国家科学技术学术著作出版基金
NFAPST

An Introduction to Markov Decision Processes

马尔可夫决策过程引论

胡奇英 刘建庸 著

西安电子科技大学出版社
http:// www.xduph. com

□国家科学技术学术著作出版基金资助出版

马尔可夫决策过程引论

胡奇英　刘建庸　著

西安电子科技大学出版社
2000

内 容 简 介

马尔可夫决策过程是研究马氏型序贯(动态)决策问题的工具。本书提供了处理离散时间、连续时间、半马氏等三类基本马氏决策过程模型的一般化方法。在此基础上，本书研究了状态部分可观察、多目标、带约束条件等一般化马氏决策过程以及处于随机变化环境中的马氏决策过程。本书最后还提供了马氏决策过程在排队/通信系统控制、生产/存贮系统控制、系统最优更换/维修、质量控制、序贯搜索、柔性制造系统控制等方面的应用例子。

本书可作为运筹学、管理科学、自动控制、通信、制造自动化等专业的大学生与研究生的教材，也可作为有关领域科技人员的参考书。

图书在版编目(CIP)数据

马尔可夫决策过程引论/胡奇英，刘建庸著. —西安：西安电子科技大学出版社，2000.7
ISBN 7-5606-0830-2
Ⅰ. 马… Ⅱ. ① 胡… ② 刘… Ⅲ. 马尔可夫决策 Ⅳ. 0225-62

中国版本图书馆 CIP 数据核字(2000)第 17538 号

责任编辑 陈宇光 戚文艳
出版发行 西安电子科技大学出版社(西安市太白南路 2 号)
电 话 (029)8227828 邮编 710071
http://www.xduph.com E-mail：xdupfxb@pub.xaonline.com
经 销 新华书店
印 刷 西安电子科技大学印刷厂
版 次 2000 年 7 月第 1 版 2000 年 7 月第 1 次印刷
开 本 787 毫米×1092 毫米 1/16 印张 17.75
字 数 451 千字
印 数 1～2 000 册
定 价 28.00 元
ISBN 7-5606-0830-2/O・0040

前　言

马尔可夫决策过程(Markov Decision Processes，简记为 MDP，也称马尔可夫决策规划或马尔可夫控制系统等)是研究一类随机序贯决策问题的理论。所谓随机序贯决策问题，是指在一系列相继的或连续的时刻(称之为决策时刻)点上作出决策，在每个决策时刻点，决策者根据观察到的状态从可用的若干个决策中选择一个；将决策付诸实施后，系统将获得与所处状态和所采取决策有关的一项报酬(或费用等)并影响系统在下一决策时刻点所处的状态。系统在下一决策时刻点处的状态是随机的。在这一新的决策时刻点上，决策者要观察系统所处的新的状态(即收集新的信息)并采取新的决策，如此一步一步进行下去。在每一决策时刻采取的决策都会影响下一决策时刻系统的运行(状态，决策)，并以此影响将来。决策的目的是使系统的运行在某种意义下(称为准则)达到最优。马尔可夫决策过程就是研究这种马尔可夫型随机序贯决策问题的一门学科，是(确定性)动态规划与马尔可夫过程相结合的产物。它不像动态规划那样以 Bellman 的“最优化原理”作为研究的出发点，而是从一些简单的、易于验证的条件(或公理)出发来严格证明“最优化原理”。MDP 既是随机运筹学的一门分支，也是应用概率的一门分支，同时，作为马尔可夫型系统最优控制的理论，它亦属于随机系统最优控制的范畴。MDP 与近年来兴起的计算机集成制造系统中的系统理论——离散事件动态系统理论密切相关。实际上，它是随机型离散事件动态系统的唯一的动态控制方法，与离散事件动态系统的逻辑控制方法也有着密切的关系。

MDP 中的一些概念可以说在 1960 年之前已经产生，但 1960 年 Howard 所著的《动态规划与马尔可夫过程》一书奠定了 MDP 作为独立学科的基础，1962 年 Blackwell 的文章则为在这一领域进行进一步的研究提供了动力。MDP 中基本的模型有离散时间马氏决策过程、连续时间马氏决策过程和半马氏决策过程，在此基础上，考虑更为接近实际问题的模型有状态部分可观察的、多目标的、自适应的、带约束条件的以及作者近年提出的随机环境 MDP。从准则函数来说，有折扣准则、平均准则、期望总报酬准则、加权准则、折扣矩最优准则、样本路径准则等等。

马氏决策过程的应用领域十分广泛，这些领域有：生产存贮系统、系统更换/维修、制造系统的调度控制、计算机/通信网络系统控制、动态资产定价、广告优化、商品与服务的定价、质量控制、序贯搜索、水资源管理、森林管理、航空订票、高速公路管理等等。

MDP 产生至今已近 40 年，国内的研究自中科院应用数学所已故研究员、作

者的导师董泽清老师(1978)至今已有20余年，已有的研究内容相当丰富，应用领域也十分广泛。但仍有需进一步研究、探讨和开垦的应用领域。

作者在三项国家自然科学基金的资助下，先后对MDP的多个方面进行了系统的研究，与合作者一起已发表和待发表的论文近80篇，主要的工作包括以下几个方面：(1) 在基本模型的离散时间MDP方面，运用初等方法进行了系统的讨论。对折扣准则，证明了其最优方程有定义时就成立，提出了一种无界报酬条件，其中准则函数值空间中的范数不再有限，以至通常的算子方法失效，在最优策略性质与算法等方面也作了系统的研究；对平均准则，研究了Bellman最优性原理、最优方程、最优不等式，对它们成立的条件作了充分的弱化。(2) 在一般化马氏决策过程的状态部分可观察、多目标、带约束条件等方面作了较为系统和深入的研究；提出和研究了模型中的参数随环境的变化而变化的一类新模型——随机环境MDP模型，系统地研究了具有折扣准则和总报酬准则的随机环境连续时间MDP模型、半MDP模型等的结构、性质、模型的逼近、算法等，从而在很大的程度上推广了MDP的研究领域和应用领域。(3) 研究了MDP中模型间的转换关系，将复杂难处理的模型转换为较简单易处理的模型，从而可将后者中的大量结果直接推广到前者中去，为研究复杂模型开辟了一条简捷的途径，如对连续时间马氏决策过程、半马氏决策过程、随机环境马氏决策过程等，同时将折扣矩最优准则转化为一系列的折扣准则。(4) 研究了MDP在存贮系统最优控制、设备最优更换/维修 、bandit问题等中的应用，得到了具有简单结构的易于实施的最优策略。

本书第1章到第5章研究离散时间马氏决策过程。其中第1章介绍基本模型，第2章介绍有限阶段期望总报酬准则，它与动态规划中的情况比较类似，其重点放在若干应用例子上。第3章系统地介绍了折扣准则，包括基本内容方面的折扣准则最优方程和(ε)最优平稳策略的存在性、性质，以及求解(ε)最优平稳策略的算法，讨论了罗朗级数展开以及Blackwell准则，并在最后一节讨论了非可数决策空间的情形，这对解决后面的一些应用问题是必不可少的。这些内容大部分是对报酬函数无界时给出的，读者如果理解上有一定困难，在阅读时可假定是有限的，此时很多结论将成为明显的事实。本章所用的方法都是很基本的，大多数时候只用到四则运算。第4章在第3章的基础上进一步将折扣准则中的结论推广到期望总报酬准则，所用的方法是将它们作统一的处理。本章的研究内容是作者在写作本书的过程当中完成的，其中使用的方法仍是初等的，具有高等数学知识的读者都能看懂。第5章讨论了平均准则最优方程和多链时的最优方程系，它们的性质、与(ε)最优策略的关系等，并顺便讨论了与其相关的强平均准则、样本路径准则等。在算法方面，我们讨论了策略迭代法、逐次逼近法以及线性规划法；最后讨论了最优方程的弱化——最优不等式，介绍了国内外(包括作者)在此

方面的最新研究成果。在讨论多链时要用到马氏链中的状态分类、极限分布等知识，因此，本章的一些内容要比前几章略微复杂一些。

第 6、7 章分别讨论半马氏决策过程和连续时间马氏决策过程，在建立相应的模型之后，我们用转换的方法将它们化成离散时间马氏决策过程。这样，前面几章中的大部分结果可直接推广到这两章中去。

以上为马氏决策过程的最为基本的内容，在它们的基础上，可跳过第 8、9 章而直接阅读有关应用的两章：第 10 章和第 11 章。

第 8 章讨论一般化马氏决策过程，它们是比前面介绍的基本模型更接近于实际、更为一般化的模型。这些模型包括部分可观察模型、带约束条件的模型、多目标模型、摄动模型等。第 9 章讨论了作者提出的随机环境中的连续时间马氏决策过程、半马氏决策过程和混合马氏决策过程模型，这是考虑环境对系统有影响时的模型。

第 10、11 两章讨论了马氏决策过程在计算机集成制造系统调度与控制(包括排队系统与排队网络的最优控制、柔性制造单元调度、生产/存贮系统最优控制、系统最优更换/维修等)、通信系统等控制中的应用，以及在质量控制、序贯搜索等方面的应用。

本书的第 8 章由刘建庸撰写，其余部分由胡奇英撰写，全书由胡奇英统稿。

本书中的定理、推论、引理、定义以及条件编号采用如下方式：定理 l.m.n 表示第 l 章第 m 节的第 n 个定理，而定理 m.n 表示本章第 m 节的第 n 个定理，推论、引理、定义编号类似。对于文中提及的参考文献，[m.n]表示第 m 章的参考文献 n，而[n]则表示本章参考文献 n。

在本书的撰写过程中，伍从斌先生和张继红先生都提出了宝贵的意见。西安电子科技大学出版社编辑陈宇光和戚文艳老师为本书增色很多，总编李荣才教授也给予了大力支持。作者在此对他们表示衷心的感谢。

作者感谢国家自然科学基金会多年来对我们研究工作的资助，感谢国家科学技术学术著作出版基金委员会对本书出版的资助。

作者期望本书能对我国马氏决策过程的研究和应用有所促进。

作者

1999 年 5 月

目　　录

第 1 章

引　论

1.1　离散时间马尔可夫决策过程模型

本节介绍最为基本的离散时间马尔可夫决策过程(Discrete Time Markov Decision Processes，简记为 DTMDP)的模型。为方便起见，我们假定在时刻点 $n=0, 1, 2, \cdots$ 处观察系统，这里 n 可取有限多个值，如 $n=0, 1, 2, \cdots, N$，也可取所有的非负整数。一个 DTMDP 模型由如下的五重组组成：

$$\{S, A(i), p_{ij}(a), r(i, a), V, i, j \in S, a \in A(i)\} \tag{1.1.1}$$

其中各元的含义如下：

(1) S 是系统所有可能的状态所组成的非空的状态集，有时也称之为系统的状态空间，它可以是有限的、可列的集或任意非空集。本书中假定 S 是可数集(即有限或可列)。我们用小写的字母 i，j，k 等(或加上上、下标)来表示状态。

(2) 对状态 $i \in S$，$A(i)$ 是在状态 i 处可用的决策集，它是非空的；当不特别指出时，本书中亦假定它是可数集。实际上，当 $A(i)$ 为非空集并赋以一定的拓扑结构后，本书中的绝大部分结论仍成立。书中常用 a 来表示决策。

(3) 当系统在决策时刻点 n 处于状态 i，采取决策 $a \in A(i)$ 时，则系统在下一决策时刻点 $n+1$ 时处于状态 j 的概率为 $p_{ij}(a)$，它与决策时刻 n 无关。称 $p=\{p_{ij}(a), i, j \in S, a \in A(i)\}$ 为系统的状态转移概率族，于是，对 $i \in S$，$a \in A(i)$，有 $\sum_{j \in S} p_{ij}(a)=1$，即 $\{p_{ij}(a), j \in S\}$ 为一随机向量。范围稍广一点的是 $\{p_{ij}(a), j \in S\}$ 为次随机向量的情形，即 $\sum_{j \in S} p_{ij}(a) \leqslant 1$。在大多数情形下，这两种情形可统一考虑。

(4) 当系统在决策时刻点 n 处于状态 i，采取决策 $a \in A(i)$ 时，系统于本阶段获得的报酬为 $r(i, a)$。如果记 $\Gamma=\{(i, a) \mid i \in S, a \in A(i)\}$，则 r 是定义在 Γ 上的一个广义函数 $r: \Gamma \to [-\infty, +\infty]$，于是我们常称 r 为报酬函数。r 可以只取有限实值，也可取广义实值。当 $r(i, a) \leqslant 0$ 时，它表示的实际上是费用。$r(i, a)$ 的含义随具体应用问题的不同而有所不同。

(5) V 为准则(Criterion)函数(也称为目标(Objective)函数)，可分为期望总报酬的(包括折扣的，正的和负的等)和平均的等多种。详细的定义将在后面给出。

注 1.1　(1) 模型的时间可以是离散变化的，也可以是连续变化的；可以在有限时间段上取值，也可以在无限时间段上取值；S，$A(i)(i \in S)$ 均可以是有限集或至少有一个是可列

集，甚至任意非空集；状态转移概率族 p 可以是时齐的(即与阶段数 n 无关)，也可以是非时齐的，或更广的半马氏或漂移的；p 也可是未知的、状态信息延迟的、部分可观察的；报酬函数 r 可以是有界的、各种各样无界的，甚至与系统的历史有关的；准则函数也可是各种各样的。这几种成分各取一种即得到一个相应的马氏决策过程模型。

(2) 我们在本书中之所以假定 S 和 $A(i)$均是可数集，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和决策空间所带来的测度论方面的纷扰。因为当 S 或 $A(i)$为非可数无穷集时，许多叙述的合理性应予以证明。

(3)在实际应用时，状态 i 处可用的决策 a 与状态 $j(\neq i)$处可用的决策 a，所代表的实际意义可以完全不同，a 仅代表在可用的决策集中编号为 a 的决策。

为方便起见，本章中将 DTMDP 简记为 MDP。从以上的定义可以看出，MDP 的历史由相继的状态和决策组成，其形式为

$$h_n = (i_0, a_0, i_1, a_1, \cdots, i_{n-1}, a_{n-1}, i_n), \ n \geqslant 0$$

其中 $i_k \in S$ 和 $a_k \in A(i_k)$分别表示在第 k 个观察时刻点系统所处的状态和采取的决策($k=0, 1, 2, \cdots, n-1$)，$i_n \in S$ 为系统当前所处的状态。称 h_n 为系统到 n(即第 n 个决策时刻)时的一个历史，其全体记为 H_n。若使用记号 Γ，则有 $H_n = \Gamma^n \times S$。

系统的一个策略 π 是指一个序列 $\pi=(\pi_0, \pi_1, \cdots)$，当系统到 n 时的历史为 h_n 时，该策略则按$A(i_n)$ 上的概率分布 $\pi_n(\cdot \mid h_n)$采取决策。策略全体记为 Π，并称之为 MDP 的策略空间。如果 π 满足条件：

$$\pi_n(\cdot \mid h_n) = \pi_n(\cdot \mid i_0, i_n), \quad h_n \in H_n, \ n \geqslant 0$$

则称 π 是半马氏(Semi-Markov)策略，其全体记为 Π_{sm}。进而，如果

$$\pi_n(\cdot \mid h_n) = \pi_n(\cdot \mid i_n), \quad h_n \in H_n, \ n \geqslant 0$$

与历史完全无关，则称 π 为随机马氏策略，其全体记为 Π_m。在随机马氏策略下，系统在 n 时所采取的决策仅仅依赖于所处的决策时刻 n 和状态 i_n。进一步地，对 $\pi \in \Pi_m$，如果还有 $\pi_n(\cdot \mid i_n)=\pi_0(\cdot \mid i_n)$与 n 无关，则称之为随机平稳策略，简记为 π_0^∞，其全体记为 Π_s。

定义决策函数集 $F = \underset{i}{\times} A(i)$。$f \in F$ 可看作为从 S 到 $\bigcup_{i \in S} A(i)$的一个映射，它满足条件 $f(i) \in A(i)$，$i \in S$。因此称 f 为决策函数。对 $\pi \in \Pi_m$，如果 π_n 均是退化的，即有 $f_n \in F$ 使 $\pi_n(f_n(i) \mid i)=1 (i \in S)$，则称此 π 为(确定性)马氏策略，并记为$(f_0, f_1, \cdots)$，其全体记为 Π_m^d。进而，称 $\pi=(f, f, \cdots)$(简记为 f^∞)为(确定性)平稳策略，其全体记为 Π_s^d。自然，Π_s^d 与 F 之间有一一对应关系。为记号简单起见，我们将随机平稳策略 π_0^∞ 简记为 π_0，将平稳策略 f^∞简记为 f。这样，我们将决策函数 f 与平稳策略 f^∞等同看待。为方便起见，也常将 f 称为平稳策略。

从上述定义可以看出，各类策略集间有如下的关系：

$$\Pi_s^d \subset \Pi_s, \quad \Pi_m^d \subset \Pi_m \subset \Pi_{sm} \subset \Pi$$

在上面所定义的策略中，平稳策略是形式最简单的策略。但在实际应用问题中，还要考虑更简单的策略，这将在有关内容中讨论。

1.2 报酬过程与准则函数

对 $n \geqslant 0$，我们用 X_n，Δ_n 分别表示在时刻 n 系统所处的状态和采取的决策，显然，它们都

是依赖于策略 π 的随机变量，从而 $(X_0, \Delta_0, X_1, \Delta_1, \cdots)$ 为一随机序列。对给定的策略 π，我们用 $\mathscr{L}(\pi)$ 来表示策略 π 下的这一随机序列。容易看出，其概率规律完全由初始状态概率分布（即 X_0 的概率分布）、转移概率族及策略 π 所确定，其状态空间为 $(S\times\bigcup_i A(i))^\infty$，其中 $X_n\in S$，$\Delta_n\in\bigcup_i A(i)$。还可看出，它与 MDP 模型(1.1.1)式中的报酬函数和准则函数无关。

为了表示随机序列 $\mathscr{L}(\pi)$ 中的概率转移规律与所采用策略 π 的关系，我们用 $P_\pi(E)$ 表示 $\mathscr{L}(\pi)$ 中事件 E 的概率。

根据策略 π 的定义以及马尔可夫性(即无后效性)的含义，容易猜想，对于一般的策略 π，$\mathscr{L}(\pi)$ 不是一个马氏过程，仅当 π 是马氏策略时，$\mathscr{L}(\pi)$ 才是一个马氏过程。对此，我们有下述定理。

定理 2.1 对策略 $\pi=(\pi_0, \pi_1, \cdots)$，

(1) 若 $\pi\in\Pi_m$ 为随机马氏策略，则 $\mathscr{L}(\pi)$ 是一非时齐马氏链，其 n 时的转移概率为

$$P_\pi\{X_{n+1}=j, \Delta_{n-1}=b|X_n=i, \Delta_n=a\}=p_{ij}(a)\pi_{n+1}(b|j), \quad (i,a),(j,b)\in\Gamma \tag{1.2.1}$$

进而，$\mathscr{L}(\pi)$ 的状态子序列 $\mathscr{L}_S(\pi)=\{X_0, X_1, \cdots\}$ 亦是一个非时齐马氏链，其 n 时的状态转移概率为

$$P_\pi\{X_{n+1}=j|X_n=i\}=\sum_{a\in A(i)}\pi_n(a|i)p_{ij}(a), \quad i,j\in S \tag{1.2.2}$$

(2) 若 $\pi=\pi_0^\infty\in\Pi_s$ 为随机平稳策略，则 $\mathscr{L}(\pi)$ 是一时齐马氏链，其状态转移概率为

$$P_{\pi_0}\{X_{n+1}=j, \Delta_{n+1}=b|X_n=i, \Delta_n=a\}=p_{ij}(a)\pi_0(b|j), \quad (i,a),(j,b)\in\Gamma \tag{1.2.3}$$

进而，$\mathscr{L}_S(\pi)=\{X_0, X_1, \cdots\}$ 亦是一个时齐马氏链，其状态转移概率为

$$P_{\pi_0}\{X_{n+1}=j|X_n=i\}=\sum_{a\in A(i)}\pi_0(a|i)p_{ij}(a), \quad i,j\in S \tag{1.2.4}$$

证明 (1) 对任一 $n\geqslant 0$ 及历史 $h_n=(i_0, a_0, i_1, a_1, \cdots, i_{n-1}, a_{n-1}, i)\in H_n$，$a\in A(i)$，$(j,b)\in\Gamma$，由 $\pi\in\Pi_m$ 的马氏性及概率乘法公式可知

$$\begin{aligned}
&P_\pi\{X_{n-1}=j, \Delta_{n+1}=b|h_n, \Delta_n=a\}\\
&\quad=P_\pi\{X_{n+1}=j|h_n, \Delta_n=a\}P_\pi\{\Delta_{n+1}=b|h_n, \Delta_n=a, X_{n+1}=j\}\\
&\quad=p_{ij}(a)\pi_{n+1}(b|(h_n, a, j))\\
&\quad=p_{ij}(a)\pi_{n+1}(b|j)\\
&\quad=P_\pi\{X_{n+1}=j|X_n=i, \Delta_n=a\}P_\pi\{\Delta_{n+1}=b|X_{n+1}=j\}\\
&\quad=P_\pi\{X_{n+1}=j, \Delta_{n+1}=b|X_n=i, \Delta_n=a\}
\end{aligned}$$

即 $\mathscr{L}(\pi)$ 为非时齐的马氏链且(1.2.1)式成立。

对于 $\mathscr{L}_S(\pi)$，设 $n\geqslant 0$，$i_0, i_1, \cdots, i_{n-1}, i, j\in S$，由全概率公式及 $\pi\in\Pi_m^d$ 的定义知

$$\begin{aligned}
&P_\pi\{X_{n+1}=j|X_0=i_0, X_1=i_1, \cdots, X_n=i\}\\
&\quad=\sum_a P_\pi\{\Delta_n=a|X_0=i_0, X_1=i_1, \cdots, X_n=i\}\\
&\qquad\cdot P_\pi\{X_{n+1}=j|X_0=i_0, X_1=i_1, \cdots, X_n=i, \Delta_n=a\}\\
&\quad=\sum_a\pi_n(a|i)p_{ij}(a)
\end{aligned}$$

因此 $\mathscr{L}_S(\pi)$ 为一非时齐马氏链且(1.2.2)式成立。

(2) 由(1)求得的转移概率表达式即可知道。 □

对于策略 $\pi=(f_0, f_1, \cdots)\in \Pi_m^d$，$\mathscr{L}(\pi)$中的 Δ_n 是 X_n 的一个函数：

$$\Delta_n = f_n(X_n), \quad n \geqslant 0$$

因此当 X_n 确定时，Δ_n 也就确定了，即 Δ_n 的随机性完全由 X_n 引起。

在 MDP 中，对任一策略 π，与随机序列 $\mathscr{L}(\pi)$相关的还有另一个随机序列

$$\mathscr{R}(\pi) = \{R_0, R_1, R_2, \cdots\}$$

其中 $R_n=r(X_n, \Delta_n)$是系统在时刻 n 时获得的报酬，故我们称之为报酬过程，但 $\mathscr{R}(\pi)$并不是一个独立的随机序列，它是依赖于 $\mathscr{L}(\pi)$的。因此，也称 $\mathscr{L}(\pi)$为带报酬的随机序列，有时也称$\{X_0, \Delta_0, R_0, X_1, \Delta_1, R_1, X_2, \Delta_2, R_2, \cdots\}$为报酬过程。

对于一般的随机序列，要确定它的转移概率，研究其稳态分布等性质，但对于带报酬的随机序列 $\mathscr{L}(\pi)$，我们要研究的主要是与 $\mathscr{R}(\pi)$相关的某些数字特征(如数学期望，方差等)，并用其来比较策略的优劣。

我们在下面假定所述的数学期望都是存在的，例如当报酬函数一致有界时。

在策略 π 下的数学期望用 $E_\pi\{\cdot\}$表示，简记

$$P_{\pi,i}\{\cdot\} = P_\pi\{\cdot \mid X_0 = i\}, \quad E_{\pi,i}\{\cdot\} = E_\pi\{\cdot \mid X_0 = i\}$$

于是在 π 下，$n=0$ 时从初始状态 i 出发，在 n 时刻获得的期望报酬为

$$E_{\pi,i}\{r(X_n, \Delta_n)\} = \sum_{(j,a)\in\Gamma} P_{\pi,i}\{X_n = j, \Delta_n = a\} r(j, a) \tag{1.2.5}$$

下面介绍马氏决策过程中常用的准则。

1. 有限阶段总报酬准则

对 $N\geqslant 0$，策略 π 下的 N 阶段期望总报酬定义为

$$V_N(\pi, i) = \sum_{n=0}^{N-1} E_{\pi,i}\{r(X_n, \Delta_n\}, \quad i \in S \tag{1.2.6}$$

它表示使用策略 π，在 0 时从状态 i 出发的条件下，系统直到 $N-1$ 时所获得的期望总报酬。用 $\boldsymbol{V}_N(\pi)$表示第 i 个分量为 $V_N(\pi, i)$的列向量，当 S 可列时，$\boldsymbol{V}_N(\pi)$为可列维向量。

2. 折扣准则

有些问题，难于确定所考察系统的有效期有多长。有的问题，即使知道系统的有效期，只要单位时间长度取得足够短，总的阶段数仍很大，这促使我们要考虑无限阶段问题。由于长期期望总报酬

$$\sum_{n=0}^{\infty} E_{\pi,i}\{r(X_n, \Delta_n\}$$

往往不收敛或为无穷大，如 $r(i, a)\equiv 1$ 时为无穷大，因此，用长期期望总报酬作为准则就不一定有意义，需要附加一定的条件，如报酬函数非负或非正。为了克服这一点，我们引进一个称之为折扣因子的常数 $\beta\in(0, 1)$，其含义是：阶段 n 时获得的单位报酬仅值 $n-1$ 时的 β，从而仅值 0 时的 β^n。于是系统在周期 n 所获得的报酬 $r(X_n, \Delta_n)$折算到时刻 0 的值为 $\beta^n r(X_n, \Delta_n)$；在策略 π 下，从初始状态 i 出发的折算到时刻 0 的第 n 阶段的期望报酬为 $E_{\pi,i}\{\beta^n r(X_n, \Delta_n)\}$。当 $\beta=1$ 时即为无折扣时的情形。

策略 π 下的无限阶段期望折扣总报酬定义为

$$V_\beta(\pi, i) = \sum_{n=0}^{\infty} E_{\pi,i}\{\beta^n r(X_n, \Delta_n)\}, \ i \in S \tag{1.2.7}$$

记 $V_\beta(\pi)$ 为相应的列向量。

折扣准则具有一定的经济意义：基于经济上的利率(设为 ρ)，现在的一元钱与将来的一元钱不能等同看待。现在的一元钱存入银行，经若干周期后就大于一元钱，为了克服各周期收入的这种“不一致性”，应引入折扣因子 $\beta=1/(1+\rho)$，按复利计算。这就给出折扣准则函数。

从(1.2.7)式可以看出，折扣准则函数中，愈是周期小的收益看得愈重要。因此，折扣准则在本质上是一个前面有限阶段的准则函数，后面无限多个阶段不起多大作用(收敛级数的尾项趋于 0)。同样的原因，在有限阶段期望总报酬准则中也可考虑折扣。另外，当折扣因子 $\beta=1$ 时，折扣准则就成了无限阶段上的期望总报酬准则。因此，期望总报酬准则可以归到折扣准则中。

3. 平均准则

平均准则定义为

$$V(\pi, i) = \lim_{N\to\infty} \frac{1}{N+1} V_N(\pi, i), \quad i \in S \tag{1.2.8}$$

它表示策略 π 下从初始状态 i 出发长期运行每周期的平均期望报酬。记 $\boldsymbol{V}(\pi)$ 为相应的列向量。平均准则是无限阶段随机动态系统最优控制中的常用准则。当需要最优化的阶段足够长(实际问题中总是有限的)，而且较短阶段并不比较长阶段更重要时，就可使用平均准则。自然，此准则不适用于金融领域，因为对于资金来说，现在花的钱比将来花的更值钱，但对于相对较快地进入“稳态”运行的系统来说(如通信网络)，采用极限的时间平均报酬(费用)准则是合理的。

在数学处理上，在折扣准则下有非常完美的结果，但在平均准则下只在有限状态集和决策集的情形下有完美的结果。当状态集或决策集可列时，已有不少例子说明求解的困难性。这其中的问题较为复杂，从而发展出多种多样的方法。

对于平均准则，有限阶段平均期望报酬的极限不一定对所有的策略 π 和状态 i 存在，因此定义中取成下极限。下极限总存在，它表示在策略 π 下的最“坏”可能的渐近平均期望报酬，是一种“悲观”测度。与之相反，若在定义中取成上极限，则它也总存在，并且表示在策略 π 下的最“好”可能的渐近平均期望报酬，是一种“乐观”测度。

在折扣准则中，将来的报酬按折扣率 $\beta(0<\beta<1)$ 进行折扣，于是当阶段足够大时，所获报酬的作用越来越小，因此，折扣准则实质上只注重于考虑近期行为。与此相反，对于平均准则，近期所获报酬不起作用，它注重于长期的稳定行为。因此，折扣准则和平均准则是所考虑准则中的两个极端。

在最优控制问题中，准则的引入是为了比较策略的优劣，我们的目的是根据某一给定的准则在所有允许的策略中寻找一个使准则函数值达到最优的策略。对于函数值越小越优的情形(如 $r(i, a)$ 表示费用时)，只要将 $r(i, a)$ 改为 $-r(i, a)$ 就可将问题转化为最大化。所以本书中假定函数值越大越优。在所讨论的问题中，自然要求准则函数对所有的策略 π 和初始状态 i 是存在的。对于折扣准则，定义最优值函数为

$$V_\beta(i) = \sup\{V_\beta(\pi, i) \mid \pi \in \Pi\}, \quad i \in S$$

上式常常写为向量形式 $\boldsymbol{V}_\beta = \sup\{\boldsymbol{V}_\beta(\pi) \mid \pi \in \Pi\}$，本书中约定向量形式的上确界是依分量分别取的。进一步地，当 $V_\beta(i)$ 均有限时，对常数 $\varepsilon \geqslant 0$ 及状态 $i \in S$，如果策略 π^* 满足：

$$V_\beta(\pi^*, i) \geqslant V_\beta(i) - \varepsilon$$

则称 π^* 是在状态 i 处的 β 折扣 ε 最优策略；当上式对所有的 i 均成立时，称 π^* 为 β 折扣 ε 最优策略。β 折扣 0 最优策略就简称为 β 折扣最优策略。在不会引起混淆的情况下，我们略去“β 折扣”。当最优值函数 $V_\beta(i)$ 不一定有限时，ε 最优策略的定义与上述有所不同，具体的将在第 4 章中给出。

总报酬准则和平均准则下，最优的定义与上述完全类似。

在 MDP 中研究的准则已有近 20 种，如矩最优准则、Blackwell 准则以及样本路径平均准则(将在第 5 章中讨论)。样本路径平均准则是平均准则在样本路径下的形式，当 $X_0=i$ 时它定义为

$$J(\pi, i) = \overline{\lim_{N\to\infty}} \frac{1}{N+1} \sum_{n=0}^{N} r(X_n, \Delta_n), \quad i \in S \tag{1.2.9}$$

因此，$J(\pi, i)$ 是一个广义实值随机变量。

对于样本路径平均准则，最优的定义如下：称策略 π^* 是样本路径平均准则最优的(或称几乎处处平均准则最优的)，如果存在一个常数 c^*，使得

$$J(\pi^*, i) = c^*, \qquad P_{\pi^*, i}\text{-a. s.}, \quad i \in S$$
$$J(\pi, i) \leqslant c^*, \qquad P_{\pi, i}\text{-a. s.}, \quad i \in S$$

此时我们称 c^* 为样本路径最优平均值。容易看出，这种准则与前面介绍的其它准则有很大的差别。

文献[9]研究了任意数值准则的 Borel 状态空间 MDP，此准则将平均准则，折扣准则，总报酬准则作为特例。文献[10]，[11]等中讨论了其它多种准则，以及各种准则之间的关系。

1.3 历史

追根溯源，MDP 可归根到文献[26]中的序贯分析(Sequential Analysis)和统计决策函数(StatiStical Decision Function)。在 40 年代末 50 年代初关于序贯对策论的研究中已涉及到 MDP 的一些本质概念，MDP 可看作只有一个对策者的对策问题。文献[3]，[4]，以及提出了随机动态规划基本机制并使用了压缩映射方法的文献[24]等在与 MDP 有关的方面作出了特别的贡献。其中讨论最多的是有限阶段，对此，向后归纳法就可进行较为完美的处理。但有限阶段与无限阶段是相当不同的。MDP 早期的工作还出现于经济学研究中[1]。

Howard 的书[37]奠定了 MDP 作为一个独立研究学科的基础，书中研究了折扣准则和平均准则，以及值迭代法和策略迭代法。Howard 是第一个研究平均准则的作者，他所提出的策略迭代法是在算法方面的第一个里程碑。另外，他证明了对于有限状态集和决策集，其策略迭代法所得到的平稳策略在平稳策略范围内是最优的。文献[5]和[25]分别独立地证明了策略迭代法所得到的平稳策略在整个策略范围内也是最优的。

Blackwell 在文献[36]中对理论方面进行了开创性的研究，对有限(状态和决策的)折扣 MDP 得到了许多重要的结果，同时还提出了研究平均准则的折扣因子消失法(Vanishing Discount Approach)：将平均准则作为折扣准则当折扣因子趋于 1(即折扣因子的作用消失)时的极限，证明了存在一个当折扣因子充分接近于 1 时均为折扣最优的平稳策略，这种类型的最优现在称之为 Blackwell 最优。在文献[14]的 4.6 中使用 Tauberian 定理也讨论了折扣准则

与平均准则间的关系，这种方法首先是由文献[12]在证明随机对策问题平均准则最优平稳策略的存在性时提出的，文献[5]将之用于证明 Blackwell 最优策略的平均准则最优性。当状态集或决策集非有限时，Blackwell 最优策略不一定存在。实际上，此时的平均最优策略也不一定存在(见第 5 章例 1.1)。平均准则比折扣准则要涉及到更多的马氏过程的性质，对它的研究要复杂一些，同时，其内容也更为丰富一些。文献[2]等在一定条件下证明了存在最优策略。文献[6]研究了可数状态集、有限决策集的 MDP，提出了平均准则最优方程(ACOE)，其作用与折扣准则中的最优方程相同。文献[21]将 Blackwell 的折扣因子消失法用于从折扣最优方程获得 ACOE。1989 年，文献[23]更进一步提出了平均准则最优不等式的概念，其作用与ACOE 相同，而条件更弱。目前，这方面的工作仍在不断进行。

最初提出报酬函数时，都假定为有界的，文献[13]、[19]、[27]、[28]等研究了折扣准则的无界报酬条件，文献[29]提出了更弱的且在其函数值空间上范数不再有限的新条件，从而不能用传统的算子方法研究。文献[32]将无界报酬条件推广到非时齐情形。

相应于马氏过程中的离散时间马氏链、连续时间马氏过程和半马氏过程，在 MDP 中亦有离散时间马氏决策过程(DTMDP)、连续时间马氏决策过程(Continuous Time MDP，简记为 CTMDP)和半马氏决策过程(Semi-MDP，SMDP)等。文献[37]中已经提出了 CTMDP，文献[20]进一步研究了有限状态和决策集条件下的 CTMDP，文献[17]、[18]则研究了可数状态和有限决策集的情形。

SMDP 是由文献[15]和文献[16]各自独立地将 MDP 推广到半马氏过程而建立的。之后，文献[7]、[8]、[19]等分别作了相当多的讨论。作为 DTMDP 的推广，本质上，SMDP 又可转化为 DTMDP 来讨论[22]，因此有关的研究都是相互交叉进行的。

除了以上介绍的三类 MDP 模型之外，还有更一般、更接近于实际问题的各种模型，如状态部分可观察模型、多目标模型、自适应模型、带约束条件的模型、摄动的模型等等，我们将它们统称为一般化 MDP，这方面的综述可见文献[31]。文献[30]、[33]、[34]、[35]等还提出了随机环境 MDP 和混合 MDP，这是一类参数随环境的变化而变化的 MDP。

依据模型参数是否与时间有关，我们可将 MDP 分为时齐模型和非时齐模型。对于非时齐模型，亦有与以上介绍的时齐模型同样或相近的内容需研究。

马氏决策过程这一学科的称呼方法有许多种，如马尔可夫决策规划(Markov Decision Programming)、马尔可夫决策问题(Markovian Decision Problems)、随机动态规划(Stochastic Dynamic Programming)、序贯决策过程(Sequential Decision Processes)、受控马尔可夫过程(序列，链)(Controlled Markov Processes(sequences，chains))等。有些作者(如文献[19])则干脆称之为动态规划。从目前国内外所发表的文章来看，用得较多的是“马尔可夫决策过程”这一术语，它强调的是一种动态的“决策过程”。作者也将本书冠以“马尔可夫决策过程引论”。好在无论是“马尔可夫决策过程”还是“马尔可夫决策规划”，它们的英文首字母缩写均为 MDP。所以只要写出 MDP，也就不会引起混淆了。

参考文献

[1] Arrow K J，Harris T and Marshak J. Optimal inventory policy. Econometrica，1951，19：250～272

[2] Bather J. Optimal decision procedures for finite Markov chains，Part 1：examples. Part 2：communicating systems. Part 3：general convex systems. Adv. Appl. Prob.，1973，5：328～339；521～540；541～553