Hou Zhenting Guo Xianping

Markov Decision Processes
Markov Decision Processes
Markov Decision Processes

马尔可夫决策过程

● 侯振挺 郭先平著
● 湖南科学技术出版社

ARKOV DECISION PROCESSES

马尔可夫决策过程

● 侯振挺 郭先平著 ● 湖南科学技术出版社

Markov Decision Processes

马尔可夫决策过程

侯振挺　郭先平著

湖南科学技术出版社

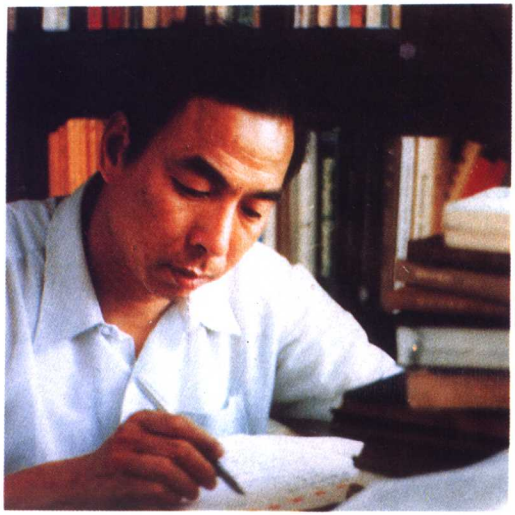
国家自然科学基金
湖南省自然科学基金资助项目

马尔可夫决策过程

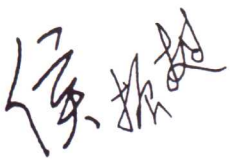
著　　者：侯振挺　郭先平
责任编辑：胡海清
出版发行：湖南科学技术出版社
社　　址：长沙市展览馆路 66 号
印　　刷：湖南省新华印刷二厂
厂　　址：邵阳市双坡岭
邮　　编：422001
（印装质量问题请直接与本厂联系）
经　　销：湖南省新华书店
出版日期：1998 年 3 月第 1 版第 1 次
开　　本：850mm×1168mm　1/32
印　　张：12.25
插　　页：8
字　　数：324000
印　　数：1—700
征订期号：地科 235—1
书　　号：ISBN 7—5357—2272—5/O・160
定　　价：28.00 元

谨以此书献给

我们的朋友已故的董泽清教授



侯振挺，我国著名数学家，长沙铁道学院教授，博士生导师，1936年生，河南密县人，1960年毕业于唐山铁道学院，对马尔可夫过程理论的发展作出了卓越贡献，发表论文70余篇，专著5本，自1978年来，培养博士生10名，硕士生36名。他曾获1978年国际戴维逊奖。1982年以来，他获得过国家及省、部级奖励10多项，1984年被评为“国家级有突出贡献的科技专家”。他是第五、六、七、八届全国人大代表，全国劳动模范。近年来，他和他的学生们对马尔可夫决策过程进行了一系列深入研究，并提出了有广阔应用前景的马尔可夫骨架过程新概念，进而奠定了这类过程的理论基础。



郭先平，男，副教授，1964 年生于湖南省桑植县。1987 年和 1990 年，他先后于湖南师范大学获理学学士和硕士学位，1996 年获长沙铁道学院理学博士学位，现为中山大学概率论与数理统计专业的博士后研究人员，湖南师范大学数学研究所副所长。他主要从事概率论和运筹学，特别是其中的马尔可夫决策过程的研究，已发表学术论文 20 余篇。

序
PREFACE

马尔可夫(简称马氏)决策过程(Markov Decision Processes，简记为MDP)由哈瓦德(Howard，R. A)于1960年提出，至今不到40年，但它在理论与应用两方面都得到了迅速而深入的发展. 在国际上已有不少专著问世，并各具特色. 在我国研究MDP及其应用的同志也越来越多，成果也很出色. 但至今尚没有一本关于MDP方面的书出版. 这无疑影响到国内在这方面的研究. 基于此，我们不揣冒昧地写成这本书，希望能对有关同志有所裨益.

董泽清教授是我国开展MDP研究的引路人. 董泽清教授在世的时候，他及其他几个朋友曾不止一次地希望我将研究Q过程的成果和方法用于MDP的研究. 当时由于种种原因未能实现，但却一直颇感遗憾和不安. 后来，郭先平同志来到我这里攻读博士学位. 他在攻读硕士学位期间就开始研究MDP，并有一些成果. 攻读博士学位期间，仍以MDP研究为主，并在MDP的几乎所有方面都做了许多很好的工作. 在他的促进下，我也逐渐对MDP有了兴趣，后来与他一起作研究工作. 现在将我们关于MDP的研究成果整理成书出版，以了却我多年的心愿，也算是对已故董泽清教授的纪念. 去年郭先平同志已投身前辈梁之舜先生门下从事博士后研究，与我常有书信、电话往来，有时也见面切磋叙旧，昔日师生已成了今日忘年之交。我想这算得上一个教学相长的小例吧!

本书由郭先平执笔，共6篇计16章，第1篇介绍MDP的基

本模型；第 2 篇介绍离散时间可数状态 MDP；第 3 篇介绍离散时间 Borel 状态空间非平稳 MDP；第 4 篇介绍连续时间可数状态 MDP；第 5 篇介绍连续时间可数状态半马氏决策过程；第 6 篇介绍 MDP 的应用. 本书除了 MDP 基本概念、模型的极少数定理外，其余内容都是作者多年研究 MDP 获得的成果，不少尚属首次发表. 这些成果大部分属于郭先平同志，若干稍有代表性的工作如非平稳 MDP 的时齐化，连续时间 MDP 的折扣目标以及半马氏过程的折扣目标等部分是我和郭先平同志共同完成的。其中有关连续时间 MDP 折扣目标这方面的成果还包含有张汉君教授的工作.

在本书付梓之际，我们要感谢国家自然科学基金委员会和湖南省自然科学基金委员会对我们研究工作的资助. 同时我们要感谢长沙铁道学院领导和科研所邹捷中、肖果能、刘再明、张汉君、李俊平、袁成桂、方小斌等同志的支持；我们还要感谢王寿仁、王梓坤、丁夏畦、马志明、梁之舜、严士健、胡迪鹤、戴永隆、严加安、杨向群、何声武、吴荣、陈木法、程侃、刘文、司徒荣、邓永录、黄志远、汪嘉冈、潘一民等前辈和同行对我们工作的帮助和好评.

我们要特别感谢清华大学林元烈教授，他以异常的耐心，十分详细地阅读了原稿，提出了许多非常宝贵的修改意见.

由于作者水平有限，书中错误、缺点在所难免，欢迎大家指正.

侯振挺

1997 年 7 月 1 日

目录
CONTENTS

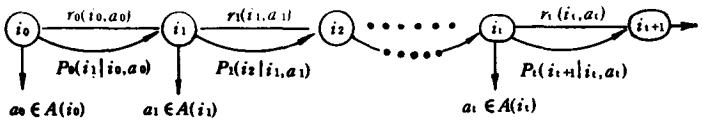
绪论

MARKOV
DECISION
PROCESSES

动态规划研究的是多阶段决策问题，但其决策结果的确定性要求又不得不限制它的应用范围. 而决策分析的产生又正好来自于决策结果的不确定性. 取长补短，将二者结合起来进行考虑，在理论上讲是再自然不过了，这便孕育着马尔可夫（简称马氏）决策过程（Markov Decision Processes，简记为 MDP）基本概念的产生. 然而 MDP 的产生仍有它的实际背景. 例如：考察有限容量 M 的存贮系统，我们周期地检查某物品的存贮量 x_t（称为系统于第 t 周期的状态，$t=0, \cdots$），每次检查之后，可供选择的决策是该物品的允许定货量 a_t（即 $a_t \in [0, M-x_t]$），当决策者选取决策 a_t（即定货 a_t）之后，系统于第 $t+1$ 周期的状态 x_{t+1}（即第 $t+1$ 周期开始时的存贮量）并不是第 t 周期开始的存贮量 x_t 与该周期的定货量 a_t 的简单相加，它还受顾客于此周期内对该物品的需求量 ξ_t 所影响，即 $x_{t+1} = \max\{0, x_t + a_t - \xi_t\}$，显然，$\xi_t$ 是随机的（其分布记为 μ_t），从而 x_{t+1} 是不确定的，且该系统从第 t 周期到第 $t+1$ 周期的状态转移概率 P_t 为：$P_t(B|x,a) = \int_B \max\{0, x + a - s\}\mu_t(\mathrm{d}s)$. 对于该系统，在每个周期都要计划定货量（即作决策），但所作决策的结果是不确定的. 所涉及的经济结构有：新增物品的单位定货费 h，库存物品的单位存贮费 m，销售物品的单位价格 q. 若用 $r_t(x,a)$ 表示 系统于第 t 周期处于状态 x，采取决

策 a 的平均获利，则有 $r_t(x,a)=\int[q\cdot\min(s,x+a)-ha-m(x+a)]\mu_t(\mathrm{d}s)$. 决策者的目的是制定定购方案，使得系统在某种目标 W 下的效益最好. 因此，决策者必须认真地研究他们的方案. 这些研究导致了人们对能为复杂过程提供最优决策的方法论感兴趣. 从而促使了 MDP 的产生并推动了其理论和实际应用的发展.

若令 $S=[0,M]$，$A(x)=[0,M-x]$，$x\in S$，则上述系统的数学模型为 $\{S,(A(x),x\in S),(P_t),(r_t),W\}$. 像这样，由状态空间 S，允许行动（决策）空间 $A(x)(x\in S)$，状态转移概率族 P_t，报酬函数 r_t，目标函数 W 所组成的五重组 $\{S,(A(x),x\in S),(P_t),(r_t),W\}$ 称为 MDP 的一个模型. 于是可知离散时间 MDP 模型中诸因素的相互关系图为：



MDP 的目标为：在各个决策时刻选取决策（即制定策略），使系统运行的全过程在某种目标 W 下达到最好.

MDP 模型按其相继决策时间的特征和模型中的诸因素，可分为许多类型. 例如：可分为平稳或非平稳的，离散时间、连续时间的，状态空间是可数或非可数的，折扣目标、平均目标或期望总报酬目标的，等等. 它们中的各种组合构成多种类型的 MDP. 然而所有 MDP 模型的主要研究内容又可归结为三点：

（1）最优策略存在的条件；

（2）最优策略的有效算法；

（3）特殊模型的具体应用.

近半个世纪来，MDP 的理论和应用的研究已有丰富的成果，其实际应用的范围已深入到经济、控制、管理、生物、医疗、交通运输、工程计划、期货买卖、设备维修、生产存贮、环境保护、水库发电等众多领域. 自从 1960 年著名数学家哈瓦德 (Howard)[127] 首次提出 MDP 以来，MDP 的各种模型不断出现，并从各种角度