确定理论与优化丛书

Certainty Theory and Optimization Series

实用马尔可夫决策过程

Applied Markov Decision Processes

刘克 编著

Liu Ke

清华大学出版社

不确定理论与优化丛书

Uncertainty Theory and Optimization Series

实用马尔可夫决策过程

Applied Markov Decision Processes

刘克 编著

Liu Ke

清华大学出版社

北京

内容简介

马尔可夫决策过程是研究随机环境下多阶段决策过程优化问题的理论工具.在过去的几十年中,随着生态科学、经济理论、通讯工程以及众多学科中需要考虑不确定因素和序列决策问题的大量新模型的涌现,进一步刺激了马尔可夫决策过程在理论上和应用领域中长足的发展.

本书从简单的例子开始,介绍了马尔可夫决策过程的基本概念、决策过程以及一些常用的基本理论.还介绍了多种最优准则,包括有限阶段准则、折扣准则、平均准则、权重报酬准则、概率准则等.从模型角度考虑了有限状态空间、可数状态空间和一般 Borel 状态空间;从决策时间上来说,考虑了离散时间、连续时间和半马氏决策时刻问题.本文还介绍了大量的应用实例以及建模方法.本书可作为高年级大学生和研究生教材,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

实用马尔可夫决策过程/刘克编著.—北京:清华大学出版社,2004.11
(不确定理论与优化丛书)
ISBN 7-302-09506-X

Ⅰ.实…　Ⅱ.刘…　Ⅲ.马尔可夫决策—研究　Ⅳ.O225

中国版本图书馆 CIP 数据核字(2004)第 094191 号

出 版 者:清华大学出版社　　**地　　址**:北京清华大学学研大厦
http://www.tup.com.cn　　**邮　　编**:100084
社 总 机:010-62770175　　**客户服务**:010-62776969
责任编辑:王海燕
印 刷 者:北京市世界知识印刷厂
装 订 者:北京市密云县京文制本装订厂
发 行 者:新华书店总店北京发行所
开　　本:170×230　**印张**:12.25　**字数**:226 千字
版　　次:2004 年 11 月第 1 版　2004 年 11 月第1 次印刷
书　　号:ISBN 7-302-09506-X/O・406
印　　数:1～3000
定　　价:22.00 元

序言

PREFACE

在过去的几十年中，马尔可夫（简称马氏）决策过程（Markov decision processes，MDP）的理论和应用得到了长足的发展.作为20世纪50年代产生的运筹学的一支，马氏决策过程的模型已经在生态科学、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.

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.马氏决策过程的模型由决策时刻、系统状态、行动、报酬和转移概率组成.在一个状态选取一个行动会产生一个报酬，并且通过转移概率函数决定下一个决策时刻的状态.策略是一些规定，即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规定.决策者就是要在某种意义下选取最优的策略.这样一个模型的分析应该包括：

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开讨论.

本书共分为8章.第1章从一些决策的例子出发，抽象出一般的决策过程，并且给出一些概念的基本定义.第2章针对有限阶段的MDP模型讨论了最优策略的存在条件，给出具体的计算方法.第3章考虑了无限阶段的折扣模型，针对有限状态的折扣MDP，建立最优方程，给出多种计算方法，最后将这些结论推广到比较一般的状态空间和行动集合的情形.第4章讨论平均准则模型，对单链结构和多链结构分别进行讨论.第5章对一些非标准的准则进行讨论，其中包括权重准则和概率准则等几种非标准的准则模型.第6章考

虑连续时间的两种模型和半马氏决策过程的两种模型. 第 7 章和第 8 章针对两个具体的实际问题，即空集装箱的调配和人力资源管理这两个问题，建立 MDP 的模型，给出了具体的步骤，最后给出实际的计算结果.

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，特别是本书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，这样可以为那些应用工作者提供方便的建模思想，能够拓宽读者的思维. 本书需要读者熟悉一些数学分析、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.

刘克

2004 年 5 月

一些常用的符号和缩写

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
N_+	全体正整数
Z	全体非负正整数
$\mathcal{A}$	A 上的 σ-代数
$\mathbb{R}$	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
$L(\pi), L_S(\pi)$	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
$V_a(i,\pi)$	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
$u_a(i,\pi)$	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
$\mathrm{sp}(v)$	向量 v 的跨度
H_t, H_∞	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

人们在日常生活中经常会做一些决策，当然，做决策时即要考虑到当前的效果，又要照顾到长远的利益. 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. 比如说长跑运动员，要根据需要跑的距离而合理的分配自己的体力，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，还要考虑为将来做决策创造机会. 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 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. 之后，该行动会对系统的运行有两个影响：一个是产生了一个即得的报酬或**费用**，而另一个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相同的问题，即选取 $t+1$ 时刻的决策. 过程就是这样循环下去，不同的只是此时的状态可能是一个新的状态，而且可采用的行动集合随着状态的变化而改变了.

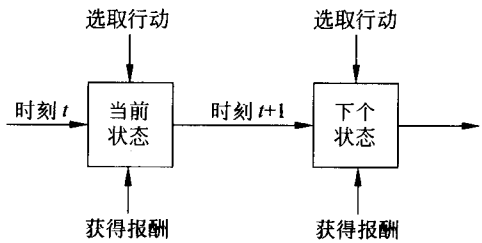

图 1.1　决策过程的图示

我们可以把这个序列决策过程的关键列举出来：

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

一般来讲，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.

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：决策者得到即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有可能出现. 针对每个不同的状态，决策者会选取不同的行动，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的那些状态和在那些状态上行动的选取. 我们把在将来任意可能的状态上选取行动的规则称之为**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.

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，即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

下面我们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关键因素.

例 1.1　机器最优维修策略问题.

等周期(如一天)的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，机器可处于两个状态：正常运行(记作 $i=1$)和出了故障(记作 $i=2$). 在任一个周期，如果机器正常运行可得收益 10 元，到下一个周期初，仍处于正常运行的概率为 0.7，发生故障的概率为 0.3. 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. 如果用 $p(j|i,a)$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；$r(i,a)$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 1.1.

表 1.1　转移概率和报酬

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

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的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. 随机运筹学中最早也是最关心的问题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. 在太平洋地区有 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，并且周期的为这 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，要开销固定的费用，它包括仓库的占用费和商店的上架费. 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手上的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产品量. 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，订货量达到一个目标水平，否则就不订货. 如果考虑了约束条件，上面的策略不一定是最优的.

较好的库存控制会有效的降低费用，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(October 20,1992,p. C24)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 等人的文章[71]，也可以参见 Puterman 的书[135].

美国 Arizona 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. 当然了，如何分配这些资金成为了 ADOT 的核心工作. 1978 年，他

们和旧金山的 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，来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 州当年公路维修预算的 1/3. 在后来的 4 年，这一系统为 ADOT 比预计的总费用节省了 10100 万美元. 后来这个系统被 Kansas 州、芬兰和沙特阿拉伯等地区和国家使用. 这个模型还被应用到桥梁和管道的管理中. 下面我们详细介绍公路管理模型.

公路管理系统是一个长期的多阶段动态模型，它所提供的管理策略是在保证公路质量的条件下极小化长期年平均费用. 为建立这一模型，Arizona 的高速公路网络被划分为 7400 个一英里长的路段，并根据每个路段的类型、交通密度以及地域环境，被分成 9 类情况中的一种. 对每种类型，分别建立了动态模型. 这些模型反映了路段的条件，就此条件下能够采取的维修行动，每年的期望磨损情况和每种可行的维修对公路条件的改善情况等. 还确定了每种维护行为的开销. 当然确定系统状态、可用行为、费用以及在不同行为下的状态到状态的转移的确不是一件容易的工作，这需要数据、路况条件模型、统计分析以及相关的专家意见等.

这里只考虑沥青混凝土路段的管理模型而且每年决策一次. 系统的状态描述了该路段的条件：粗糙度（分为 3 个水平），路面破损的百分比（分为 3 个水平），前一年到现在的破损变化（分为 3 个水平），以及维修指标①（分为 5 个水平）. 因此，所有可能的状态有 135 个. 除去一些不可能的组合，共有 120 个状态.

行为是可用的公路修复活动，它覆盖了从不需要费用的一般维护到最昂贵的重铺路面. 这些行为的描述和相应的费用列入了表 1.2. 对每个状态来说差不多有 6 个可用的行动可以选取.

表 1.2　复原行为和相关费用

状态指标	行　为　描　述	费用/美元/$yd^2$②
1	线路维护	0
2	封闭层	0.55
3	沥青混凝土涂层	0.75
4	沥青混凝土涂层＋沥青橡胶	2.05
5	沥青混凝土涂层＋热翻路	1.75
6	1.5in③ 沥青混凝土	1.575

① 这里的维修指标是指：最近的一次维修方式和维修后延续到现在的时间.

② $1yd^2=0.836\ 127\ 36m^2$.

③ 1in=25.4mm.

续表

状态指标	行 为 描 述	费用/美元/yd^2
7	1.5in 沥青混凝土＋沥青橡胶	2.875
8	1.5in 沥青混凝土＋热翻路	2.575
9	2.5in 沥青混凝土	2.625
10	2.5in 沥青混凝土＋沥青橡胶	3.925
11	2.5in 沥青混凝土＋热翻路	3.625
12	3.5in 沥青混凝土	3.675
13	3.5in 沥青混凝土＋沥青橡胶	4.975
14	3.5in 沥青混凝土＋热翻路	4.675
15	4.5in 沥青混凝土	4.725
16	5.5in 沥青混凝土	5.775
17	重铺(相当于 6in 沥青混凝土)	6.3

费用包括复原行为费用(见表 1.2)和年线路维护费用(见表 1.3,其中修复行为中的 RM 表示线路维护,SC 表示封闭层,OT 表示其他行为).年维护费用是依赖于路况的:除了维护行为以外,“封闭层”行为也依赖于路的粗糙度和破损情况,而其他的行为年维护费用是常数.这里的费用是依赖于所考虑的路段类型,对于其他不同的路段类型,费用会不一样.

表 1.3 年线路维护费用

修复之后的状态		修复行为	费用/美元/yd^2
粗糙度	破损		
120(±45)	5(±5)	RM	0.066
120(±45)	20(±10)	RM	0.158
120(±45)	45(±15)	RM	0.310
120(±45)	无要求	SC	0.036
210(±45)	5(±5)	RM	0.087
210(±45)	20(±10)	RM	0.179
210(±45)	45(±15)	RM	0.332
210(±45)	无要求	SC	0.057
300(±45)	5(±5)	RM	0.102
300(±45)	20(±10)	RM	0.193
300(±45)	45(±15)	RM	0.346
300(±45)	无要求	SC	0.071
无要求	无要求	OT	0.036

转移概率表示出了在维护行为下的路况年度变化情况,它是由历史数据估计出来的,而且是在描述状态的那些指标是相互独立的条件下得到的.实际上状

态的转移变化不大，大约有97%的元素是0.

准则是在可接受的路况状态和不可接受的路况状态的一个比例的约束下极小化费用. 比如说：ADOT的策略是至少80%的高速路面粗糙度不能超过165in/mile①，同时粗糙度超过256in/mile的比例不能大于5%. 而破损度要求的约束也是类似的.

这个模型恰恰是一个受约束的马氏决策过程问题，是用**线性规划**求解的. 这个模型不仅提供了解，而且还能用于预算和常规策略之间的协调. 具体的求解过程太占篇幅，这里就不列举了. 但是要说的是在约束马氏决策过程中，有些解会没有什么意义. 因为增加了约束条件以后，最优策略可能是随机策略. 比如，在路段的某种破损情况下，最优行为是以40%的概率铺1.5in的沥青混凝土，以60%的概率铺2.5in的沥青混凝土. 这一点在我们的问题中没有受到影响，原因是路段很多，可以将40%的路段铺1.5in的沥青混凝土，其余的这类路段铺2.5in的沥青混凝土.

利用这个模型，求出了最优策略并加以应用，为ADOT节省了很多费用. 在谈到原因时，Golabi[71]等人总结道：过去的行为太过保守了，通常总是铺5in的沥青混凝土……. 公路管理系统得到的结果其保守性要小得多，比通常铺5in的沥青混凝土的情况还要差的状态，建议只铺3in.

1.3 马氏决策过程的定义与记号

总结1.2节的各个例子，我们发现马氏决策过程的主要成分包括：决策周期、状态、行动、转移概率和报酬. 作为决策者，所面对的问题就是抓住影响所控制的概率系统的机会，也就是适时的做出系列的行动的选择，以期达到决策者心目中的某种准则的优化. 由于受控制的系统在持续发展，过去的决策通过状态的转移影响到今天的决策. 一般来讲一步最优的选择不是最好的决策，必须要考虑系统将来状态上的预期机会和费用. 下面我们给出详细描述.

1.3.1 决策时刻与周期

选取行动的时间点被称为**决策时刻**，并用T记所有决策时刻的点集. T是非负实直线上的子集，可以是有限点集、可列无限点集或者是连续集合. 在T为离散的情况下，决策都是在决策时刻做出. 当T连续的时候比较复杂，可以连续的做决策，也可以在某些随机点（某些事件发生时）上做决策（例如排队模型的顾

① 1mile=1609.344m.

客到达或离去的时刻)或者由决策者选择时机做决策等.

对于离散时间问题,两个相邻的决策时刻被称为**决策周期**或者**阶段**. 图 1.2 示意了决策时刻和阶段的概念.

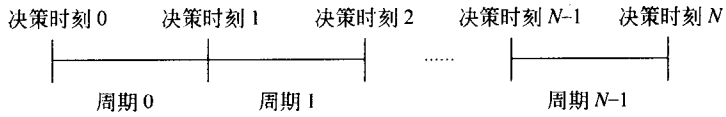

图 1.2 离散决策时刻与周期

我们把**有限阶段**的决策时刻集记为 $T=\{0,1,2,\cdots,N\}$,**无限阶段**的决策时刻集记为 $T=\{0,1,2,\cdots\}$.

1.3.2 状态与行动集

在每个决策时刻,对系统的描述就是**状态**. 记系统的所有可能状态为 S,也称为状态空间. 如果在任一个决策时刻,决策者观察到的状态是 $i\in S$,他可以在这个状态 i 的**可用行动**集 $A(i)$ 中选取行动 a,其中 $A(i)$ 也称为行动空间. 令 $A=\bigcup_{i\in S}A(i)$,并且假定 S 和 $A(i)$ 都不依赖于时刻 t. 状态集合 S 和行动集合 $A(i)$ 可以是任意的有限集合、可数的无限集合、有限维欧氏空间的紧致子集或者是完备可分度量空间上的博雷尔(Borel)子集. 除非特别声明,我们总考虑 S 和 $A(i)$ 都是离散的情况(有限或可数无限).

行动的选取可以是确定性的选取一个,也可以在多个可用的行动中随机性的选取. 我们记 $\mathrm{Dis}(A(i))$ 为 $A(i)$ 的博雷尔子集上的所有概率分布,$\mathrm{Dis}(A)$ 为 A 的博雷尔子集上的所有概率分布. 随机选取行动就是选取一个概率分布 $q(\cdot)\in\mathrm{Dis}(A(i))$,其中选取行动 a 的概率是 $q(a)$. 如果这个分布是退化的,就是确定性的选取行动.

状态空间 S 和行动空间 $A(i)$ 也可以一般化为依赖于时间 t 的情形,但对于大部分应用这样做并不适合. 就理论上来讲只需要令 $S=\bigcup_{t\in T}S_t$ 以及令 $A(i)=\bigcup_{t\in T}A_t(i)$,再对下面定义的转移概率和报酬函数做相应的修正,就转化为标准的马氏决策模型. 有时为了符号的简化,我们可以令 $A(i)\equiv A$,这对理论研究没有什么影响,只会在应用时对问题的理解造成困难.

1.3.3 转移概率和报酬

任意一个决策时刻,在状态 i 采取行动 $a\in A(i)$ 之后,有两个结果:

1) 决策者获得报酬 $r(i,a)$;