运筹与管理科学丛书 22

马尔可夫决策过程理论与应用

刘 克　曹 平　编著

科 学 出 版 社

运筹与管理科学丛书 22

马尔可夫决策过程理论与应用

刘　克　曹　平　编著

科　学　出　版　社

北　京

内 容 简 介

本书从马氏决策的一般理论出发，介绍了马氏决策的基本概念，给出了决策过程的表述方法并介绍了不同准则条件下的基本理论，还给出了作者对一些实际问题的研究心得，为读者提供参考. 本书在《实用马尔可夫决策过程》一书的基础上增加了 Bandit 过程、部分可观察过程、软件可靠性建模分析以及大规模计算方法等章节，为读者提供更为宽阔的视野.

本书可作为高等院校高年级大学生和研究生的教材，也可作为运筹学、管理科学、信息科学、系统科学以及计算机科学和工程领域的学者和技术人员的参考书.

图书在版编目(CIP)数据

马尔可夫决策过程理论与应用/刘克，曹平编著. —北京：科学出版社，2015.1

(运筹与管理科学丛书；22)

ISBN 978-7-03-043123-3

Ⅰ. ①马… Ⅱ. ①刘… ②曹… Ⅲ. ①马尔可夫决策 Ⅳ. ①O225

中国版本图书馆 CIP 数据核字 (2015) 第 017453 号

责任编辑：赵彦超／责任校对：钟 洋
责任印制：张 倩／封面设计：王 浩

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

安泰印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2015 年 2 月第 一 版 开本：720×1000 1/16
2015 年 2 月第一次印刷 印张：18 1/4
字数：360 000

定价：98.00 元

(如有印装质量问题，我社负责调换)

《运筹与管理科学丛书》序

运筹学是运用数学方法来刻画、分析以及求解决策问题的科学. 运筹学的例子在我国古已有之, 春秋战国时期著名军事家孙膑为田忌赛马所设计的排序就是一个很好的代表. 运筹学的重要性同样在很早就被人们所认识, 汉高祖刘邦在称赞张良时就说道:“运筹帷幄之中, 决胜千里之外.”

运筹学作为一门学科兴起于第二次世界大战期间, 源于对军事行动的研究. 运筹学的英文名字 Operational Research 诞生于 1937 年. 运筹学发展迅速, 目前已有众多的分支, 如线性规划、非线性规划、整数规划、网络规划、图论、组合优化、非光滑优化、锥优化、多目标规划、动态规划、随机规划、决策分析、排队论、对策论、物流、风险管理等.

我国的运筹学研究始于20世纪50年代, 经过半个世纪的发展, 运筹学研究队伍已具相当大的规模. 运筹学的理论和方法在国防、经济、金融、工程、管理等许多重要领域有着广泛应用, 运筹学成果的应用也常常能带来巨大的经济和社会效益. 由于在我国经济快速增长的过程中涌现出了大量迫切需要解决的运筹学问题, 因而进一步提高我国运筹学的研究水平、促进运筹学成果的应用和转化、加快运筹学领域优秀青年人才的培养是我们当今面临的十分重要、光荣, 同时也是十分艰巨的任务. 我相信,《运筹与管理科学丛书》能在这些方面有所作为.

《运筹与管理科学丛书》可作为运筹学、管理科学、应用数学、系统科学、计算机科学等有关专业的高校师生、科研人员、工程技术人员的参考书, 同时也可作为相关专业的高年级本科生和研究生的教材或教学参考书. 希望该丛书能越办越好, 为我国运筹学和管理科学的发展做出贡献.

袁亚湘

2007 年 9 月

前　言

马氏决策过程 (Markov decision processes) 的基本数学模型框架是 20 世纪 50 年代由美国学者 Howard 教授最早提出的, 在一定程度上与 Bellman 提出的动态规划 (dynamic programming) 模型有些类似. 经过大量学者的深入研究, 它已经成为随机优化的主要组成部分, 也被人们称为受控马尔可夫链 (controlled Markov chain)、随机控制问题 (stochastic controlled problem)、马氏决策规划 (Markov decision programming) 等等.

马氏决策过程是研究多阶段决策问题的一种方法, 其基本思想是着重于决策过程的形象化描述, 包括决策时刻、系统状态、行动、报酬和转移概率等组成因素. 如果直接观测决策过程的一个阶段, 就会发现每个阶段都具有如下的逻辑特征: 在观察到系统的一个状态后, 选取一个决策行动, 因此产生一个报酬并且通过转移概率函数决定下一个决策阶段的状态. 所以, 马氏决策的基本模型需要描述系统的状态、决策者能够采用的决策行为、与决策行为相关的收益或者费用, 以及系统状态在决策行为的干预下状态发生转移的规律等等. 当然, 决策的目的是要优化决策者关心的目标函数, 使得目标函数极大化的系列决策方案构成最优策略. 除此以外, 对整个系统, 还要进一步分析如下内容:

(1) 提供一些条件以保证存在易于操作的最优策略;

(2) 确定如何辨别出这些策略;

(3) 寻求得到这些策略的有效算法;

(4) 建立这些算法的收敛性质.

实际上, 策略的比较分析强烈地依赖于准则的不同. 因此, 本书将根据不同的准则分别讨论.

到目前为止, 马氏决策过程的理论和应用得到了长足的发展, 已经在生态科学、经济理论、通信工程以及众多学科中得到了应用, 而这些新的应用也为其带来了丰富的理论结果.

本书是在 2004 年出版的《实用马尔可夫决策过程》的基础上, 参考了 2009 年出版的《摄动马尔可夫决策与哈密尔顿圈》的部分内容, 以及总结了我们在中国科学院大学授课的经验修订而成的. 与《实用马尔可夫决策过程》的不同之处有: 在第 2 章中增加了有限阶段的部分可观察马氏决策部分; 在第 3 章中增加了吉廷斯指标问题; 增加了第 9 章软件测试的最优发布问题和第 10 章大规模问题的近似算法; 我们还在前六章增加了一些练习题, 并在附录中给出了参考答案.

本书的写作有三个目的：一个是为理论研究者提供参考, 为高等院校有关专业的高年级大学生和研究生提供教材; 另一个是希望本书的内容能够引起管理者、计算机科学工作者、经济学家、应用数学家、控制与通信工程方面的工作者、信息科学与工业工程等方面的学者和技术人员的兴趣; 最后是想通过增加的练习题, 为读者熟悉这些理论方法提供帮助.

本书特别地利用大量篇幅介绍了一些问题是如何被建立为马氏决策过程模型并求解的, 这样可以为应用工作者提供方便的建模思想, 能够拓广读者的思维. 本书需要读者适当地熟悉一些数学分析、线性代数、概率论、随机过程和线性规划等方面的知识, 不过作者力求语言浅显易懂, 对繁杂的证明只给出证明的思路, 并且注明参考文献, 便于感兴趣的读者进一步学习.

本书的写作得到了国家自然科学基金 (11271356, 71390334, 71401159) 和 973 项目 (2010CB731400) 的部分资助, 在此作者表示衷心的感谢.

同时, 感谢所有被本书直接或间接引用其文献资料的同行学者们.

最后, 感谢作者的研究生范萌萌、杨雅婷、白玉真、谢新丽、杨珂和蔡爽等所做的大量工作.

刘 克 曹 平

2014 年 7 月

常用符号表

MDP, MDPs	马氏决策, 马氏决策过程
$A, A(i)$	行动空间，行动集
i, j	系统状态
S	状态空间，状态集
$\mathrm{Dis}(A)$	集 A 上的概率分布集合
$\mathcal{Z}$	全体正整数
$\mathcal{Z}_0$	全体非负整数
$\mathcal{A}$	A 上的 σ- 代数
$\Re$	实数集合
$\lVert\cdot\rVert$	范数
$\lVert\cdot\rVert_w$	以 w 为权重的上界范数
T_f, T_π, T, L	算子
Ł(π), Ł$_S(\pi)$	策略 π 诱导的状态行动过程和状态过程
$\mathrm{ext}(X)$	集合 X 的支撑集合
σ_π^2	策略 π 的稳定状态方差
$V_N(i,\pi)$	策略 π 的 N 阶段期望总报酬函数
$V_\beta(i,\pi)$	策略 π 的折扣期望总报酬函数
$V_\alpha(i,\pi)$	半马氏模型中策略 π 的折扣期望总报酬函数
$\bar{V}(i,\pi)$	策略 π 的平均报酬函数
$V(i,\pi)$	策略 π 的折扣权重报酬函数
$\omega(i,\pi)$	策略 π 的折扣与平均权重报酬函数
$F_n^\pi(i,x)$	策略 π 的终达目标最小风险有限阶段目标函数
$F_\infty^\pi(i,x)$	策略 π 的终达目标最小风险无限阶段目标函数
$G_n^\pi(i,x)$	策略 π 的首达目标最小风险有限阶段目标函数
$G_\infty^\pi(i,x)$	策略 π 的首达目标最小风险无限阶段目标函数
$u_\alpha(i,\pi)$	策略 π 的连续时间折扣报酬函数
$U(i,\pi)$	策略 π 的连续时间平均报酬函数

$sp(v)$	向量 v 的跨度
$H_t,\ H_\infty$	历史集合
$\Xi^\pi(\alpha)$	关于策略 π 和初始分布 α 的状态–行动极限平均频率的集合
$\Xi_s(\alpha)$	随机平稳策略类的状态–行动极限平均频率的集合
$\Xi_s^d(\alpha)$	平稳策略类的状态–行动极限平均频率的集合
$\Xi_m^d(\alpha)$	马氏策略类的状态–行动极限平均频率的集合
$\Xi_m(\alpha)$	随机马氏策略类的状态–行动极限平均频率的集合
$(\Xi)^c$	欧氏空间中子集 Ξ 的闭凸包
Π	策略类, 最一般的策略集合
Π_m	随机马氏策略类
Π_m^d	确定性马氏策略类, 或简称马氏策略
Π_s	随机平稳策略类
Π_s^d	确定性平稳策略类, 或简称平稳策略
Π_0	与目标值无关的策略全体
F	决策函数集合, 在不混淆的情况下也表示平稳策略类

目　　录

第1章　引　　论

做决策是人们在日常生活和生产实践中经常遇到的问题. 人们也总希望做出的决策能够达到最优的效果. 事实上, 人们在做决策的时候需要考虑很多影响决策效果的因素, 如当前决策立即显现出的效果、当前决策行为对长远利益的影响等等. 因此, 做决策不是孤立的, 也就是说今天的决策会影响到明天, 而明天的决策会影响到将来. 如果不顾及对将来的影响而只考虑当前的利益做决策 (即采用近视眼策略), 从长远的角度来看, 通常效果不会很好. 比如说长跑运动员, 要根据需要跑的距离而合理分配自己的体力, 以避免尚未跑完全程就筋疲力尽.

本书描述和研究了在不确定环境下的一类序列决策模型, 决策者不仅要考虑决策结果的即时效应, 还要考虑到为将来做决策创造机会. 看上去这个模型比较直观且不复杂, 但是它的应用极其广泛, 而且产生了丰富的数学理论. 这一章主要通过一些例子来说明做决策过程中的关键因素、这些因素之间的关系以及决策过程的动态表现, 然后给出马氏决策过程的一般记号与定义, 最后叙述了马氏决策过程的发展简史和一些比较有影响的相关书籍.

1.1　序列决策模型

我们用图 1.1.1 描述多阶段决策过程的一个完整步骤. 在时刻 t, 控制系统的决策者观察到系统当前所处的状态, 并根据这个状态从可行的决策行为集合中选取一个决策行为 (我们称选择一个行动). 之后, 该行动会对系统的后续运行产生两方面的影响：一方面是产生了一个既得的报酬或费用, 而另一方面是系统的状态会按照与这个行动有关的一个概率规律在下个阶段即在 $t+1$ 时刻转移到一个新的状态. 这时决策者面临着与开始时 (即时刻 t) 相同的问题, 也就是选取 $t+1$ 时刻的决策行为. 依此循环下去, 不同的只是在不同时刻系统的状态可能是不同的, 而且可采用的行动集合随着状态的变化也会发生相应的变化.

我们可以把这个序列决策过程的关键因素列举出来：

- 所有的决策时刻点集.
- 系统的所有可能的状态集合.
- 可以采用的全体行动集合.
- 与状态和行动相关联的既得报酬或费用集合.
- 与状态和行动相关联的转移概率的集合.

一般来说, 我们总认为决策者在开始做决策的时候这些量是已经知道的.

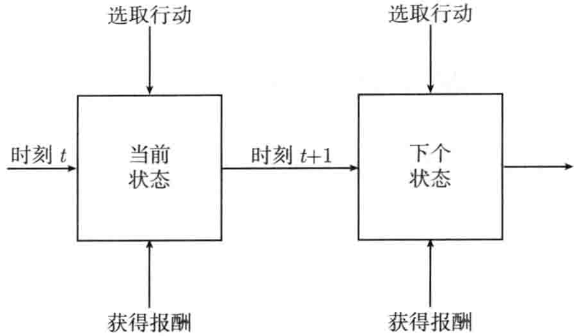

图 1.1.1　决策过程的图示

这样, 我们就可以描述一个不确定的序列决策过程. 在每一个决策时刻, 系统的状态为决策者提供了选取行动的一切必要信息, 其中包括在这个状态上的有效的行动集合. 作为选取行动的结果, 有两件事情发生: 决策者得到既得报酬和系统的状态依照一定的概率规律在下一个决策时刻转移到一个可能的新状态, 当然报酬和转移概率都是依赖于当时的状态和在这个状态上决策者选取的行动的. 这个过程随着时间的推移, 决策者可以得到一个报酬序列.

从另一个角度来看, 在每个决策时刻, 系统可能的每一个状态在决策过程中都是有可能出现的. 因此, 针对每个有可能的不同的状态, 决策者要选取与状态相对应的行动, 我们把在一个特定的决策时刻在每个可能的状态上选取行动的原则称为决策规则. 决策规则不仅依赖于当前状态, 而且还有可能依赖于以前所经过的那些状态和在那些状态上所选取的行动. 我们把在将来任意决策时刻的任意可能状态上选取行动的规则称为策略. 一个策略实际上就是一个决策规则的序列. 因此一个策略产生了一个报酬序列. 而序列决策问题就是要在第一个决策时刻之前就预先选好策略, 使得报酬序列的某个函数值 —— 准则在这个策略下达到最大. 准则的选取要由决策者权衡各方面的利弊而决定. 比较常用的准则有折扣期望报酬准则和平均报酬准则等.

表面上看, 上述的描述似乎是一个静态的表述. 事实上, 如果决策状态考虑得非常完整 (事先预料到所有可能出现的状态), 就可以做到动态地随着过程的发展 (即过程的状态变化) 采用适当的决策行为, 达到动态调控系统的目的.

本书讨论一类特殊的序列决策问题 —— 马氏决策过程模型, 其特点是可采用的行动集、既得报酬和转移概率只依赖于当前的系统状态和选取的行动, 与过去的历史无关. 尽管看上去似乎有些过于受限制, 其实这种模型已经涵盖了大部分的序列决策模型. 如果读者熟悉扩充状态空间的方法, 会更加深刻地认识到这一点.

1.2 马氏决策过程的例子

下面给出几个例子说明马氏决策过程模型的动态过程, 特别是决策过程中的主要成份.

例 1.1 *机器最优维修策略问题.*

等周期 (如一天) 地观察一台运行的机器, 以初始观察到的运行情况作为机器这一周期的状态. 根据运行情况, 假设机器可处于两个状态: 正常运行 (记作 $i=1$) 和出了故障 (记作 $i=2$). 在任何一个周期, 如果机器正常运行决策者可以获得收益 10 元, 同时到下一个周期初, 机器仍处于正常运行的概率为 0.7, 发生故障的概率为 0.3. 当机器处于正常运行状态时, 决策者可以采取的行动只有一个, 即继续生产 (记为 a_1). 如果机器处于状态 2(出了故障), 决策者有两个行动可供选择: 一个是快修 (记为 a_2), 费用是 5 元 (即收益为 -5 元), 而该时段能将系统修复为正常运行状态的概率为 0.6; 另一个是常规修理 (记为 a_3), 费用是 2 元, 且在该时段能修复的概率为 0.4. 用 $p(j|i,a)$ 表示 t 时刻观察到的系统状态是 i, 选用行动 a, 于 $t+1$ 时刻转移到状态 j 的概率; 用 $r(i,a)$ 表示在时刻 t 观察到的状态为 i 并选用行动 a 所获得的报酬, 则把上面的数据整理为表 1.2.1.

表 1.2.1 转移概率和报酬

状态 (i)	可用行动 (a)	转移概率 $p(j\|i,a)$ $j=1$	$j=2$	报酬/元 $r(i,a)$
1	a_1	0.7	0.3	10
2	a_2	0.6	0.4	-5
	a_3	0.4	0.6	-2

需要解决的问题是: 在各个周期初, 根据决策者观察到系统实际运行的状态后, 应该如何选取行动才能使整个考察期内的收益最大.

例 1.2 *库存管理.*

马氏决策过程的模型被广泛应用到库存控制问题中, 其实库存问题也是马氏决策过程最早应用的领域之一. 这些应用的范围从单一产品订货点的确定到多产品、多中心供货的网络控制, 应有尽有. 在 20 世纪 90 年代末兴起的供应链管理(supply chain management) 中, 马氏决策过程也是有用的工具. 随机运筹中最早也是最关心的问题之一就是在参数的各种假设下最优策略所具有的形式. 我们给出这种类型的一个应用.

通过当地的代理商, 加拿大 Tire 公司运作着一个为全加拿大提供汽车的供应链. 在太平洋地区的 21 个商店由一个管理集团管理. 为这 21 个商店提供后援库存

的中心仓库在 British Columbia 州的 Burnaby, 该仓库存有 29000 余件产品, 并且周期地为这 21 个商店分别供货以保证每个商店维持安全的库存量.

库存量补充的时间是随着商店的规模而变化的. 作为一个“小”的商店, 每种产品的库存量一周盘点一次, 并且根据盘点的库存量 (手上的现货) 决定该种产品的进货量, 进货 3 天内可以到达. 对每一件订货, 要开销一个固定的费用, 包括仓库的占用费和商店的上架费等. 另外, 还要开销固定的 (与订货量无关的) 订货费用和在商店里的每日保存费. 商店的管理者还要求用手上现货满足需求的比例不能低于 97.5%.

考虑一个商店的单一产品库存问题, 可以用一个马氏决策过程来确定其最优的订货点和最优的订货量. 决策时刻是每周的盘点时间, 系统的状态是盘点时商店里的产品库存量. 在给定的状态下, 可以采用的行动就是中心仓库能够为这个商店提供的可能的产品数量. 状态的转移概率依赖于决策者的订货量和下一周的需求量. 一个决策规则就是订货的数量, 它是盘点时库存量的函数; 一个策略就是由这样的一个函数序列组成, 它可以告诉决策者任何时候如果系统处于任一个状态时, 决策者应该如何做决定 (即确定其订货量). 决策者要寻找一个订货策略, 在保证满足顾客需求的概率不低于某给定值的条件下使长期的平均费用达到最小.

在这个问题中, 最优策略应该具有这样的性质: 易于操作而且不随时间变化. 如果没有满足顾客需求的这种概率约束, 人们证明了存在具有这样性质的最优策略: 在商店的库存水平低于某个固定的界限时, 订货到一个目标水平, 否则就不订货. 这种策略也被称为 (s, S) 策略. 如果考虑了约束条件, 则上面的策略不一定是最优的.

较好的库存控制会有效降低费用, 这一点绝没有被过分夸大. 正如 Britain's Cadbury-Schweppes PCL 的总裁 Graham Day 先生在 *The Globe and Mail* (1992) 中写道: “我相信任何具有库存的企业最容易省钱的地方就是库存的极小化.”

例 1.3 高速公路管理问题.

美国高速公路的实际案例是马氏决策过程的成功案例之一. 这个例子取材于 Golabi 等[81] 的文章, 也可以参见文献 [152].

美国 Arizona 州的交通部门 (简记为 ADOT) 管理着 7400 英里的公路网络. 直到 20 世纪 70 年代中期, 它的基本工作是新路网络的建造. 当 Arizona 的公路网络基本建成时, ADOT 的主要工作转变为公路的维护. 从 1975 年到 1979 年, 高速公路的维护费用从 2500 万美元增加到 5200 万美元, 翻了一番, 从趋势上看还要继续增加. 当然, 如何分配这些资金成为了 ADOT 的核心工作. 1978 年, ADOT 和旧金山的 Woodward-Clyde 咨询公司合作开发了基于马氏决策过程的公路管理系统, 用以分配这有限的资源并保证公路维护的质量. 1980 年, 利用这一系统的第一年, 就节省了 1400 万美元, 而且路的质量没有任何下降, 这几乎是 Arizona 州当年公路维