动态规划
确定性和随机模型

[美] D.P.柏塞克斯 著
李人厚 韩崇昭 译

西安交通大学出版社

外国教材精选

动态规划 确定性和随机模型

[美] D. P. 柏塞克斯 著

李人厚 韩崇昭 译

西安交通大学出版社

内 容 简 介

本书介绍了动态规划的概念、算法及其应用。对不完全状态信息情况下的动态规划问题，书中作了详尽的叙述，尤其对无限时域动态规划及其应用的阐述较为精辟，不多见于其它文献。作者对问题的表述尽可能深入浅出，而对动态规划在不同领域中的应用，以统一的数学方法予以处理，很有特色。

本书适合作为自动控制、系统工程、计算机科学与工程以及应用数学等专业大学本科生、研究生的教材或参考书，也可供工程技术人员参考。

Dynamic Programming: Deterministic and Stochastic Models

D. P. Bertsekas

Prentice-Hall, 1987

动态规划 确定性和随机模型

[美] D. P. 柏塞克斯 著

李人厚 韩崇昭 译

*

西安交通大学出版社出版

（西安市咸宁路28号）

西安电子科技大学出版社印刷厂印装

陕西省新华书店发行 各地新华书店经售

*

开本 787×1092 1/16 印张 18.5 字数 448 千字

1990 年 1 月第 1 版 1990 年 1 月第 1 次印刷

印数：1—1 000

ISBN 7-5605-0308-×/O·57 定价：4.10 元

《外国教材精选》总序

近十年来，我国高等学校教材建设在经历了从无到有、巩固提高的过程之后，目前正进入向高质量、高层次、多品种发展的欣欣向荣，百花争艳时期。现在，教材建设仍是高等学校教学改革的重要方面，这里也存在一个改革开放的问题。在这种形势下，精选国外一些有影响、有特色、特别是世界上著名大学现用的优秀教材翻译出版，无疑将对我国当前教材建设起到借鉴、促进和填补某些学科空白的积极作用。为此，西安交通大学出版社决定组织翻译出版一套《外国教材精选》系列书。

外国教材专业面广、类型繁多、层次各异，我们这套系列书在选题时以专业面较广，内容新颖或具有明显特色的教材为目标，具体原则如下：

1. 列选的教材不限国别及语种，以便博采众长。

2. 国外著名经典性教材，多次修订重版经久不衰者。

3. 最新出版，为国外著名大学所采用，有独特风格、体系，能反映国外教育动向，可供借鉴者。

4. 反映最新科技成果，能填补国内某学科教材空缺者。

根据我校具体情况，《外国教材精选》系列书将以电类教材(含电力、电子、计算机与信息科学)为主。今后随着形势的发展和需要，再进一步组织其他学科的国外先进教材翻译出版。

我们期望这套系列书，不仅是高等学校的学生和教师的良师益友；而且对已在生产科研第一线的广大科技工作者的知识更新，吸取国外科技新成果方面也大有裨益。

这套《外国教材精选》虽然从搜求原著、遴选、翻译、审校等方面都做了较细致的工作，但从浩如烟海的外国教材中精选少数形成一套系列书，对我们毕竟还是一种尝试。书源还不够充分，经验也感不足，缺点在所难免，诚挚地希望读者予以指正。

西安交通大学《外国教材精选》编委会

1988年6月

译 者 序

动态规划自贝尔曼(Bellman)开创以来，在理论研究和实际应用两个方面都有着重大的进展，它已成为动态系统优化的重要手段之一。

呈现在读者面前的这套教材是柏塞克斯(D.P.Bertsekas)教授十多年来在美国斯坦福大学、伊里诺斯大学、麻省理工学院等著名大学中讲课内容的结晶，是一本很具特色的优秀著作。作者以完整的理论体系阐述了确定性和随机性动态规划的基本理论，并把它在自动控制、人工智能、系统工程等学科中的应用融会贯通。书中不仅对有限时域、完全状态信息的动态规划作了系统的论述，而且对无限时域、不完全状态信息的动态规划问题也作了较严格的探讨，这是一般参考书中所很少涉及的。

本书在论述中，尽量避免采用繁琐的数学推导，作者往往从实际工程应用例子入手，其中包括数学发展史上许多著名的例题，引人入胜地介绍了动态规划的基本概念和方法，深入浅出，脍炙人口，书中各章带有启发性的习题，更会引起读者对动态规划的广泛兴趣。所以，本书很适合作教材，适用于工科大学自动控制、计算机科学、系统工程、应用数学等专业的本科生和研究生。对这些领域中的工程技术人员，也无疑是一本很好的参考书。

书中最后部分引入了无限时域问题的一些主要算法，如逐次逼近法，策略迭代法，和自适应集结法。这些都是作者最近几年来对动态规划理论和方法的新贡献，也是研究动态规划的热门课题。

愿本书的翻译出版对推动动态规划的学术研究和工程应用有所裨益。

在翻译过程中，对原著的个别印刷错误和笔误作了修正。限于译者的水平，译文中的错误和不妥之处在所难免，敬希广大读者不吝赐教。

游兆永教授对本书的译稿进行了认真的审阅，并提出许多宝贵意见，谨表感谢。

李人厚　　韩崇昭

一九八八年九月

于西安交通大学

原 著 序

本书是根据14年来在斯坦福大学、伊里诺斯大学和麻省理工学院讲授动态规划和随机控制的课程发展而来的。本书的目的是从工程、运筹学、以至经济学和应用数学的某些方面，提供适合于广大读者的关于这个专题的统一处理方法。例如，我们同时处理现代控制理论中众所周知的随机控制问题；运筹学中普遍涉及的马尔科夫决策问题；以及在计算机科学中经常强调的组合问题。书中用大量不同类型的例题来阐明理论，其中许多例题本身还包含重要的应用。这些例题可以按类概括而彼此独立，所以教师可有所侧重地应用其中一类，使课程适合于自己的对象。

很好掌握概率导论和本科生数学是本书在教学基础方面的要求。这包括一学期的概率论与普通微积分、实分析、向量矩阵代数、以及几乎所有本科生在第四学年都要学习的基础优化理论。在本书附录中提供了这些材料的综述。关于动态系统理论、优化或控制的先修课程或背景材料无疑对读者会有帮助，我们认为本材料的内容是合理地自成体系的。

动态规划是能用基本分析充分阐明和解释的一种概念上简单的技术。然而在数学上如要严格地全面处理，则随机动态规划需要测度概率论方面的复杂知识。我决定绕开这些复杂的数学，用一般的假定来进行分析。它仅当所探讨的概率空间为可数时才认为是严格的。对此问题的数学上的严格处理，在我和 Steven Shreve 合著的专著《随机最优控制：离散时间情况》(Stochastic Optimal Control: The Discrete Time Case, Academic Press, 1978)中予以完成。这本专著是对本教材的补充，并为本书中某些未十分成熟的专题提供了坚实的基础。

我衷心感谢对本书作出贡献的许多个人和院校。在与 Steven Shreve 一起完成1978年的专著时，使我加深了对专题的理解。在此期间完善了处理无限时域问题的几个证明和结果，现已成为本教材内容的一部分。Michael Caramanis, Lennart Ljung 和 John Tsitsiklis 讲授了本书的修订本，而且提供了一些实质性评论和习题。这些年来，在与几位有才干的助教交往中得益匪浅。这些交往中，特别要提到 Paris Canellakis, Panos Constantopoulos 和 John Tsitsiklis。很多同事提供了有价值的见解和信息，特别是David Castanon 和 Krishna Pattipati。(美国)国家科学基金会(NSF)支持了第五章所述关于无限时域问题的研究，麻省理工学院以其令人鼓舞的教学和研究环境，成为实现该项工作的理想场所。

D.P.柏塞克斯

(Dimitri P. Bertsekas)

目　录

第一章　动态规划算法

§1.1　基本问题

本书研究分阶段进行决策的情况。虽然每次决策所产生的结果无法充分预见，但在下次决策之前可以进行观测。其目标是对某个代价函数——被认为是希望结果的一个数学表达式求极小。

这种问题的关键是不能用孤立的观点看待决策。决策者必须在希望当前代价低与未来难以避免的高代价可能性之间进行平衡。动态规划把握了这一思想，它在每个阶段选择一个使当前阶段代价之和最小，而且对未来阶段也可期望是最好的决策。

范围很广的一类问题可以按这种方式来处理，而本书尽力保持这一主导思想，不因与问题结构不相干的假设而造成混乱。因此，本节对有限阶段(有限时域)动态系统的最优控制问题，建立了一种应用范围很广的模型。前四章我们遇到的都是这种模型。其无限时域形式将是后三章研究的主题。

基本问题的两个主要特征决定了其结构：(1)一个作为基础的离散时间动态系统；(2)对整个时间相加的一个代价泛函。动态系统具有如下形式

$$x_{k+1}=f_k(x_k,\ u_k,\ w_k),\quad k=0,\ 1,\ \cdots,\ N-1,$$

其中

k：离散时间标号；

x_k：系统状态，总合了与未来优化有关的过去的信息；

u_k：在已知状态 x_k 的情况下，在时刻 k 所选择的控制或决策变量；

w_k：随机参数(也称为扰动或噪声)；

N：时域或施加控制的次数。

代价泛函在下述意义上具有相加性，即在每个时刻 k，引起代价 $g_k(x_k,\ u_k,\ w_k)$，而沿任意系统样本轨线的总代价为

$$g_N(x_N)+\sum_{k=0}^{N-1}g_k(x_k,\ u_k,\ w_k),$$

其中 $g_N(x_N)$是在过程结束时所产生的终端代价。然而，因为 w_k 的存在，代价函数一般是一个随机变量，优化无意义。所以我们把问题按如下方式来表述，即要选择控制 u_0，u_1，$\cdots$，u_{N-1}，使得期望代价

$$E\left\{g_N(x_N)+\sum_{k=0}^{N-1}g_k(x_k,\ u_k,\ w_k)\right\}$$

为极小，式中是相对于所涉及随机变量的联合分布求期望。

稍后我们会对上面所用术语给出更精确的定义。首先我们通过几个例题提供某些带方向性的问题。

库存控制例题

考虑在N个时间周期的每一个开始时刻订购一定数量货物以满足随机需求的问题。我们记

x_k：表示在第k个周期开始时刻具有的存货量；

u_k：表示在第k个周期开始时刻的订货量(且立即发货)；

w_k：表示第k个周期的需求量，具有给定的概率分布。假定w_0，…，w_{N-1}为独立的随机变量，且假定超额的订货被拖欠，一旦具有库存立即予以满足。这样，存货量按离散时间(或差分)方程演变：

$$x_{k+1}=x_k+u_k-w_k,$$

其中负的存货量相应于拖欠的需求(见图 1.1)。

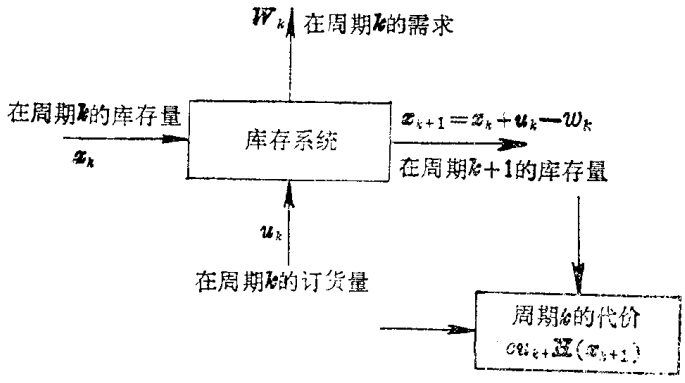

图 1.1　库存控制例题。在周期k的存货量(状态)x_k，在周期k的订货量(控制)u_k，以及在周期k的需求量(随机扰动)w_k，利用差分方程$x_{k+1}=x_k+u_k-w_k$，决定了下一周期$k+1$的存货量，以及第k个周期的代价。

每个周期k所造成的代价由两部分组成：(1)采购价格cu_k，其中c是订购单位货物的价格；(2)代价$H(x_{k+1})$，它表示在周期结束时刻正的存货量$x_{k+1}>0$(对超库存的保管费用)，或负的存货量$x_{k+1}<0$(不满足需求的缺货代价)所引起的惩罚。利用方程$x_{k+1}=x_k+u_k-w_k$，可以写出周期k的代价函数为

$$cu_k+H(x_k+u_k-w_k),$$

而且N个周期的总期望代价为

$$E\left\{\sum_{k=0}^{N-1}cu_k+H(x_k+u_k-w_k)\right\}.$$

我们的目标是选择合适的订货量u_0，…，u_{N-1}，在自然约束$u_k\geqslant 0$，$k=0$，…，$N-1$条件下，使上述代价函数为极小。一种可能性是在时刻 0 选择所有的订货量u_0，…，u_{N-1}，而不必等到知道后来的需求水平。然而比较好的选择是将订货u_k推迟到时刻k，此时可以知道当时的存货水平。这种运行模式涉及信息的收集，以及基于有用信息而作出的序贯决策，这在动态规划中是至关重要的。这就意味着，我们对选择库存订货的最优数值并不真正感兴趣，感兴趣的是寻求一个最优规则，以便在每个周期k，对每个可能发生的存货量x_k，选择

订货量 u_k。这就是一个"行动对策略"的特性。数学上，这个问题就是寻求一个函数序列μ_k，$k=0$，…，$N-1$，把存货量 x_k 映射到订货量 u_k，使得总的期望代价为极小。μ_k 的意义是，对每个 k 和 x_k 的可能值，

$\mu_k(x_k)$ = 在时刻 k，当存货量为 x_k 时，应该订货的总量。

序列 $\pi=\{\mu_0, \mu_1, \cdots, \mu_{N-1}\}$ 也称为一个控制律或一个策略。对每个这样的 π，在固定的初始存货量 x_0 情况下相应的代价函数为

$$J_\pi(x_0)=E\left\{\sum_{k=0}^{N-1} c\mu_k(x_k)+H[x_k+\mu_k(x_k)-w_k]\right\},$$

我们的目标是对固定的 x_0，对所有容许的 π 使 $J_\pi(x_0)$ 为极小。在§2.2节将会证明，合理选择代价函数 H，最优订货规则具有如下形式

$$\mu_k(x_k)=\begin{cases} S_k-x_k, & \text{若 } x_k<S_k, \\ 0, & \text{若 } x_k\geqslant S_k, \end{cases}$$

其中 S_k 是由问题的数据确定的适当的阈值。换句话说，当存货量低于阈值 S_k 时，订货只要使存货量达到 S_k 就行了。

上面的例题说明建立基本问题的主要组成部分：

1. 具有如下形式的离散时间系统

$$x_{k+1}=f_k(x_k, u_k, w_k),$$

其中 f_k 是某个函数；此例中 $f_k(x_k, u_k, w_k)=x_k+u_k-w_k$。

2. 独立的随机参数 w_k。容许 w_k 的概率分布依赖于 x_k 和 u_k。这将具有普遍性；在这个例子中，我们可以认为需求量 w_k 受当前存货水平影响。

3. 控制约束。在此例中 $u_k\geqslant 0$。在一般情况下，约束集合将依赖于 x_k 和时标 k，即$u_k\in U_k(x_k)$。为了理解约束如何依赖于 x_k，可在库存的例题中想象一种情况，即存在一个能调节存货量的上界 B，使 $u_k\leqslant B-x_k$。

4. 具有如下形式的相加性代价函数

$$E\left\{g_N(x_N)+\sum_{k=0}^{N-1} g_k(x_k, u_k, w_k)\right\},$$

其中 g_k，$k=0$，…，N 是某些函数；在前面的例题中，

$g_N(x_N)=0$，而 $g_k(x_k, u_k, w_k)=cu_k+H(x_k+u_k-w_k)$。

5. 对控制律的优化。即对每个 k 和 x_k 的可能值，选择 u_k 的规则。

在前面的例题中，状态 x_k 是实数。在另外的情况下，状态是一个 n 维向量。然而，也有可能状态取值于一个离散集合，如整数集合，或者甚至是一个有限集合。

当存货量以整单位来计量时(如汽车)，其中每一个量就是 x_k，u_k 或 w_k 的一个有效分数，此时库存问题以离散形式描述更自然。于是，把所有整数集而不是实数集取为状态空间更合适。当然，系统方程和每个周期代价函数的形式保持不变。

在另一些系统中，状态本质上是离散的，不存在问题的连续对应部分。这样的系统往往根据状态间的转移概率可方便地加以确定。我们需要知道的是 $p_{ij}(u,k)$，它定义为在 k 时刻，给定当前的状态 x_k 为 i，控制 u_k 选为 u 情况下，下一个状态 x_{k+1} 为 j 的概率，即

$$p_{ij}(u, k)=P\{x_{k+1}=j \mid x_k=i, u_k=u\}.$$

[如果系统是平稳的，即以前的概率与 k 无关，可去掉自变量 k，用 $p_{ij}(u)$ 代替 $p_{ij}(u,k)$。]这种系统可以换一种形式，按如下离散时间系统方程来描述。

$$x_{k+1}=w_k,$$

式中随机参数 w_k 的概率分布为

$$P\{w_k=j\,|\,x_k=i,\ u_k=u\}=p_{ij}(u,\ k).$$

根据所碰到的情况，可以按爱好用差分方程或用转移概率来描述系统。我们将用例题来阐明这些思想。

排队例题

考虑可容纳 n 个顾客的房间，且在 N 个时间周期内运行的一个排队系统(见图 1.2)。假定对顾客的服务只能在一个周期的开始(或终止)时刻开始(或终止)。在一个时间周期内，m 个顾客到达的概率 p_m 给定，而且在两个不同周期的到达数是相互独立的。顾客一经发现系统客满就离去，而且以后不再进来。系统提供两种服务：快服务和慢服务，而且分别具有每周

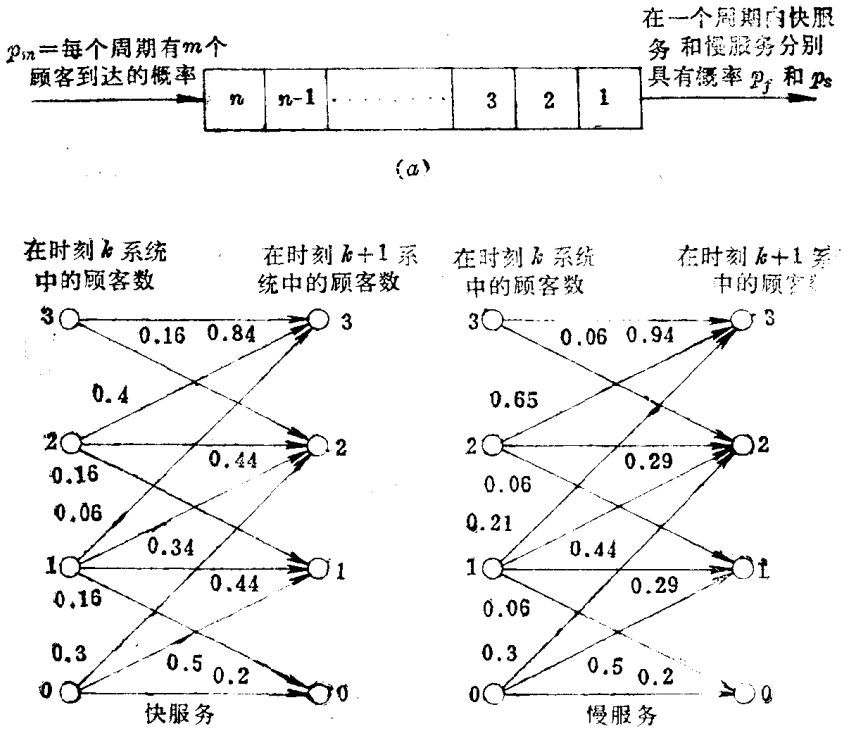

图 1.2　能容纳 n 个顾客房间的排队系统。在任一时间周期，服务可在快慢之间转换，使得顾客等待和服务代价之和为极小：(a) 具有能容纳 n 个顾客房间及两种服务的排队系统；(b) 快和慢服务的转移概率图。假设的数据为 $n=3$，$p_0=0.2$，$p_1=0.5$，$p_2=0.3$，对 $m>0$，$p_m=0$，且 $q_f=0.8$，$q_s=0.3$。

期价格 c_f 和 c_s。服务可在每个周期的开始时刻在快和慢之间转换。如果在某一周期提供快(慢)服务，在该周期开始时刻接受服务的顾客，将以概率 q_f (相应地，q_s) 在这个周期末终止服务，而与顾客已被服务的周期数以及系统中的顾客数无关($q_f>q_s$)。对每一个周期，系

统中有 i 个顾客，就有代价 $c(i)$。同样在最后一个周期末，有 i 个顾客留在系统时，有终端代价 $c(i)$。问题是选择每个周期所提供的服务类型，它作为周期开始时系统中顾客数目的函数，使得 N 个周期的总期望代价为极小。

这里把周期开始时系统中的顾客数作为状态，而把所提供的服务种类作为决策变量（控制）是合适的。那么，每个周期的代价是 $c(i)$ 加上 c_f 或 c_s（取决于提供快服务或慢服务）。我们来推导系统的转移概率。当在一个周期开始时系统为空，下一个状态是 j 的概率与提供的服务种类无关。当 $j<n$ 时，它等于 j 个顾客到达的给定概率

$$p_{0j}(u_f)=p_{0j}(u_s)=p_j,\quad j=0,\ 1,\ \cdots,\ n-1,$$

当 $j=n$ 时，它等于 n 个或更多顾客到达的概率：

$$p_{0n}(u_f)=p_{0n}(u_s)=\sum_{m=n}^{\infty}p_m .$$

当系统中至少有一个顾客（$i>0$）时，则有

$$p_{ij}(u_f)=0,\qquad \text{如果 } j<i-1,$$

$$p_{i(i-1)}(u_f)=q_f p_0,$$

$$\begin{aligned}p_{ij}(u_f)&=P\{j-i+1\text{ 到达，服务完毕}\}\\&\quad+P\{j-i\text{ 到达，服务未完毕}\}\\&=q_f p_{j-i+1}+(1-q_f)p_{j-i},\quad \text{如果 } i-1<j<n-1,\end{aligned}$$

$$p_{i(n-1)}(u_f)=q_f\sum_{m=n-i}^{\infty}p_m+(1-q_f)p_{n-1-i},$$

$$p_{in}(u_f)=(1-q_f)\sum_{m=n-i}^{\infty}p_m .$$

当提供慢服务时，转移概率也用这些公式表示，只是分别用 u_s 和 q_s 代替 u_f 和 q_f。

有时转移概率表示在一张图上，图中的连线代表各种状态间的转移，这就是所谓转移概率图，或简称转移图。图 1.2 表示了一个特殊情况，其中 $n=3$，$p_0=0.2$，$p_1=0.5$，$p_2=0.3$；对 $m>2$，$p_m=0$，且 $q_f=0.8$，$q_s=0.3$。

在下面建立方程的过程中，假定状态 x_k 取值于称为状态空间的某集合 S_k。我们将不要求 S_k 是一个有限集，或一个 n 维向量空间。动态规划令人鼓舞的方面是它的可用性与状态空间的性质关系甚少（虽然其有效性肯定依赖于 S_k）。为此，对 S_k 不作任何假定，这样处理起来方便些。以后，这些假设确会变成严重的障碍。我们同样允许 u_k 和 w_k 分别从某些非特定的空间 C_k 和 D_k 中取值。

基本问题

给定离散时间动态系统

$$x_{k+1}=f_k(x_k,\ u_k,\ w_k),\quad k=0,\ 1,\ \cdots,\ N-1, \tag{1.1}$$

式中状态 x_k 是空间 S_k 的元，控制 u_k 是空间 C_k 的元，而随机“扰动” w_k 就是空间 D_k 的元。控制 u_k 被限定在给定的 C_k 的非空子集 $U_k(x_k)$ 中取值，该子集依赖于当前的状态 x_k［对所有 $x_k\in S_k$ 和 k，$u_k\in U_k(x_k)$］。随机扰动 w_k 由概率测度 $P_k(\cdot\,|x_k,\ u_k)$ 来表征，它明显依赖于 x_k 和 u_k，但与先前的扰动 w_{k-1}，$\cdots$，w_0 的值无关。我们考虑控制律类（也称为策略），

它由函数序列 $\pi=\{\mu_0, \mu_1, \cdots, \mu_{N-1}\}$ 构成. 其中 u_k 把状态 x_k 映射到控制 $u_k=\mu_k(x_k)$，并且对所有 $x_k\in S_k$，使得 $\mu_k(x_k)\in U_k(x_k)$。这样的控制律称为容许控制律。

给定一个初态 x_0，问题是寻求一个容许控制律 $\pi=\{\mu_0, \mu_1, \cdots, \mu_{N-1}\}$，使代价泛函

$$J_\pi(x_0)=\underset{\substack{w_k\\ k=0,\cdots,N-1}}{E}\left\{g_N(x_N)+\sum_{k=0}^{N-1}g_k[x_k, \mu_k(x_k), w_k]\right\} \tag{1.2}$$

为极小，并受如下系统方程约束

$$x_{k+1}=f_k[x_k, \mu_k(x_k), w_k],\ k=0, 1, \cdots, N-1. \tag{1.3}$$

代价函数 g_k，$k=0, 1, \cdots, N$ 是给定的。

对给定的初态 x_0，最优控制律 π^* 就是使相应代价为极小

$$J_{\pi^*}(x_0)=\min_{\pi\in\Pi}J_\pi(x_0),$$

式中 Π 是所有容许控制律的集合。相应于 x_0 的最优代价记为 $J^*(x_0)$，即

$$J^*(x_0)=\min_{\pi\in\Pi}J_\pi(x_0).$$

我们把 J^* 看作为把每个初态 x_0 赋给为最优代价 $J^*(x_0)$ 的一个函数，称之为最优代价函数，或最优值函数。

[为了方便从事数学方面的读者，我们指出，上述方程中 min 表示数集 $\{J_\pi(x_0)\mid\pi\in\Pi\}$ 的最大下界(或下确界)。把上式写成 $J^*(x_0)=\inf_{\pi\in\Pi}J_\pi(x_0)$，更符合通常数学上所使用的表示方法。然而，(如在附录 B 中讨论的那样)，即使下确界不能达到，我们发现用 min 代替 inf 还是更方便些，这样少一些麻烦，也不会引起任何混淆。]

信息在基本问题中的作用

我们在前面已指出，策略 $\{\mu_0, \mu_1, \cdots, \mu_{N-1}\}$ 可以看成是一个计划，在每个时刻，对每个在该时刻可能出现的状态，它规定了所应施加的控制。这种运行模式隐含着信息的收集，认清这一点十分重要。由控制器接收到的信息，是每个时刻当前的状态值，而且在控制过程中直接利用这信息。因为 k 时刻的控制通过函数 μ_k 依赖于当前的状态 x_k(参见图 1.3)。这个信息的可用性确实具有重要意义。如果这一信息不利用，控制器就不能适应不期望的状态值，其结果使代价受到不利的影响。例如，在前面考虑过的库存控制问题中，在每个周期 k 开始时可用的信息是仓库的存货量 x_k。显然，对于库房经理来说，这个信息十分重要，他要根据当前存货量 x_k 的多少来调节订购量 u_k。

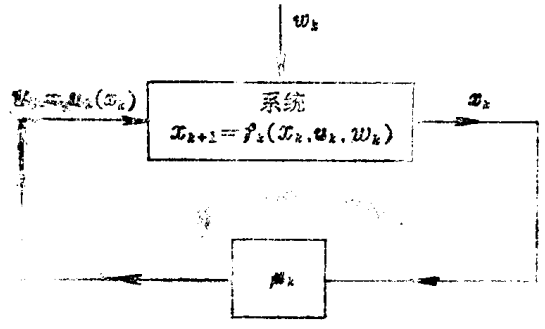

图 1.3 基本问题中的信息收集。在每个时刻 k，控制器观测到当前的状态 x_k，并且施加依赖于该状态的控制 $u_k=\mu_k(x_k)$。

然而应该注意，状态信息的可用性不会带来坏处，但也可能不会带来好处。譬如，在确定性控制问题中，不存在随机扰动，给定初态和控制序列，就可预报未来的状态。所以对所有控制序列 $\{u_0, u_1, \cdots, u_{N-1}\}$ 的优化，犹如对所有容许策略的优化一样，获得相同的最优代价。甚至在某些随机控制问题中也有相同的结论(见习题13)。这就引出一个有关的专题，假

定没有忘记任何信息，控制器实际上知道以前的状态和控制 x_0，u_0，…，x_{k-1}，u_{k-1}，以及当前的状态 x_k。于是提出一个问题：利用整个系统历史的策略是否优于只利用当前状态的策略？答案是否定的(见[B23])。其直觉上的理由是：对一给定的问题，和给定的时刻 k 和状态 x_k，所有未来的期望代价明显地只依赖于 x_k，而与以前的历史无关。

建立基本问题的理论限制

对动态规划算法进行开发之前，我们试图澄清不是建立在坚实数学基础上的某些方面的问题。这是数学上严密性的问题之一，本质上是高技术。非数学专业的读者不必关心这些问题，可以跳过本节其余部分而不会失去内容的连贯性。

首先，一旦采用容许控制律$\{\mu_0, \mu_1, \cdots, \mu_{N-1}\}$，对于每个阶段 $k=0,1,\cdots, N-1$，就会有如下事件序列：

1. 控制器观测到 x_k，并施加 $u_k=\mu_k(x_k)$。
2. 按照给定的概率测度 $P_k(\cdot\,|x_k, \mu_k(x_k))$，产生扰动 w_k。
3. 引起代价 $g_k[x_k, \mu_k(x_k), w_k]$，并加到以前的代价中去。
4. 按照系统方程

$$x_{k+1}=f_k[x_k, \mu_k(x_k), w_k].$$

产生下一个状态 x_{k+1}。如果这是最后一个阶段 ($k=N-1$)，将终点代价 $g_N(x_N)$ 加到前面的代价中去。否则，增加 k，下一阶段重复相同的事件序列。

这个过程是已经定义的，并且以精确的概率项来表达。然而，需要把代价函数

$$g_N(x_N)+\sum_{k=0}^{N-1} g_k[x_k, \mu_k(x_k), w_k]$$

看作为具有完全确定期望值的有定义的随机变量时，事情就复杂了。概率论的框架要求对每个$\{\mu_0, \mu_1, \cdots, \mu_{N-1}\}$，定义一个作为基础的概率空间，即一个集合 Ω，Ω 中事件的集类，以及这些事件上的概率测度。而且，按附录C，代价函数必须是这一空间上合适定义的随机变量(用测度概率论的术语说，即由概率空间到实轴的一个可测函数)。为满足这一要求，对函数 f_k，g_k 和 μ_k 就要有附加的(可测性)假定，并且有必要对空间 S_k，C_k 和 D_k 引入附加的结构。而这些假说可能限制容许控制律类，因为函数 μ_k 要受限制以满足附加的(可测性)要求。

因此，除非规定了这些附加的假设和结构，问题就不能合适地建立。另一方面，对一般的状态、控制和扰动空间，严格建立基本问题已经超出了导论性教科书的数学框架，这里不予考虑(见[B23])。不过，这些困难主要是技术上的，不会严重影响所得的基本结果。因此，如同所有导论性书本和大多数有关这一专题的杂志文献一样，我们认为用非正式的推导和论证方法来处理这些问题是比较合适和方便的。

然而，我们还要强调，在扰动空间 D_k，$k=0, 1, \cdots, N-1$ 为可数集的假设条件下，所有上述数学上的困难都不存在了。因为在这种情况下，只要附加假设：代价函数 (1.2) 中所有项的期望值存在，且对任一策略 π 是有限的，就能对此问题提供一个合理的框架。

当 D_k 可数时，一种简便的做法是把代价函数中的所有期望值，按照 D_k 元素的概率重写成无限和的形式。另一个办法是把代价 $J_\pi(x_0)$ 写成

$$J_\pi(x_0)=\underset{x_1,\cdots,x_N}{E}\left\{g_N(x_N)+\sum_{k=0}^{N-1}\tilde{g}_k[x_k,\ \mu_k(x_k)]\right\},\tag{1.4}$$

式中

$$\tilde{g}_k[x_k,\ \mu_k(x_k)]=\underset{w_k}{E}\{g_k[x_k,\ \mu_k(x_k),\ w_k]|x_k,\ \mu_k(x_k)\},$$

它相对于定义在可数集 D_k 上的概率分布 $P_k(\cdot|x_k,\ \mu_k(x_k))$取期望值。那么，可以把基本概率空间取作 $\tilde{S}_1$，$\tilde{S}_2$，…，$\tilde{S}_N$ 的笛卡尔积，其中

$$\tilde{S}_1=\{x_1\in S_1|x_1=f_0[x_0,\ \mu_0(x_0),\ w_0],\ w_0\in D_0\},$$

$$\tilde{S}_{k+1}=\{x_{k+1}\in S_{k+1}|x_{k+1}=f_k[x_k,\ \mu_k(x_k),\ w_k],$$

$$x_k\in\tilde{S}_k,\ w_k\in D_k\},\ k=1,\ 2,\ \cdots,\ N-1.$$

集合 $\tilde{S}_k$ 是加上控制律 $\{\mu_0,\ \mu_1,\ \cdots,\ \mu_{N-1}\}$ 时，在时刻 k 所有能达到的状态所构成的 S_k 的子集。D_0，D_1，…，D_{N-1} 均为可数集的事实，保证了集合 $\tilde{S}_1$，…，$\tilde{S}_N$ 也均为可数的(因为可数集合的任意可数集类的并是可数集)。现在，系统方程(1.3)，概率分布 $P_k(\cdot|x_k,\ \mu_k(x_k))$，初态 x_0，和控制律$\{\mu_0,\ \mu_1,\ \cdots,\ \mu_{N-1}\}$，定义了可数集 $\tilde{S}_1\times\tilde{S}_2\times\cdots\times\tilde{S}_N$ 上的一个概率分布，而式(1.4)是对于这一分布定义的数学期望。

总之，只有当扰动空间 D_0，…，D_{N-1} 是可数集时，才能严格建立基本问题。没有 D_k 的可数性，读者应理解以后所得结果和结论基本上是正确的，但数学上陈述不严密。事实上，当讨论无限时域问题时(需要更高的严密性)，我们将使可数性假设明朗化。我们注意到，水平较高的读者，严格地导出第二章和第三章中特殊应用问题的大多数结果，不会有多大困难。如在本章注记和习题12中所阐述的那样，是能够做到这一点的。

§1.2 动态规划算法

动态规划(DP)技术建立在非常简单的概念上，即最优性原理。这个名称起源于贝尔曼，他为 DP 的普及作出了一系列贡献，并且使其转换成系统性的工具。概略地说，最优性原理表述了如下不言而喻的事实。

令 $\pi^*=\{\mu_0^*,\ \mu_1^*,\ \cdots,\ \mu_{N-1}^*\}$ 是基本问题的最优控制律，考虑在时刻 i 处于状态 x_i，并希望从时刻 i 到时刻 N 的“余留代价”(Cost-to-go)

$$E\left\{g_N(x_N)+\sum_{k=i}^{N-1}g_k[x_k,\ \mu_k(x_k),\ w_k]\right\},$$

为极小的子问题。且假定利用 π^*时，状态 x_i 以正概率出现。那么，对此子问题，截控制律 $\{\mu_i^*,\ \mu_{i+1}^*,\ \cdots,\ \mu_{N-1}^*\}$是最优的。

最优性原理的直觉判断是非常简单的。如果截控制律按所说的那样不是最优，那么一旦到达x_i，我们就可对此子问题切换到一个最优策略，从而可进一步减小代价。用汽车旅行作比喻，假定我们已经找到了从洛杉矶到波士顿的最快路径，而且这一路径通过芝加哥。最优性原理道出了显而易见的真理：这个路径中从芝加哥到波士顿的一段，也一定是从芝加哥开始到波士顿结束旅行的最快路径。

最好用一个例题来介绍 DP 算法。

库存控制例题(续)

考虑上一节的库存控制例题。以下是从最后一个时间周期开始，时间上反向倒推，确定最优库存订货量策略的过程。

周期 $N-1$　假定在周期 $N-1$ 开始时，存货量为 x_{N-1}。显然，不管过去情况如何，仓库经理应订购存货 $u_{N-1}^{*}=\mu_{N-1}^{*}(x_{N-1})$，它对 u_{N-1} 使得最后一个时间周期的订购、存货和缺货代价之和为极小。而代价等于

$$\underset{w_{N-1}}{E}\{cu_{N-1}+H(x_{N-1}+u_{N-1}-w_{N-1})\}.$$

我们用 $J_{N-1}(x_{N-1})$ 表示最后一个周期的最优代价，则

$$J_{N-1}(x_{N-1})=\min_{u_{N-1}\geqslant 0}\underset{w_{N-1}}{E}\{cu_{N-1}+H(x_{N-1}+u_{N-1}-w_{N-1})\}.$$

从本质上说，J_{N-1} 是存货量 x_{N-1} 的函数。对于每个 x_N，J_N 可用解析方法或数值方法(在这种情况下，可用一个表格在计算机中存贮函数 J_N)计算。在计算 J_{N-1} 的过程中，我们得到最后一个周期的最优库存订货量策略 $\mu_{N-1}^{*}(x_{N-1})$。这里 $\mu_{N-1}^{*}(x_{N-1})\geqslant 0$，它对于每个 x_{N-1} 的值使上述方程的右边达到极小。

周期 $N-2$　假定在周期 $N-2$ 的开始，存货量为 x_{N-2}。现在，仓库经理显然应订购 $u_{N-2}=\mu_{N-2}^{*}(x_{N-2})$，它并非只使周期 $N-2$ 的期望代价为极小，而是使

(周期 $N-2$ 的期望代价)+(在周期 $N-1$ 采用最优策略时，周期 $N-1$ 的期望代价)

为极小。这就等于

$$\underset{w_{N-2}}{E}\{cu_{N-2}+H(x_{N-2}+u_{N-2}-w_{N-2})\}+\underset{w_{N-2}}{E}\{J_{N-1}(x_{N-1})\}.$$

利用系统方程 $x_{N-1}=x_{N-2}+u_{N-2}-w_{N-2}$，上式后一项可以写成

$$E_{w_{N-2}}\{J_{N-1}(x_{N-2}+u_{N-2}-w_{N-2})\}.$$

这样，给定状态 x_{N-2}，最后两个周期的最优代价 $J_{N-2}(x_{N-2})$ 由下式给出

$$J_{N-2}(x_{N-2})=\min_{u_{N-2}\geqslant 0}\underset{w_{N-2}}{E}\{cu_{N-2}+H(x_{N-2}+u_{N-2}-w_{N-2})+J_{N-1}(x_{N-2}+u_{N-2}-w_{N-2})\}.$$

对每个 x_{N-2} 再来计算 $J_{N-2}(x_{N-2})$。同样也计算最优订货策略 $\mu_{N-2}^{*}(x_{N-2})$。

周期 k　类似地，在周期 k 对初始库存 x_k，仓库经理应订购 u_k，使得

(周期 k 的期望代价)+(在周期 $k+1$，…，$N-1$ 采用最优策略时，这些周期的期望代价)

为极小。用 $J_k(x_k)$ 表示最优代价，则有

$$J_k(x_k)=\min_{u_k\geqslant 0}\underset{w_k}{E}\{cu_k+H(x_k+u_k-w_k)+J_{k+1}(x_k+u_k-w_k)\},\tag{1.5}$$

实际上这就是该问题的动态规划方程。

函数 $J_k(x_k)$ 表示由周期 k 开始，初始库存为 x_k 时，对于余留周期的最优期望代价。这些函数是从周期 $N-1$ 开始，到周期 0 结束，按时间反向递推计算得到的。$J_0(x_0)$ 的值是 0 时刻，初始库存为 x_0 时，过程的最优期望代价。在计算时，最优库存策略 $\{\mu_0^{*}(x_0),\mu_1^{*}(x_1),$