摄动马尔可夫决策与哈密尔顿圈

Perturbed Markov Decision Processes and Hamiltonian Cycles

刘 克 著

中国科学技术大学出版社

当代科学技术基础理论与前沿问题研究丛书

摄动马尔可夫决策与哈密尔顿圈

Perturbed Markov Decision Processes and Hamiltonian Cycles

中国科学技术大学出版社

内 容 简 介

马氏决策过程是一个非常有用的决策分析工具，已经成功的用于解决很多实际问题. 利用马氏决策过程的建模思想，可以将一些离散数学中的传统问题描述为特殊的马氏决策过程加以考虑. 通过优化这些特殊的马氏决策过程，不仅可以为解决这些传统问题提供新的思路，而且还可以促进马氏决策过程本身理论的发展. 但是，在研究这类特殊马氏决策过程时，只有引入摄动因素才能有效的处理问题，所以我们还介绍了马氏决策的摄动理论. 本书的内容包括一些基本的马氏决策过程知识，主要集中在有限状态和有限行动的马氏决策过程上. 然后介绍了有关马氏决策过程的摄动理论. 最后，利用前面的内容，比较详细的介绍了摄动马氏决策与哈密尔顿圈之间的关系和近些年的最新研究成果，提出了一些这个领域里人们现在最为感兴趣的研究问题.

本书适用于三种读者，一个是希望利用马氏决策过程建立有效的模型来分析决策行为的读者，通过前四章的阅读可以了解基本的分析工具，后面的阅读可以使读者获得建立具体模型并进行分析的一些技巧；二是为希望利用这个随机优化的工具研究离散数学或者其他相关科学里的问题的读者提供思路；最后，对于希望发展马氏决策过程理论的读者，可以了解这方面的动态，尽快介入这方面的前沿研究领域.

图书在版编目(CIP)数据

摄动马尔可夫决策与哈密尔顿圈/刘克著. —合肥：中国科学技术大学出版社，2009.4

(当代科学技术基础理论与前沿问题研究丛书；中国科学技术大学校友文库)

“十一五”国家重点图书

ISBN 978-7-312-02241-8

Ⅰ.摄… Ⅱ.刘… Ⅲ.①马尔可夫决策②哈密尔顿圈 Ⅳ.O225，O157.5

中国版本图书馆 CIP 数据核字(2009)第 036449 号

出版发行	中国科学技术大学出版社
	地址 安徽省合肥市金寨路 96 号，邮编：230026
	网址 http://press.ustc.edu.cn
印　　刷	合肥晓星印刷有限责任公司
经　　销	全国新华书店
开　　本	710mm×1000mm 1/16
印　　张	21.75
字　　数	390 千
版　　次	2009 年 4 月第 1 版
印　　次	2009 年 4 月第 1 次印刷
印　　数	1—2000 册
定　　价	58.00 元

总　序

侯建国
(中国科学技术大学校长、中国科学院院士、第三世界科学院院士)

大学最重要的功能是向社会输送人才. 大学对于一个国家、民族乃至世界的重要性和贡献度, 很大程度上是通过毕业生在社会各领域所取得的成就来体现的.

中国科学技术大学建校只有短短的五十年, 之所以迅速成为享有较高国际声誉的著名大学之一, 主要就是因为她培养出了一大批德才兼备的优秀毕业生. 他们志向高远、基础扎实、综合素质高、创新能力强, 在国内外科技、经济、教育等领域做出了杰出的贡献, 为中国科大赢得了“科技英才的摇篮”的美誉.

2008 年 9 月, 胡锦涛总书记为中国科大建校五十周年发来贺信, 信中称赞说: 半个世纪以来, 中国科学技术大学依托中国科学院, 按照全院办校、所系结合的方针, 弘扬红专并进、理实交融的校风, 努力推进教学和科研工作的改革创新, 为党和国家培养了一大批科技人才, 取得了一系列具有世界先进水平的原创性科技成果, 为推动我国科教事业发展和社会主义现代化建设做出了重要贡献.

据统计, 中国科大迄今已毕业的 5 万人中, 已有 42 人当选中国科学院和中国工程院院士, 是同期 (自 1963 年以来) 毕业生中当选院士数最多的高校之一. 其中, 本科毕业生中平均每 1000 人就产生 1 名院士和七百多名硕士、博士, 比例位居全国高校之首. 还有众多的中青年才俊成为我国科技、企业、教育等领域的领军人物和骨干. 在历年评选的“中国青年五四奖章”获得者中, 作为科技界、科技创新型企业界青年才俊代表, 科大毕业生已连续多年榜上有名, 获奖

总人数位居全国高校前列. 鲜为人知的是, 有数千名优秀毕业生踏上国防战线, 为科技强军做出了重要贡献, 涌现出二十多名科技将军和一大批国防科技中坚.

为反映中国科大五十年来人才培养成果, 展示毕业生在科学研究中的最新进展, 学校决定在建校五十周年之际, 编辑出版《中国科学技术大学校友文库》, 于 2008 年 9 月起陆续出书, 校庆年内集中出版 50 种. 该《文库》选题经过多轮严格的评审和论证, 入选书稿学术水平高, 已列为"十一五"国家重点图书出版规划.

入选作者中, 有北京初创时期的毕业生, 也有意气风发的少年班毕业生; 有"两院"院士, 也有 IEEE Fellow; 有海内外科研院所、大专院校的教授, 也有金融、IT 行业的英才; 有默默奉献、矢志报国的科技将军, 也有在国际前沿奋力拼搏的科研将才; 有"文革"后留美学者中第一位担任美国大学系主任的青年教授, 也有首批获得新中国博士学位的中年学者…… 在母校五十周年华诞之际, 他们通过著书立说的独特方式, 向母校献礼, 其深情厚意, 令人感佩!

近年来, 学校组织了一系列关于中国科大办学成就、经验、理念和优良传统的总结与讨论. 通过总结与讨论, 我们更清醒地认识到, 中国科大这所新中国亲手创办的新型理工科大学所肩负的历史使命和责任. 我想, 中国科大的创办与发展, 首要的目标就是围绕国家战略需求, 培养造就世界一流科学家和科技领军人才. 五十年来, 我们一直遵循这一目标定位, 有效地探索了科教紧密结合、培养创新人才的成功之路, 取得了令人瞩目的成就, 也受到社会各界的广泛赞誉[1].

成绩属于过去, 辉煌须待开创. 在未来的发展中, 我们依然要牢牢把握"育人是大学第一要务"的宗旨, 在坚守优良传统的基础上, 不断改革创新, 提高教育教学质量, 早日实现胡锦涛总书记对中国科大的期待: 瞄准世界科技前沿, 服务国家发展战略, 创造性地做好教学和科研工作, 努力办成世界一流的研究型大学, 培养造就更多更好的创新人才, 为夺取全面建设小康社会新胜利、开创中国特色社会主义事业新局面贡献更大力量.

是为序.

2008 年 9 月

前　言

在研究随机环境下的序贯决策优化问题中, 一个有效的分析工具就是马尔可夫决策过程 (Markov Decision Processes), 通常也简称为马氏决策过程. 马氏决策过程的理论在最近的几十年中得到了长足的发展. 作为从 20 世纪 50 年代产生的运筹学的一个分支, 马氏决策过程的模型已经在生态科学、经济理论、通讯工程以及众多学科中得到了广泛的应用, 而这些新的应用也为其带来了丰富的理论研究内容. 例如: 近期利用摄动马氏决策过程的相关理论, 分析离散数学方面的一些核心问题, 为用随机化技术解决确定性离散问题提供了手段.

马氏决策过程也被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. 马氏决策过程的模型由决策时刻、系统状态、行动、报酬和转移概率所组成. 在一个状态上选取一个行动就会产生相应的报酬并且通过转移概率函数决定下一个决策时刻的状态. 策略是一些规则, 这些规则可以告诉决策者任一个决策时刻在任一个状态上是如何选取行动的. 决策者就是要在某种意义下选取最优的策略. 这样一个模型的的分析应该包括以下几个重要环节:

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.

事实上, 策略的比较分析强烈的依赖于准则的不同.

这本书主要介绍了两个方面的研究工作: 一个是马氏决策过程的理论及其摄动问题. 在介绍了一般的马氏决策过程理论模型之后, 我们还介绍了一些最新的相关进展. 特别的, 我们专门介绍马氏决策过程的摄动问题. 在应用决策模型

的时侯, 人们首先就要确定模型的参数, 在这个过程中会导致参数的误差出现. 而这些误差对模型以及相关结果产生的影响是人们十分关心的问题, 摄动分析就是研究模型在扰动下的行为表现. 我们将分别讨论各种准则在摄动下对问题的影响, 例如有: 折扣模型、平均模型、权重准则模型等等.

另一方面的工作就是将离散数学中的一类经典问题, 诸如哈密尔顿圈问题、旅行商问题等等嵌入到凸域上的、可处理的分析问题中去, 使得问题可能得到解决. 很明显, 这些经典问题的主要困难是来自于问题定义域的离散性. 将原始的确定性问题的关键元素赋予概率解释之后, 就可以获得扩展解域的凸化结构. 以哈密尔顿圈问题或者旅行商问题为例, 可以建立一种技术将其嵌入到单摄动的马氏决策过程中去. 其主要思想就是将子图解释为由确定性策略 (如果有, 就包含哈密尔顿圈) 为顶点所构成的凸多面体空间中的元素, 即为随机平稳策略所对应.

我们主要从理论和算法两个方面着手考虑哈密尔顿圈或者旅行商问题, 揭示了图论的理论结构、概率代数和相应的马尔可夫链之间的一些关系, 包括首次返回时间的矩、访问节点的极限频率、用于分析马尔可夫链的某些矩阵的谱等等. 我们还列出了一些尚未解决的开问题, 以供读者欣赏和研究.

本书的写作有两个目的: 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; 另一个目的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、控制与通讯工程方面的工作者、信息科学与工程等方面的学者和技术人员的兴趣.

通过本书的介绍, 可以为那些应用工作者提供方便的建模思想, 能够拓广读者的思维. 本书需要读者适当的熟悉一些数学分析、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, 使读者不失去兴趣.

本书的写作受到了国家基金和创新群体基金的部分资助, 在此作者表示非常的感谢.

主要符号表

MDP, MDPs	马尔可夫决策, 马氏决策过程, 马尔可夫决策过程
MDP_ϵ, MDPs_ϵ	在 ϵ 摄动下的马氏决策过程
$A, A(i)$	行动空间, 行动集
i, j	系统的状态
S	状态空间, 状态集
$Dis(A)$	集 A 上的概率分布集合
$\mathcal{Z}$	全体正整数
$\mathcal{Z}_0$	全体非负正整数
$\mathcal{A}$	A 上的 $\sigma-$ 代数
$\Re$	实数集合
$\|\|\cdot\|\|$	范数
$\|\|\cdot\|\|_w$	以 w 为权重的上界范数
T_f, T_π, T, L	算子
(π), $_S(\pi)$	策略 π 诱导的状态行动过程和状态过程
$\text{supp}(X)$	集合 X 的支撑集合
H_t, H_∞	历史集合
$V_N(i,\pi)$	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

$\bar{V}(i,\pi)$	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
$\mathbb{X}^{\pi}(\alpha)$	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
$\mathbb{X}_s(\alpha)$	随机平稳策略类的状态 – 行动极限平均频率的集合
$\mathbb{X}_s^d(\alpha)$	平稳策略类的状态 – 行动极限平均频率的集合
$\mathbb{X}_m^d(\alpha)$	马尔可夫策略类的状态 – 行动极限平均频率的集合
$\mathbb{X}_m(\alpha)$	随机马尔可夫策略类的状态 – 行动极限平均频率的集合
$(\mathbb{X})^c$	欧氏空间中子集 $\mathbb{X}$ 的闭凸包
Π	策略类, 最一般的策略集合
Π_m	随机马尔可夫策略类
Π_m^d	确定性马尔可夫策略类, 或简称马氏策略
Π_s	随机平稳策略类
Π_{ds}	双随机平稳策略类
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
Π_{SUS}	简单最终随机平稳策略类
Π_{US}	最终随机平稳策略类
Π_{UD}	最终平稳策略类
$\boldsymbol{H}$	偏差矩阵
$\boldsymbol{P}^*$	转移概率矩阵 $\boldsymbol{P}$ 的极限平均矩阵, 也称为 $\boldsymbol{P}$ 的 Cesaro 极限矩阵
$\mathfrak{M}$	随机平稳策略空间到对偶规划解空间的映射
$\mathfrak{M}^{-1}$	对偶规划解空间到随机平稳策略空间的映射

HC	哈密尔顿圈
HCP	哈密尔顿圈问题
TSP	旅行售货商问题
$\mathcal{L}(f)$	基本矩阵的第 $(1,1)$ 分量, 也叫 $\mathcal{L}$ 函数
$\mathcal{D}$	双随机矩阵集合
$\Delta(N)$	哈密尔顿间隙

目　次

第二部分 摄动 MDP 与哈密尔顿圈

第一部分

马氏决策过程与摄动

第 1 章　绪　论

人们在做决策的时候, 不仅要考虑做决策当前的效果, 也要照顾到所做的决策对长远利益的影响. 正像一个长跑运动员, 他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. 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, 从长远的角度来看, 效果不会很好.

本书涉及的马尔可夫决策过程是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, 还要考虑为将来继续做决策创造机会, 也就是要考虑这次选择决策后对将来发展过程的影响. 看上去这个模型似乎不复杂, 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决策的过程和动态, 然后给出马尔可夫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 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, 并根据这个状态选取一个行动. 其后, 该行动会对系统的运行产生两个影响: 一个是产生了一个即得的报酬或费用, 而