

第 1 章 动态规划算法

生活只能靠回味来理解，但必须向前生活。

——Kierkegaard(克尔凯郭尔，丹麦哲学家——译者注)

1.1 概 述

本书处理分阶段决策的情形。每个决策的后果未必完全可预测，但可以在做出下一次决策前在某种程度上进行预测。目标是 최소화某种费用——对不希望出现的结果的一种数学表示。

这些情形的关键点是不能孤立地看待决策，因为必须权衡所期待的较低当前费用与不可取的高的未来费用。动态规划技术刻画了这种权衡。在每个阶段，假设在剩余阶段采用最优决策，这种技术基于当前费用和期待的后续费用之和对决策进行排序。

动态规划可以处理非常广泛且多样的实际问题。在本书中，我们试着不让对问题结构的无关假设影响主要思想的简洁性。为此，我们在本节建立广泛可用的有限阶段（有限时段）动态系统的最优控制模型。这一模型将在之后六章中使用；其无限阶段模型将是本书最后一章及第 II 卷的主题。

我们的基本模型有两个主要特征：(1) 潜在的离散时间动态系统；(2) 在时间上可加的费用函数。这一动态系统表达了某些变量，即系统的“状态”，的演化受到在离散时间点上决策的影响。系统具有如下形式

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

其中

k : 离散时间索引；

x_k : 系统的状态且综合了与未来优化相关的历史信息；

u_k : 将在 k 时刻选中的控制或者决策变量；

w_k : 随机参数（也称为扰动或者噪声，取决于上下文）；

N : 施加控制的阶段数或者次数；

f_k : 描述系统的函数，并且确定状态更新的机制。

费用函数是可加的，即在 k 时刻产生的费用，记作 $g_k(x_k, u_k, w_k)$ ，随时间累加。总费用为

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

其中 $g_N(x_N)$ 是在过程最后产生的末端费用。然而，因为 w_k 的存在，一般而言费用是随机变量且难以优化。于是我们将这个问题建模成对期望费用的优化

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

其中的数学期望针对所涉及的随机变量的联合分布计算。这里的优化对象是控制 $u_0, u_1, \dots, u_{N-1}$ ，但需要一些资质认定；基于对当前状态 x_k 的一些知识（其精确价值或者其他相关信息）挑选每个控制 u_k 。

很快将给出对于上面使用的术语的更精确的定义。我们首先通过例子给出一些介绍。

例 1.1.1（库存控制）

考虑如下问题，在 N 个阶段的每个时刻订购一定量的某种货物以（大致）满足一个随机的需求，同时最小化所产生的费用的数学期望值。引入如下变量：

x_k ：在第 k 阶段开始时的库存水平；

u_k ：在第 k 阶段开始时的订购（并立即送到）的货物量；

w_k ：按给定概率分布在第 k 阶段的需求。

我们假设 $w_0, w_1, \dots, w_{N-1}$ 是独立随机变量，未满足的需求将被推迟直至有额外的库存就即刻满足。所以，库存水平按照如下的离散时间方程演化

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

其中负库存代表被推迟的需求（见图 1.1.1）。

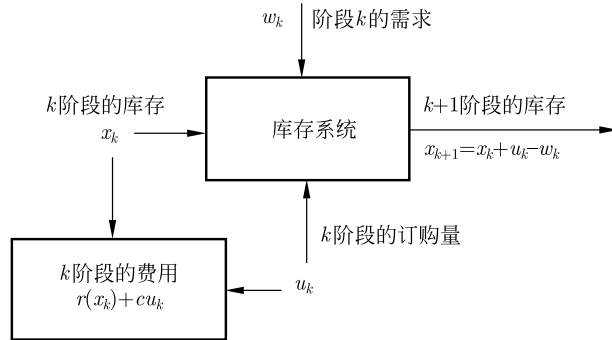


图 1.1.1 库存控制的例子。在阶段 k ，当前库存（状态） x_k 、订购的库存（控制） u_k 和需求（随机扰动） w_k 确定了费用 $r(x_k) + cu_k$ 和下一阶段的库存 $x_{k+1} = x_k + u_k - w_k$

在第 k 阶段出现的费用包括两部分：

(a) 费用 $r(x_k)$ 代表对正库存 x_k 的惩罚（为保持剩余库存的费用），或者对负库存 x_k 的惩罚（对未满足的需求的短缺费用）。

(b) 购买费用 cu_k ，其中 c 是每单位订购量的费用。

还有一个末端费用 $R(x_N)$ 表示在 N 个阶段结束时剩余的库存水平 x_N 对应的费用。所以， N 个阶段的总费用是

$$E \left\{ R(x_N) + \sum_{k=0}^{N-1} (r(x_k) + cu_k) \right\}$$

在自然约束 $u_k \geq 0, \forall k$ 之下，我们希望通过恰当地选择订购量 $u_0, \dots, u_{N-1}$ 让这一费用达到最小。

在这一点上我们需要区分对这一费用的闭环和开环最小化。在开环最小化中我们在 0 时刻一次性确定所有的订购量 $u_0, \dots, u_{N-1}$ ，而不去等着看后续的需求水平。在闭环最小化中我们尽可能推迟确定订购量 u_k ，直至最后一个可能的时刻（ k 时刻）已知当前的库存水平 x_k ，再确定 u_k 。其思想是推迟到 k 时刻再确定订购量 u_k 不仅没有任何惩罚，而且我们可以利用在 0 时刻与 k 时刻之间变得可用的信息（在过去阶段中的需求与库存水平）。

闭环优化在动态规划中占有核心重要性，也是我们在本书中将着重考虑的。所以，在我们的基本模型中，在各阶段进行决策，同时在各阶段之间采集将用于提高决策质量的信息。这在最终优化问题的结构上的影响相当深远。特别地，在闭环库存优化中我们对找到订购量的最优数值解不感兴趣，我们感兴趣的是找到为在每个阶段 k 的库存水平 x_k 的每个可能出现的值确定订购量 u_k 的最优规则。这是“动作与策略”的区别。

数学上，在闭环库存优化中，我们希望找到一系列函数 $\mu_k, k = 0, \dots, N-1$ ，将库存 x_k 映射到订购量 u_k 以最小化费用的数学期望值。 μ_k 的意义在于，对于 k 和 x_k 的每个可能取值，

$$\mu_k(x_k) = \text{在 } k \text{ 时刻库存为 } x_k \text{ 时应该订购的量}$$

序列 $\pi = \{\mu_0, \dots, \mu_{N-1}\}$ 将被称为策略或者控制律。对每个 π ，固定的初始库存 x_0 对应的费用是

$$J_\pi(x_0) = E \left\{ R(x_N) + \sum_{k=0}^{N-1} (r(x_k) + c\mu_k(x_k)) \right\}$$

我们想对给定的 x_0 在所有满足问题约束条件的 π 上让 $J_\pi(x_0)$ 最小。这是典型的动态规划问题。我们将在后续章节中在不同形式下分析这一问题。例如，我们将在 4.2 节中展示对于适当选择的费用函数，最优订购策略具有如下形式

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

其中 S_k 是由问题的数据决定的适当的阈值水平。换言之，当库存水平低于阈值 S_k 时，只需订购足够量将库存涨回至 S_k 。

前面的例子阐明了问题模型的主要要素：

(a) 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；例如在库存例子中，我们有 $f_k(x_k, u_k, w_k) = x_k + u_k - w_k$ 。

(b) 独立随机参数 w_k 。这将通过让 w_k 的概率分布依赖于 x_k 和 u_k 来生成；在库存的例子中，我们可以设想一种情形，其中需求水平 w_k 受当前库存水平 x_k 影响。

(c) 控制约束；在本例中，我们有 $u_k \geq 0$ 。通常，约束集依赖 x_k 和时间 k ，即 $u_k \in U_k(x_k)$ 。为在库存的例子中看出约束如何依赖 x_k ，设想一种情形其中库存水平应在某上界 B 之内，于是有 $u_k \leq B - x_k$ 。

(d) 如下形式的加性费用

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

其中 g_k 是某个函数; 在库存例子中, 我们有

$$g_N(x_N) = R(x_N), \quad g_k(x_k, u_k, w_k) = r(x_k) + cu_k$$

(e) 在 (闭环) 策略上的优化, 这里的策略指对 k 及 x_k 的每个可能取值选择 u_k 的规则。

离散状态和有限状态问题

在前面的例子中, 状态 x_k 是连续实变量, 易于设想在高维情形下的推广, 其中状态是 n 维实向量。然而, 状态也可能从离散集合中取值, 比如从整数中取值。

在一个版本的库存问题中, 库存水平用整数计量 (比如汽车的辆数), 每个单位量都是 x_k 、 u_k 或者 w_k 的不可忽略的一部分。此时自然出现了离散的观点。于是将所有整数构成的集合视作状态空间比将实数集视作状态空间更加合适。当然, 系统方程和每阶段的费用保持不变。

一般而言, 在许多情形下状态的取值自然是离散的, 而没有对应的连续版本。在这些情形中使用状态之间的转移概率更加方便。我们需要知道的是 $p_{ij}(u, k)$, 这是在 k 时刻当前状态是 i 且选中的控制是 u 的条件下, 下一个状态是 j 的概率, 即

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

这类状态转移也可以用离散时间系统方程的形式描述

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

其中随机变量 w_k 的概率分布是

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

反过来, 给定如下形式的离散状态系统

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

以及 w_k 的概率分布 $P_k(w_k | x_k, u_k)$, 我们可以提供等价的转移概率描述。对应的转移概率给定如下

$$p_{ij}(u, k) = P_k\{W_k(i, u, j) | x_k = i, u_k = u\}$$

其中 $W_k(i, u, j)$ 是如下集合

$$W_k(i, u, j) = \{w | j = f_k(i, u, w)\}$$

所以一个离散状态的系统可以等价地用差分方程或者转移概率描述。取决于给定的问题, 在符号上或者数学上用一种描述可能比用另一种描述更加方便。

下面的例子展示了离散状态问题。第一个例子涉及确定性问题, 即问题中没有随机的不确定性。在这样的过程中, 当在给定状态下选定控制后, 下一个状态完全确定。即对任意状态 i 、控制 u 和时刻 k , 转移概率 $p_{ij}(u, k)$ 对单个状态 j 为 1, 对所有其他状态为 0。其他三个例子涉及随机问题, 从给定状态下的给定控制出发导致的下一个状态事先不能确定。

例 1.1.2 (确定性调度问题)

假设为了生产某种产品, 需要在一台机器上执行四种操作。这些操作记作 A 、 B 、 C 和 D 。我们假设操作 B 仅能在操作 A 执行之后进行, 操作 D 仅能在操作 B 执行之后进行 (所以序列 $CABD$ 是允许的, 但是序列 $CDBA$ 是不允许的。)。给定从任意操作 m 转移到任意其他操作 n 的设置费用 C_{mn} 。对于从操作 A 或者 C 开始分别还有相应的初始费用 S_A 和 S_C 。一个序列的费用是与之相关联的设置费用之和; 例如, 操作序列 $ACDB$ 的费用为

$$S_A + C_{AC} + C_{CD} + C_{DB}$$

我们可以将这个问题视作由三个决策构成的序列，即对前三个操作的选择（由前三个操作可确定最后一个操作）。可以将已经执行的操作构成的集合作为状态，初始状态是虚拟的，对应于决策过程的开始。这一问题可能的状态与决策对应的状态转移示于图 1.1.2 中。这里的问题是确定性的，即在给定状态下，每个控制对应于唯一确定的状态。例如，在状态 AC 下采用操作 D 将确定地导致状态 ACD ，同时有费用 C_{CD} 。具有有限状态的确定性问题可被方便地表示为如图 1.1.2 所示的转移图。最优解对应于一条路径，从初始状态出发并终止于终止时刻的某个状态，其弧费用与末端费用之和最小。我们将在第 2 章系统地学习这类问题。

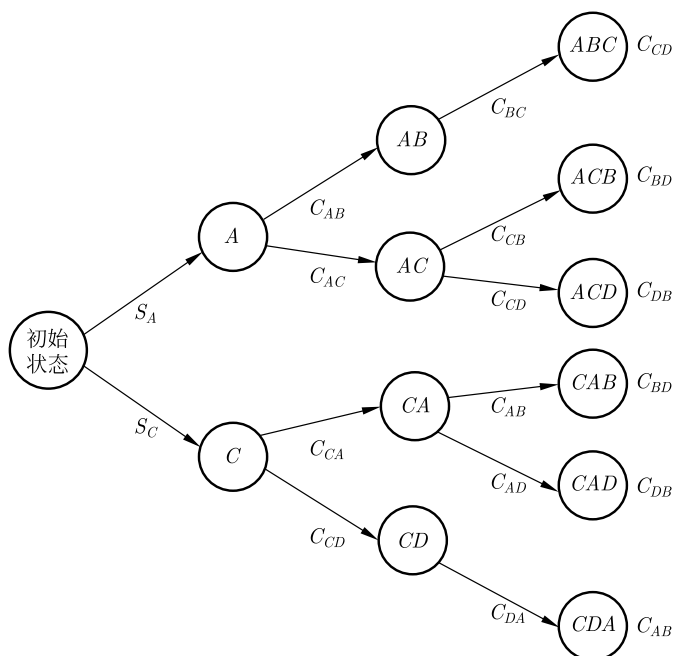


图 1.1.2 例 1.1.2 的确定性调度问题的转移图。图中的每条弧对应于一次决策，从某个状态开始（该弧的起点）到某个状态结束（该弧的终点）。对应的费用展示在弧旁。最后一个操作的费用表示为末端费用并示于图的末端节点旁

例 1.1.3（机器替换）

考虑在 N 个阶段高效操作一台机器的问题，机器可以处于 n 个状态中的任意一个，分别表示为 $1, 2, \dots, n$ 。我们用 $g(i)$ 表示当机器在状态 i 时每个阶段的操作费用，我们假设

$$g(1) \leq g(2) \leq \dots \leq g(n)$$

这里隐含的意思是状态 i 比状态 $i+1$ 好，状态 1 对应于机器处于最好状态。

在操作的时间段中，机器的状态可能变得更差或者保持不变。我们于是假设转移概率为

$$p_{ij} = P\{\text{下一个状态是 } j | \text{当前状态是 } i\}$$

满足

$$p_{ij} = 0, \text{ 若 } j < i$$

我们假设在每个阶段的开始知道机器的状态且必须从如下两个选项中进行选择：

- (a) 让机器在其当前状态下再工作一个阶段。
- (b) 以费用 R 维修机器并将其带回最佳状态 1。

我们假设一旦维修，机器可以保证在状态 1 停留一个阶段。在后续阶段，会以概率 p_{1j} 转移到状态 $j > 1$ 。

所以这里的目标是决定在哪些状态依然值得支付费用维修机器，从而获得更小的未来的维护费用。注意这一决策也将受我们所在的阶段影响。例如，当剩下的阶段不多时，我们倾向于不维修机器。

这个问题中系统的演化过程可以表示为图 1.1.3。这些图刻画了在每个控制下不同状态对之间的转移概率，并被称为转移概率图或者更简单地称为转移图。注意对每个控制有不同的图；在当前情形下有两个控制（维修或者不维修）。

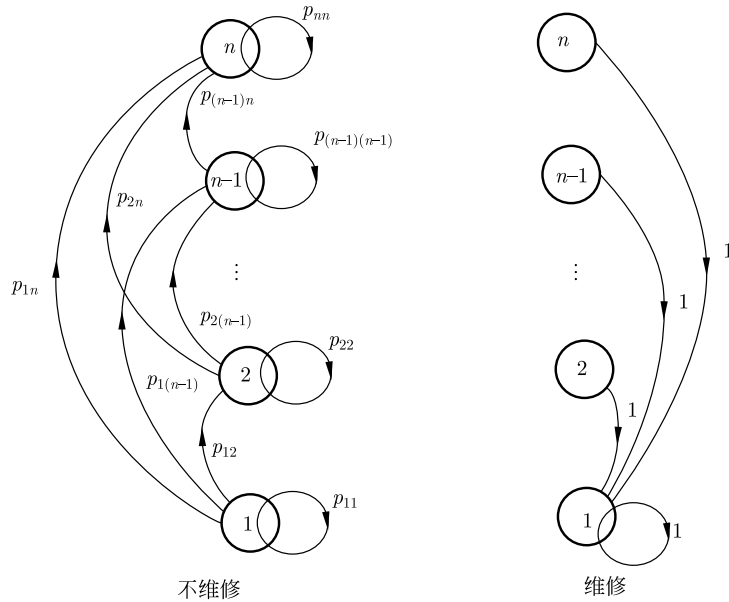


图 1.1.3 机器替换例子。两个可能控制（维修或不维修）中每一个对应的转移概率图。在每个阶段和状态 i ，维修的费用是 $R + g(1)$ ，不维修的费用是 $g(i)$ 。末端费用是 0

例 1.1.4（队列控制）

考虑一个有 n 个顾客空间的排队系统在 N 个时段的运行。我们假设顾客的服务仅可以在一个阶段的开始（结束）而开始（结束），系统每次仅可以服务一个顾客。给定在一个阶段中有 m 个顾客到达的概率 p_m ，在两个不同阶段到达的顾客数量独立。到达时若发现系统是满的，顾客将直接离开而不会尝试稍后再来。这个系统提供两种服务，快速和慢速，对应的每阶段的费用分别是 c_f 和 c_s 。在每个阶段的开始可以在快速和慢速服务之间切换。在快速（慢速）服务下，在一个阶段开始正在接受服务的顾客将在阶段结束时以概率 q_f （相应的 q_s ）完成服务，这一概率与该顾客已经被服务的阶段数和在系统中的顾客数独立（ $q_f > q_s$ ）。当系统中有 i 个顾客时每阶段的费用是 $r(i)$ 。在最后一个阶段结束时若系统中有 i 个顾客，则还有末端费用 $R(i)$ 。

问题是如何在每个阶段按照系统中顾客的数量选择服务的类型让 N 个阶段的总费用的数学期

望最小。我们期待当在队列中的顾客数量 i 大时, 最好用快速服务, 而问题是如何确定 i 的取值。

这里适合用每个阶段开始时系统中的顾客数量 i 为状态, 以所提供的服务类型为控制。那么, 取决于所提供的是快速或者慢速服务, 每个阶段的费用是 $r(i)$ 加上 c_f 或者 c_s 。我们下面推导系统的转移概率。

当在阶段开始系统为空时, 下一状态为 j 的概率与所提供的服务类型独立。当 $j < n$ 时这一概率等于有 j 个顾客到达的概率,

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

当 $j = n$ 时这一概率等于有 n 个或者更多顾客到达的概率,

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

当在系统中至少有一个顾客时 ($i > 0$), 我们有

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

$$p_{ij}(u_f) = q_f p_0, \text{ 若 } j = i-1$$

$$\begin{aligned} p_{ij}(u_f) &= P\{j-i+1 \text{ 个到达, 服务结束}\} + P\{j-i \text{ 个到达, 服务未结束}\} \\ &= q_f p_{j-i+1} + (1-q_f) p_{j-i}, \text{ 若 } i-1 < j < n-1 \end{aligned}$$

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

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

采用慢速服务时对应的转移概率也可以将这些公式中的 u_f 和 q_f 分别替换为 u_s 和 q_s 来获得。

例 1.1.5 (国际象棋策略优化)

一位棋手将要与对手下两局国际象棋, 想让他赢棋的概率最大。每局比赛有如下两种可能结果:

- (a) 两位棋手之一赢棋 (赢家得 1 分, 输家得 0 分)。
- (b) 和棋 (每人得 1/2 分)。

如果在两局棋后得分为 1-1 平手, 那么比赛进入突然死亡模式, 即选手们继续比赛直到其中一方首先赢棋 (同时赢得整场比赛)。棋手有两种下棋风格, 可以在每场比赛中任意选择两者之一, 与他在之前比赛中选择的风格独立。

- (1) 胆小风格, 平局概率为 $p_d > 0$, 输棋概率为 $(1-p_d)$ 。
- (2) 胆大风格, 赢棋概率为 p_w , 输棋概率为 $(1-p_w)$ 。

所以, 在每局棋中, 胆小风格不会赢钱, 而胆大风格不会平局。棋手想找到选择风格的策略让赢得整场比赛的概率最大。注意一旦比赛进入突然死亡模式, 棋手应当采用胆大风格, 因为在胆小风格下他最多只能延长比赛时间, 从而可能输掉整场比赛。于是, 棋手只有两个选择, 即在前两局棋中风格的选择。所以, 我们可以将问题建模成包括两个阶段, 将前两个阶段每个阶段开始的可能得分作为状态, 如图 1.1.4 所示。初始状态为初始得分 0-0。对两个不同控制 (比赛风

格) 中的每一个, 棋局的转移概率也示于图 1.1.4。在末端状态有费用: 若 2-0 或者 1.5-0.5 赢棋则费用为 -1 , 若 0-2 或者 0.5-1.5 输棋则费用为 0 , 若 1-1 和棋则费用为 $-p_w$ (因为在突然死亡模式下赢棋的概率为 p_w)。注意为了让赢得比赛的概率 P 最大, 我们必须让 $-P$ 最小。

该问题有个特点很有趣。您可能想, 如果 $p_w < 1/2$, 即使用最优的策略比赛, 选手赢得比赛的机会也会小于 50-50, 因为无论他采用哪种比赛风格, 其输棋的概率都大于赢棋的概率。然而, 事实并非如此, 因为棋手可以根据当前得分调整比赛风格, 但是他的对手没有这种选择。换言之, 棋手可以使用闭环策略, 稍后将看到在由动态规划算法确定的最优策略下, 他赢得比赛的机会可以大于 50-50, 前提是 p_w 大于阈值 $\bar{p}$, 这个阈值依赖于 p_d 的取值, 并且满足 $\bar{p} < 1/2$ 。

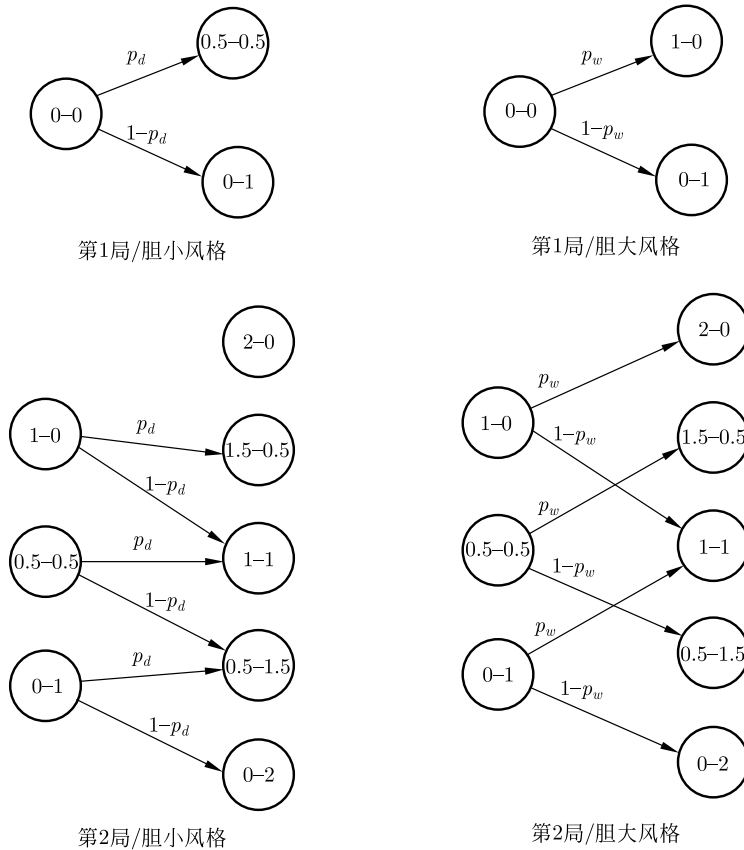


图 1.1.4 国际象棋比赛的例子。两个可能控制 (胆小或者胆大的走棋) 中的每一个对应的转移概率图。这里注意状态空间在每个阶段不同。赢棋的最终比分 2-0 和 1.5-0.5 的末端费用是 -1 , 输棋的最终比分 0-2 和 0.5-1.5 的末端费用是 0 , 平局比分 1-1 的末端费用是 $-p_w$

1.2 基本问题

我们现在建立有限阶段的在随机不确定下的通用决策问题。我们称这个问题为基本问题, 这是本书的中心内容。我们将在前六章中讨论基于动态规划求解这个问题的方法, 将在最后一章和本书第 II 卷中将我们的分析推广到涉及无限阶段的版本。

基本问题是非常具有一般性的。特别地，我们将不要求状态、控制或者随机参数取有限个值或者属于 n -维向量空间。动态规划的一个惊人方面是其适用性几乎不依赖于状态、控制和随机参数空间的特点。因此，可以方便地在不假设这些空间的结构的前提下推进讨论；稍后这些假设会成为严肃的问题。

基本问题

给定离散时间动态系统

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

其中状态 x_k 是空间 S_k 的元素，控制 u_k 是空间 C_k 的元素，随机“扰动” w_k 是空间 D_k 的元素。

控制 u_k 被限制在给定的非空子集 $U(x_k) \subset C_k$ 中取值，并依赖于当前状态 x_k ；即对所有的 $x_k \in S_k$ 和 k ，有 $u_k \in U_k(x_k)$ 。

随机扰动 w_k 由概率分布 $P_k(\cdot|x_k, u_k)$ 描述，可明确依赖于 x_k 和 u_k 但并不依赖于之前扰动 $w_{k-1}, \dots, w_0$ 的取值。

我们考虑由一系列函数构成的一类策略（也被称为控制律），

$$\pi = \{\mu_0, \dots, \mu_{N-1}\}$$

其中 μ_k 将状态 x_k 映射到控制 $u_k = \mu_k(x_k)$ 且对所有的 $x_k \in S_k$ 满足 $\mu_k(x_k) \in U_k(x_k)$ 。称这样的策略为可接受的。

给定初始状态 x_0 和可接受的策略 $\pi = \{\mu_0, \dots, \mu_{N-1}\}$ ，状态 x_k 和扰动 w_k 是随机变量，其分布由如下系统方程定义

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

所以，对于给定的函数 $g_k, k = 0, 1, \dots, N$ ， π 从 x_0 开始的期望费用是

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

其中期望值在随机变量 w_k 和 x_k 上取值。最优策略 π^* 能最小化这一费用；即

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

其中 Π 是所有可接受的策略构成的集合。

注意最优策略 π^* 与一个固定的初始状态 x_0 相关联。然而，这个基本问题和动态规划有趣的一面是通常能够找到策略 π^* 同时对于所有的初始状态是最优的。

将依赖于 x_0 的最优费用记作 $J^*(x_0)$ ，即

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

一种有用的视角是将 J^* 视作函数为每个初始状态 x_0 分配最优费用 $J^*(x_0)$ ，并将其称为最优费用函数或者最优值函数。^①

^① 为了方便喜欢数学上严格的读者，我们注意到在之前的方程中，“min”表示数值集合 $\{J_\pi(x_0) | \pi \in \Pi\}$ 的最大下界（或者下确界）。与通常的数学习惯更一致的符号是 $J^*(x_0) = \inf_{\pi \in \Pi} J_\pi(x_0)$ 。然而（正如在附录 B 中所讨论的），我们发现使用“min”而不是“inf”更加方便，即使在下确界达不到的情形下也是如此。这不易分散注意力，也不会导致歧义。

信息的角色与价值

我们之前注意到开环最小化与闭环最小化的区别，前者在 0 时刻一次性选定所有的控制 $u_0, u_1, \dots, u_{N-1}$ ，后者选择策略 $\{\mu_0, \mu_1, \dots, \mu_{N-1}\}$ 在 k 时刻基于当前状态 x_k 的信息施加控制 $\mu_k(x_k)$ （见图 1.2.1）。通过闭环策略，可以取得更低的费用，本质上是利用了额外的信息（当前状态的取值）。费用的降低可称为信息的价值而且可能确实很显著。如果没有这一信息，控制器不能针对不可预期的状态取值适当调整，结果会影响费用。例如，在前一节的库存控制例子中，在每个阶段 k 的开始变得可用的信息是库存水平 x_k 。显然，这一信息对于库存管理员非常重要，他希望基于当前库存水平 x_k 是否较高或者较低来调整订购量 u_k 。

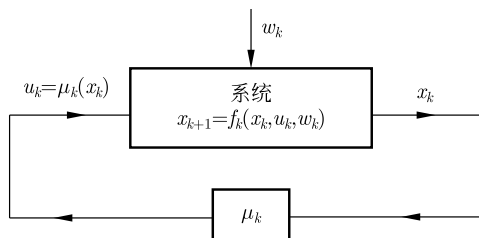


图 1.2.1 基本问题中的信息获取。在每个时刻 k 控制器观测当前状态 x_k 并施加依赖于那个状态的控制 $u_k = \mu_k(x_k)$

为了说明恰当使用信息带来的好处，让我们考虑前一节中的国际象棋比赛例子。在那里，在比赛中两局棋的每一局，棋手可以选择胆小风格（平局和输棋的概率分别为 p_d 和 $1 - p_d$ ），或者胆大策略（赢棋和输棋的概率分别为 p_w 和 $1 - p_w$ ）。假设棋手选择当且仅当领先时使用胆小策略，如图 1.2.2 所示；我们将在下一节中看到这一策略是最优的，假设 $p_d > p_w$ 。那么在第一局比赛之后（他在其中使用胆大风格），比分以概率 p_w 是 1-0，以概率 $1 - p_w$ 为 0-1。在第二局中，他在前一种情形中使用胆小风格，在后一种情形中使用胆大风格。所以在两局比赛后，赢得整场

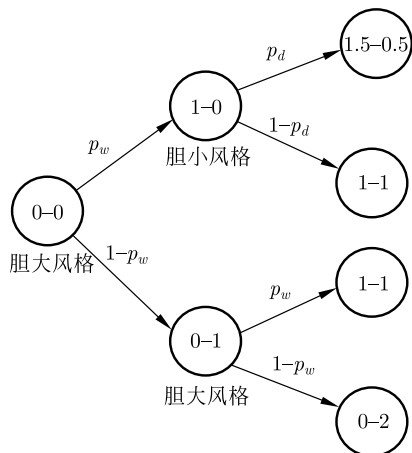


图 1.2.2 例 1.2.1 中用于获得高于 50-50 比例赢得国际象棋比赛的策略及相关的转移概率的示意图。玩家当且仅当他的得分领先时选择胆小风格