略"；全体如此策略所成之集称为"随机马氏策略类"，记作 Π_m。一个随机马氏策略 $\pi=\{\pi_0, \pi_1, \pi_2, \cdots\}$，如果它的每个 π_t 均是一个退化概率分布，则称它为"（决定性）马氏策略"，全体马氏策略所成之集称为"马氏策略类"，记作 Π_m^d。

定义在 S 上的映像 f，映 i 入 $A(i)$，即 $f(i) \in A(i)$，$i \in S$，则称 f 为一"决策函数"。全体决策函数所成之集，记作 F，不难看出一个马氏策略 $\pi=\{\pi_0, \pi_1, \pi_2, \cdots\}$ 必存在一串 $f_n \in F$，使得 $\pi=\{f_0, f_1, f_2, \cdots\}$。

一个随机马氏策略 $\pi=(\pi_0, \pi_1, \pi_2, \cdots)$，如果它的每个 π_t 均与 t 无关，即 $\pi=\{\pi_0, \pi_0, \pi_0, \cdots\}$，则称它为"随机平稳策略"；全体如此策略所成之集称为"随机平稳策略类"，记作 Π_s。一个随机平稳策略 $\pi=\{\pi_0, \pi_0, \pi_0, \cdots\}$，如果 π_0 是退化的，即存在一个 $f \in F$，使 $\pi_0 \equiv f$，则称它为"平稳策略"，全体平稳策略所成之集称为"平稳策略类"，记作 Π_s^d。平稳策略 $\pi=\{f, f, f, \cdots\}$ 有时写作 f^∞。显然平稳策略是一特殊马氏策略。

从上面的定义，我们不难看出有如下结论：

1) $\Pi^d \subset \Pi$，$\Pi_m^d \subset \Pi_m$，$\Pi_s^d \subset \Pi_s$；

2) $\Pi_s^d \subset \Pi_m^d \subset \Pi^d$，$\Pi_s \subset \Pi_m \subset \Pi$；

3) Π_s^d 与 F 包含的元素数目是相等的。

有些作者（如 Blackwell［1962］，mine & Osaki［1970］）仅在 Π_m^d 类上研究了所提的模型。有些作者（如 Howard［1960］）在更小的 Π_s^d 类上研究了所提的模型。一