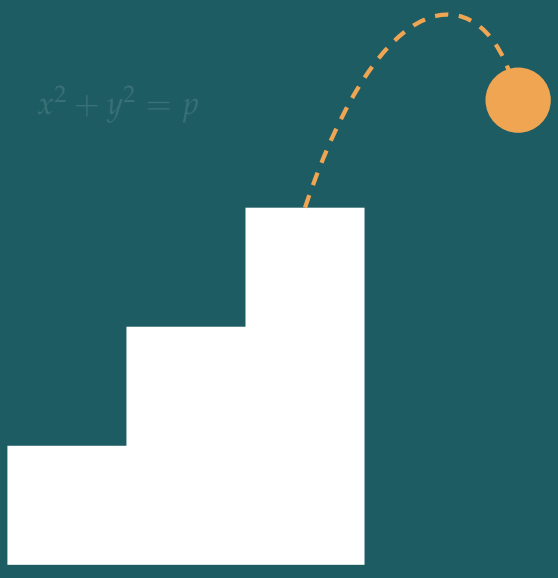
STEP 备试 · 大学数学

Beyond STEP

以 STEP 历届试题为门径
通往大学数学



KEITH McNULTY

Beyond STEP

Past STEP questions as a doorway
to university mathematics

Keith McNulty

Beyond STEP: Past STEP questions as a doorway to university mathematics

First published online November 2026. This version: 2026 年 11 月 18 日.

Everything in this book that is the author's own work, including the text, guided solutions, figures, exercises and answers, is dedicated to the public domain under the Creative Commons CC0 1.0 Universal licence: you may copy, adapt and share it for any purpose, without asking permission. See `creativecommons.org/publicdomain/zero/1.0`.

The STEP questions discussed in this book are the copyright of Cambridge University Press & Assessment. They are not reproduced here: each section links to its question at an official source, the STEP Support Programme's question database or OCR's past papers.

The book is also available free online, with printable worksheets for every section, at `beyondstep.keithmcnulty.org`. Corrections are welcome as issues at `github.com/keithmcnulty/step_book_mcnulty`.

本版本由机器从英文翻译而成，尚未经以中文为母语的数学教师审阅。如有不清楚之处，概以英文原文为准。

目录

导言 1

词汇表 11

数学符号 23

1 数论与证明 **33**

1.1 质数工厂 39

1.2 质数分布得有多稀疏? 47

1.3 最小反例 53

1.4 e 是无理数，e^2 也是 59

1.5 无穷递降法 65

1.6 两个平方数之和 71

1.7 因数分解失效之时 77

1.8 计数有理数 83

1.9 三次曲线的切线 89

回顾 95

练习答案 99

2 数列、级数与收敛 **103**

2.1 “最终”是甚么意思 109

2.2 作为极限的黄金比例 115

2.3 一个发散级数与它的收敛近亲 121

2.4 $\pi^2/6$ 127

2.5 韦达乘积 133

2.6 算术几何平均 139

2.7 六位还是六百位 145

2.8 平方再减二 151

2.9 逐步逼近 e 的分数 159
回顾 165
练习答案 169

3 函数、不等式与严谨 173

3.1 找出错处 179
3.2 倒过来的 AM–GM 185
3.3 凸性的威力 191
3.4 以面积证明不等式 197
3.5 π^e 还是 e^π? 203
3.6 斜率的界 209
3.7 $\pi^2 \leq 10$ 215
3.8 由加法公式决定的函数 221
3.9 创造一个新函数 227
回顾 233
练习答案 237

4 严谨的微积分 241

4.1 甚么是积分? 247
4.2 由一个积分建立三角学 253
4.3 22/7 大于 π 259
4.4 沃利斯、Beta 与二分之一的阶乘 265
4.5 大二学生之梦 271
4.6 没有公式的积分 277
4.7 变分法 283
4.8 切球形面包 289
4.9 Holditch 定理 295
回顾 301
练习答案 305

5 复数 309

5.1 梅钦公式 315

5.2 单位根筛选法 321

5.3 因式分解 $z^{2n}+1$ 327

5.4 正多边形中的距离 333

5.5 名不副实的有理角 339

5.6 费马点 345

5.7 作为映射的复指数函数 351

5.8 圆盘中隐藏的距离 357

回顾 363

练习答案 367

6 代数与结构 371

6.1 以单位根为根的多项式 377

6.2 高斯的正十七边形 383

6.3 以 $\sqrt{2}+\sqrt{3}+\sqrt{5}$ 为根的多项式 389

6.4 解四次方程 395

6.5 从公理出发的微分法 401

6.6 子群的并集 407

6.7 四元数群 413

6.8 由奇异矩阵组成的群 419

回顾 425

练习答案 429

7 线性代数与几何 433

7.1 对矩阵不成立的法则 439

7.2 只伸缩，不转向 445

7.3 甚么才算行列式 451

7.4 回到起点的矩阵 457

7.5 三维空间中的旋转 463
7.6 双曲旋转与相对论 467
7.7 永恒不变的量 473
7.8 四个平方数 479
7.9 只有五种柏拉图立体 485
回顾 491
练习答案 495

8 微分方程与动力系统 499

8.1 当解不唯一 505
8.2 不求解而估界 511
8.3 收获种群 517
8.4 盘旋趋向圆 523
8.5 格力犬追逐格力犬 529
8.6 量子振子 535
8.7 可以解出的混沌 541
8.8 周期三蕴含混沌 549
回顾 557
练习答案 561

9 计数与机会 565

9.1 直线分割平面 571
9.2 小型版的拉姆齐定理 577
9.3 为篱笆桩着色 583
9.4 把数列看成函数 589
9.5 凑零钱 595
9.6 奇怪的骰子 601
9.7 相等的幂和 607
9.8 没有人拿到自己的恤衫 613

9.9 生日问题 619
9.10 总能找赎 625
9.11 集齐一套 631
9.12 随机子集 637
9.13 做了手脚的骰子 643
回顾 649
练习答案 653

10 守恒定律、碰撞与引力 659

10.1 牛顿摆 665
10.2 叠球效应 671
10.3 火箭与枪炮 677
10.4 一个分子与一个活塞 683
10.5 有限时间内的无限次反弹 689
10.6 二体化为一体 695
10.7 球壳内的引力 701
10.8 排成三角形的三颗恒星 707
回顾 713
练习答案 717

11 振动、转动与运动参考系 721

11.1 共振 727
11.2 简正模式 733
11.3 真实的摆 739
11.4 等时曲线 745
11.5 加速火车中的重力 751
11.6 后旋 757
11.7 杆中的辛普森法则 763
11.8 倾斜的平衡 769

11.9 安全抛物线 775
11.10 会朝你飞回来的球 781
11.11 外野手如何接球 787
11.12 橡皮绳上的蚂蚁 793
11.13 铲雪车问题 799
11.14 是抛物线，不是悬链线 805
回顾 811
练习答案 815

12 概率 821

12.1 精确地取条件 827
12.2 基本比率谬误 833
12.3 “随机”是甚么意思? 837
12.4 两两独立，却非相互独立 843
12.5 最后一位乘客的座位 849
12.6 多一枚硬币 855
12.7 淘汰赛中谁会遇上谁 859
12.8 Penney 游戏 863
12.9 用不公平的硬币作公平的决定 869
12.10 把绳头系起来 875
12.11 邮票、杯子与 e 881
12.12 何时停止寻找 885
12.13 三人决斗 891
12.14 混合策略 895
12.15 灭绝还是爆发 901
12.16 顺向与逆向的马可夫链 907
12.17 巴拿赫的火柴盒 913
12.18 无记忆性 919
12.19 平面上的随机点 925

12.20 蒲丰投针 . . . 931
12.21 折断一根棒 . . . 935
12.22 没有平均值的分布 . . . 941
12.23 由概率得出斯特灵公式 . . . 947
12.24 中心极限定理 . . . 953
12.25 众多之中的最大值 . . . 959
回顾 . . . 965
练习答案 . . . 969

13 统计与现实世界中的随机性 979

13.1 解读新闻标题 . . . 983
13.2 计数田鼠 . . . 989
13.3 德国坦克问题 . . . 995
13.4 平均数还是中位数? . . . 1001
13.5 问尴尬的问题 . . . 1007
13.6 累积起来的误差 . . . 1013
13.7 制造相关性 . . . 1019
回顾 . . . 1025
练习答案 . . . 1029

引言

> 我无法靠这本书教会你数学，但我相信，只要你肯下苦功，一点一点地凿开这些题目，终会让你内在的那位数学家显露出来。
>
> — Stephen Siklos，*Advanced Problems in Mathematics*

这是一本关于难题的书，也是一本关于难题背后有甚么的书。全书 136 节，每一节都由一道 STEP 历届试题开始。但题目只是入口。几乎每一道好的 STEP 题目背后，都有一片你将会在大学里正式学到的数学，而本书的目的，就是带你穿过这道门，让你看看门后的风景。

STEP 是甚么，为何存在

STEP，即 Sixth Term Examination Paper，是一组在中学最后一年夏天应考的数学考试。它由剑桥大学（University of Cambridge）于 1987 年设立，作用是协助决定谁适合到那里修读数学：剑桥数学系发出的取录通知，几乎全部以 STEP 成绩为条件，多年来其他大学亦有采用，当中以华威大学（Warwick）最为人熟知。起初 STEP 设有多个科目的试卷，但到了 2001 年，就只剩下数学一科。在它大部分的历史里，考试共有三份试卷，即 I、II 和 III；Paper I 最后一次开考是在 2019 年，现时考生应考 Paper II 和 Paper III。这项考试多年来由剑桥大学考评部（Cambridge Assessment）主办，自 2024 年起改由 OCR 负责。

STEP 试卷与学校的考试截然不同。卷上不是十道必答的短题，而是十多道长题，考生从中选答六题，而一位实力强的考生只要在四题上取得实质进展，成绩已经非常好。题目不会告诉你每一部分占多少分，也很少要你做一些你练习过的东西。它们要你把已经懂得的数学，用在全新的地方。

对 STEP 的认识无人能及的 Stephen Siklos，曾这样形容它的用途。它是一道门槛：事实证明，STEP 的成绩能相当准确地预示学生在大学数学学位课程中的表现（据他所述，STEP 名次与剑桥一年级考试名次的相关系数达 0.63，以这类数据的标准而言非常高）。它是一种准备：STEP 所奖励的思考方式，正是大学本科数学的思考方式。它也是对动力的考验：准备 STEP 要下真功夫，一个不愿付出这份努力的学生，大概也不会享受要求严格的数学学位课程。

Stephen Siklos

谈 STEP 而不谈 Stephen Siklos，是不可能的。他是剑桥的数学家，曾是霍金（Stephen Hawking）的研究生，研究早期宇宙和引力波的数学，在应用数学及理论物理系任讲师，其后出任耶稣学院（Jesus College）的高级导师（Senior Tutor），再担任该学院的 President。1987 年 STEP 数学科创办时，他出任首席考官（Senior Examiner），并担任此职三十二年，直至 2019 年 8 月离世为止。这些年间，他草拟题目，“把别人的题目修整成形”（这是他自己的说法，他说这件事长久以来都是他兴趣和乐趣的来源），逐一审阅每份试卷，并督导评卷工作。他也是 STEP 复活节学校（STEP Easter School）和 STEP 支援计划（STEP Support Programme）背后的推手，这两个项目为学校无法提供协助的学生免费提供帮助。有一篇讣闻说，在令剑桥数学入学试向不同学校背景的学生开放这件事上，没有人比他做得更多。

他的著作 *Advanced Problems in Mathematics: Preparing for University*（《高等数学难题：为大学作准备》）由 Open Book Publishers 在网上免费出版，至今仍是认识 STEP 的最佳入门书，我极力推荐。他在书中写下的两点，塑造了本书的面貌。第一点是他对甚么才算一道好的 STEP 题目的看法：题目的第一部分向你展示一个陌生的想法，其后的部分则要你在更难的情境中再次运用它，以证明你确实明白；又或者题目的答案取决于一个要你自己去发现的参数；又或者题目由特殊情况开始，以一般结果作结。正是这类题目才会引领你到达某处，而本书的题目大多属于这几类。第二点是他坚持解答应该是“跳板，而不是羽毛床”。一份解答，如果你未曾先自己与题目搏斗过就去读，是学不到任何东西的。

我与 STEP 的相遇

我第一次接触 STEP 是在 1990 年代初，当时我的数学老师给了我几份试卷，作为额外的挑战。我做这些试卷是为了乐趣，而不是为了取录资格，也从未真正应考，但我记得那些试卷难得出奇：题目看似与我学过的任何东西都毫无关系，一道题就写满一页，我能够起步，却很少能做完。我也记得，在终于做通了的那些时刻，我感到自己看见了一些学校数学从未让我看见的东西。

许多年后，我从另一边回到 STEP：协助学生准备应考。与他们一起做过数百道题目之后，我对这些试卷有了不同的看法。并非每一道题都是瑰宝。坦白说，有些题目只是漫长的耐力测试：一整页要小心处理的代数，一段要求很高却属例行公事的计算，到最后既没有多少洞见，也没有多少乐趣。耐力固然重要，大学数学也需要它，但令 STEP 与别不同的并不是这些题目。

令它与别不同的是，在 1987 年以来所出的 1,500 多道题目之中，散落着数百道真正优美的题目。要解释我的意思，最好的方法就是给你看几道。

几颗瑰宝

以下是你稍后会在书中遇到的七道题目，我特意选来展示本书涵盖的范围：两道来自组合数学（combinatorics），两道来自动力系统（dynamical systems），两道来自力学（mechanics），一道来自概率（probability）。每一道都由你能想像得到的东西开始，以你可能难以置信的结果作结，并通往一片在大学里会认真研究的数学。

藏在混乱中的秩序。在纸上画六个点，把每两点用一条线连起来，合共十五条线，每条线随你喜欢涂成金色或银色。现在试试避免出现三边同色的三角形。你每一次都会失败。五个点是做得到的（把五边形的边涂成金色，五条对角线涂成银色），但六个点就不可能，无论你的涂法多么巧妙。1989 年 STEP I 的一道题目要你证明这件事，而且只给一个提示：看看从其中一点出发的五条线。其中三条必定同色，由此出发，证明只需寥寥数行。线的涂色方法共有 $2^{15} = 32\,768$ 种，而这个论证一次过处理了全部情况，连一个个案也不用检查。

这个结果是拉姆齐定理（Ramsey's theorem）最小的情况。这条定理由拉姆齐（Frank Ramsey）于 1928 年证明，它说的是：彻底的无序是不可能的；任何足够大的结构，无论拼凑得多么杂乱无章，都必定包含一小块完美的秩序。多大才算“足够大”？要保证有四个点彼此全以同一颜色相连，你需要 18 个点，这是 1955 年证明的事实。至于五个点，没有人知道：答案介乎 43 与 46 之间，而上界到 2024 年才获证明。艾狄胥（Paul Erdős）曾开玩笑说，如果外星人威胁要毁灭地球，除非我们说出这个数，我们就应该动员所有电脑和数学家去找；但如果他们问的是六个点的答案，我们就应该设法消灭外星人。你会在大学的组合数学或图论（graph theory）课程中遇到拉姆齐理论，而推动这个 STEP 证明的鸽巢原理（pigeonhole principle），则几乎在你修读的每一门课都会出现（Section 9.2）。

奇怪的骰子。掷两颗普通骰子，总和为 7 的机会最大，总和为 2 和 12 的机会最小，各总和的掷法数目依次为大家熟悉的 $1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1$。用正整数为一对骰子编号，还有没有别的方法，使每个总和出现的机会完全相同？看来不大可能，但这样的方法恰好有一个：一颗骰子写上 $1, 2, 2, 3, 3, 4$，另一颗写上 $1, 3, 4, 5, 6, 8$。这就是 *Sicherman* 骰子，以发现它们的 George Sicherman 命名，并因马丁·加德纳（Martin Gardner）于 1978 年的介绍而广为人知。在任何只取决于所掷总和的游戏中，你大可用它们换掉普通骰子，而永远不会有人察觉。

1993 年的一道题目从另一端切入这个想法。它给你一个表，列出用两颗编号古怪的骰子掷出每个总和的方法数目，然后要你推算出骰子的编号，就像侦探根据遗下的痕迹重组案情一样。该节接着介绍令整个课题变得简单的想法：把每颗骰子变成一个多项式（polynomial），每一个点数为 k 的面对应一项 x^k。掷两颗骰子相当于把它们的多项式相乘，因为 $x^i \times x^j = x^{i+j}$ 正好把两面的数字相加。于是，寻找总和分布相同的新骰子，就变成用另一种方式把一个多项式因式分解，情况有点像 36 可以拆成 4×9 而不是 6×6。又因为多项式与整数一样，分解成不能再拆的因子时只有一种方式，这个因式分解同时证明

了 Sicherman 骰子是唯一的另类选择。把计数问题编码成多项式，即*生成函数*（generating function），这个技巧是组合数学中最强大的工具之一，而背后的唯一分解（unique factorisation），则是大学第一门代数课的中心主题（Section 9.6）。

有公式的混沌。取一个介乎 0 与 1 之间的数，把它乘以 4，再乘以一减去它本身，然后不断重复：$u_{n+1} = 4u_n(1-u_n)$。这就是*逻辑斯谛映射*（logistic map），一个粗略的模型，描述一个不断增长但受粮食供应限制的种群。这条规则简单得不能再简单，但它产生的数却跳来跳去，看不出任何规律，而两个小数点后十位都相同的初始值，通常在数十步之内就会给出完全不同的数列（sequence）。生物学家 Robert May 于 1976 年在一篇著名的论文中引起大家对它的注意，论文题为“Simple mathematical models with very complicated dynamics”（具有非常复杂动态的简单数学模型），它自此成为混沌（chaos）的标准例子，通常藉电脑来探索。

2014 年的一道题目揭示了一件出人意表的事：这个数列有一条精确的公式。如果初始值是 $\sin^2\theta$，那么 $u_n = \sin^2(2^n\theta)$，而你只需用倍角公式（double angle formulae）和归纳法（induction）就能证明。（乌拉姆（Stanisław Ulam）和冯·诺伊曼（John von Neumann）在 1940 年代已经知道这条公式。）你也许以为有了精确的公式，混沌就会被驯服。事实上，公式反而把混沌解释得一清二楚。每一步都把角加倍，所以角上任何微小的不确定性每一步都会加倍，而初始值的误差平均而言也以同样的速度增长：五十步之后，误差已增大了约一千万亿倍。把这个角表示成 π 的分数倍，再写成二进制（binary），就看得格外分明：映射的每一步都把二进制数字向左移一位，并丢掉第一位，所以数列是把初始值的数字逐个读出来，而要预测一百步之后的情况，你就要知道初始值准确至一百个二进制位。同一条公式还能找出这个映射的所有循环（cycle），并显示每一种长度的循环都存在，它们稠密地挤在 $[0,1]$ 之内，而且每一个都不稳定。这就是动力系统这门学问，你会在大学的微分方程（differential equations）课程中遇到它，其后还会在专门讲混沌的课程中再遇上（Section 8.7）。

周期三意味着混沌。取函数 $f(x) = 7 - 2|x|$，由一个数开始，不断把答案代回去。由 1 开始，得到 $1, 5, -3, 1, 5, -3, \ldots$：一个长度为三的循环。2025 年的一道题目要你寻找这类循环，方法是描绘 f、$f(f(x))$ 和 $f(f(f(x)))$ 的图像，找出它们与直线 $y = x$ 相交的位置。每多应用一次 f，图像中的锯齿数目就加倍，而该节把这个观察化为一个证明，说明这个函数具有*每一种*长度的循环：二、三、四、一百、一百万。

这并不是这个函数独有的特点。1975 年，李天岩（Tien-Yien Li）和 James Yorke 证明了任何具有长度为三的循环的连续函数（continuous function），都必定具有其他每一种长度的循环，论文的题目令人难以抗拒：“Period three implies chaos”（周期三意味着混沌）。证明所需的不过是介值定理（intermediate value theorem），也就是每当你论证一个变号的连续函数必定穿过零时已经在用的定理，再加上一幅巧妙的图，显示函数把哪些区间拉伸覆盖到哪些区间之上。后来人们才发现，乌克兰数学家沙可夫斯基（Oleksandr Sharkovsky）早在 1964 年已证明了一个更强的结果：他找到了正整数的一种奇特排序，

以 $3, 5, 7, 9, \ldots$ 开始，使得任何一种长度的循环，都会迫使排在它后面的每一种长度的循环出现。介值定理是大学分析（analysis）课程中最先严格证明的结果之一；这道题目显示了它能做到多少事情（Section 8.8）。

单车气泵为何会发热。替单车轮胎打气，然后摸一摸气泵：它是暖的。挤压气体会令它升温，柴油引擎正是这样不靠火花就点燃燃料。但为甚么呢？2002 年的一道题目这样回答：追踪单独一个空气分子，看它在气泵末端与正被推入的活塞之间来回弹跳。分子每次从推进中的活塞反弹，都会加快一点点，就像网球被移动中的球拍击中一样。题目要你准确描述一次反弹，把它化成每次碰撞时气缸长度的递归关系（recurrence），然后解出这个递归关系。

解答揭示了一条隐藏的定律：分子的速率乘以气缸的长度，几乎不变。把长度减半，速率就加倍，而由于温度量度的是分子的能量，气体便会升温。像这样在系统缓慢改变时几乎保持不变的量，称为*绝热不变量*（adiabatic invariants），而只要再多下一点功夫，同一个论证就能导出气体的绝热压缩定律，你会在大学的热力学中学到它。这个想法影响深远。费米（Enrico Fermi）于 1949 年提出，宇宙射线正是以这种方式获得巨大的能量，即在太空中移动的磁场之间反复弹跳；而埃伦费斯特（Paul Ehrenfest）则于 1916 年指出，绝热不变量正是量子力学必须加以量子化的量。一道关于单车气泵的题目，原来触及统计物理、天体物理和量子理论的开端（Section 10.4）。

会回来的球。把一个球轻轻向前抛，它会稳定地远离你，直至落地。常识告诉我们事情总是如此：球一直向前移动，又怎会靠近你呢？但只要抛得够陡，与水平所成的角大于约 $70.5°$，怪事就会发生。在飞行的某一段，当球下降时，它会离你*更近*，尽管它在水平方向上仍然远离你。它的高度下降得比水平距离增长得更快。

2019 年的一道题目要你证明这件事，并求出临界角的准确值。所用的方法你日后会一用再用：不对距离本身求导，而对距离的平方求导，这样就避开了平方根；所得的导数（derivative）原来是时间的二次式（quadratic），于是距离会否减少的问题，就变成这个二次式有没有实根的问题。它的判别式（discriminant）给出了答案，而临界角原来就是正弦值恰好等于 $\frac{2\sqrt{2}}{3}$ 的那个角。这一课的意义远远超出抛体运动。许多“显然”递增的量其实并非如此，而在大学里，通过求出导数的正负号来证明某个量是单调（monotonic）的，是证明不等式、证明数列收敛（converge），以及证明微分方程的解趋于稳定的关键一步（Section 11.10）。

最后一位乘客。一班列车全部满座。第一位上车的乘客不理会自己的车票，随机选了一个座位坐下。其后每一位乘客，如果自己的座位空着就坐自己的座位，否则就在空座位中随机选一个。最后一位上车的乘客发现自己的座位空着的机会有多大？如果有一百个座位，各种可能性的分支令人眼花缭乱：一位被占了座位的乘客可以占去另一位的座位，后者又占去另一位的，形成一条任意长度的链。直接计算看来毫无希望。然而，无论座位数目是多少，答案都恰好是 $\frac{1}{2}$。

2021 年的一道题目一步一步地得出这个答案：它要你求出两个和三个座位时的答案，再按第一位乘客所坐的位置取条件（conditioning）来建立递归关系，然后用归纳法证明答案永远是 $\frac{1}{2}$。这是一个完全妥当的证明，但它令你不禁要问，答案为甚么如此简单，而该节接着给出一个一句话的论证来解释。每当有乘客要随机选择时，第一位乘客的座位和最后一位乘客的座位所受的对待完全相同；在其中一个被占之前，整个过程无法分辨它们。所以最后剩下的空座位是两者之一的机会均等，而它必定是两者之一，因为其他每个座位，在最后一位乘客上车之前，都已被它的主人或某位被占了座位的乘客坐了。像这样利用对称（symmetry）的论证，以一个观察取代好几页的计算，贯穿整个大学概率论（Section 12.5）。

综合来看，这类题目触及数学学位一年级的每一个主要范畴：数论（number theory）、分析、微积分（calculus）、复数（complex numbers）、代数、线性代数（linear algebra）、微分方程、组合数学、力学、概率与统计。

本书的用途

这个观察正是本书的缘起。大多数的 STEP 备试，都以通过考试为目标，这十分合理：学习技巧，操练试卷，锻炼速度和耐力。本书也以此为目标，而读完它会令你更擅长应付 STEP。但它真正的目的并不在此，而在于帮助你心怀大学数学去准备 STEP：把每道题目看成一份邀请，而不是一个要清除的障碍，并且循着这份邀请一直走到尽头。

在 STEP 中表现最好、其后在大学里如鱼得水的学生，几乎总是那些为数学本身而热爱数学的人。他们享受难题终于被攻破的那一刻。他们想知道一个技巧为何行得通，而不仅仅是知道它行得通。当答案简单得令人起疑，例如 $\frac{1}{2}$ 或 $\frac{1}{e}$ 或 $\pi^2/6$，他们会怀疑背后另有更深的道理，并且动身去找。本书想培养的正是这些习惯。每一道题目之所以入选，都是因为它通往一个发现：一个比题目本身更大的想法、定理或思考方式。

思考方式与定理同样重要。大学数学要求你用学校数学很少用到的方式思考：去证明而不是去计算，用反证法（proof by contradiction）或无穷递降法（infinite descent）来论证，用两种方法数算同一样东西，找出一个永不改变的量，构造一个反例（counterexample），在不把某样东西找出来的情况下证明它存在。STEP 题目训练的正是这些习惯，而且往往不加明言。本书的每一节都会点明该节核心的思维技巧，好让你再次遇上时能够认出它。

本书的编排

本书共有十三章，按题目所通往的大学数学范畴编排，而不是按题目所出自的 STEP 试卷编排。

- **纯粹数学**（第 1 至 9 章）：数论与证明；数列、级数（series）与收敛；函数、不等式与严谨性；正正式式的微积分；复数；代数与结构；线性代数与几何；微分方程与动力系统；以及计数与机会，当中亦取材自概率卷的组合题。

- **力学**（第 10 和 11 章）：守恒定律、碰撞与引力；以及振动、转动与运动参考系，最后以一组出人意表的轨迹作结。
- **概率与统计**（第 12 和 13 章）：概率，由关于取条件的清晰思考到中心极限定理（central limit theorem）；以及现实世界中的统计与随机性。

每章以一篇短文开始：这个范畴为何重要，你在大学里会在其中遇到甚么，以及贯穿全章的三个大问题。每章以一篇总结收结，当中会重温这些问题，并附上所用方法的工具箱、通往其他章节的指引，以及入读大学前的阅读建议。每章末尾载有练习的简短答案。第 1 章之前有一个词汇表，收录书中那些在学校不大可能接触过的术语，还有一个列出书中所用数学符号的表；如果有哪个词或哪个符号难倒了你，就到那里查一查。

每一节都依循同一个模式。它以*题目*开始：一条连结，通往该道 STEP 题目在官方来源的原文，与当年所出的一字不差。*引子*解释这道题目为何比表面看来更有趣。*动笔之前*给你第一个立足点。*逐步解题*是分阶段的引导式解答，中间穿插着以*问问自己*和*停下来试试*为题的方框，那些都是一位好老师会在那一刻向你提出的问题；它们是用来回答的，不是用来跳过的。*回顾*挑出关键的几步。*背后的数学*是该节的核心：题目背后的大学数学，它的历史和它的惊喜。*思维技巧*点明这道题目所锻炼的思考习惯。*延伸练习*提供后续的题目，当中常有相关的 STEP 题目，而*延伸阅读*则推介一些书籍和文章，程度是实力强的中学生或大学一年级生应付得来的。

本书详细讨论了 136 道题目，但 STEP 所出的题目超过 1,500 道。网站为其中每一道都提供了详细解答，由样本试卷直至 2025 年。它们的写作精神与本书一致，但篇幅短得多。每一份解答都给出解开题目的关键思路、逐部分的完整解答、几项你可以对答案作出的验算、容易犯的错误，以及该题所触及的大学数学的指引。如果某道题目在本书中有完整的讨论，它的解答会注明，并附上连结。

有几节用到现行 STEP 课程纲要以外的课题，例如群（groups）、转动惯量（moments of inertia）和矩母函数（moment generating functions）。收录它们，是因为那些题目好得不容割舍，而每一节都会给出它所需要的事实。

学生如何使用本书

- **先做题目。**从该节开首的连结打开题目，阅读时把它放在手边，因为解答会不时提到题目的各个部分和用字。每道题目至少认真地花半小时，然后才往下读。卡住了并不代表失败；学习正是在这里发生的。如果你完全卡住了，先读*动笔之前*，然后再试。
- **回答方框里的问题。**当正文向你提问时，请停下来，在纸上作答。这些引导式解答，是写给手里拿着铅笔的读者看的。
- **不要跳过“背后的数学”各节。**它们不在考试范围之内，却是这本书存在的理由，也是你日后会记得的东西。

- **各节可按任何次序阅读。**每一节都自成一体，而交叉引用会指引你到别处的相关想法。跟着你的好奇心走。
- **先做练习，再对答案。**每章末尾的答案很简短：只给出结果和关键的一步，而不是完整的解答过程。
- **把详细解答当作跳板。**用历届试卷练习时，先认真尝试每道题目，然后才打开网站上的解答，而且只读到足以让你重新起步的地方为止。如果关键思路已经足够，就关上解答，自己把题目做完。
- **善用工作纸。**网站为每一节提供可列印的工作纸，载有题目的连结、依次排列的提示、*背后的数学*部分和练习，方便你离开屏幕来做。
- **在本书以外多读。**Siklos 的书、STEP 支援计划和历届试卷本身，全部都可以免费取得。每一节的延伸阅读，都是挑选入读大学前已能读懂的材料。

教师如何使用本书

- **作为延伸教材的来源。**每一节都是一堂自成一体的课：题目、一条贯穿题目的引导路径，以及做完之后一个值得一去的地方。各项思维技巧很适合用作一连串课堂的主题。
- **作为讨论指引。***问问自己*和*停下来试试*的方框，是全班或小组讨论的现成引子。先提出这些问题，然后才揭示下一阶段。
- **配合工作纸使用。**每张工作纸都载有题目的连结、依次排列的提示、*背后的数学*部分和后续练习；学生可以自行决定要看多少个提示。
- **用于 STEP 备试。**书中收录了全部三份试卷的题目，由样本试卷直至 2025 年，其中不少在练习中附有相关题目。各章的总结和工具箱有助学生看到不同课题之间的联系。
- **给没有 STEP 支援的学生。**本书的写法，是要让一个有决心的学生能够独力读完。如果你的学校不提供 STEP 备试，本书可以承担其中大部分的工作。

版权与勘误

本书中凡属我本人创作的内容，包括正文、引导式解答、插图、练习和答案，均按照 Creative Commons CC0 1.0 Universal 授权条款贡献予公有领域：你可以为任何目的复制、改编和分享，无须征求许可。STEP 题目本身的版权属 Cambridge University Press & Assessment 所有，不在该授权条款的涵盖范围之内。因此，本书并没有转载这些题目。取而代之，每一节都连结到题目的官方来源：样本试卷和 1987 至 2018 年的试卷，连结到 STEP 支援计划的题目资料库；2019 年起的试卷，则连结到 OCR 提供下载的各年试卷。用到其他 STEP 题目的练习，也以同样的方式连结。

我已尽力确保书中的数学正确无误，但这样篇幅的一本书必定会有错误。如果你发现错误，

无论多么细微,请到本书的GitHub 储存库 `github.com/keithmcnulty/step_book_mcnulty` 以 issue 的形式提出。收到勘误后，本书的网站会随即更正。

动笔之前的最后几句话

Siklos 以米高安哲罗（Michelangelo）的大卫像的故事，为他那本书的引言作结：这座雕像是用一块数十年来无人问津的大理石雕成的。他说，他无法教会你数学，但下苦功一点一点地凿开题目，会让你内在的那位数学家显露出来。这番话我无法说得更好。本书为你提供大理石、一些工具，以及几幅图画，让你看看别人在石头里找到了甚么。凿石的功夫，就要靠你自己了。我希望你乐在其中，一如我编写这本书时那样乐在其中。

Keith McNulty
2026 年 11 月

词汇表

大学数学有自己的一套词汇，其中不少你大概从未见过。本词汇表收录了书中那些你在中学不大可能遇过的用语，每个词附上简短的解释，并在有帮助的地方指出哪一节用到或详细说明了这个概念。这些解释只为帮你入门，并非精确的定义；凡有严格定义的词，所指的那一节都会给出。

有几个词，例如*集合*、*函数*和*证明*，你早已熟悉，但大学里使用它们时比中学更为讲究，所以也一并收录。词条内以*斜体*标示的词，本身另有词条。

A

绝热不变量（Adiabatic invariant） 当系统缓慢改变时几乎保持不变的量。一个分子在缓慢压缩的气缸内来回弹跳时，它的速率乘以气缸长度几乎固定不变（Section 10.4）。

代数数（Algebraic number） 某个整数系数（可正、可负、可为零，但不全为零）多项式方程的根。$\sqrt{2}$、i 和 $\frac{1}{2}(1+\sqrt{5})$ 都是代数数；不是代数数的数称为*超越数*（transcendental）（Section 6.3）。

分析（Analysis） 为微积分奠定严格基础的数学分支：极限、连续性、微分、积分和无穷级数，全都精确地定义，并从基本原理出发加以证明。第 2 至 4 章是这门学问的入门。

角动量（Angular momentum） 动量在转动中的对应量。对于绕轴自转的物体，它等于*转动惯量*（moment of inertia）乘以角速度；只要没有外来的转动效应，角动量便守恒（Section 11.6）。

算术几何平均（Arithmetic–geometric mean） 把一对数反复换成它们的算术平均和几何平均，由此得到两个数列，它们的共同极限就是算术几何平均。其收敛速度快得惊人（Section 2.6）。

结合律（Associative） 若对所有 a、b、c 都有 $(a*b)*c = a*(b*c)$，则运算 $*$ 满足结合律，括号因而可以省去。数的加法和乘法都满足结合律；减法则不然。

渐近（Asymptotic） 若两个式子之比当 $n \to \infty$ 时趋向 1，则称它们渐近相等，记作 $f(n) \sim g(n)$。斯特灵公式（Stirling's formula）指出 $n! \sim \sqrt{2\pi n}\,(n/\mathrm{e})^n$，尽管两边之差会无止境地增大（Section 12.23）。

公理（Axiom） 不加证明而视为出发点的命题，其他一切都由它推导出来。大学代数有

很大部分都是从一张简短的公理清单出发，探究由此可以推出甚么（Section 7.3、Section 6.5）。

B

基础步骤（Base case） *归纳法*（induction）证明的第一步：验证命题对最小的值成立，通常是 $n = 0$ 或 $n = 1$。其后的*归纳步骤*（inductive step）则证明：命题对某个值成立，便对下一个值也成立。

基本比率谬误（Base-rate fallacy） 只凭证据判断某事的概率（probability），却忽略它本来有多常见，这种错误就是基本比率谬误。如果某种疾病足够罕见，即使准确的测试得出阳性结果，它是虚惊一场的可能性也可以比真正患病更大；实际是否如此，取决于疾病有多罕见和测试有多准确（Section 12.2）。

基（Basis） 一组向量，使得空间中每个向量都能以唯一的方式写成它们的组合。在平面上，**i** 和 **j** 构成一个基；任何一对不平行的向量也是如此。一个基所含向量的数目就是空间的*维数*（dimension）（Section 6.3）。

贝叶斯定理（Bayes' theorem） 把一个条件概率倒转过来的法则：$\mathrm{P}(A \mid B) = \mathrm{P}(B \mid A)\,\mathrm{P}(A)/\mathrm{P}(B)$。它告诉你怎样根据证据修正原有的判断（Section 12.2）。

偏差（Bias） 若一个*估计量*（estimator）在大量样本上平均而言偏高或偏低，它便是有偏的；若它的期望值恰好等于所估计的量，它便是*无偏*（unbiased）的（Section 13.3）。

双射（Bijection） 一一对应：一种把一个集合的每个成员与另一个集合的恰好一个成员配对的规则，而且两边都没有剩余。两个有限集合大小相同，恰好就是它们之间存在双射的时候；无限集合的大小正是这样定义的（Section 1.8、Section 9.10）。

二元运算（Binary operation） 把两个对象结合而得出第三个对象的规则，例如数的加法、矩阵的乘法或函数的复合。

界（Bound） 一个数集或数列的上界，是其中没有任何数超过的数；下界则是其中没有任何数低于的数。若数列两者皆有，便称为*有界*（bounded）（Section 2.2）。

分支过程（Branching process） 一种族群模型，其中每个个体各自独立地产生随机数目的后代。核心问题是这个家族会否绝后，而答案来自*概率生成函数*（probability generating function）（Section 12.15）。

C

变分法（Calculus of variations） 研究这样一类问题的学问：未知的是整个函数或整条曲线，要选取它使某个量尽可能小或尽可能大，例如最短的路径、最快的滑道、围出最

大面积的曲线（Section 4.7）。

悬链线（Catenary） 悬挂的链条所呈的形状，即 $y = a\cosh(x/a)$ 的图像。它看似抛物线，其实不是（Section 11.14）。

柯西分布（Cauchy distribution） 一种连续分布，图像呈钟形，但尾部极厚，以致没有平均值。从中抽取样本，样本的平均值并不比单独一个观测值可靠（Section 12.22）。

中心极限定理（Central limit theorem） 这个定理指出：大量独立、同分布而方差有限的随机量之和，经适当平移和缩放后，近似服从正态分布，不论那个共同的分布是甚么。对于分布各不相同的随机量也有相应的版本，条件是其中没有少数几个量主导整个和。它解释了为何钟形曲线无处不在（Section 12.24）。

混沌（Chaos） 确定性系统的一种行为：它对起点极为敏感，以致实际上无法作长期预测——两个非常接近的起点，会导向截然不同的未来（Section 8.7、Section 8.8）。

封闭（Closed） 若把集合中两个成员结合起来，所得的总是集合中的另一个成员，则该集合在这个运算下封闭。偶数在加法下封闭；奇数则不然。

交换律（Commutative） 若 $a * b = b * a$ 恒成立，则运算 $*$ 满足交换律。数的乘法满足交换律；矩阵的乘法和*四元数*（quaternions）的乘法则不满足（Section 7.1、Section 6.7）。

完备性（Completeness） 实数没有空隙这一性质：每个有上界的递增数列都收敛于某个实数。有理数并无这个性质，因为一列分数可以逐步逼近 $\sqrt{2}$。*分析*有很大部分建基于此（Section 2.2）。

同余（Congruence） $a \equiv b \pmod{n}$，读作“a 与 b 模 n 同余”，意思是 a 和 b 除以 n 所得的余数相同；等价地说，n 整除 $a - b$。用同余式进行的运算称为*模算术*（modular arithmetic）（Section 1.5）。

猜想（Conjecture） 人们相信为真但尚未证明的命题。

守恒定律（Conservation law） 指出某个量（例如能量、动量或角动量）在系统演变过程中保持不变的命题。有了守恒定律，往往不必追踪细节便能解决问题（Section 10.1）。

连续（Continuous） 若输入的微小改变只引致输出的微小改变，则函数是连续的；通俗地说，就是笔不离纸也能画出它的图像。精确地说，若当 $x \to a$ 时 $\mathrm{f}(x) \to \mathrm{f}(a)$，则 f 在 a 连续（Section 3.1、Section 8.8）。

反证法（Contradiction, proof by） 先假设命题不成立，再由此推出不可能的结论，这样的证明称为反证法。经典的例子是 $\sqrt{2}$ 为无理数的证明（Section 1.3）。

逆否命题（Contrapositive） “若 P 则 Q”的逆否命题是“若非 Q 则非 P”。两个命题在逻辑上等价，而证明逆否命题往往较为容易。它与*逆命题*（converse）“若 Q 则 P”并

不相同，后者很可能不成立。

收敛（Converge） 若数列各项能够任意接近某个固定的数，并一直保持接近，则数列收敛，该数就是它的*极限*（limit）。若级数的*部分和*（partial sums）收敛，则级数收敛。不收敛的数列或级数称为*发散*（diverge）（Section 2.1、Section 2.3）。

凸（Convex） 若连接函数图像上任意两点的弦都位于图像之上或与图像重合，则函数是凸的，例如 x^2 或 e^x。许多不等式，包括 AM–GM 不等式，都是凸性的推论（Section 3.3）。

推论（Corollary） 由刚证明的*定理*（theorem）很快便能得出的结果。

相关系数（Correlation） 一个介乎 -1 与 1 之间的数，量度两个随机量之间的关系有多接近直线关系。它是*协方差*（covariance）除以两个标准差所得的值（Section 13.7）。

可数（Countable） 若集合的成员可以排成第一个、第二个、第三个……这样列下去（可能永无止境），则该集合是可数的。整数是可数的，而令人意外的是，分数也是可数的；实数则不可数（Section 1.8）。

反例（Counterexample） 显示某个一般性命题不成立的单一例子。不论有多少例子支持某个论断，一个反例已足以把它推翻。

协方差（Covariance） $\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$，量度两个随机量如何一同变化。独立的随机量协方差为零，但协方差为零并不代表独立（Section 12.4、Section 13.7）。

循环（Cycle） 反复施行某个函数时依次到访的一组点：$x_1 \mapsto x_2 \mapsto \cdots \mapsto x_n \mapsto x_1$。数 n 是它的*周期*（period），而这些点称为*周期点*（periodic points）（Section 8.8）。

摆线（Cycloid） 滚动的轮子边缘上一点所描出的曲线。把它上下倒转，它既是最速降线，也是无论从哪里滑下所需时间都相同的曲线（Section 11.4）。

分圆多项式（Cyclotomic polynomial） 以全部*本原*（primitive）n 次*单位根*（roots of unity）为根、且别无其他根的多项式 $\Phi_n(x)$。它的系数都是整数，而且*不可约*（irreducible）；$x^n - 1$ 等于 n 的所有因数 d 所对应的 $\Phi_d(x)$ 之积：例如 $x^4 - 1 = (x-1)(x+1)(x^2+1)$（Section 6.1）。

D

错排（Derangement） 把一列东西重新排列，使得没有任何一项留在原来的位置（Section 9.8）。

导子（Derivation） 作用于函数（或其他对象）的运算 D，满足 $D(f+g) = Df + Dg$ 和乘积法则 $D(fg) = f\,Dg + g\,Df$。求导是标准的例子，而它的许多性质单凭这两条法则便可推出（Section 6.5）。

行列式（Determinant） 由方阵算出的一个数，量度相应的变换把面积或体积放大多少倍。它等于零，恰好就是矩阵没有逆矩阵的时候（Section 7.3）。

可微（Differentiable） 若函数的图像在某点有明确的切线，即定义导数的那个极限存在，则函数在该点可微。每个可微函数都是*连续*的，但反之不然：$|x|$ 在 0 连续但不可微。

维数（Dimension） 一个基所含向量的数目，即*向量空间*（vector space）中独立方向的数目。

发散（Diverge） 见*收敛*。

动力系统（Dynamical system） 描述系统状态如何随时间改变的规则，可以是逐步改变（由 $x_{n+1} = \mathrm{f}(x_n)$ 定义的数列），也可以是连续改变（微分方程）。第 8 章以此为主题。

E

特征值、特征向量（Eigenvalue, eigenvector） 若对某个数 λ 有 $\mathbf{M}\mathbf{v} = \lambda\mathbf{v}$，则非零向量 $\mathbf{v}$ 是矩阵 $\mathbf{M}$ 的特征向量，而该数就是它的特征值：这个变换只把 $\mathbf{v}$ 拉伸，而不把它转动（Section 7.2）。

包络线（Envelope） 与一族曲线中每一条都相切的曲线。以固定速率抛出的球，其所有轨迹的包络线就是*安全抛物线*（safety parabola），即球所能到达的区域的边界（Section 11.9）。

平衡（Equilibrium） 系统不再改变的状态。若微小的扰动始终保持微小，平衡便是*稳定*（stable）的；若扰动还会逐渐消失，便是*渐近稳定*（asymptotically stable）的，有摩擦的摆就是这样；没有摩擦而垂直下垂的摆是稳定的，但轻轻一推便会令它永远摆动下去。若有些微小的扰动会增大，平衡便是*不稳定*（unstable）的（Section 8.3、Section 11.8）。

估计量（Estimator） 根据数据计算出某个未知量（例如族群的大小或分布的平均值）的估计值的规则（Section 13.3）。

欧拉示性数（Euler characteristic） 对于有 V 个顶点、E 条棱和 F 个面的多面体，指 $V - E + F$ 这个数；没有洞的多面体，这个数永远是 2（Section 7.9）。

存在性证明（Existence proof） 证明某物存在，却不一定说明怎样把它找出来的证明。*鸽巢原理*（pigeonhole principle）和*介值定理*（intermediate value theorem）都能给出存在性证明。

期望值（Expectation） 随机量 X 的平均值，记作 $\mathrm{E}(X)$：即多次独立重复试验的长期平均。期望值具有可加性，$\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$，不论 X 与 Y 是否独立（Section 9.11）。

极端原理（Extremal principle） 一种策略：考察某类对象中最小、最大、最先或最后的一个，并证明它具有某种性质，从而得出所求的结果或导出矛盾（Section 1.3）。

F

域（Field） 可以进行加、减、乘、除（除数不为零）而且所有惯常法则都成立的数系。有理数、实数和复数都是域；模质数的整数也是。整数不是域，因为并非总能相除（Section 6.3）。

域扩张（Field extension） 包含一个较小的域的较大的域，例如所有形如 $a + b\sqrt{2}$（其中 a、b 为有理数）的数，它们包含了有理数。域扩张的*次数*（degree），就是把它看成较小的域上的*向量空间*时的*维数*（Section 6.3）。

不动点（Fixed point） 满足 $\mathrm{f}(x) = x$ 的点 x，即函数让它留在原处的点。在图像上，不动点就是 $y = \mathrm{f}(x)$ 与 $y = x$ 相交之处（Section 2.2、Section 8.8）。

参考系（Frame of reference） 量度位置和运动时所采取的观点。*惯性*（inertial）参考系是牛顿（Newton）第一定律成立的参考系；在加速的参考系中，例如正在加速的火车上，看起来会多出一些力（Section 11.5）。

函数（Function） 一种规则，把一个集合（*定义域*（domain））的每个成员对应到另一个集合（*上域*（codomain））的恰好一个成员。在大学里，函数并不等同于公式：许多重要的函数没有公式，而有些公式并不定义函数。

函数方程（Functional equation） 以函数为未知量的方程，例如 $\mathrm{f}(x+y) = \mathrm{f}(x)\mathrm{f}(y)$（Section 3.8）。

代数基本定理（Fundamental theorem of algebra） 每个次数至少为 1 的复系数多项式都有一个复数根，因而可以在复数范围内完全分解成一次因式。

G

伽玛函数（Gamma function） 把阶乘推广至非整数的函数 Γ：对整数 n 有 $\Gamma(n+1) = n!$，而且 $\Gamma(\frac{1}{2}) = \sqrt{\pi}$，在某种意义上这是说 $(-\frac{1}{2})! = \sqrt{\pi}$（Section 4.4）。

高斯整数（Gaussian integer） 形如 $a + b\mathrm{i}$ 的复数，其中 a 和 b 为整数。它们的表现与普通整数十分相似，有自己的质数，也有唯一分解（Section 1.6）。

生成函数（Generating function） 以某个数列各项为系数的幂级数 $a_0 + a_1x + a_2x^2 + \cdots$。对级数所作的运算（例如乘法）对应于对数列所作的有用运算（Section 9.4）。*概率生成函数*是对取整数值的随机量的各个概率这样做（Section 9.6、Section 12.15）；*矩生成函数*（moment generating function）则是它的近亲（Section 12.24）。

群（Group） 一个集合连同一个*二元运算*，这个运算满足*结合律*，有*单位元*（identity），并且每个元素都有*逆元*（inverse）。一个图形的所有对称构成一个群，非零复数在乘法下也构成一个群。群是描述对称的语言（Section 6.6、Section 6.7）。

H

双曲几何（Hyperbolic geometry） 一种几何，其中过已知直线外一点有无穷多条直线永不与该直线相交，而三角形的内角和小于 π。配上适当的距离量度方法的圆盘，就是它的一个模型（Section 5.8）。

I

单位元；恒等式（Identity） 运算 $*$ 的单位元是这样的元素 e：对每个 a 都有 $e * a = a * e = a$。加法的单位元是 0，乘法的是 1，矩阵乘法的是单位矩阵 $\mathbf{I}$。另一个意思是恒等式，即对变量的每个值都成立的等式，例如 $\cos^2\theta + \sin^2\theta \equiv 1$。

当且仅当（If and only if） “P 当且仅当 Q”，常写作“P iff Q”或 $P \iff Q$，意思是 P 和 Q 要么同时为真，要么同时为假。证明它需要两个论证，每个方向各一个。

独立（Independent） 若知道其中一个事件已经发生并不改变另一个事件的概率，则两个事件独立：$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$。三个或更多事件可以两两独立，却并非整体独立（Section 12.4）。

归纳法（Induction） 证明某个命题对每个整数 n 都成立的方法：先证明它对第一个值成立（*基础步骤*），再证明只要它对某个值成立，便对下一个值也成立（*归纳步骤*）。在*强归纳法*（strong induction）中，归纳步骤可以使用命题对所有较小的值都成立这一点，而不只是对前一个值（Section 1.2）。

无穷递降法（Infinite descent） 反证法的一种形式：证明某个整数问题的任何一个解都会引出一个更小的解，从而得到一列无穷递减的正整数，而这是不可能的（Section 1.5）。

介值定理（Intermediate value theorem） 若*连续*函数取到 a 和 b 这两个值，它便取到两者之间的每一个值。因此，变号的连续函数必定有根（Section 8.8）。

不变量（Invariant） 在某个过程进行期间，或在某种变换之下，保持不变的量。要证明某事不可能，或者不必追踪每一步便解决问题，寻找不变量是最有力的方法之一（Section 7.7、Section 10.4）。

逆元；反函数（Inverse） 对于具有*单位元* e 的运算，元素 a 的逆元是满足 $a * b = b * a = e$ 的元素 b。反函数则把函数的作用还原：$\mathrm{f}^{-1}(\mathrm{f}(x)) = x$。

无理数（Irrational） 不能写成分数 $\frac{p}{q}$（其中 p 和 q 为整数，$q \neq 0$）的实数，例如 $\sqrt{2}$、π 或 e（Section 1.4）。

不可约（Irreducible） 若多项式不能写成两个次数较低、系数属同一类的多项式之积，则它是不可约的。$x^2 - 2$ 在有理数范围内不可约，但在实数范围内可分解为 $(x - \sqrt{2})(x + \sqrt{2})$（Section 6.1）。

同构（Isomorphic） 若两个结构（例如两个群）之间存在一个双射，使它们的运算完全对应，则这两个结构同构：它们是同一个结构，只是标签不同（Section 6.7、Section 6.8）。

迭代（Iteration） 反复施行同一个函数，于是 $x_1 = f(x_0)$，$x_2 = f(x_1)$，如此类推。数列 x_0、x_1、x_2、… 称为 x_0 的轨道（orbit）（Section 8.7）。

L

大数定律（Law of large numbers） 这个定理指出：只要期望值存在，一个随机量多次独立重复所得的平均值便趋向它的期望值（Section 12.22）。

引理（Lemma） 主要为了证明某个更大的定理而先行证明的结果。

极限（Limit） 数列或函数任意接近、并一直保持接近的那个值。把这个想法精确化，是十九世纪数学的伟大成就之一，也是分析的起点（Section 2.1）。

极限环（Limit cycle） 系统运动中的一个闭合回路，邻近的运动会螺旋地趋近（或远离）它，例如心跳或时钟的稳定节奏（Section 8.4）。

线性映射（Linear map） 向量空间之间满足 $T(\mathbf{u}+\mathbf{v}) = T(\mathbf{u}) + T(\mathbf{v})$ 和 $T(\lambda\mathbf{u}) = \lambda T(\mathbf{u})$ 的函数 T。每个线性映射都可以用矩阵表示，而线性代数就是研究线性映射的学问（Section 7.2）。

逻辑斯谛映射（Logistic map） 规则 $u_{n+1} = ru_n(1-u_n)$，是受食物供应限制的族群的简单模型，也是混沌的标准例子（Section 8.7）。

M

马可夫链（Markov chain） 一种在各个状态之间逐步转移的随机过程，其中下一步的概率只取决于当前的状态，而与过程怎样到达这个状态无关（Section 12.16）。

最大似然法（Maximum likelihood） 估计未知量的一种方法：选取使观测所得数据出现的概率最大的那个值（Section 13.2）。

无记忆性（Memoryless） 若不论你已经等了多久，余下的等候时间的分布都与开始时相同，则这个随机等候时间具有无记忆性。指数分布是唯一具有无记忆性的连续分布（Section 12.18）。

最小多项式（Minimal polynomial） 以某个已知代数数为根、系数为有理数而首项系数为 1 的多项式之中，次数最低的一个。$\sqrt{2}$ 的最小多项式是 $x^2 - 2$（Section 6.3）。

模算术（Modular arithmetic） 见同余。

转动惯量（Moment of inertia） 质量在转动中的对应量：量度要改变物体绕某轴转动的快慢有多困难。它取决于质量离轴有多远，而不只取决于质量有多少（Section 11.7）。

单调（Monotonic） 若数列或函数从不下降（递增）或从不上升（递减），则它是单调的。有界的单调数列必定收敛（Section 2.2、Section 11.10）。

N

必要及充分（Necessary and sufficient） 若没有 P 则 Q 不可能成立（$Q \Rightarrow P$），则 P 是 Q 的必要条件；若 P 保证 Q 成立（$P \Rightarrow Q$），则它是充分条件。既必要又充分的条件与 Q 等价。

简正模态（Normal modes） 系统的一些特殊振动形态，其中每个部分都以相同的频率同步运动。系统的任何振动都是它的简正模态的组合（Section 11.2）。

O

阶（Order） 群中元素 g 的阶，是使 g^n 等于单位元的最小正整数 n：即必须把它施行多少次才回到起点（Section 7.4）。若不存在这样的 n，例如整数加法群中的数 1，则称 g 的阶为无限（infinite order）。有限群的阶是它所含元素的数目。

正交（Orthogonal） 即垂直。在大学里，这个词的用法推广至向量以外：若在适当的范围上 $\int \mathrm{f}(x)\mathrm{g}(x)\,\mathrm{d}x = 0$，则两个函数 f 和 g 可称为正交（Section 8.6）。

P

参数（Parameter） 在某一个问题之内固定不变、但在不同问题之间可以改变的量，例如逻辑斯谛映射中的 r。许多 STEP 题目都会问答案怎样随参数而变。

部分和（Partial sum） 级数首 n 项之和 $a_1 + a_2 + \cdots + a_n$。若部分和收敛，则级数收敛（Section 2.3）。

鸽巢原理（Pigeonhole principle） 若把多于 n 件物件放进 n 个盒子，则必有某个盒子装有至少两件。它虽然简单，却能证明一些出人意表的存在性结果（Section 9.2）。

泊松过程（Poisson process） 一种模型，描述在时间或空间中以恒定的平均率随机而独立地发生的事件：放射性衰变、电话来电、天上的星星（Section 12.19、Section 12.18）。

本原单位根（Primitive root of unity） 一个 n 次单位根，其各次幂恰好给出全部 n 个 n 次单位根；等价地说，对任何较小的 m，它都不是 m 次单位根（Section 6.1）。

概率密度函数（Probability density function） 对于连续的随机量，指这样的函数 f：它在某个区间上的积分，就是随机量落在该区间内的概率。

证明（Proof） 从定义、公理和已经证明的结果出发，说明某个命题为真的完整逻辑论证。在大学里，即使一段计算得出正确答案，除非每一步都有理据，否则不算证明。

Q

四元数（Quaternions） 哈密顿（William Rowan Hamilton）于 1843 年发现的数系，它把复数加以扩充，包含 -1 的三个平方根，记作 i、j、k，满足 $i^2 = j^2 = k^2 = ijk = -1$。四元数的乘法不满足*交换律*。它们用于描述三维空间中的旋转（Section 6.7、Section 7.5）。

R

随机变量（Random variable） 其值由随机试验的结果决定的量，例如掷十次硬币所得正面的次数。

随机游走（Random walk） 由接连的随机步伐组成的路径，例如一枚棋子按掷硬币的结果向左或向右移动一格（Section 9.10、Section 12.17）。

快度（Rapidity） 狭义相对论中量度快慢的一个量；与速度不同，当一个运动与另一个同方向的运动合成时，快度只需直接相加。若 c 为光速，则速度 v 对应的快度为 $\operatorname{artanh}(v/c)$（Section 7.6）。

递推关系（Recurrence relation） 以数列中较前的项表示每一项的等式，例如斐波那契（Fibonacci）数所满足的 $F_{n+1} = F_n + F_{n-1}$。

共振（Resonance） 系统受到周期性的推动，而推动的频率等于或接近系统本身的固有频率时所产生的强烈反应，就像配合秋千的节奏去推秋千上的小孩那样（Section 11.1）。

黎曼积分（Riemann integral） 积分的精确定义：把积分定为许多细长矩形面积之和的极限（当这个极限存在时）（Section 4.1）。

严谨性（Rigour） 一种论证的标准：每一步都要有理据，不会因为某事看来显然、或者在你试过的例子中都成立，便视为理所当然（第 3 章）。

单位根（Root of unity） 对某个正整数 n 满足 $z^n = 1$ 的复数 z。全部 n 次单位根是 $e^{2\pi ik/n}$，其中 $k = 0、1、\ldots、n-1$，它们在单位圆上构成一个正 n 边形（Section 5.2）。

S

样本空间（Sample space） 一个随机试验所有可能结果组成的集合。事件是它的子集（Section 12.1）。

集合（Set） 一批对象（称为它的*元素*（elements））合起来看成的单一整体。元素的次序无关重要，重复与否也无关重要：$\{1,2,3\}$ 和 $\{3,1,2,2\}$ 是同一个集合。

简谐运动（Simple harmonic motion） 加速度与相对某固定点的位移成正比、并且指向该点的运动，$\ddot{x} = -\omega^2 x$。它的解是 $x = A\cos(\omega t + \varepsilon)$，而周期 $2\pi/\omega$ 与摆动的幅度无关（Section 11.4）。

稳定、不稳定（Stable, unstable） 见*平衡*。

斯特灵公式（Stirling's formula） 近似式 $n! \approx \sqrt{2\pi n}\,(n/\mathrm{e})^n$，其百分误差随 n 增大而趋向零（Section 12.23）。

子群（Subgroup） *群*的一个子集，它在同一运算下本身也是一个群（Section 6.6）。

对称性论证（Symmetry argument） 这样的一种论证：由于问题中没有任何东西能把两个量区分开来，所以它们必定相等。一个好的对称性论证可以取代好几页的计算（Section 12.5）。

T

定理（Theorem） 已获证明的重要命题。较次要的结果常称为*命题*（propositions），在通往定理的途中证明的称为*引理*，而很快便能得出的结果则称为*推论*。

迹（Trace） 方阵主对角线上各元素之和。与*行列式*一样，把矩阵改用另一个*基*来表示时，它并不改变；它还等于各*特征值*之和（Section 7.4、Section 7.7）。

超越数（Transcendental） 不是*代数数*的数：它不是任何整数系数（不全为零）多项式方程的根。e 和 π 都是超越数，不过要证明这一点并不容易（Section 1.4）。

三角不等式（Triangle inequality） $|a+b| \le |a| + |b|$，对实数、复数或向量都成立：三角形的一边永远不会比另外两边加起来更长。

U

唯一分解（Unique factorisation） 每个大于 1 的整数都只能以一种方式写成质数之积(不计因数的次序）这一性质。它并不如看起来那么显然：在有些数系中它并不成立（Section 1.7）。

唯一性定理（Uniqueness theorem） 指出某个问题只有一个解的定理。对于微分方程，当方程的性质不够良好时，唯一性可能失效，以致同一个起点会导向几个不同的未来（Section 8.1）。

V

向量空间（Vector space） 一批称为向量的对象，它们可以相加，也可以乘以数，并遵从平面上箭头所满足的那些熟悉法则。多项式、函数以及许多微分方程的解都构成向量

空间，所以线性代数的方法对它们都适用（Section 6.3）。

W

良定义（Well-defined） 若某个对象或规则的定义确实指明了单独一个毫不含糊的东西，例如不取决于过程中所作的任意选择，它便是良定义的。

良序原理（Well-ordering principle） 每个非空的正整数集合都有最小的成员。它与*归纳法*原理等价，而每一个以“取最小的反例”开始的论证，背后都是这个原理（Section 1.3）。

不失一般性（Without loss of generality） 常简写为 WLOG 的一句用语，意思是论证只考虑其中一种情况，因为其他情况除了要重新标记之外并无分别。“不失一般性，设 $a \leq b$”的意思是：如果 $a > b$，把两者的名称对调便可。

数学符号

数学写作往往以寥寥几个记号承载丰富的意思，而大学讲师所用的许多符号，在中学课本里甚少出现。下面各表按主题列出本书用到的符号，并说明每个符号的含义和读法。入读大学后不出几个星期，其中大部分都会变得习以为常；在此之前，这里可供你随时查阅。

有两项惯例值得一开始就知道。第一，本书依照 STEP 的做法，把某些具有固定含义的字母印成正体而非斜体：e 表示自然对数的底，i 表示 $\sqrt{-1}$，$\mathrm{d}y/\mathrm{d}x$ 中的 d，以及函数的名称，例如 $\mathrm{f}(x)$ 和 $\mathrm{P}(A)$。斜体字母则表示变量。第二，同一个符号在不同场合有时会有不同的意思；$\equiv$ 和 $\sim$ 是最常见的例子，表中会在相应位置指出。

逻辑与证明

符号	含义	例子或读法
$P \Rightarrow Q$	P 蕴含（implies）Q：若 P 成立，则 Q 亦成立	$x = 2 \Rightarrow x^2 = 4$；“$P$ 蕴含 Q”
$P \Leftarrow Q$	P 由 Q 所蕴含	“只要 Q 成立，P 就成立”
$P \iff Q$	P 当且仅当（if and only if）Q：两者互相蕴含	$x^2 = 4 \iff x = \pm 2$；亦写作“iff”
$\forall$	对所有、对每一个	$\forall x \in \mathbb{R},\ x^2 \geq 0$
$\exists$	存在	$\exists n \in \mathbb{Z}$ 使得 $n^2 = 9$
$\neg P$	非 P，即 P 的否定（negation）	“$\forall x,\ P(x)$”的否定是“$\exists x$ 使得 $\neg P(x)$”
$:=$	定义为	$\mathrm{f}(x) := x^2 + 1$
■	证明完毕；亦写作 QED	源自拉丁文 *quod erat demonstrandum*，意即“这就是所要证明的”

集合与数系

符号	含义	例子或读法
$\{\dots\}$	由所列元素组成的集合（set）	$\{1,2,3\}$
$\{x:\dots\}$	所有满足 ... 的 x 所组成的集合；亦写作 $\{x \mid \dots\}$	$\{x:x^2<2\}$；“满足 $x^2<2$ 的 x 所组成的集合”
$x\in A$	x 是 A 的元素（element）	$\sqrt{2}\in\mathbb{R}$；“$\sqrt{2}$ 属于 $\mathbb{R}$”
$x\notin A$	x 不是 A 的元素	$\sqrt{2}\notin\mathbb{Q}$
$A\subseteq B$	A 是 B 的子集（subset）：A 的每个元素都在 B 之内	$\mathbb{Z}\subseteq\mathbb{Q}$
$A\cup B$	并集（union）：属于 A 或 B（或同时属于两者）的元素	$\{1,2\}\cup\{2,3\}=\{1,2,3\}$
$A\cap B$	交集（intersection）：同时属于 A 和 B 的元素	$\{1,2\}\cap\{2,3\}=\{2\}$
$A\setminus B$	属于 A 但不属于 B 的元素	$\{1,2,3\}\setminus\{2\}=\{1,3\}$
A'	补集（complement）：所有不属于 A 的元素	亦写作 A^{c} 或 $\bar{A}$
$\varnothing$	空集（empty set），不含任何元素	$\{x\in\mathbb{R}:x^2<0\}=\varnothing$
$\lvert A\rvert$	有限集 A 的元素个数	$\lvert\{a,b,c\}\rvert=3$
$A\times B$	所有数对 (a,b) 所组成的集合，其中 $a\in A$，$b\in B$	$\mathbb{R}\times\mathbb{R}$ 就是平面，亦写作 $\mathbb{R}^2$
$\mathbb{N}$	自然数（natural numbers）$1,2,3,\dots$	有些作者把 0 也算在内
$\mathbb{Z}$	整数（integers）$\dots,-2,-1,0,1,2,\dots$	源自德文 *Zahlen*，意即“数”
$\mathbb{Q}$	有理数（rational numbers），即分数	源自 *quotient* 一字
$\mathbb{R}$	实数（real numbers）	
$\mathbb{C}$	复数（complex numbers）	
$\mathbb{H}$	四元数（quaternions）	以哈密顿（Hamilton）命名（Section 6.7）
$\mathbb{Z}[\mathrm{i}]$	形如 $a+b\mathrm{i}$ 的数，其中 $a,b\in\mathbb{Z}$	高斯整数（Gaussian integers）（Section 1.6）
$[a,b]$	闭区间（closed interval）：$a\le x\le b$	包含端点
(a,b)	开区间（open interval）：$a<x<b$	不包含端点；亦可表示数对或点

数、整除性与近似

符号	含义	例子或读法
$a \mid b$	a 整除（divides）b：b 是 a 的整数倍	$3 \mid 12$；“3 整除 12”
$a \nmid b$	a 不整除 b	$5 \nmid 12$
$a \equiv b \pmod{n}$	a 与 b 除以 n 所得的余数相同	$17 \equiv 2 \pmod{5}$；“17 与 2 模 5 同余（congruent）”
$\equiv$	（不带模时）恒等于：对所有值都成立	$(x+1)^2 \equiv x^2 + 2x + 1$
$\lfloor x \rfloor$	下取整（floor）：$\le x$ 的最大整数	$\lfloor 3.7 \rfloor = 3$，$\lfloor -3.7 \rfloor = -4$
$\lceil x \rceil$	上取整（ceiling）：$\ge x$ 的最小整数	$\lceil 3.2 \rceil = 4$
$n!$	n 的阶乘(factorial),即 $1 \times 2 \times \cdots \times n$	按惯例 $0! = 1$
$\binom{n}{r}$	二项式系数（binomial coefficient）$\frac{n!}{r!\,(n-r)!}$	读作“n choose r”；亦写作 ${}^{n}\mathrm{C}_{r}$
$\approx$	约等于	$\pi \approx 3.1416$
$\sim$	渐近于（asymptotic to）：两者之比趋向 1	$n! \sim \sqrt{2\pi n}\,(n/\mathrm{e})^n$（但另见概率部分）
$\propto$	与……成正比	$v \propto 1/d$
$\pm, \mp$	正负；负正	$\cos(A \pm B) = \cos A \cos B \mp \sin A \sin B$
$0.\overline{01}$	循环数字	$0.\overline{01} = 0.010101\ldots$

函数

符号	含义	例子或读法
$\mathrm{f} : A \to B$	f 是由集合 A 到集合 B 的函数（function）	$\mathrm{f} : \mathbb{R} \to \mathbb{R}$
$x \mapsto \dots$	x 被送到 ...	$x \mapsto x^2 - 2$；“x 映射到 $x^2 - 2$”
$\mathrm{f} \circ \mathrm{g}$	复合（composition）：先作 g，再作 f	$(\mathrm{f} \circ \mathrm{g})(x) = \mathrm{f}(\mathrm{g}(x))$
f^{-1}	f 的反函数（inverse function）	并非 $1/\mathrm{f}$
f^n	在本书中，通常指把 f 施行 n 次	$\mathrm{f}^3(x) = \mathrm{f}(\mathrm{f}(\mathrm{f}(x)))$；但 $\sin^2 x$ 表示 $(\sin x)^2$
$\lvert x \rvert$	模（modulus），或绝对值	$\lvert -3 \rvert = 3$
$\exp x$	指数函数（exponential function）e^x	
$\ln x$	自然对数（natural logarithm），以 e 为底	在大学里常写作 $\log x$
sec, cosec, cot	$1/\cos$, $1/\sin$, $1/\tan$	
arcsin, arccos, arctan	反三角函数	亦写作 $\sin^{-1}$ 等
sinh, cosh, tanh	双曲函数（hyperbolic functions）：$\sinh x = \frac{1}{2}(\mathrm{e}^x - \mathrm{e}^{-x})$，$\cosh x = \frac{1}{2}(\mathrm{e}^x + \mathrm{e}^{-x})$，$\tanh = \sinh / \cosh$	读作“shine”（或“sinch”）、“cosh”和“than”（或“tanch”）
artanh	tanh 的反函数	$\operatorname{artanh} x = \frac{1}{2} \ln \frac{1+x}{1-x}$
max, min	当中最大者、最小者	$\max(3, 7) = 7$
$\Gamma(x)$	伽玛函数（gamma function）	$\Gamma(n+1) = n!$（Section 4.4）

求和、极限与微积分

符号	含义	例子或读法
$\sum$	求和	$\sum_{k=1}^{n} k = \frac{1}{2}n(n+1)$
$\prod$	求积	$\prod_{k=1}^{n} k = n!$
∞	无穷大：并不是一个数，而是“没有界限”的简写	$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$
$x \to a$	x 趋向 a	当 $n \to \infty$ 时，$\frac{1}{n} \to 0$
$\lim$	极限（limit）	$\lim_{x\to 0} \frac{\sin x}{x} = 1$
$\mathrm{f}'(x), \mathrm{f}''(x)$	一阶及二阶导数（derivative）	读作“f dash”或“f prime”
$\dot{x}, \ddot{x}$	对时间的导数	$\dot{x} = \frac{\mathrm{d}x}{\mathrm{d}t}$；读作“$x$ dot”、“x double dot”
$\frac{\partial f}{\partial x}$	偏导数（partial derivative）：对 x 求导，其他变量保持不变	读作“partial f by partial x”
$\int_a^b$	由 a 到 b 的定积分（definite integral）	
$\oint$	沿闭合回路的积分	$\oint p\,\mathrm{d}q$（Section 10.4）
$\delta x, \Delta x$	x 的微小变化；x 的变化	
ε, δ	分析学中惯常用来表示细小正数的记号	“对每个 $\varepsilon > 0$，都存在 $\delta > 0$ 使得 ...”

复数

符号	含义	例子或读法
i	-1 的一个平方根	工程师写作 j
$\operatorname{Re} z, \operatorname{Im} z$	实部与虚部	若 $z = 3 + 4\mathrm{i}$，则 $\operatorname{Re} z = 3$，$\operatorname{Im} z = 4$
$z^*, \bar{z}$	共轭复数（complex conjugate）	$(3+4\mathrm{i})^* = 3 - 4\mathrm{i}$；两种记法都有人使用
$\lvert z \rvert$	模：与 0 的距离	$\lvert 3 + 4\mathrm{i} \rvert = 5$
$\arg z$	辐角（argument）：由正实轴量起的角	$\arg \mathrm{i} = \frac{\pi}{2}$
$\mathrm{e}^{\mathrm{i}\theta}$	$\cos\theta + \mathrm{i}\sin\theta$	$\mathrm{e}^{\mathrm{i}\pi} = -1$
ω	惯常表示 1 的一个复立方根（cube root of unity）	$\omega = \mathrm{e}^{2\pi\mathrm{i}/3}$，$1 + \omega + \omega^2 = 0$

向量与矩阵

符号	含义	例子或读法
$\mathbf{a}$, $\overrightarrow{AB}$	向量（vector）；由 A 到 B 的向量	手写时作 $\underline{a}$
$\hat{\mathbf{n}}$	单位向量（unit vector），长度为 1	读作“n hat”
$\lvert\mathbf{a}\rvert$	$\mathbf{a}$ 的长度（大小）	
$\mathbf{a}\cdot\mathbf{b}$	纯量积（scalar product），又称点积（dot product）	$\mathbf{a}\cdot\mathbf{b}=\lvert\mathbf{a}\rvert\lvert\mathbf{b}\rvert\cos\theta$
$\mathbf{a}\times\mathbf{b}$	向量积（vector product），又称叉积（cross product）	同时垂直于 $\mathbf{a}$ 和 $\mathbf{b}$
$\mathbf{M}$	矩阵（matrix）	
$\mathbf{I}$	单位矩阵（identity matrix）	对每个 $\mathbf{v}$ 都有 $\mathbf{Iv}=\mathbf{v}$
$\mathbf{M}^{-1}$	逆矩阵（inverse matrix）	$\mathbf{M}\mathbf{M}^{-1}=\mathbf{I}$
$\det\mathbf{M}$	$\mathbf{M}$ 的行列式（determinant）	$\det\begin{pmatrix}a & b\\ c & d\end{pmatrix}=ad-bc$
$\operatorname{tr}\mathbf{M}$	迹（trace）：对角线上各元素之和	$\operatorname{tr}\begin{pmatrix}a & b\\ c & d\end{pmatrix}=a+d$
λ	惯常表示特征值（eigenvalue）	$\mathbf{Mv}=\lambda\mathbf{v}$

概率与统计

符号	含义	例子或读法
$\mathrm{P}(A)$	事件 A 的概率（probability）	
$\mathrm{P}(A \mid B)$	已知 B 发生时 A 的概率	$\mathrm{P}(A \mid B) = \mathrm{P}(A \cap B)/\mathrm{P}(B)$；读作“$A$ given B”
$\mathrm{E}(X)$	X 的期望值（expectation），即平均值	亦写作 μ
$\mathrm{Var}(X)$	X 的方差（variance）	$\mathrm{E}(X^2) - \mathrm{E}(X)^2$；亦写作 σ^2
$\mathrm{Cov}(X, Y)$	协方差（covariance）	$\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$
$\mathrm{Corr}(X, Y)$	相关系数（correlation）	$\mathrm{Cov}(X, Y)/\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}$；亦写作 ρ
$X \sim \ldots$	X 的分布为 ...	$X \sim \mathrm{B}(16, \frac{1}{2})$；“$X$ 的分布为 ...”
$\mathrm{B}(n, p)$	二项分布（binomial distribution）	n 次试验中成功的次数
$\mathrm{N}(\mu, \sigma^2)$	正态分布（normal distribution），平均值为 μ，方差为 σ^2	
$\mathrm{f}(x), \mathrm{F}(x)$	概率密度函数（probability density function）；累积分布函数（cumulative distribution function）	$\mathrm{F}(x) = \mathrm{P}(X \leq x)$
$\mathrm{G}(t)$	概率生成函数（probability generating function）	$\mathrm{G}(t) = \mathrm{E}(t^X)$
$\mathrm{M}_X(\theta)$	矩生成函数（moment generating function）	$\mathrm{M}_X(\theta) = \mathrm{E}(\mathrm{e}^{\theta X})$
$\bar{x}$	样本的平均值	$\bar{x} = \frac{1}{n}\sum x_i$
$\hat{\theta}$	未知量 θ 的估计值	读作“theta hat”

希腊字母

希腊字母在数学中随处可见，而且往往有约定俗成的含义。下表列出各字母的小写和大写形式（与罗马字母外形相同的大写字母，例如 alpha 的大写 A，不会用作独立的符号，故此从略）、名称，以及一些常见用途。

字母	名称	常见用途
α	alpha	角；方程的根
β	beta	角；方程的根；贝塔函数（beta function）
γ, Γ	gamma	欧拉（Euler）常数 $\gamma \approx 0.5772$；伽玛函数 Γ；在相对论中，$\gamma = 1/\sqrt{1 - v^2/c^2}$
δ, Δ	delta	微小的量；Δ 表示变化，或表示三角形、判别式（discriminant）
ε, ϵ	epsilon	微小的正数量；在力学中表示相位角
ζ	zeta	黎曼（Riemann）zeta 函数，当 $s > 1$ 时 $\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$（以更高深的方法可推广至 $s = 1$ 以外的其他值）
η	eta	效率；微小的量
θ, Θ	theta	角；统计学中的未知参数
ι	iota	甚少使用
κ	kappa	曲率；常数
λ, Λ	lambda	特征值；参数，例如泊松（Poisson）分布的率；波长
μ	mu	平均值；摩擦系数
ν	nu	频率
ξ, Ξ	xi	变量或随机量
π, Π	pi	$\pi \approx 3.14159$；Π 表示求积
ρ	rho	密度；相关系数；与原点的距离
σ, Σ	sigma	标准差；Σ 表示求和
τ	tau	时间；力矩
υ, Υ	upsilon	甚少使用
ϕ, φ, Φ	phi	角；黄金比 $\phi = \frac{1}{2}(1 + \sqrt{5})$；快度（rapidity）；$\Phi$ 表示标准正态分布函数
χ	chi	χ^2（卡方，chi-squared）分布
ψ, Ψ	psi	量子力学中的波函数
ω, Ω	omega	角速度或角频率；1 的立方根；Ω 表示样本空间

力学

符号	含义	例子或读法
g	重力加速度	在地球表面约为 $9.8\,\mathrm{m\,s^{-2}}$
G	万有引力常数	两个质量之间的力为 Gm_1m_2/r^2
e	恢复系数（coefficient of restitution）（以斜体印刷，切勿与 e 混淆）	沿碰撞线方向，分离速率 $= e\times$ 接近速率
ω	角速度；角频率	简谐运动（simple harmonic motion）中 $\ddot{x} = -\omega^2 x$
I	转动惯量（moment of inertia）	转动动能 $\frac{1}{2}I\omega^2$（Section 11.7）
c	光速	约为 $3 \times 10^8\,\mathrm{m\,s^{-1}}$
$\hbar$	普朗克（Planck）常数除以 2π	读作“h bar”（Section 8.6）

章

1

数论与证明

整数是数学中最古老的对象，却至今仍藏着数学中一些最深奥的秘密。

本章内容

1.1	质数工厂	STEP II 2002
1.2	质数分布得有多稀疏?	STEP III 2016
1.3	最小反例	STEP III 2015
1.4	e 是无理数，e^2 也是	STEP III 2024
1.5	无穷递降法	STEP III 2022
1.6	两个平方数之和	STEP II 2023
1.7	因数分解失效之时	STEP I 2016
1.8	计数有理数	STEP III 2020
1.9	三次曲线的切线	STEP I 2009
A	练习答案	第 99 页

为甚么是整数?

> 数学是科学的皇后，而数论（*number theory*）是数学的皇后。
>
> — 高斯（Carl Friedrich Gauss）

整数是每个人最先接触到的数学对象。你还未识字，就已经学会数数。然而，关于整数的最简单的问题，却可以难得出奇。每一个大于 2 的偶数，是否都是两个质数（prime）之和？没有人知道：哥德巴赫（Christian Goldbach）在 1742 年提出这个问题，电脑已经验证到 4×10^{18} 以外，但至今仍然没有人能够证明。像 11 和 13，或 101 和 103 这样相差 2 的质数对，是否有无穷多对？这同样没有人知道。不过在 2013 年，一位几乎寂寂无名的数学家张益唐（Yitang Zhang）震惊了全世界：他证明了有无穷多对质数，彼此相差不超过 70 百万，而这个数字后来已被降低至 246。费马（Pierre de Fermat）在 1637 年于书页边的空白处随手写下：当 $n \geq 3$ 时，$x^n + y^n = z^n$ 没有正整数解。人们花了超过 350 年，并动用了有史以来最精深的一些数学，才终于由怀尔斯（Andrew Wiles）在 1994 年证明他是对的。

问题连小孩子都听得懂，答案却可能要花上几个世纪——正是这种组合令数论叫人欲罢不能。这也是为甚么数论一直是证明（proof）的训练场。整数有无穷多个，你不可能逐一检查每个情况，于是就必须找出一个能一举涵盖所有情况的论证。你在往后的数学生涯中会用到的证明技巧，大部分最初都是在有关整数的问题上磨练出来的。

纯粹、无用，却支撑着互联网

1940 年，剑桥数学家哈代（G. H. Hardy）颇为自豪地写道：数论是所有数学中最纯粹的，与日常生活相距太远，不可能有任何实际用途，因而也不可能有任何有害的用途。他错得不能再错了。每当你在网上购物、发送加密信息或为手机解锁，你都在依靠数论。RSA 系统（1977 年）所依赖的事实是：把两个大质数相乘很容易，但要从它们的乘积还原出这两个质数，却似乎难得无从入手。现代的系统所用的，正是你会在本章末遇到的那些曲线上的算术。十八、十九世纪的“无用”数学，成了二十一世纪的安全保障。

你在大学会遇到甚么

如果你在大学修读数学，数论会是你最先到达的地方之一，也会是你最后到达的地方之一。

- **在第一个学期**，你几乎一定会修读一门名为*数与集合*（Numbers and Sets）或*数学基础*（Foundations）之类的课程。它会教你怎样正式地书写证明，而其中大部分素材都来自数论：归纳法（induction）、欧几里得算法（Euclidean algorithm）、模算术（modular arithmetic）、费马小定理（Fermat's Little Theorem）、中国余数定理（Chinese Remainder Theorem），以及一个令人惊讶的事实——有些无穷集合比另一些无穷集合更大。
- **在群论（group theory）中**（通常在第一年修读），你会发现许多数论事实，原来都是同一条关于对称的定理——拉格朗日定理（Lagrange's Theorem）——投下的影子。
- **在第二年**，一门*数论*课程通常会涵盖二次互反律（quadratic reciprocity，即高斯的"黄金定理"）、连分数（continued fractions）和佩尔方程（Pell's equation）。
- **再往后**，你还可以走得更远：*代数数论*（algebraic number theory），其中高斯整数（Gaussian integers）等新的数系挽救了唯一分解（unique factorisation）；*解析数论*（analytic number theory），它运用微积分和复分析（complex analysis）来数算质数；以及*椭圆曲线*（elliptic curves）的理论，它既是费马最后定理（Fermat's Last Theorem）证明背后的支柱，也是七道悬赏一百万美元的千禧年大奖难题（Millennium Prize Problems）之一的根源。

贯穿本章的问题

本章的九道 STEP 题目，围绕着三个大问题打转。做题时请把它们记在心上；我们会在本章结尾回到这些问题。

1. **质数有多少个？它们是怎样分布的？**人人都知道质数有无穷多个。但它们越往后变得有多稀疏？甚么样的论证才能告诉我们答案？
2. **怎样证明一个数不是分数？**要说明一个数是有理数（rational），只需举出一个例子。要说明它是无理数（irrational），却要排除无穷多种可能。这到底是怎样做到的？
3. **哪些方程有整数解？当我们改变量系时，又会发生甚么事？**有时我们能证明某个方程完全没有解。有时要了解整数，最好的方法却是走出整数的范围，进入复数（complex numbers），或走到一条曲线上。

关于证明的一章

在这三个问题底下，还贯穿着第四个主题：*怎样证明*。每一节之所以入选，部分原因是它展示了一种不同的证明方法的实际运用；合起来，它们就是一套相当适合大学一年级生

的工具。

§	课题	思考方式
1.1	费马数（Fermat numbers）	一个需要无穷多个质数的无穷族
1.2	质数的乘积	强归纳法（strong induction）；隐藏在二项式系数（binomial coefficients）中的信息
1.3	再谈 $\sqrt{2}$	“取最小的那一个”：良序原理（well-ordering principle）
1.4	e 与 e^2	把一个整数严格地困在 0 与 1 之间
1.5	不可能的方程	无穷递降法（infinite descent）；模 n 运算
1.6	两个平方数之和	转换场景：复数
1.7	一个奇怪的数系	质疑你视为理所当然的事
1.8	把分数逐一列出	构造双射（bijection）；反向的递降
1.9	立方数与曲线	用几何解决算术问题

怎样阅读本章

每一节都以一道 STEP 题目的连结开始，题目与当年考试所出的完全一样。请先自己尝试，至少做半小时，然后才往下读。一道你曾经苦苦思索过的题目，会比一道你只是读过的题目令你学到多得多。当你往下读时，问问自己方框是邀请你在答案出现之前先停下来想一想，而停一停，试一试方框则标示着关键思路已在你触手可及之处的时刻。之后，每一节的背后的数学部分会把目光越过考试，望向你在大学将会遇到的数学。

各节的深度大致上逐步递增，但它们基本上互相独立。如果某一节未能吸引你，就先跳过，稍后再回来。

1.1 质数工厂

题目

STEP II 2002，第 3 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/02-s2-q3`

引子

你几乎肯定早已知道质数（prime）有无穷多个。欧几里得（Euclid）约在公元前 300 年证明了这件事，这也是每个人最早接触到的证明之一。那么，STEP 的题目为甚么要你再证明一次？

因为这个证明的运作方式完全不同。欧几里得的论证是说：*给我任何一张有限的质数清单，我都能指出它并不完整。*而这个证明则说：*这里有一个预先写好、明确给出的数列（sequence），当中每一个数都含有一个其他各数都没有的质数。*它是一部制造新质数的机器，而它之所以行得通，全靠一条简洁优美的恒等式（identity），题目会一步一步引领你走向它。

途中你还会遇上一些连费马（Fermat）本人也被骗倒的数。

动笔之前

先把题目从头到尾读一遍，留意它的结构。它分为四个阶段：

1. 计算一些例子；
2. 找出（其实是题目告诉你）一个规律，并用这些例子验证；
3. 证明这个规律在一般情况下成立；
4. 两次运用这个规律：先推导出关于公因数（common factor）的结论，再推导出关于质数的结论。

这种“例子 → 猜想 → 证明 → 推论”的结构，正是许多数学研究的实际做法，STEP 的

题目也经常依循这个模式。遇到这种结构时，开头的部分通常不只是热身：你日后需要的直觉，正是在那里建立的。

> **想一想**　最后一行问的是*无穷多个质数*，中间问的是*公因数*。“任何两个费马数（Fermat number）都没有公因数”和“质数有无穷多个”之间，究竟有甚么关系？

请记住这个问题，我们稍后会回来讨论。

逐步解题

阶段 1　熟习这些数

要小心记号：2^{2^n} 是指 $2^{(2^n)}$，而不是 $(2^2)^n$。指数（exponent）本身就是 2 的幂，所以这些数增长得*非常*快。

$$F_0 = 2^1 + 1 = 3, \qquad F_2 = 2^4 + 1 = 17,$$
$$F_1 = 2^2 + 1 = 5, \qquad F_3 = 2^8 + 1 = 257.$$

> **想一想**　这四个数有甚么值得留意的地方？

它们全是质数。（验证 257：它不能被 $2, 3, 5, 7, 11$ 或 13 整除，而 $17^2 = 289 > 257$，所以可以就此停下。）下一个数 $F_4 = 2^{16} + 1 = 65537$ 也是质数。皮埃尔．德．费马在 1640 年正正注意到这一点，并猜想（conjecture）*所有*这种形式的数都是质数。请把这件事记在心里，稍后我们会看看他是否猜对了。

阶段 2　验证恒等式

现在就较小的 k 验证 $(*)$：

- $k = 1$：左边只有 $F_0 = 3$，而 $F_1 - 2 = 3$。✓
- $k = 2$：$F_0 F_1 = 3 \times 5 = 15$，而 $F_2 - 2 = 15$。✓
- $k = 3$：$F_0 F_1 F_2 = 15 \times 17 = 255$，而 $F_3 - 2 = 255$。✓

> **想一想**　看看由 $k = 2$ 到 $k = 3$ 的那一步。我们把 15 乘以 17，得到 255。用费马数来表达，就是计算了 $(F_2 - 2) \times F_2$，结果得到 $F_3 - 2$。这是巧合吗？

这个观察就是整个证明的缩影。让我们看看它为甚么应该成立。

阶段 3　证明恒等式

停一停，试一试 继续读下去之前，试试直接证明

$$(F_k - 2)\,F_k = F_{k+1} - 2.$$

提示：把所有东西都用 2 的幂来表示。$F_k - 2$ 是甚么样子的?

我们有 $F_k - 2 = 2^{2^k} - 1$ 和 $F_k = 2^{2^k} + 1$，所以

$$(F_k - 2)\,F_k = \left(2^{2^k} - 1\right)\left(2^{2^k} + 1\right).$$

想一想 这是一个形如 $(a-1)(a+1)$ 的乘积。它等于甚么?

它是平方差（difference of two squares）$a^2 - 1$。取 $a = 2^{2^k}$，则

$$a^2 = \left(2^{2^k}\right)^2 = 2^{2\cdot 2^k} = 2^{2^{k+1}},$$

所以

$$(F_k - 2)\,F_k = 2^{2^{k+1}} - 1 = F_{k+1} - 2. \tag{†}$$

这就是整部机器的引擎。留意平方的作用：它把指数加倍，而 2^k 加倍后就是 2^{k+1}。费马数的定义与平方差的因式分解（factorisation）配合得天衣无缝。

这样一来，归纳法（induction）几乎是水到渠成。

想一想 在归纳步骤中，我们究竟要证明甚么?

假设 $(*)$ 对某个 $k \geq 1$ 成立，即

$$F_0 F_1 \cdots F_{k-1} = F_k - 2,$$

而我们想证明它对 $k+1$ 也成立：

$$F_0 F_1 \cdots F_{k-1} F_k = F_{k+1} - 2.$$

把假设的等式两边同乘以 F_k，再运用 (†)：

$$F_0 F_1 \cdots F_{k-1} F_k = (F_k - 2)\,F_k = F_{k+1} - 2.$$

我们已经验证了 $k = 1$ 的情况，所以根据归纳法，$(*)$ 对所有 $k \geq 1$ 都成立。 ■

想一想 我们能否完全不用归纳法?

可以，而且值得看看怎样做。由 $F_0 - 2 = 1$ 出发，反复运用 (†)：

$$F_k - 2 = (F_{k-1} - 2)F_{k-1} = (F_{k-2} - 2)F_{k-2}F_{k-1} = \cdots = (F_0 - 2)\,F_0 F_1 \cdots F_{k-1}.$$

由于 $F_0 - 2 = 1$，证明便完成了。这种“逐层展开”的做法，其实就是把同一个归纳法完整地写出来，不过有些人会觉得这样更容易看明白。

阶段 4　没有公因数

我们想证明：若 $m \neq n$，则 F_m 与 F_n 没有大于 1 的公因数。

> **想一想**　假设某个数 d 同时整除 F_m 和 F_n，其中 $m < n$。F_m 在 $(*)$ 的哪个位置出现？

在 $(*)$ 中取 $k = n$：

$$F_0 F_1 \cdots F_m \cdots F_{n-1} = F_n - 2.$$

由于 $m < n$，F_m 是左边其中一个因数。所以若 d 整除 F_m，则 d 整除整个左边；也就是说，d 整除 $F_n - 2$。

> **想一想**　现在 d 既整除 F_n，又整除 $F_n - 2$。它还必定整除甚么？

它们的差：d 整除 $F_n - (F_n - 2) = 2$。所以 $d = 1$ 或 $d = 2$。

> **想一想**　d 可以是 2 吗？

不可以：每个费马数都是 2 的幂加 1，所以是奇数。因此 $d = 1$，即任何两个不同的费马数都没有大于 1 的公因数。 ■

数学家会说费马数*两两互质*（pairwise coprime）。

阶段 5　无穷多个质数

现在回到我们一开始提出的问题。

> **想一想**　每个大于 1 的整数都至少有一个质因数（prime factor）。从每个费马数中各选一个质因数。对于你选出的这些质数，你能说些甚么？

设 p_n 为任意一个整除 F_n 的质数。假如其中两个相等，例如 $p_m = p_n$ 而 $m \neq n$，那么这个质数就是 F_m 与 F_n 的一个大于 1 的公因数，但我们刚刚证明了这是不可能的。所以质数

$$p_0,\ p_1,\ p_2,\ p_3,\ \ldots$$

全都*互不相同*。费马数有无穷多个，所以不同的质数也有无穷多个。 ■

留意我们从来不需要知道每个费马数是被*哪一个*质数整除，甚至不需要知道费马数本身是不是质数。我们只需要知道它们永远没有公因数。每个费马数都必须“用掉”至少一个其他费马数不能用的质数。

n	$F_n = 2^{2^n} + 1$	质因数分解
0	3	质数
1	5	质数
2	17	质数
3	257	质数
4	65 537	质数
5	4 294 967 297	$641 \times 6\,700\,417$
6	18 446 744 073 709 551 617	$274\,177 \times 67\,280\,421\,310\,721$

图 1.1.1 最初几个费马数。费马猜想它们全是质数，但欧拉（Euler）在 1732 年找到了 F_5 的因数 641，而 F_4 之后至今仍未发现任何费马数是质数。即使如此，恒等式 $(*)$ 表明不同费马数的因数永不重叠：每一行都带来新的质数。

回顾

值得停下来看看这道题成功的关键何在，因为同样的想法会在本书中多次重现。

- **例子并不是装饰。**看出 $15 \times 17 = 255$ 其实是“$(F_2 - 2) \times F_2 = F_3 - 2$”，正是关键思路。当题目给你一些简单情况时，要问每一步在做甚么，而不只是看数字是否吻合。
- **一条恒等式完成了所有工作。**一切都源自 $(a-1)(a+1) = a^2 - 1$。看似艰深的 STEP 题目，往往只是由一小段简单的代数推动。
- **最后的论证着眼于结构，而非大小。**我们证明了质数有无穷多个，却没有找出任何一个新的质数。我们证明的是：要构成这个无穷的数族，必须用上无穷多个质数。

背后的数学

费马猜对了吗？

没有。1732 年，莱昂哈德．欧拉证明了

$$F_5 = 2^{32} + 1 = 4\,294\,967\,297 = 641 \times 6\,700\,417.$$

欧拉并不是靠运气找到 641 的。他证明了 F_n 的任何质因数都必定具有一种非常特殊的形式（当 $n = 5$ 时，它必定是 64 的倍数加一），从而大幅减少了需要考虑的可能。你可以在练习 4 中亲自证明这一点。

自此以后，凡是有人成功检验过的费马数，由 F_5 到 F_{32}，结果全都是合成数（composite）。就目前所知，F_0 至 F_4 可能是仅有的费马质数，但没有人能够证明这一点。这是一个货

真价实的未解难题。

妙就妙在我们的证明完全不受影响。不论是质数还是合成数，费马数依然能产生无穷多个质数。

费马质数与正多边形作图

费马质数会在一个相当意想不到的地方出现。古希腊人懂得用直尺和圆规作出等边三角形和正五边形，而请留意 $3 = F_0$ 和 $5 = F_1$。其后两千年间，没有人能作出其他边数为质数的正多边形（regular polygon）。到了 1796 年，年仅 19 岁的卡尔．弗里德里希．高斯（Gauss）示范了怎样作出*正十七边形*，而 $17 = F_2$。

高斯更进一步：正 N 边形可以用直尺和圆规作出，当且仅当 N 是 2 的幂乘以若干个*互不相同的费马质数*。（Pierre Wantzel 在 1837 年补全了这个证明。）所以哪些正多边形可以作图，取决于费马质数有多少个。你会在 Section 6.2（STEP II 1991，第 4 题）遇到高斯正十七边形背后的关键计算，而完整的来龙去脉则属于*伽罗瓦理论*（Galois theory），那是大学本科数学课程的精华之一。

一个定理，多种证明

数学家为甚么要不厌其烦地把同一个定理证明多于一次？因为不同的证明会揭示不同的东西。这个证明一般归功于克里斯蒂安．哥德巴赫（Goldbach）在 1730 年写给欧拉的一封信，它表明：*任何一个由大于 1 的整数组成、各项两两互质的无穷数列，都必然带来无穷多个质数*。费马数只是其中一个方便的例子。（练习 3 给出另一个例子。）

它给出的结论也比欧几里得的证明更多。质数 $p_0, p_1, \ldots, p_{n-1}$ 是 n *个不同的质数*，每个都不超过 F_{n-1}。所以不大于 $2^{2^{n-1}} + 1$ 的质数至少有 n 个。这是一个关于质数如何分布的*定量陈述*。它非常弱（实际情况要好得多），但它是迈向十九世纪数学其中一项伟大成果的第一步，那就是*质数定理*（Prime Number Theorem），它精确地描述了质数怎样变得越来越稀疏。Section 1.2 会踏出下一步。

思维技巧

从结构性质出发建立论证

要证明 A 类对象有无穷多个，可以找出无穷多个 B 类对象，而每个 B 类对象都需要一个专属于自己的 A 类对象。

这里 A = 质数，B = 费马数，而“需要一个专属于自己的”是指“有一个不与

其他任何一个共用的质因数”。同样的模式也出现在计数论证、线性代数（linear algebra）（无穷多个线性独立的向量必然导致无穷维）和分析（analysis）之中。

你也练习了*归纳法*，尤其是养成追问“*由 k 到 $k+1$ 的那一步实际上做了甚么？*”的习惯。这个问题往往比形式上的范本更有用。

延伸练习

1 **欧几里得的版本。**写出欧几里得对“质数有无穷多个”的证明。它在甚么意义上是“反证法”（by contradiction），而费马数的证明又在甚么意义上是“构造性”（constructive）的？

2 **末位数字。**证明当 $n \geq 2$ 时，费马数 F_n 的末位数字是 7。

3 **西尔维斯特（Sylvester）数列。**定义 $a_1 = 2$ 及 $a_{n+1} = a_n^2 - a_n + 1$，于是数列的开头几项是 $2, 3, 7, 43, 1807, \ldots$

(a) 证明 $a_{n+1} - 1 = a_1 a_2 \cdots a_n$。

(b) 由此推导出各项两两互质，从而给出“质数有无穷多个”的另一个证明。

(c) 较难 证明 $\dfrac{1}{a_1} + \dfrac{1}{a_2} + \cdots + \dfrac{1}{a_n} = 1 - \dfrac{1}{a_{n+1} - 1}$。

4 **欧拉的妙着。***本题需要用到费马小定理（Fermat's Little Theorem）：若 p 是质数且 $p \nmid a$，则 $a^{p-1} \equiv 1 \pmod{p}$。*假设质数 p 整除 F_n。

(a) 证明 $2^{2^n} \equiv -1 \pmod{p}$，从而证明 $2^{2^{n+1}} \equiv 1 \pmod{p}$。

(b) 设 d 为满足 $2^d \equiv 1 \pmod{p}$ 的最小正整数。证明 d 整除 2^{n+1} 但不整除 2^n，并由此得出 $d = 2^{n+1}$。

(c) 利用费马小定理，证明 d 整除 $p - 1$。由此推导出 $p \equiv 1 \pmod{2^{n+1}}$。

(d) 当 $n = 5$ 时，这表示 F_5 的任何质因数都具有 $64k + 1$ 的形式。验证 641 确实如此。小于 641 而又具有这种形式的质数有多少个？

5 **质数的一个上界。**利用*一个定理，多种证明*中的论证，证明第 n 个质数不超过 $2^{2^{n-1}} + 1$。当 $n = 5$ 时，这个上界有多准确？

延伸阅读

▶ **Martin Aigner & Günter M. Ziegler, *Proofs from THE BOOK***（Springer），第 1 章“Six proofs of the infinity of primes”（质数无穷性的六个证明）。书中除了这个证明外，还

收录了欧几里得的证明和另外四个证明，其中一个用上了拓扑学（topology）。这是入读大学前最值得一读的书之一。

- **Harold Davenport, *The Higher Arithmetic*** （Cambridge University Press），第 1–2 章。一本文笔优美的入门书，介绍质数与因数分解，最适合在入读大学前的暑假阅读。
- **Michal Křížek, Florian Luca & Lawrence Somer, *17 Lectures on Fermat Numbers*** （Springer，2001）。关于费马数的一切，由欧拉的分解到高斯的正多边形，应有尽有。不妨翻翻开首几章。
- **Ian Stewart, *Why Beauty Is Truth*** （Basic Books），关于高斯的一章。以一般读者为对象，讲述正十七边形的故事。

1.2 质数分布得有多稀疏?

题目

STEP III 2016，第 5 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/16-s3-q5`

引子

在 Section 1.1 中，我们已看到质数（prime）有无穷多个。但“无穷多个”是个很粗略的说法。质数确实愈来愈稀疏：100 以下有 25 个质数，但 900 与 1000 之间只有 14 个。它们究竟会稀疏到甚么程度?

这道题要证明的结论听来相当平凡：把不超过 n 的所有质数相乘，所得的积小于 4^n。然而，这个看似平平无奇的不等式（inequality），正是回答“有多稀疏?”的第一步实质进展，也是数学史上最著名的简短证明之一的关键引理（lemma）；该证明出自年轻的保罗·艾狄胥（Paul Erdős）之手，时为 1932 年。

动笔之前

记号 $P_{r,s}$ 看来有点繁琐，但背后的意思很简单：把某个范围内的质数相乘。务必先弄懂题中的例子。$P_{3,7}$ 用的是大于 3 而不超过 7 的质数，即 5 和 7，所以它等于 35。

> **想一想** 第 (iv) 部分要求你对所有 $n \geq 2$ 证明一个命题。第 (iii) 部分则假设该命题对不超过 $2m$ 的每一个 k 都成立。这是哪一种归纳法（induction）?

这是强归纳法（strong induction）：要证明 n 的情况，你可以假设前面所有情况都成立，而不仅是紧接在前的那一个。请把这一点记在心里；题目之所以这样编排，原因正在于此。

逐步解题

阶段 1　二项式系数不会太大

想一想　在 $(1+x)^{2m+1}$ 的二项展开式（binomial expansion）中代入 $x=1$，会得到甚么？

你会得到

$$2^{2m+1} = \binom{2m+1}{0} + \binom{2m+1}{1} + \cdots + \binom{2m+1}{2m+1}.$$

各项都是正数，所以每一项都小于 2^{2m+1}。这还不够好；我们想要的是 2^{2m}，即它的一半。

想一想　帕斯卡三角形（Pascal's triangle）是对称的。和式中还有哪一项等于 $\binom{2m+1}{m}$？

由对称性，$\binom{2m+1}{m} = \binom{2m+1}{m+1}$，而这是和式中两个不同的项（因为 $m \neq m+1$）。所以

$$2\binom{2m+1}{m} = \binom{2m+1}{m} + \binom{2m+1}{m+1} < 2^{2m+1},$$

这里是严格不等式，因为和式中还有其他正项，例如 $\binom{2m+1}{0} = 1$。两边除以 2，即得 $\binom{2m+1}{m} < 2^{2m}$。 ■

阶段 2　哪些质数整除这个二项式系数？

把这个二项式系数（binomial coefficient）写出来：

$$\binom{2m+1}{m} = \frac{(2m+1)!}{m!\,(m+1)!} = \frac{(m+2)(m+3)\cdots(2m+1)}{m!}.$$

想一想　设 p 为质数，且 $m+1 < p \le 2m+1$。p 整除分子吗？它有可能整除分母吗？

分子 $(m+2)(m+3)\cdots(2m+1)$ 的因子当中就有 p 本身，所以 p 整除分子。分母 $m!$ 是若干个数的积，而这些数全都小于 p；由于 p 是质数，如果它不能整除其中任何一个数，就不能整除它们的积。因此 p 整除分子而不整除分母，于是 p 整除整个二项式系数（别忘记，它是一个整数）。

停一停，试一试　我们现在知道，这个范围内的每一个质数都整除 $\binom{2m+1}{m}$。为甚么由此可知它们的积 $P_{m+1,2m+1}$ 也整除它？（小心：即使 4 和 6 都整除某个数，24 也不一定整除它。）

分别在于，我们的质数是*相异质数*。如果相异质数 $p_1, p_2, \ldots$ 各自整除 N，那么它们的积也整除 N：写 $N = p_1 N_1$；这时 p_2 整除 $p_1 N_1$ 但不整除 p_1，所以它整除 N_1；如此类推。（这里用到一个事实：如果一个质数整除某个积，它必整除其中一个因子。这个事实值得我们稍加敬重；Section 1.7 会展示一个世界，在那里与它相应的命题并不成立。）

所以 $P_{m+1,2m+1}$ 整除 $\binom{2m+1}{m}$，而正数的因数不会大于该数本身，故

$$P_{m+1,\,2m+1} \le \binom{2m+1}{m} < 2^{2m}. \qquad \blacksquare$$

阶段 3 把积拆开

想一想 $P_{1,2m+1}$ 是不超过 $2m+1$ 的所有质数的积，怎样用我们已经掌握的积把它拼出来?

把不超过 $2m+1$ 的质数分成两批：不超过 $m+1$ 的，以及大于 $m+1$ 的：

$$P_{1,2m+1} = P_{1,m+1} \times P_{m+1,2m+1}.$$

由第 2 阶段，第二个因子小于 $2^{2m} = 4^m$。至于第一个因子，我们可以假设 $P_{1,k} < 4^k$ 对由 2 至 $2m$ 的每一个 k 都成立，而 $k = m+1$ 正在这个范围之内（因为由 $m \ge 1$ 可得 $2 \le m+1 \le 2m$）。所以

$$P_{1,2m+1} < 4^{m+1} \times 4^m = 4^{2m+1}. \qquad \blacksquare$$

强归纳法正是在这里大派用场：要处理 $2m+1$，我们需要的是 $m+1$ 的情况，它的大小约为一半，而不是紧接在前的那个情况。

阶段 4 所有 n

想一想 第 (iii) 部分处理了奇数。对于偶数 $n = 2m+2$，情况又如何？它是质数吗?

先验证起始情况：$P_{1,2} = 2 < 16$ 及 $P_{1,3} = 6 < 64$。现在假设 $P_{1,k} < 4^k$ 对由 2 至 $n-1$ 的所有 k 都成立，其中 $n \ge 4$。

- 若 n 是偶数，则 n 不是质数（因为它大于 2），所以不超过 n 的质数与不超过 $n-1$ 的质数完全相同。因此 $P_{1,n} = P_{1,n-1} < 4^{n-1} < 4^n$。
- 若 n 是奇数，写 $n = 2m+1$，其中 $m \ge 2$。这时由第 (iii) 部分直接得出 $P_{1,n} < 4^n$。

由强归纳法，$P_{1,n} < 4^n$ 对每一个 $n \ge 2$ 都成立。 $\blacksquare$

回顾

- **二项式系数蕴藏算术信息。**$\binom{2m+1}{m}$ 是单独一个数，它不大（小于 4^m），却必须被许多质数整除。把许多质数挤进一个不大的数里，就说明这些质数不可能太多。
- **当问题会把自己减半时，强归纳法是最自然的工具。**$2m+1$ 的情况取决于 $m+1$ 的情况。
- **小的情况须另行验证。**归纳要到 $n \geq 4$ 才能开始运作，所以 $n = 2$ 和 3 必须亲手计算。

背后的数学

从 4^n 到质数定理

对结果取对数：$\ln P_{1,n} < n \ln 4$，即

$$\sum_{p \leq n} \ln p < (\ln 4)\, n \approx 1.39\, n.$$

在 1850 年代，切比雪夫（Pafnuty Chebyshev）证明了正是这一类的界（bound），而且上界和下界俱备，从而说明 $\sum_{p \leq n} \ln p$ 的增长与 n 成正比。高斯（Gauss）在大约 15 岁时，已从数表中猜到不超过 n 的质数个数大约是 $n / \ln n$。这后来成为*质数定理*（Prime Number Theorem），最终在 1896 年由阿达马（Jacques Hadamard）和德拉瓦莱普桑（Charles de la Vallée Poussin）运用复分析（complex analysis）证明。它的对数形式是

$$\sum_{p \leq n} \ln p \sim n,$$

所以不超过 n 的质数之积大约是 e^n。我们的界 4^n 其实相差不远：图 1.2.1 把三者一并画出。

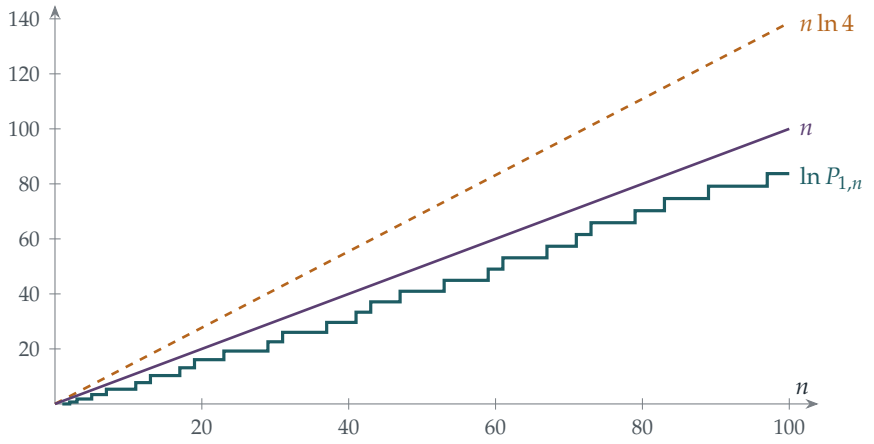


图 1.2.1 不超过 n 的所有质数之积的对数（蓝绿色），在每个质数 p 处跳升 $\ln p$。本节证明它始终低于 $n \ln 4$（琥珀色虚线）；质数定理则指出它的增长与 n 相若（紫红色）。

伯特兰公设

1845 年，伯特兰（Joseph Bertrand）猜想：对每一个 $n \geq 1$，在 n 与 $2n$ 之间总有一个质数。这就是伯特兰公设（Bertrand's postulate）。切比雪夫在 1852 年证明了它。八十年后，年仅 19 岁的艾狄胥找到了一个优美的初等证明。它的关键材料，就是本题的引理，再加上仔细考察哪些质数整除中间的二项式系数 $\binom{2n}{n}$。Martin Aigner 和 Günter Ziegler 把它选为 *Proofs from THE BOOK* 的第 2 章，该书汇集了数学中最优雅的证明。

接下来的方向

质数的精确分布，由黎曼 ζ 函数（Riemann zeta function）

$$\zeta(s) = 1 + \frac{1}{2^s} + \frac{1}{3^s} + \cdots,$$

的零点（zero）所支配。你会在 Section 2.4 中遇见这个函数朴素得多的一面，那里有 $\zeta(2) = \pi^2/6$。黎曼猜想（Riemann Hypothesis）预测了这些零点的确切位置；它若成立，便能把质数定理中的误差限定得紧无可紧。它也许是数学中最著名的未解难题，也是千禧年大奖难题（Millennium Prize Problems）之一。

思维技巧

强归纳法，以及从巧妙的对象中提取信息

如果 n 的情况取决于远小于 $n-1$ 的情况，就一次过假设前面所有情况都成立。

这里的另一个想法较为微妙，也极具数论（number theory）的特色：要证明某个范围内所有质数的某种性质，就找出一个对象（这里是一个二项式系数），使得这些质数每一个都必须整除它，而你又能证明它不会太大。

延伸练习

1 **感受一下数字。**对 $n = 10, 20, 30$ 计算 $P_{1,n}$，并把它与 4^n 及 e^n 比较。（这里可以使用计算机。）

2 **中间的二项式系数。**证明 $\binom{2m}{m}$ 是 $(1+1)^{2m}$ 展开式的 $2m+1$ 项中最大的一项，并由此推出 $\binom{2m}{m} \geq \dfrac{4^m}{2m+1}$。

3 **质数不会太多。**设 $\pi(n)$ 为不超过 n 的质数个数。介乎 $\sqrt{n}$ 与 n 之间的每个质数都大于 $\sqrt{n}$，而它们的积至多为 $P_{1,n} < 4^n$。由此推出

$$\pi(n) - \pi(\sqrt{n}) < \frac{2\ln 4}{\ln n}\, n,$$

从而得出 $\pi(n) < 2.8\, n/\ln n + \sqrt{n}$。把这个结果与质数定理的 $\pi(n) \approx n/\ln n$ 比较。

4 **验证伯特兰公设。**对 $n = 1, 2, \ldots, 50$ 验证伯特兰公设，即在每个情况下找出一个介乎 n 与 $2n$ 之间的质数。你能否利用一张简短的质数表有效率地完成，使表中每个质数都小于前一个的两倍？

延伸阅读

- **Aigner & Ziegler, *Proofs from THE BOOK***，第 2 章“Bertrand's postulate”（伯特兰公设）。
- **Marcus du Sautoy, *The Music of the Primes***（Fourth Estate）。一部讲述质数定理与黎曼猜想历史的普及读物。
- **G. J. O. Jameson, *The Prime Number Theorem***（Cambridge University Press）。严谨而平易近人，为大学本科生而写；开首几章只需要高级程度（A-level）的微积分。

1.3 最小反例

题目

STEP III 2015，第 5 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/15-s3-q5`

引子

你大概见过 $\sqrt{2}$ 是无理数（irrational）的证明：假设 $\sqrt{2} = a/b$ 为最简分数（lowest terms），两边平方，推出 a 是偶数，继而推出 b 也是偶数，于是得出矛盾（contradiction）。这是个漂亮的证明，但它依赖一个关于偶数的技巧，而这个技巧看不出可以怎样推广。换成 $\sqrt[3]{2}$ 又如何？$\sqrt{7}$ 呢？

本题给出另一个证明，它的核心想法简单得近乎荒谬：看看最小的那一个。在本章里，同一个想法还会改头换面地出现好几次。

动笔之前

这是一道不寻常的 STEP 题目：它给你一个证明的骨架，要你补上当中欠缺的理据。大学生正是这样学习阅读数学的。课本里的证明常常说“显然”或“由此可得”，而你的部分工作，就是查证事实是否真的如此。

想一想 第 4 步说“定义 k 为 S 中最小的正整数”。我们凭甚么可以这样做？假如 S 是一个由正实数组成的集合，这样做还可以吗？

任何由正整数组成的非空集合都有最小元素。这称为良序原理（well-ordering principle），它与数学归纳法（mathematical induction）原理等价。对于由正实数组成的集合，这个原理并不成立：集合 $\{x : x > 0\}$ 就没有最小元素。所以第 4 步既依赖第 3 步（集合必须非空），也依赖 S 由整数组成这一点。

逐步解题

阶段 1　第 3 步：S 非空

想一想　若 $\sqrt{2} = a/b$，其中 a 和 b 是正整数，你能写出 S 的一个元素吗？

若 $\sqrt{2} = a/b$，则 $b\sqrt{2} = a$，是一个整数。所以 b 属于 S，S 非空。

阶段 2　第 5 步：一个更小的元素

现在 k 是 S 的最小元素，所以 $k\sqrt{2}$ 是整数，记为 j。

停一停，试一试　要证明 $(\sqrt{2}-1)k$ 属于 S，你必须检查三件事。是哪三件？

我们需要 $(\sqrt{2}-1)k$ 满足：(a) 是整数；(b) 是正数；(c) 乘以 $\sqrt{2}$ 后得到整数。

- (a) $(\sqrt{2}-1)k = k\sqrt{2} - k = j - k$，是两个整数之差。
- (b) $\sqrt{2} > 1$，所以 $\sqrt{2} - 1 > 0$，因此 $(\sqrt{2}-1)k > 0$。
- (c) $(\sqrt{2}-1)k \cdot \sqrt{2} = 2k - k\sqrt{2} = 2k - j$，是整数。

所以 $(\sqrt{2}-1)k \in S$。

阶段 3　第 6 步：矛盾

想一想　$\sqrt{2} - 1$ 有多大？

由于 $1 < \sqrt{2} < 2$，我们有 $0 < \sqrt{2} - 1 < 1$，所以 $(\sqrt{2}-1)k < k$。我们找到了 S 的一个元素，它比 S 的最小元素 k 还要小。这是不可能的，所以第 1 步的假设不成立：$\sqrt{2}$ 是无理数。 ■

留意我们从头到尾都不需要“最简分数”，也不需要任何关于奇偶数的事实。整个证明只建基于三个事实：$\sqrt{2} \cdot \sqrt{2}$ 是整数；$\sqrt{2}$ 严格地介乎两个连续整数之间；整数具有良序性。

阶段 4　立方根：把两个数联系起来

设 $c = 2^{1/3}$，则 $c^2 = 2^{2/3}$ 且 $c^3 = 2$。

想一想　若 c 是有理数（rational），c^2 是否也是？反过来，若 c^2 是有理数，你能否只用乘法和除法回到 c？

若 c 是有理数，则 $c^2 = c \times c$ 也是有理数。反之，若 c^2 是有理数，则 $2/c^2 = c^3/c^2 = c$

也是有理数。所以 $2^{1/3}$ 是有理数，当且仅当 $2^{2/3}$ 是有理数。 ■

阶段 5 改造论证

想一想 在第 (i) 部分，关键在于 $(\sqrt{2})^2 = 2$ 是整数，所以把 S 的元素乘以 $\sqrt{2}$ 两次，就会回到整数。对于 $c = 2^{1/3}$，你需要乘三次。集合 S 应该怎样修改?

假设 c 是有理数，则 c^2 也是。定义

$$S = \{n \text{ 为正整数} : nc \text{ 和} nc^2 \text{ 都是整数}\}.$$

- **S 非空。**若 $c = a/b$ 且 $c^2 = a'/b'$，则 bb' 属于 S。
- **设 k 为它的最小元素**，并考虑 $(c-1)k$。我们有 $1 < c < 2$（因为 $1^3 < 2 < 2^3$），所以 $0 < (c-1)k < k$。
- **它是整数：**$(c-1)k = kc - k$。
- **乘以 c：**$(c-1)k \cdot c = kc^2 - kc$，是整数。
- **乘以 c^2：**$(c-1)k \cdot c^2 = kc^3 - kc^2 = 2k - kc^2$，是整数。

所以 $(c-1)k \in S$，而且比 k 小：矛盾。因此 $2^{1/3}$ 是无理数，而根据第 4 阶段，$2^{2/3}$ 也是无理数。 ■

回顾

- **“取最小的那一个”是一种证明技巧。**假设反例存在，取其中最小的一个，再由它造出一个更小的。这称为*极端原理*（extremal principle），或称*最小反例*（minimal counterexample）论证。
- **集合 S 替我们记帐。**我们不必周旋于分子和分母之间，而是把所有“好”的乘数收进同一个集合。
- **推广，就是追问证明真正行得通的原因。**对于 $\sqrt{2}$，需要乘以 $\sqrt{2}$ 一次；对于 $\sqrt[3]{2}$，则需要乘两次。看出这一点，我们就清楚知道该怎样重新定义 S。

背后的数学

一个画得出来的证明

第 (i) 部分背后有一幅优美的图，出自 Stanley Tennenbaum。若 $\sqrt{2} = a/b$，则 $a^2 = 2b^2$：一个边长为 a 的正方形，面积等于两个边长为 b 的正方形面积之和。如图 1.3.1 所示，把两个边长为 b 的正方形放在边长为 a 的正方形的两个对角。它们重叠之处是一个小正方形，另外还有两个小正方形没有被盖住。由于面积相等，重叠部分必定等于两个未被盖住的正方形之和，于是得到一对更小而具有同样性质的整数。整数对不可能无止境地愈变愈小，所以这样的正方形并不存在。这对较小的整数正是 $(2b - a,\ a - b)$，也就是我们那一步 $(\sqrt{2} - 1)k$ 的化身。

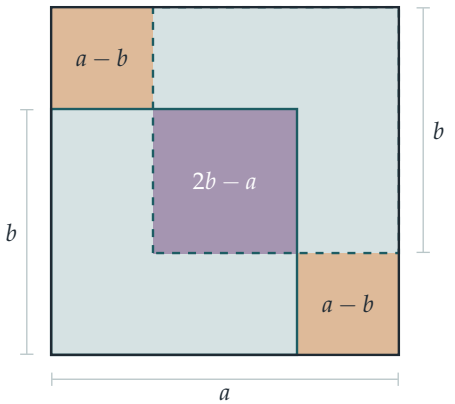


图 1.3.1　Tennenbaum 的证明。两个边长为 b 的正方形（蓝绿色）放在边长为 a 的正方形的两个对角。若 $a^2 = 2b^2$，则被计算了两次的面积（紫红色的重叠部分）必定等于没有计算到的面积（两个琥珀色的角落），所以 $(2b - a)^2 = 2(a - b)^2$：一个更小的解。

良序、归纳与递降

良序原理与归纳法原理在逻辑上等价：两者可以互相推导。到了大学，你会看到两者都被用作自然数的基础。在 Section 1.5 中，我们会以另一种形式运用同一个原理，那就是费马（Fermat）的*无穷递降法*（method of infinite descent）：不存在由正整数组成的无穷严格递减数列。

它能走多远？

这个论证可以证明，只要 N 不是完全平方数（perfect square），$\sqrt{N}$ 就是无理数（见练习）；更一般地，一个整数的 n 次方根要么是整数，要么是无理数。到了大学，这会变成一个关于*代数整数*（algebraic integers）的命题：若一个有理数满足某个整数系数的首一多项式（monic polynomial）方程，例如 $x^3 - 2 = 0$，它就必定是整数。你会在 Section 6.3 再次遇到这个想法。

相传毕达哥拉斯学派（Pythagoreans）因为希帕索斯（Hippasus）泄露了 $\sqrt{2}$ 是无理数而把他淹死，这个传说多半并不属实。属实的是，这个发现迫使希腊数学家把对数的研究与对长度的研究分开；这道裂痕要等到十九世纪，待戴德金（Dedekind）构造出实数，才得以弥合。

思维技巧

极端原理

如果你想证明某件事永远不会发生，就假设它会发生，选出最小（或最大、或最简单）的例子，再用它造出一个更小的例子。

决定“最小”是甚么意思，正是需要创意的一步。在这里，它是使 $n\sqrt{2}$ 成为整数的最小的 n。

延伸练习

1. **其他平方根。**改造这个论证，证明 $\sqrt{3}$ 是无理数。哪一个数担当了 $\sqrt{2}-1$ 的角色？
2. **所有非平方数。**设 N 为一个不是完全平方数的正整数，并设 m 为满足 $m < \sqrt{N} < m+1$ 的整数。考虑 $(\sqrt{N} - m)k$，从而证明 $\sqrt{N}$ 是无理数。
3. **找出漏洞。**试把这个论证套用于 $\sqrt{4}$。究竟是哪一步行不通？
4. **Tennenbaum 的图。**利用图 1.3.1 证明：若 $a^2 = 2b^2$，则 $(2b-a)^2 = 2(a-b)^2$。再用代数方法验证这个恒等式。

延伸阅读

- **Daniel Velleman, *How to Prove It***（Cambridge University Press）。一本全面的证明入门书，内容包括归纳法和良序原理。

- **R. P. Burn, *A Pathway into Number Theory***（Cambridge University Press）。全书由精心编排次序的问题组成，是修读大学一年级课程前的理想准备。
- **John Conway & Joseph Shipman, "Extreme proofs I: the irrationality of $\sqrt{2}$"**（*Mathematical Intelligencer*，2013）。同一个事实的多个证明，比较得妙趣横生。

1.4 e 是无理数，e^2 也是

本节会用到两道题目。第一道出自 1997 年，证明 e 是无理数（irrational number）。第二道出自 2024 年，把同一个想法用得更巧妙，证明 e^2 是无理数，而这原来确实困难得多。

热身题

STEP III 1997，第 7 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/97-s3-q7`

主题目

STEP III 2024，第 7 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2024 年 STEP 试卷（zip 文件）内的 STEP III 试卷：

`www.ocr.org.uk/Images/725753-2024-past-papers.zip`

引子

在 Section 1.3 中，我们证明了 $\sqrt{2}$ 是无理数，用的是 $\sqrt{2}$ 满足一条简单方程 $x^2 = 2$ 这个事实。但就我们所知，e 并不满足任何这样的方程。我们手上只有它的级数（series）：

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots .$$

一个级数怎能告诉你某个数不是分数?

答案是数论（number theory）中最有用的想法之一：整数（*integer*）不可能严格介乎 0 与 1 之间。如果你能证明某个量是整数，同时又严格介乎 0 与 1 之间，就得出矛盾（contradiction）。本节的一切，都是这个主题的变奏。

动笔之前

想一想　假设 $e = p/q$。把 e 乘以 q 会得到整数。那么乘以 $q!$ 又会对 e 的级数的每一项有甚么影响?

只要 $k \le q$，把 $\dfrac{1}{k!}$ 乘以 $q!$ 都会得到整数。因此 $q!\,e$ 可分成整数部分（首 $q+1$ 项）和“尾部”（tail，即其后的所有项）。若 $e = p/q$，则 $q!\,e$ 本身是整数，所以尾部也必定是整数。我们的计划，就是证明尾部严格介乎 0 与 1 之间。

逐步解题

阶段 1　热身题：e 是无理数

级数 b_n 是等比级数（geometric series），首项和公比都是 $\frac{1}{n+1}$，所以

$$b_n = \frac{1/(n+1)}{1 - 1/(n+1)} = \frac{1}{n}.$$

想一想　逐项比较 a_n 和 b_n。哪一个较大?

两者的首项相等。其后，$(n+1)(n+2) > (n+1)^2$，而一般来说，当 $k \ge 2$ 时 $(n+1)(n+2)\cdots(n+k) > (n+1)^k$，所以 a_n 其后的每一项都小于 b_n 的对应项。因此 $0 < a_n < b_n = 1/n$。

现在看看 a_n 究竟是甚么。把 e 的级数乘以 $n!$：

$$n!\,e = \underbrace{\frac{n!}{0!} + \frac{n!}{1!} + \cdots + \frac{n!}{n!}}_{\text{整数}} + \underbrace{\frac{n!}{(n+1)!} + \frac{n!}{(n+2)!} + \cdots}_{=\,a_n}.$$

由于 $0 < a_n < 1$，$n!\,e$ 的整数部分就是第一个括号，而 a_n 就是小数部分（fractional part）：$a_n = n!\,e - [n!\,e]$。

停一停，试一试　现在完成证明。假设 $e = p/q$。你应该选取哪个 n 值?

取 $n = q$。那么 $q!\,e = (q-1)!\,p$ 是整数，所以它的小数部分 a_q 是 0。但 $a_q > 0$。矛盾：e 是无理数。 ∎

这个证明出自傅里叶（Joseph Fourier），约于 1815 年提出。

阶段 2　为何 e^2 较难

想一想 试把同一个想法用于 e^2。它的级数是 $\sum 2^k/k!$。当你乘以 $n!$ 再看尾部时，哪里出了问题?

尾部现在是 $\dfrac{2^{n+1}}{n+1}+\dfrac{2^{n+2}}{(n+1)(n+2)}+\cdots$，而 2 的幂把一切都破坏了：尾部不再细小。乘以 $n!$ 对分子中的 2^k 毫无帮助。

2024 年这道题的妙着，是完全避开 e^2。若 $e^2=p/q$，则

$$qe=\frac{p}{e}.$$

现在我们有两个数 e 和 $1/e$，各自都有一个阶乘（factorial）级数，而且没有麻烦的 2 的幂：

$$e=\sum_{k=0}^{\infty}\frac{1}{k!},\qquad \frac{1}{e}=\sum_{k=0}^{\infty}\frac{(-1)^k}{k!}.$$

阶段 3 两个尾部

第 (i) 部分正正就是热身题：$0<f(n)<1/n$。

至于第 (ii) 部分，$g(n)$ 是一个各项的大小递减的交错级数（alternating series）。

想一想 对于交错级数 $t_1-t_2+t_3-\cdots$，其中 $t_1>t_2>t_3>\cdots>0$，为何它的和介乎 0 与 t_1 之间?

把各项两两配对：$(t_1-t_2)+(t_3-t_4)+\cdots$ 是若干个正括号之和，所以是正数。换另一种方式配对：$t_1-(t_2-t_3)-(t_4-t_5)-\cdots$ 是 t_1 减去若干个正括号，所以小于 t_1。这里 $t_1=\frac{1}{n+1}$,而各项递减,因为每一项都是前一项除以一个大于 1 的数。所以 $0<g(n)<\frac{1}{n+1}$。

阶段 4 又见整数

把 e 乘以 $(2n)!$，然后分出尾部，做法与热身题完全一样：

$$(2n)!\,e=(\text{整数})+f(2n).$$

所以 $(2n)!\,e-f(2n)$ 是整数。现在对 $1/e$ 做同样的事：

$$\frac{(2n)!}{e}=\sum_{k=0}^{2n}(-1)^k\frac{(2n)!}{k!}+\sum_{k=2n+1}^{\infty}(-1)^k\frac{(2n)!}{k!}.$$

想一想 尾部的首项（即 $k=2n+1$ 的一项）是甚么符号？尾部与 $g(2n)$ 相比又如何?

尾部首项的 $k = 2n+1$ 是奇数，所以该项是负数。尾部是

$$-\frac{1}{2n+1} + \frac{1}{(2n+1)(2n+2)} - \cdots = -g(2n).$$

所以 $\dfrac{(2n)!}{e} + g(2n)$ 是整数。 ■

这就是题目用 $2n$ 而不用 n 的原因：乘数是偶数的阶乘时，$1/e$ 的尾部恰好是 $-g$，符号可以预知。

阶段 5　矛盾

假设 $qe = p/e$。把等式两边乘以 $(2n)!$，并利用第 4 步：

$$q\big(\text{整数} + f(2n)\big) = p\big(\text{整数} - g(2n)\big).$$

移项整理后，$qf(2n) + pg(2n)$ 等于若干整数的组合，所以它是整数。这就是第 (iv) 部分。

> **停一停，试一试**　现在利用第 3 步的上下界。当 n 非常大时，$qf(2n) + pg(2n)$ 会怎样？

$f(2n)$ 和 $g(2n)$ 都是正数，所以 $qf(2n) + pg(2n) > 0$。而且

$$qf(2n) + pg(2n) < \frac{q}{2n} + \frac{p}{2n+1} < \frac{p+q}{2n}.$$

选取 $n > (p+q)/2$。那么 $qf(2n) + pg(2n)$ 就是一个严格介乎 0 与 1 之间的整数。这是不可能的，所以 e^2 是无理数。 ■

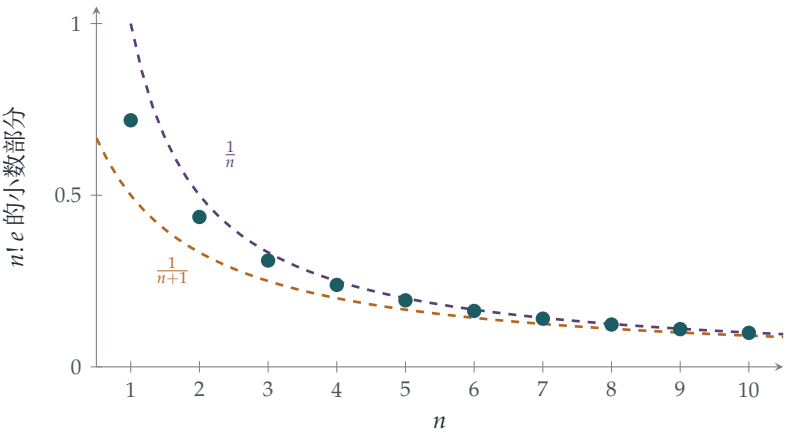


图 1.4.1　尾部 $a_n = n!\,e - [n!\,e]$（青绿色），即 $n!\,e$ 的小数部分。它总是严格介乎 $\frac{1}{n+1}$ 与 $\frac{1}{n}$ 之间：永不为零，但小于 1。假如 e 是分数 $\frac{p}{q}$，那么对每个 $n \geq q$，$n!\,e$ 都会是整数，青绿色的点就会是 0。

回顾

- **推动整个证明的是“没有整数严格介乎 0 与 1 之间”。**两个证明都构造出一个必定是整数的量，然后把它困住。
- **阶乘令尾部变小。**乘以 $n!$ 把级数的开首部分变成整数，剩下一个像 $1/n$ 那样缩小的尾部。
- **正面进攻失败时，就重新整理。**我们没有正面处理 $e^2 = p/q$，而是把它改写成 $qe = p/e$，把困难分摊到两个较易处理的数上。
- **细节很重要。**选用 $2n$ 而非 n，固定了尾部的符号。许多证明的成败，都取决于这样的一个小细节。

背后的数学

e 有多“无理”？

刘维尔（Joseph Liouville）在 1840 年证明了 e^2 是无理数，而且事实上它不满足任何整数系数的二次方程。1873 年，埃尔米特（Charles Hermite）走得远得多：他证明了 e 是*超越数*（transcendental number），意思是它不是任何整数系数多项式的根。他的证明正是上述论证的精密版本：构造一个整数，再证明它严格介乎 0 与 1 之间。九年后，林德曼（Ferdinand von Lindemann）改造了埃尔米特的方法，证明 π 是超越数，终于证实以尺规“化圆为方”这个古老问题是不可能做到的。

我们仍未知道的事

尽管这些方法威力强大，一些简单得令人尴尬的问题却仍未解决。$e + \pi$ 是无理数吗？$e\pi$ 呢？没有人知道，虽然两者之中至少有一个必定是超越数（练习 4）。$\zeta(3) = 1 + \frac{1}{8} + \frac{1}{27} + \cdots$ 是无理数吗？是的：阿培里（Roger Apéry）在 1978 年证明了这一点，震惊了整个数学界。$\zeta(5)$ 呢？没有人知道。

逼近

这些证明背后有一个更深刻的想法：如果一个数能被分数逼近得*太好*，它就是无理数。e 的级数的部分和（partial sum）都是分数，而且很快就非常接近 e；一个分数不可能被其他分数逼近得那么好。这就是*丢番图逼近*（Diophantine approximation）的起点。它是数论的一个分支，你可以在 Section 2.9 进一步探索，那里会用另一族积分得出更好的分数，步步逼近 e。

思维技巧

困住一个整数

要证明一个数是无理数，先假设它是分数，由此构造一个必定是整数的量，再证明这个量严格介乎 0 与 1 之间。

个中的艺术在于选择这个量。在这里，乘以阶乘把级数的大部分变成整数，剩下一个细小但为正数的尾部。

延伸练习

1 **只用 $1/e$。**假设 $1/e = p/q$，并只利用 $1/e$ 的交错级数，证明 e 是无理数。

2 $\sin 1$ **是无理数。**利用 $\sin 1 = 1 - \dfrac{1}{3!} + \dfrac{1}{5!} - \cdots$ 证明 $\sin 1$ 是无理数。（乘以一个合适的奇数的阶乘。）

3 **为何 π 不行？**解释为何这个方法不能直接用于 π。然后查阅 Ivan Niven 在 1947 年发表、只有一页的 π 是无理数的证明，当中用的是积分而不是级数。能力强的高级程度（A-level）学生应可读懂。

4 **其中一个是超越数。**较难 证明 e 和 π 是 $x^2 - (e+\pi)x + e\pi = 0$ 的根。你可以假设：若一个多项式的系数都是代数数（algebraic number），则它的根本身也是代数数；以及 π 不是代数数（林德曼）。由此推出 $e+\pi$ 和 $e\pi$ 之中至少有一个是超越数。

延伸阅读

- **Aigner & Ziegler, *Proofs from THE BOOK***，第 8 章“Some irrational numbers”（一些无理数）。正是用这些想法，证明 e、e^2、e^4 和 π^2 都是无理数。
- **Ivan Niven, *Irrational Numbers***（Mathematical Association of America）。一本篇幅不长的经典，内容由 $\sqrt{2}$ 一直谈到超越性。
- **Julian Havil, *The Irrationals***（Princeton University Press）。一本浅白易读的无理数历史。

1.5 无穷递降法

题目

STEP III 2022，第 2 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2022 年 STEP 试卷（zip 文件）内的 STEP III 试卷：

`www.ocr.org.uk/Images/696587-2022-past-papers.zip`

引子

怎样才能证明一条方程*没有*整数解？你不可能逐一试遍所有可能，因为可能性有无穷多个。费马（Pierre de Fermat）在 1630 年代找到了一个方法：证明任何一个解都会产生另一个数字*更小*的解，然后又产生一个更小的，如此无休止地继续下去。可是正整数不可能永无止境地愈变愈小。所以，从一开始就根本没有解。

费马把这个方法称为*无穷递降法*（method of infinite descent），并且深以为傲。它是 Section 1.3 中良序原理（well-ordering principle）的另一面，至今仍是证明方程无解的主要工具之一。

动笔之前

四个部分的格局都一样：假设有解，把解缩小，得出矛盾（contradiction）。问题在于*怎样*缩小。

> **想一想** 在第 (i) 部分，看看 $a^3 + 2b^3 + 4c^3 = 0$。三项之中有两项明显是偶数。这告诉你第三项有甚么性质？

每一次递降都是由这类观察开始的：某种整除性（divisibility）是你无法回避的。

逐步解题

阶段 1　第一次递降

由于 $a^3 = -2b^3 - 4c^3$，a^3 这个数是偶数，所以 a 是偶数（奇数的立方是奇数）。设 $a = 2p$。那么 $|p| = |a|/2 < |a|$，代入后得

$$8p^3 + 2b^3 + 4c^3 = 0 \quad \Longrightarrow \quad 4p^3 + b^3 + 2c^3 = 0.$$

想一想 新的方程是 $4p^3 + b^3 + 2c^3 = 0$。它的系数和原来一样，只是轮换了位置。你可以再用同一招吗？

现在 $b^3 = -4p^3 - 2c^3$ 是偶数，所以 $b = 2q$，其中 $|q| < |b|$，再除以 2 便得 $2p^3 + 4q^3 + c^3 = 0$。同样地，c 必定是偶数：$c = 2r$，其中 $|r| < |c|$，再除以 2，

$$p^3 + 2q^3 + 4r^3 = 0.$$

这正是原来的方程，只是以 p、q、r 取代了 a、b、c。每个数的大小都减半，而且没有一个是零，因为把非零的数减半不可能得到零。

停一停，试一试 把以上论证写成一个证明（proof），说明非零解并不存在。你究竟应该假设甚么是“最小”的？

在所有非零整数解之中，选取 $|a|$ 最小的一个。（只要有解存在，良序原理就容许我们这样做。）上述论证会产生另一个非零解，并且 $|p| < |a|$，与最小性矛盾。所以没有非零解。 ■

另一个等价的说法是：由一个解出发，我们可以构造出无穷多个解，它们满足 $|a| > |p| > \cdots$，即一个严格递减的无穷正整数数列，而这是不可能的。

阶段 2　换一个质数

想一想 在 $9a^3 + 10b^3 + 6c^3 = 0$ 中，奇偶性（parity）看来帮不上忙。看看系数 9、10、6。哪一个质数（prime）整除其中两个？

质数 3 整除 9 和 6，所以 3 整除 $10b^3$，因此 3 整除 b。设 $b = 3q$ 并除以 3：

$$3a^3 + 90q^3 + 2c^3 = 0.$$

现在 3 整除 $2c^3$，所以 $c = 3r$；再除以 3 得 $a^3 + 30q^3 + 18r^3 = 0$。于是 3 整除 a：$a = 3p$，再除以 3 一次，

$$9p^3 + 10q^3 + 6r^3 = 0.$$

这与原来的方程完全相同，只是换成了较小的非零数。根据同样的递降论证，方程没有非零解。 ■

阶段 3　模 3 下的平方数

想一想 每个整数都可写成 $3n$、$3n+1$ 或 $3n-1$ 的形式。它们的平方模（modulo）3 是甚么？

$(3n\pm1)^2 = 9n^2 \pm 6n + 1 = 3(3n^2 \pm 2n) + 1$，比 3 的倍数多一。而 $(3n)^2$ 是 3 的倍数。所以平方数除以 3 的余数（remainder）是 0 或 1，永远不会是 2。

现在如果 $a^2 + b^2$ 是 3 的倍数，余数的可能组合是 $0+0$、$0+1$、$1+0$ 和 $1+1$，分别得出 0、1、1、2。只有第一种是 3 的倍数，所以 a 和 b 都必定是 3 的倍数。

想一想 如果 $a^2 + b^2 = 3c^2$，而 $a = 3p$、$b = 3q$，你对 c 有甚么结论？

我们得到 $9p^2 + 9q^2 = 3c^2$，所以 $c^2 = 3(p^2 + q^2)$ 是 3 的倍数，因此 $c = 3r$。于是 $9p^2 + 9q^2 = 27r^2$，即 $p^2 + q^2 = 3r^2$：同一条方程，较小的数。又是递降。 ■

阶段 4 变奏：方程变了

想一想 在 $a^2 + b^2 + c^2 = 4abc$ 中，右边是 4 的倍数。平方数模 4 是甚么？

偶数的平方是 4 的倍数，奇数的平方则比 4 的倍数多一（$(2n+1)^2 = 4n^2 + 4n + 1$）。所以 $a^2 + b^2 + c^2$ 除以 4 的余数是 0、1、2 或 3，视乎 a、b、c 之中有多少个是奇数。由于右边是 4 的倍数，它们没有一个可以是奇数。设 $a = 2p$、$b = 2q$、$c = 2r$：

$$4(p^2 + q^2 + r^2) = 32\,pqr \quad\Longrightarrow\quad p^2 + q^2 + r^2 = 8\,pqr.$$

想一想 这并不是同一条方程：4 变成了 8。递降法还行得通吗？

窍门在这里：证明一个更强的命题。考虑所有以下形式的方程

$$x^2 + y^2 + z^2 = 2^m\,xyz \qquad \text{其中} m \geq 2.$$

对任何这样的 m，右边都是 4 的倍数，所以完全相同的论证显示 x、y、z 全是偶数。把它们减半，便得到含 2^{m+1} 的方程的一个解，而这条方程仍属这种形式。因此，由 m 方程的一个非零解，我们得到 $m+1$ 方程的一个较小的非零解。在这个方程族中任何一条方程的所有解之中，取 $|x|$ 最小的一个，便得出矛盾。特别地，$a^2 + b^2 + c^2 = 4abc$ 没有非零解。 ■

有时候，你想证明的命题太弱，不足以支撑它自身的归纳（induction），补救的方法是证明一个更一般、能够自我支撑的命题。这个技巧在大学里出奇地常见。

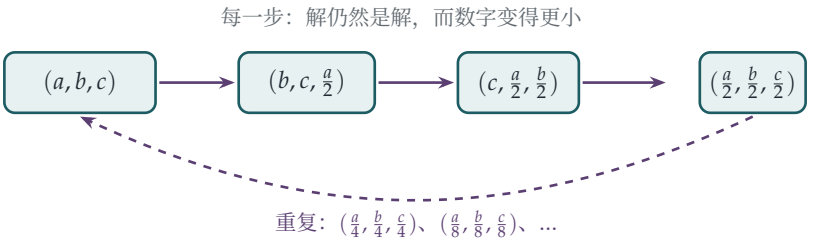


图 1.5.1　第 (i) 部分的递降。由 $a^3+2b^3+4c^3=0$ 的一个解出发，三步之后得到另一个解，每个分量都减半。无休止地重复下去，便会得出一个没有尽头的递减正整数数列 $|a|>|\frac{a}{2}|>|\frac{a}{4}|>\cdots$，而这是不可能的。所以一开始就没有解。

回顾

- **递降需要整除性的来源。**每一次，我们都找到一个质数 p（这里是 2 或 3），它必然整除所有变量，然后把它除去。
- **模 n 运算正是找出它的方法。**考察模 2、3 或 4 的余数，把“有没有任何可能行得通?”变成一个简短的有限验证。
- **如果方程变了，就推广。**第 (iv) 部分需要一整族方程，而这一族方程在递降步骤下是封闭的。

背后的数学

费马、递降法与他的最后定理

费马用无穷递降法证明了 $x^4+y^4=z^2$ 没有正整数解。在他著名的“最后定理”之中，这是他唯一留下证明的情况，它解决了 $x^n+y^n=z^n$ 在 $n=4$ 时的情形。欧拉（Euler）其后用递降法处理 $n=3$ 的情形，当中有一个漏洞，后来才被补上。递降法至今仍是现代理论的起点：关于 Section 1.9 中那类曲线的莫德尔–韦伊定理（Mordell–Weil theorem），其证明就是一个精密的递降。

局部障碍

本节的每一个证明，都是靠考察模某个数的余数来完成的。如果一条方程对某个 n 而言模 n 无解，或者没有实数解，它当然没有整数解。这类理由称为*局部障碍*（local obstructions）；到了大学，你会遇到*p 进数*（p-adic numbers），那是一种同时研究模 $p, p^2, p^3, \ldots$ 所有余数的方法。

哈塞（Hasse）与闵可夫斯基（Minkowski）有一个精彩的定理：对于像 $a^2+b^2=3c^2$ 这样的二次方程，局部障碍就是仅有的障碍——如果没有任何“局部”的理由令它无解，它就确实有有理数解。对三次方程而言，这个结论却彻底失效。Ernst Selmer 在 1951 年证明了

$$3x^3+4y^3+5z^3=0$$

模任何整数都有解，在实数中也有解，但没有非零整数解。解释“局部到整体”原理何时失效、为何失效，是现代数论的一大主题。

一条确实有解的方程

第 (iv) 部分有一个著名的近亲：$x^2+y^2+z^2=3xyz$，即*马尔可夫方程*（Markov equation）。它有无穷多个正整数解，由 $(1,1,1)$、$(1,1,2)$、$(1,2,5)$、$(1,5,13)$ 和 $(2,5,29)$ 开始，而这些解与无理数能被分数逼近到甚么程度有关。见练习 4。

思维技巧

无穷递降法

要证明一条方程无解，可证明任何一个解都会产生一个严格更小的解。这样，最小的解便不可能存在。

把它与模 n 运算配合使用，找出推动递降的整除性；如果命题在递降步骤下不能保持，就要准备把命题推广。

延伸练习

1 **STEP I 热身题。**（*STEP I 2010 第 8 题。*）证明 $a^3+3b^3=9c^3$ 的唯一整数解是 $a=b=c=0$。然后，藉考虑末位数字，证明 $p^4+2q^4=5r^4$ 也有同样的结论。

2 **只用模运算，毋须递降。**只用模 4 运算，证明 $x^2+y^2=4z+3$ 完全没有整数解。

3 **有理点。**利用第 (iii) 部分，证明圆 $x^2+y^2=3$ 上没有坐标为*有理数*的点。（试与 $x^2+y^2=2$ 对比，后者包含 $(1,1)$。）

4 **马尔可夫三元组。**证明若 (x,y,z) 满足 $x^2+y^2+z^2=3xyz$，则 $(x,y,3xy-z)$ 亦满足。由 $(1,1,1)$ 出发，生成上文列出的三元组。（这个“跳跃”技巧正好与递降相反。）

延伸阅读

- **Harold Davenport, *The Higher Arithmetic***，第 2 及第 5 章。
- **Joseph Silverman, *A Friendly Introduction to Number Theory***（Pearson）。非常易读；包括费马对 $x^4 + y^4 = z^2$ 的递降证明。
- **Simon Singh, *Fermat's Last Theorem***（Fourth Estate）。为一般读者而写的故事，由费马的页边笔记讲到怀尔斯（Wiles）的证明。

1.6 两个平方数之和

题目

STEP II 2023，第 7 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2023 年 STEP 试卷（zip 文件）内的 STEP II 试卷：

`www.ocr.org.uk/Images/707102-2023-past-papers.zip`

引子

哪些整数可以写成两个平方数之和？$5 = 1^2 + 2^2$ 和 $13 = 2^2 + 3^2$ 可以，但 3、7 和 21 却不行。这里有一个奇怪的事实：$65 = 5 \times 13$ 可以用*两种*不同的方法写成两个平方数之和，即 $1^2 + 8^2$ 和 $4^2 + 7^2$。为甚么两个这样的数相乘，会得出另一个这样的数？又为甚么会有两种写法？

答案来自一个意想不到的方向。这道题目要你放下整数，改用*复数*（complex number）来思考。在复数的世界里，两平方和变得很简单，一些看似令人摸不着头脑的算术也变得近乎机械化。

动笔之前

想一想 若 $z = a + \mathrm{i}b$，其中 a 和 b 为整数，那么 $|z|^2$ 是甚么？

它是 $a^2 + b^2$，即两个平方数之和。所以“N 是两个平方数之和”的意思正正就是“存在某个实部（real part）和虚部（imaginary part）皆为整数的复数 z，使得 $N = |z|^2$”。这样的复数称为*高斯整数*（Gaussian integer）。请牢记这个转译；整道题目都建基于此。

逐步解题

阶段 1 模相乘

想一想　以 a、b、c、d 写出 zw。关于 $|zw|^2$，你需要证明甚么？

我们有 $zw = (ac - bd) + \mathrm{i}(ad + bc)$，所以

$$|zw|^2 = (ac - bd)^2 + (ad + bc)^2 = a^2c^2 + b^2d^2 + a^2d^2 + b^2c^2,$$

因为交叉项 $-2abcd$ 和 $+2abcd$ 互相抵消。而这个式子可以因式分解（factorise）：

$$a^2c^2 + b^2d^2 + a^2d^2 + b^2c^2 = (a^2 + b^2)(c^2 + d^2) = |z|^2|w|^2.$$

取平方根，得 $|zw| = |z||w|$。 ■

现在把它看成一个关于整数的命题：

$$(a^2 + b^2)(c^2 + d^2) = (ac - bd)^2 + (ad + bc)^2. \qquad (\star)$$

两个两平方和的乘积，仍是一个两平方和，而复数更确切地告诉你那两个平方数是甚么。

阶段 2　学以致用

第 (ii) 部分：$5 = |2 + \mathrm{i}|^2$，而 $221 = 10^2 + 11^2 = |10 + 11\mathrm{i}|^2$。于是相乘：

$$(2 + \mathrm{i})(10 + 11\mathrm{i}) = 20 + 22\mathrm{i} + 10\mathrm{i} - 11 = 9 + 32\mathrm{i},$$

得出 $5 \times 221 = 9^2 + 32^2$。（验算：$81 + 1024 = 1105$。）

想一想　如果改为把 $10 + 11\mathrm{i}$ 乘以 $2 - \mathrm{i}$，会怎么样？它的模（modulus）与 $2 + \mathrm{i}$ 的相同。

$(2 - \mathrm{i})(10 + 11\mathrm{i}) = 20 + 22\mathrm{i} - 10\mathrm{i} + 11 = 31 + 12\mathrm{i}$，所以同样有 $1105 = 31^2 + 12^2$。选用 $2 + \mathrm{i}$ 还是它的共轭（conjugate），会得出不同的表示法。这就解释了引子中 65 的两种写法（图 1.6.1）：$65 = |(2 + \mathrm{i})(3 + 2\mathrm{i})|^2 = |4 + 7\mathrm{i}|^2$ 以及 $65 = |(2 - \mathrm{i})(3 + 2\mathrm{i})|^2 = |8 + \mathrm{i}|^2$。

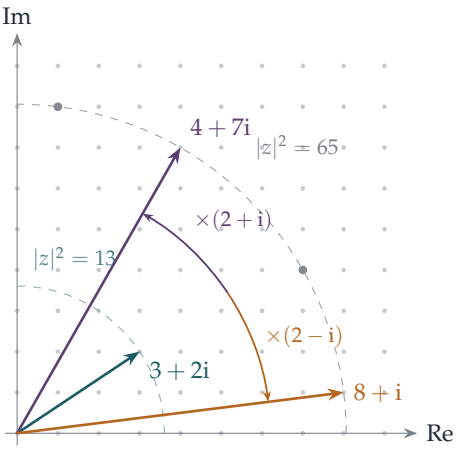


图 1.6.1 乘以 $2 \pm \mathrm{i}$ 会把长度放大 $\sqrt{5}$ 倍，并旋转约 26.6°：$2+\mathrm{i}$ 为逆时针方向，$2-\mathrm{i}$ 为顺时针方向。由圆 $|z|^2 = 13$ 上的 $3+2\mathrm{i}$ 出发，两种选择会落在圆 $|z|^2 = 65$ 上不同的格点（lattice point）。灰色的点是它们对直线 $y = x$ 的反射。

阶段 3 先作因式分解

? **想一想** 8045 的末位数字是 5。这提示了甚么？

$8045 = 5 \times 1609$。我们需要把 1609 写成两个平方数之和；试试略小于 1609 的平方数，很快便找到 $1609 = 40^2 + 3^2$。于是

$$(2+\mathrm{i})(40+3\mathrm{i}) = 77 + 46\mathrm{i}, \qquad \text{所以} \qquad 8045 = 77^2 + 46^2,$$

而共轭亦给出 $8045 = 83^2 + 34^2$。

第 (iv) 部分更快：$36 = 6^2 + 0^2$，而乘以 6 只是按比例放大：$36 \times 50805 = 612^2 + 1206^2$。

阶段 4 三种写法

? **想一想** 25 可以用多于一种方法写成两个平方数之和。若不计正负号和次序对调，有多少个高斯整数的模为 5?

$25 = 5^2 + 0^2 = 3^2 + 4^2$，分别对应 5、$3+4\mathrm{i}$ 和 $3-4\mathrm{i}$。接着看 1002082：由于 $1001^2 = 1002001$，我们有 $1002082 = 1001^2 + 9^2$。把 $1001 + 9\mathrm{i}$ 分别乘以这三个数，得：

$$\begin{aligned} 5\,(1001+9\mathrm{i}) &= 5005 + 45\mathrm{i}, \\ (3+4\mathrm{i})(1001+9\mathrm{i}) &= 2967 + 4031\mathrm{i}, \\ (3-4\mathrm{i})(1001+9\mathrm{i}) &= 3039 - 3977\mathrm{i}. \end{aligned}$$

所以三对 (r,s) 是 $(45,5005)$、$(2967,4031)$ 和 $(3039,3977)$。

阶段 5　除法

> **停一停，试一试**　第 (vi) 部分把这部机器倒过来运转。我们知道 $109 = 10^2 + 3^2$ 和 $1002037 = 1001^2 + 6^2$。怎样才能不靠搜索，就把 9193 写成两个平方数之和？

既然高斯整数相乘时模会相乘，那么相除时模就应该相除。试试

$$\frac{1001 + 6\mathrm{i}}{10 + 3\mathrm{i}} = \frac{(1001 + 6\mathrm{i})(10 - 3\mathrm{i})}{109} = \frac{10028 - 2943\mathrm{i}}{109} = 92 - 27\mathrm{i}.$$

刚好整除，所以 $9193 = 92^2 + 27^2$。（验算：$8464 + 729 = 9193$。）

> **想一想**　如果改为除以 $10 - 3\mathrm{i}$，会怎么样？

你会得到 $(1001 + 6\mathrm{i})(10 + 3\mathrm{i})/109 = (9992 + 3063\mathrm{i})/109$，这不是高斯整数，因为 9992/109 不是整数。在 109 的两个“因数”之中，只有一个能整除 $1001 + 6\mathrm{i}$，正如只有某些质数（prime）能整除一个给定的整数。高斯整数有它们自己的质数，也有它们自己的因式分解。

回顾

- **转译、求解、再译回来。**“两个平方数之和”变成了“高斯整数的模的平方”。用这种语言来说，问题就变得容易。
- **先作因式分解。**每一部分都是先把一个大数拆成较小的数，而这些较小的数的表示法很容易找到。
- **共轭带来新答案。**表示法的数目，取决于有多少种方法选择一个因数或它的共轭。

背后的数学

费马二平方和定理

哪些质数是两个平方数之和？$2 = 1^2 + 1^2$、5、13、17、29、37、41 是；3、7、11、19、23、31 则不是。规律是：除 2 以外，一个质数是两个平方数之和，当且仅当它除以 4 的余数为 1。费马（Fermat）于 1640 年提出这个结果；欧拉（Euler）于 1749 年给出证明。用高斯整数的语言来说：质数 $p \equiv 1 \pmod 4$ 会分裂（split）成 $p = (a + b\mathrm{i})(a - b\mathrm{i})$，而质数 $p \equiv 3 \pmod 4$ 则仍然是质数。例如 $5 = (2 + \mathrm{i})(2 - \mathrm{i})$，但 3 完全无法分解。

1990 年，Don Zagier 发表了二平方和定理的一个证明，它以全长只有一句而闻名。这

个证明连同解说收录于 *Proofs from THE BOOK*。

一个拥有自己质数的新数系

高斯（Gauss）发现，高斯整数的表现与普通整数极为相似。它们有除法算法（division algorithm）和欧几里得算法（Euclidean algorithm），而且每个高斯整数都可以唯一地分解成高斯质数（Gaussian prime）的乘积（不计乘以 ± 1 或 $\pm i$ 的差别）。结合 $(\star)$，你只需知道一个数的质因数分解，就能准确数出它有多少种方法写成两个平方数之和。

在大学里，这是通往代数数论（algebraic number theory）的大门：研究 $\mathbb{Z}[i]$ 或 $\mathbb{Z}[\sqrt{-5}]$ 这类数系，借此了解普通的整数。Section 1.7 说明了为甚么需要创立这门学科：并非每个这样的数系都像高斯整数那样规矩。

四个平方数及其他

每个正整数都是四个平方数之和（拉格朗日（Lagrange），1770 年）。证明依赖 $(\star)$ 的四平方版本，它源自哈密顿（Hamilton）的四元数（quaternion），情形就如 $(\star)$ 源自复数一样。你会在 Section 7.8 见到它。

思维技巧

> **转换场景**
>
> 一个问题在某个数系中显得棘手，在一个更大的数系中却可能十分自然。
>
> 两平方和在整数中很棘手，在复数中却不过是模而已。数学家经常这样扩大数系：由整数到有理数，由有理数到实数，由实数到复数。每次扩大，都会令一些旧问题变得容易。

延伸练习

1. **一个障碍。**证明没有任何形如 $4k+3$ 的数是两个平方数之和。（平方数模 4 是甚么?）
2. **两种写法。**把 65 和 85 各自用两种不同的方法写成两个平方数之和，并用高斯整数解释你的答案。
3. **四种写法。**$1105 = 5 \times 13 \times 17$。找出四种不同的方法，把 1105 写成两个正平方数之和。为甚么是四种?

4 **高斯质数。** 较难 证明若 $3 = (a + b\mathrm{i})(c + d\mathrm{i})$，其中 a, b, c, d 为整数，则其中一个因数的模为 1。（取模的平方。）这说明 3 是高斯质数。

延伸阅读

- **John Stillwell, *Elements of Number Theory***（Springer）。循序渐进讲到高斯整数及更深入的内容，并附历史背景。
- **Aigner & Ziegler, *Proofs from THE BOOK***，第 4 章，“Representing numbers as sums of two squares”（把数表示为两个平方数之和）。
- **Joseph Silverman, *A Friendly Introduction to Number Theory***，有关平方和与高斯整数的各章。

1.7 因数分解失效之时

题目

STEP I 2016，第 7 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/16-s1-q7`

引子

每个大于 1 的整数都可以写成质数(prime)的乘积,而且写法只有一种:$60 = 2 \times 2 \times 3 \times 5$,再没有别的写法。这个事实你大概已经用过无数次，熟悉得让人觉得显然成立，根本不用证明。

这道题构造了一个小小的数字世界，看上去完全合情合理，但在这个世界里，那个“显然”的事实却是错的。这是一道 STEP I 的题目，所用的数学并不艰深，但当中的教训却十分深刻。1847 年，一个著名的费马最后定理（Fermat's Last Theorem）证明尝试，正是败在这个教训之上，而代数学的一个全新分支亦因此诞生。

动笔之前

想一想 列出 S 和 T 的首几个元素。按照第 (iv) 部分的定义，S 的哪些元素是S-质数?

$S = \{1, 5, 9, 13, 17, 21, 25, 29, 33, \ldots\}$，而 $T = \{3, 7, 11, 15, 19, 23, \ldots\}$。$S$ 中的一个数如果不能写成 S 中较小元素（1 除外）的乘积，它就是 S-质数。留意 9 是 S-质数，尽管在普通整数中 $9 = 3 \times 3$，因为 3 并不在 S 之中。一个数是否“质数”，取决于你获准使用哪些数。

逐步解题

阶段 1 并集、交集与乘积

$S \cup T$ 是所有正奇数组成的集合，即两者的并集（union），因为每个奇数除以 4 的余数不是 1 就是 3。$S \cap T$ 是空集，即两者的交集（intersection）没有任何元素，因为没有一个数会有两个不同的余数。

> **想一想**　S 的一般元素应该怎样写？两个这样的元素相乘，乘积是甚么？

S 的元素是 $4a+1$，其中 $a \geq 0$，而

$$(4a+1)(4b+1) = 16ab + 4a + 4b + 1 = 4(4ab + a + b) + 1 \in S.$$

所以 S 在乘法下封闭（closed）。至于 T：

$$(4a+3)(4b+3) = 16ab + 12a + 12b + 9 = 4(4ab + 3a + 3b + 2) + 1,$$

这个数属于 S，而不属于 T。例如 $3 \times 7 = 21 \in S$。所以 T 在乘法下并不封闭。

阶段 2　属于 T 的质因数

> **想一想**　在 T 中取一个不是质数的数。它的质因数（prime factor）全是奇数。假如这些质因数全都属于 S，会有甚么后果？

这个数是奇数，所以它的质因数全是奇数，各自属于 S 或 T。假如它们全部属于 S，那么由于 S 在乘法下封闭，它们的乘积也会属于 S。但这个乘积就是我们原来的数，而它属于 T。所以至少有一个质因数属于 T。 ■

阶段 3　奇数个因数

> **想一想**　把偶数个 T 的元素乘起来，乘积会落在哪里？奇数个又如何？

每一对 T-数相乘都得到一个 S-数，而 S-数相乘仍然是 S-数，所以偶数个 T-因数的乘积属于 S。奇数个的话，乘积就是（S 中的某个数）$\times$（T 的一个元素），用同类的计算可知它属于 T。

现在取 $t \in T$。如果 t 是 T-质数，那就完成了。否则，把 t 写成 T 的元素的乘积，其中每个元素都大于 1，因而小于 t。当中任何不是 T-质数的因数都可以再分解下去，如此类推。因数每次都变小，所以这个过程必会停止（这里再次用到良序性质（well-ordering）），最后 t 就写成了 T-质数的乘积。由于乘积属于 T，因数的个数必定是奇数。 ■

阶段 4　两种分解

> **停一停，试一试**　在 S 中找一个数，它有两种真正不同的 S-质数分解。提示：利用像 9 这样的数，它们之所以是 S-质数，只因为它们在普通整数中的质因数并

不在 S 之中。

S-质数依次为 $5, 9, 13, 17, 21, 29, 33, 37, 41, 49, \ldots$。$21 = 3 \times 7$、$9 = 3 \times 3$ 和 $49 = 7 \times 7$ 全都是 S-质数，因为 3 和 7 都不在 S 之中。而且

$$441 = 21 \times 21 = 9 \times 49.$$

这就是 441 的两种不同的 S-质数分解（图 1.7.1）。唯一分解（unique factorisation）在 S 中并不成立。 ■

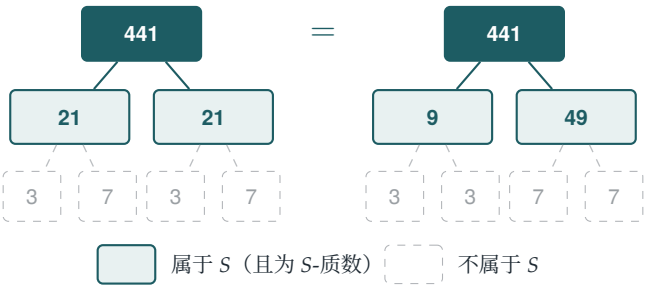


图 1.7.1 441 的两种 S-质数分解。在普通整数中，两棵树都会继续向下分解至 $3 \times 3 \times 7 \times 7$，但 3 和 7 都不在 S 之中，所以在 S 之内，分解到第二行便停止，而两种分解并不相同。

回顾

- **是否“质数”取决于所在的系统。**9 和 21 在整数中是合数（composite），在 S 中却是质数，因为分解它们所需的数并不存在。
- **封闭性很重要。**S 在乘法下封闭；T 则不然。第 (iii) 和第 (iv) 部分的论证，全靠这一点分别。
- **显然的事实也可以是一条定理。**整数的唯一分解是需要证明的。这道题说明了原因：在一个看上去非常相似的系统中，它并不成立。

背后的数学

整数有何特别之处

为甚么分解在普通整数中确实行得通？关键在于欧几里得引理（Euclid's lemma）：如果质数 p 整除乘积 ab，那么 p 整除 a 或 p 整除 b。这条引理是用辗转相除法（Euclidean algorithm）证明的，而它正正是在 S 中失效的地方。在那里，S-质数 9 整除 $21 \times 21 = 441$

（因为 $441 = 9 \times 49$），但 9 在 S 中并不整除 21。到了大学一年级，你会严谨地证明欧几里得引理，并看到唯一分解如何由它推导出来。

费马最后定理的失败证明

1847 年 3 月 1 日，拉梅（Gabriel Lamé）向巴黎科学院宣布，他已证明了费马最后定理。他的想法是利用复数单位根（roots of unity）把 $x^n + y^n$ 分解，

$$x^n + y^n = (x+y)(x+\zeta y)(x+\zeta^2 y)\cdots(x+\zeta^{n-1}y), \qquad \zeta = e^{2\pi i/n}$$

（其中 n 为奇数），然后就每个括号的质因数进行论证。刘维尔（Joseph Liouville）随即提出质疑：由 ζ 构成的数，凭甚么一定有唯一分解？数星期后，人们才知道库默尔（Ernst Kummer）早已证明这类数往往没有唯一分解。最先失效的情况是 $n = 23$。

库默尔尽力补救，发明了“理想数”（ideal numbers），这些额外的对象能够恢复唯一分解，他并借此对一大类指数证明了费马最后定理。戴德金（Richard Dedekind）把库默尔的理想数发展成理想（ideal），成为近世代数的核心概念之一。

一个值得记住的较小例子

在数系 $\mathbb{Z}[\sqrt{-5}] = \{a + b\sqrt{-5} : a, b \in \mathbb{Z}\}$ 中，

$$6 = 2 \times 3 = (1+\sqrt{-5})(1-\sqrt{-5}),$$

而这四个因数在该数系中全都是“质数”（练习 3）。试把这个例子与 Section 1.6 的高斯整数（Gaussian integers）比较，后者的分解确实是唯一的。这类数系之中哪些具有唯一分解，是一个很深刻的问题。高斯（Gauss）猜想，在形如 $\mathbb{Z}[\sqrt{-d}]$ 的数系及与其相近的数系之中，只有九个具有唯一分解。这个猜想最终由黑格纳（Heegner）、贝克（Baker）和斯塔克（Stark）在 1952 年至 1967 年间证明。

思维技巧

> **质疑你视为理所当然的事**
>
> 当一个事实看来显然得不用证明时，问问它依赖甚么，并找出一个令它失效的情境。
>
> 构造反例（counterexample），也就是构造一个令“显然”的事实不成立的世界，是弄清一个证明真正需要甚么的最佳方法之一。大学本科代数有很大部分，就是在探问整数那些熟悉的性质，有哪些能在更一般的系统中保留下来。

延伸练习

1 **另一个例子。**证明 693 有两种不同的 S-质数分解。你能找出一个制造这类例子的通用方法吗?

2 **偶数。**设 $E = \{2, 4, 6, 8, \ldots\}$。哪些数是 E-质数?证明 60 有多于一种 E-质数分解。

3 $\mathbb{Z}[\sqrt{-5}]$**。**定义*范数*（norm）$N(a + b\sqrt{-5}) = a^2 + 5b^2$。证明 $N(zw) = N(z)N(w)$。先证明没有任何元素的范数是 2 或 3，由此推导出 2、3 和 $1 \pm \sqrt{-5}$ 在 $\mathbb{Z}[\sqrt{-5}]$ 中不能再分解。

4 **欧几里得引理的应用。**较难 假设欧几里得引理成立（如果一个质数整除 ab，它就整除 a 或 b），证明每个大于 1 的整数都只有一种质数分解，因数的次序不计。

延伸阅读

- **Ian Stewart & David Tall,** ***Algebraic Number Theory and Fermat's Last Theorem***（CRC Press）。开首几章讲述拉梅、库默尔与理想的故事，程度适合好学的大学一年级学生。
- **John Stillwell,** ***Elements of Number Theory***。把整数的唯一分解与高斯整数的唯一分解并列比较。
- **Harold Edwards,** ***Fermat's Last Theorem: A Genetic Introduction***（Springer）。适合有志挑战的读者：循历史发展解说库默尔的工作。

1.8 计数有理数

题目

STEP III 2020，第 8 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2020 年 STEP 试卷（zip 文件）内的 STEP III 试卷：

`www.ocr.org.uk/Images/696589-2020-past-papers.zip`

引子

你能否把所有正分数逐一列出，使每个分数都恰好出现一次？这看来是不可能的：任意两个分数之间还有无穷多个分数，根本没有“下一个”分数可以写下来。然而，康托尔（Georg Cantor）在 1870 年代证明了这是做得到的。这表示，在某个精确的意义下，分数并不比整数多。

康托尔的列表颇为笨拙：表中既有 $\frac{1}{2}$，也有 $\frac{2}{4}$，重复的项要逐一删去。本题要构造的列表却完全没有重复，而且来自一个规则简单得出奇的数列（sequence）。这个列表由 Neil Calkin 和 Herbert Wilf 于 2000 年发表，不过数列本身早在 1858 年已由 Moritz Stern 研究过。

动笔之前

想一想 在证明任何东西之前，先计算最初十五项左右。你留意到比值 u_n/u_{n+1} 有甚么特别之处？

数列的开首几项是

$$1,\ 1,\ 2,\ 1,\ 3,\ 2,\ 3,\ 1,\ 4,\ 3,\ 5,\ 2,\ 5,\ 3,\ 4,\ 1,\ 5,\ \ldots$$

而比值 u_n/u_{n+1} 依次是

$$\frac{1}{1},\ \frac{1}{2},\ \frac{2}{1},\ \frac{1}{3},\ \frac{3}{2},\ \frac{2}{3},\ \frac{3}{1},\ \frac{1}{4},\ \frac{4}{3},\ \frac{3}{5},\ \frac{5}{2},\ \frac{2}{5},\ \frac{5}{3},\ \frac{3}{4},\ \frac{4}{1},\ \ldots$$

每个分数都以最简分数（lowest terms）的形式出现，而且没有一个重复。题目要你证明这个现象会一直持续下去，并且每个正分数迟早都会出现。

逐步解题

阶段 1　奇数项胜过偶数项

相邻两项不是 (u_{2k}, u_{2k+1})，就是 (u_{2k+1}, u_{2k+2})。

> **想一想**　试用 u_k 和 u_{k+1} 表示这两种数对。哪一项较大？

$(u_{2k}, u_{2k+1}) = (u_k,\ u_k + u_{k+1})$，而 $(u_{2k+1}, u_{2k+2}) = (u_k + u_{k+1},\ u_{k+1})$。由于每一项都是正整数（用归纳法（induction）不难证明），$u_k + u_{k+1}$ 比 u_k 和 u_{k+1} 都大。唯一的例外是第一对 $(u_1, u_2) = (1, 1)$，它不符合上述任何一种形式，因为 u_1 是直接定义的。 ■

阶段 2　相邻两项互质

> **想一想**　假设 $d > 1$ 同时整除 u_{2k} 和 u_{2k+1}，即同时整除 u_k 和 $u_k + u_{k+1}$。那么 d 还整除甚么？

d 整除 $(u_k + u_{k+1}) - u_k = u_{k+1}$，所以 d 整除相邻数对 u_k, u_{k+1}，而这一对在数列中的位置较前。另一种情况完全相同：若 d 整除 $u_k + u_{k+1}$ 和 u_{k+1}，它便整除 u_k。

因此，相邻数对的公因数会传回较前的一对，然后再传回更前的一对，如此类推。位置不可能无止境地减小（又是良序性（well-ordering）），所以 d 最终必须整除第一对 $u_1 = u_2 = 1$。当 $d > 1$ 时这是不可能的。所以每一对相邻项都互质（coprime）。 ■

阶段 3　没有数对出现两次

> **停一停，试一试**　阶段 1 告诉你一件有用的事：看到一个相邻数对 (a, b) 时，你可以判断 a 位于奇数位置还是偶数位置。怎样判断？然后，你能否还原出阶段 2 所指的、在它之前的那个数对？

若 $a < b$，则（由阶段 1）b 位于奇数位置，所以对某个 k，这个数对是 $(u_{2k}, u_{2k+1}) = (u_k, u_k + u_{k+1})$，而它的“父”数对 (u_k, u_{k+1}) 就是 $(a,\ b - a)$。若 $a > b$，这个数对是 (u_{2k+1}, u_{2k+2})，其父数对是 $(a - b,\ b)$。若 $a = b$，阶段 1 表明这个数对必定是最初的一对 $(1, 1)$。

所以除第一对外，每个相邻数对都决定了它的父数对，*以及父数对的位置*（这个结构可以看成一棵树，见图 1.8.1）：当子数对位于 $2k$ 或 $2k + 1$ 时，父数对位于位置 k。现在假设 (a, b) 出现在两个不同的位置。它们的父数对便是出现在两个位置的同一数对，而*这些*父数对的父数对亦然，如此类推。这样一直追溯上去，我们最终会在 1 以外的位置找到 $(1, 1)$，而阶段 1 已排除了这个可能。 ■

阶段 4 每个互质数对都会出现

想一想 假设某个互质数对 (a,b) 从未以相邻两项的形式出现。你应该考虑哪一个这样的数对？

在所有“缺席”的互质数对之中，取 $a+b$ 的值最小的一个：这就是极小反例（minimal counterexample）。它不可能是 $(1,1)$，因为这一对确实出现了，所以 $a \neq b$（互质而又相等，必有 $a=b=1$）。

- 若 $a>b$：数对 $(a-b,b)$ 互质（$a-b$ 和 b 的任何公因数都整除 a），而它的和 a 小于 $a+b$。由极小性，这个数对确实出现，设它位于位置 $k,k+1$。但这样便有 $u_{2k+1}=u_k+u_{k+1}=a$ 和 $u_{2k+2}=u_{k+1}=b$，所以 (a,b) 终究还是出现了。
- 若 $a<b$：同理，$(a,b-a)$ 出现在某个 $k,k+1$，于是 $(u_{2k},u_{2k+1})=(a,b)$。

无论哪种情况都得出矛盾。所以每一对互质的正整数都会以相邻两项的形式出现。 ■

（题目把第 (iv) 部分写成递降法（descent）：一个缺席的数对会产生一个更小的缺席数对。这其实是同一个论证，只是倒过来读而已。）

阶段 5 所有分数的列表

每个正有理数（rational number）都可以唯一地写成最简分数 a/b，其中 a,b 互质。由阶段 4，(a,b) 会以某个相邻数对 (u_n,u_{n+1}) 的形式出现，所以 $f(n)=a/b$：f 的值域（range）是全体正有理数。由阶段 3，它只出现一次；又由阶段 2，每个值 $f(n)$ 本身已是最简分数，所以两个不同的 n 不可能给出同一个分数。因此 f 是由正整数到正有理数的双射（bijection），并且有逆函数（inverse）。 ■

回顾

- **递降与它的逆向。**公因数和重复的数对都是往回追溯到数列的起点。缺席的数对则用极小反例处理，向前构造。
- **双射要证两半。**“每个分数都出现”（满射，onto）和“没有分数出现两次”（单射，one-to-one）是两个独立的命题，要分开证明。
- **先动手计算。**写出十五项之后，这些命题便显得可信，同时也提示了推动整个证明的父子结构。

背后的数学

Calkin–Wilf 树

把正分数排成一棵树：把 $\frac{1}{1}$ 放在顶端，并给每个分数 $\frac{a}{b}$ 两个子节点，左边是 $\frac{a}{a+b}$，右边是 $\frac{a+b}{b}$。逐行由左至右读这棵树，得到的正是本题的列表 $f(1), f(2), f(3), \ldots$（图 1.8.1）。阶段 3 说明每个分数都有唯一的父节点，阶段 4 说明每个分数都在树中。2003 年，Moshe Newman 发现连这棵树也不需要：列表中每个分数都可以由前一个分数算出，公式是

$$x_{n+1} = \frac{1}{2\lfloor x_n \rfloor - x_n + 1}.$$

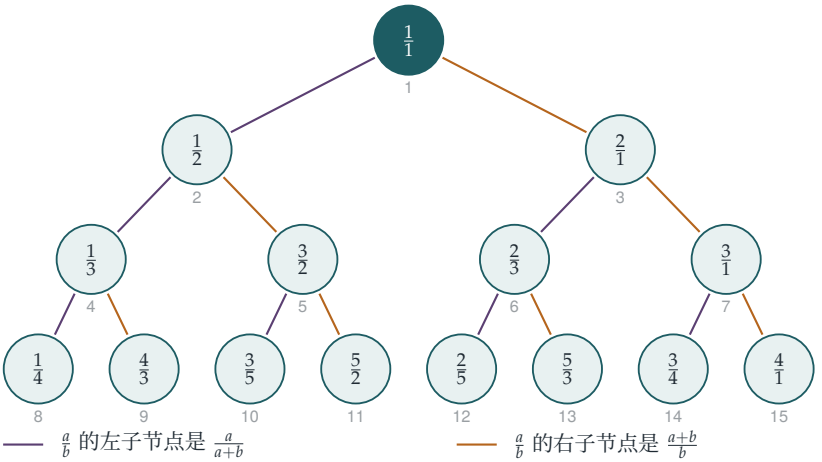


图 1.8.1　Calkin–Wilf 树的首四行。逐行由左至右读，便得到 $f(1), f(2), f(3), \ldots$；灰色小数字是位置 n。每个正分数都以最简分数的形式恰好出现一次。

不同大小的无穷

可以这样逐一列出的集合称为*可数*（countable）集合。分数是可数的，这令人意外，因为分数看来比整数多得多。在大学的第一个学期，你还会遇到康托尔的*对角论证法*（diagonal argument），它证明*实数*（real numbers）是不可数的：任何实数列表都必定有所遗漏。所以无穷确实有不同的大小，而实数所构成的无穷严格地大于有理数所构成的无穷。这是你在大学一年级课程中会遇到的最令人震惊的结果之一。

Stern 数列

数列 u_n 称为 *Stern 双原子数列*（Stern's diatomic sequence），它还有许多其他面貌。例如，u_{n+1} 等于把 n 写成 2 的幂之和、而每个幂最多用两次的方法数目。它与 *Stern–Brocot 树*以及连分数（continued fractions）关系密切，而连分数能给出无理数的最佳有理逼近。

思维技巧

构造双射

要证明两个集合大小相同，就要把它们的元素准确地一一配对：证明目标集合的每个元素都被对应到（满射），而且没有元素被对应两次（单射）。

双射是计数的基础，不论有限还是无穷，你在大学里会经常用到。在本题中，两半的证明都是把一个结构追溯回它的起点。

延伸练习

1 **找出分数。**对哪一个 n 有 $f(n) = \frac{3}{5}$？又 $f(n) = \frac{5}{3}$ 呢？利用父节点的规则倒推回 $\frac{1}{1}$。

2 **Newman 公式。**验证由 $x_1 = 1$ 开始，Newman 公式能给出列表中的首十个分数。

3 **康托尔的对角论证法。**假设你有一个列表，以小数形式列出了 0 与 1 之间的所有实数。构造一个小数，使它与列表中第 n 个数在第 n 位数字上不同。为甚么这就说明列表并不完整？（对于像 $0.0999\ldots = 0.1$ 这样的小数，需要留意甚么？）

4 **可数并集。** 较难 给出一个与正整数之间的明确双射，从而证明由全体正整数数对 (m, n) 组成的集合是可数的。（试试 $2^{m-1}(2n-1)$。）

延伸阅读

- **Neil Calkin & Herbert Wilf, "Recounting the rationals"**（*American Mathematical Monthly*，2000 年；〈重新数算有理数〉）。全文四页，完全读得懂。
- **Aigner & Ziegler, *Proofs from THE BOOK***（《数学天书中的证明》），第 17 章 "Sets, functions, and the continuum hypothesis"（集合、函数与连续统假设），当中包括 Calkin–Wilf 树。
- **Ian Stewart & David Tall, *The Foundations of Mathematics***（《数学基础》，Oxford Uni-

versity Press）。可数性、康托尔与实数，内容与大学一年级课程相若。

1.9 三次曲线的切线

题目

STEP I 2009，第 2 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/09-s1-q2`

引子

这题看来只是一道例行的坐标几何（coordinate geometry）题：求导、找切线、解三次方程。可是最后一行却要你找出一个立方数，它是另外三个立方数之和，而几何竟像变魔术一样把它变了出来。

这个窍门是过曲线上一个已知点作直线，从而找出新的点，它可追溯至公元三世纪亚历山大港的丢番图（Diophantus）。到了现代数学家手上，它发展成*椭圆曲线*（elliptic curves）的理论。这套理论位于当今数论的核心，是费马最后定理（Fermat's Last Theorem）证明的基础，也保障着全球一大部分互联网通讯的安全。

动笔之前

? **想一想** 点 $(-a, b)$ 在曲线上。为甚么？这跟立方和又有甚么关系？

代入：$(-a)^3 + a^3 + b^3 = b^3$，所以确实在曲线上。在 $y^3 = x^3 + a^3 + b^3$ 上坐标为*有理数*（rational）的点正是我们想要的：若 $x = q/s'$ 而 $y = p/s'$，去分母后便得到一条立方数之间的关系式。我们的计划是从一个容易找到的点出发，利用几何找出一个没那么明显的点。

逐步解题

阶段 1 切线

对 $y^3 = x^3 + a^3 + b^3$ 作隐函数求导（implicit differentiation）：$3y^2\dfrac{dy}{dx} = 3x^2$，所以 $\dfrac{dy}{dx} = \dfrac{x^2}{y^2}$。在 $(-a, b)$ 处斜率（gradient）为 a^2/b^2，所以切线（tangent）是

$$y - b = \frac{a^2}{b^2}(x + a) \quad\Longrightarrow\quad b^2y - b^3 = a^2x + a^3 \quad\Longrightarrow\quad b^2y - a^2x = a^3 + b^3. \qquad \blacksquare$$

阶段 2　切线与曲线再次相交之处

取 $a = 1$、$b = 2$，曲线是 $y^3 = x^3 + 9$，而在 $(-1, 2)$ 的切线是 $4y - x = 9$，即 $y = \dfrac{x+9}{4}$。代入曲线方程：

$$\frac{(x+9)^3}{64} = x^3 + 9 \quad\Longrightarrow\quad x^3 + 27x^2 + 243x + 729 = 64x^3 + 576,$$

整理后得 $63x^3 - 27x^2 - 243x - 153 = 0$，再除以 9 便得 $7x^3 - 3x^2 - 27x - 17 = 0$。 $\blacksquare$

阶段 3　解三次方程

> **想一想**　不用做任何代数运算，你已经知道这个三次方程的一个根。是哪一个？它为甚么应该出现两次？

切线在 $x = -1$ 处与曲线相切（图 1.9.1），而“相切”意味着重根（repeated root），正如抛物线的切线与抛物线相交于某个二次方程的二重根。所以 $(x+1)^2$ 应该是一个因式。验算：$7(-1)^3 - 3(1) - 27(-1) - 17 = -7 - 3 + 27 - 17 = 0$，而导数 $21x^2 - 6x - 27$ 在 $x = -1$ 处也等于零。相除后得

$$7x^3 - 3x^2 - 27x - 17 = (x+1)^2(7x - 17).$$

> **停一停，试一试**　所以切线与曲线再次相交于 $x = \frac{17}{7}$ 之处。求出 y，然后把这个点化为一条立方数之间的关系式。

$y = \dfrac{17/7 + 9}{4} = \dfrac{80/7}{4} = \dfrac{20}{7}$。点 $\left(\frac{17}{7}, \frac{20}{7}\right)$ 在 $y^3 = x^3 + 9$ 上，所以

$$\frac{20^3}{7^3} = \frac{17^3}{7^3} + 9 \quad\Longrightarrow\quad 20^3 = 17^3 + 9 \times 7^3.$$

> **想一想**　我们希望 9×7^3 是两个立方数之和。回想一下那个 9 是怎样来的。

$9 = a^3 + b^3 = 1^3 + 2^3$，所以 $9 \times 7^3 = 7^3 + 14^3$。因此

$$20^3 = 17^3 + 14^3 + 7^3.$$

验算：$8000 = 4913 + 2744 + 343$。 ■

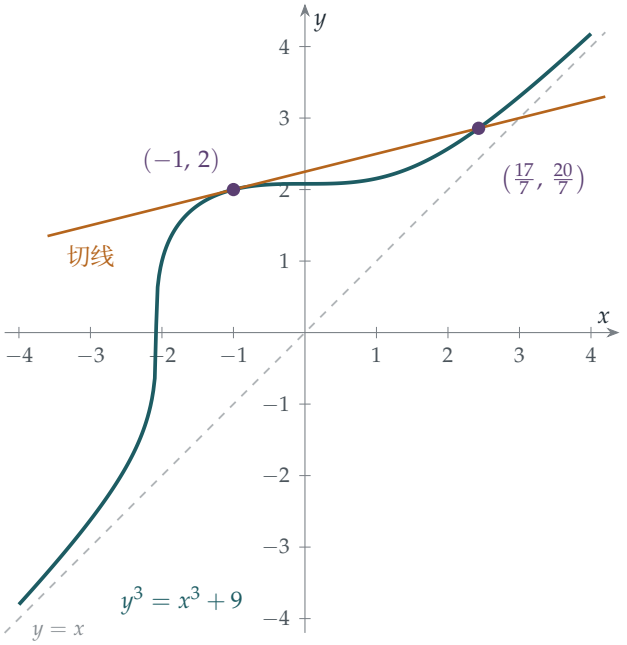


图 1.9.1 曲线 $y^3 = x^3 + 9$（蓝绿色）在远离原点处趋近直线 $y = x$。在有理点 $(-1, 2)$ 的切线（琥珀色）在该处与曲线相切，这算作二重交点，并在另一个有理点 $\left(\frac{17}{7}, \frac{20}{7}\right)$ 再与曲线相交一次。

回顾

- **相切即二重根。**这令一个没有明显根的三次方程，变成可以即时因式分解的方程。
- **新的根自动是有理数。**三个根之中一旦有两个是有理数（这里两个都是 -1），第三个也必定是有理数，因为各根之和 $\frac{3}{7}$ 是有理数。这正是整个方法的精髓。
- **几何带来了算术。**我们从没有刻意搜索立方数；是一条切线把它们找了出来。

背后的数学

弦切法

取一条三次曲线（cubic curve），例如 $y^3 = x^3 + 9$，或者更标准的形式 $y^2 = x^3 + Ax + B$。假如你知道曲线上两个坐标为有理数的点，那么过这两点的直线恰好与曲线再相交于一点，而该点的坐标也是有理数（理由见回顾中的论证）。假如你只知道一个点，就改用

该点的切线。无论哪种情况，旧的有理点都能产生新的有理点。

值得注意的是，这个“弦切”（chord–tangent）作图法可以变成一种把曲线上的点*相加*的方法，于是有理点便构成一个*群*（group），这是你在大学一年级会学到的一种代数结构。莫德尔（Louis Mordell）在 1922 年证明了这个群由有限个点生成：所有有理点，不论有多少个，都可以由一个有限集合出发，利用弦和切线构造出来。他的证明是 Section 1.5 中费马（Fermat）下降法的一个版本。

今日的椭圆曲线

这类曲线称为*椭圆曲线*（并非因为它们是椭圆，而是因为历史上它们与椭圆的弧长有关）。它们在现代数学中无处不在。

- **费马最后定理。**在 1980 年代，Gerhard Frey 留意到 $x^n + y^n = z^n$ 的一个解会产生一条非常古怪的椭圆曲线。怀尔斯（Andrew Wiles）在 1994 年证明了这样古怪的曲线不可能存在，从而证明了费马最后定理。
- **贝赫与斯维讷通－戴尔猜想（Birch and Swinnerton-Dyer conjecture）。**有理点所构成的群需要多少个生成元（generators）？在 1960 年代，有人凭借剑桥一些数学史上最早的电脑实验提出了一个猜想，它用一个函数来预测答案，而这个函数是由计数曲线在模质数之下的点数构造出来的。它是七条千禧年大奖难题（Millennium Prize Problems）之一。
- **密码学（cryptography）。**在有限数系上的椭圆曲线上，把一个点 P 自己相加 n 次是很快的，即使 n 有数百位数字也一样（把 P 反复加倍，就像你用反复平方来计算高次幂那样）。但已知 P 和结果 nP，要找出 n 却似乎极之困难：这就是*椭圆曲线离散对数问题*（elliptic-curve discrete logarithm problem）。这种不对称性保护着今日大部分的网上通讯。

同次幂之和

能写成三个正立方数之和的最小立方数是 $6^3 = 3^3 + 4^3 + 5^3$；切线法找到的是 $20^3 = 7^3 + 14^3 + 17^3$，它并非前者的倍数，而重复使用这个方法会不断产生新的恒等式（虽然有些会涉及负数）。要多少个 k 次幂才能凑成另一个 k 次幂？欧拉（Euler）在 1769 年猜想，至少需要 k 个 k 次幂才能凑成另一个 k 次幂。对于四次幂和五次幂，他错了：1966 年 Leon Lander 和 Thomas Parkin 以电脑搜索找到

$$27^5 + 84^5 + 110^5 + 133^5 = 144^5$$

而在 1988 年，Noam Elkies 利用椭圆曲线找到了四次幂的反例（counterexamples）。

思维技巧

以几何解决算术

要找出一个方程的整数解或有理数解，不妨把它看成一条曲线，然后寻找过已知点的直线。

背后的代数很简单：若一个有理系数多项式（polynomial）除了一个根之外其余的根都是有理数，那么最后一个根也是有理数。这样在图像与方程之间来回转换，是数学家所能拥有的最强大的习惯之一。

延伸练习

1 **圆上的有理点。**过 $(-1,0)$ 而斜率为 t 的直线与圆 $x^2+y^2=1$ 再次相交。证明另一个交点是 $\left(\dfrac{1-t^2}{1+t^2}, \dfrac{2t}{1+t^2}\right)$，并由此得出无穷多组毕氏三元数（Pythagorean triples）。

2 **一条落空的弦。**点 $(1,2)$ 和 $(2,1)$ 在 $x^3+y^3=9$ 上。证明过这两点的直线与曲线再没有其他交点。那个“第三点”到哪里去了？（查一查无穷远点（point at infinity）。）

3 **故技重施。**求 $y^3=x^3+9$ 在 $\left(\frac{17}{7}, \frac{20}{7}\right)$ 的切线，以及它与曲线相交的第三点。（使用电脑代数系统会有帮助；留意数字增长得有多快。）

4 **的士数（taxicab numbers）。**$1729 = 1^3+12^3 = 9^3+10^3$。求过 $(1,12)$ 和 $(9,10)$ 的直线与曲线 $x^3+y^3=1729$ 相交的第三点，并由此推导出一条形如 $p^3=q^3+r^3+s^3$ 的新恒等式。

延伸阅读

- **Joseph Silverman & John Tate, *Rational Points on Elliptic Curves***（Springer）。为大学本科生而写；第 1 章所需的知识不超出英国高考（A-level）程度。
- **Avner Ash & Robert Gross, *Elliptic Tales***（Princeton University Press）。一本普及读物，一路讲到贝赫与斯维讷通－戴尔猜想。
- **Simon Singh, *Fermat's Last Theorem***。讲述椭圆曲线如何解决一个有 350 年历史的难题。

我们学到了甚么

在本章开首，我们就整数提出了三个问题，并预告还有第四个主题贯穿其中：怎样证明（prove）一件事。现在是时候看看，九道 STEP 题目把我们带到了多远。

1. 质数有多少个？它们是怎样分布的？

在 Section 1.1 中，我们证明了质数（prime）有无穷多个，却连一个新的质数也没有找出来。我们构造了一族无穷多个的数，即费马数（Fermat numbers），它们两两之间没有公因数，所以每一个都需要一个属于自己的质数。这个证明谈的是*结构*，而不是大小。

在 Section 1.2 中，我们更进一步，证明了不超过 n 的所有质数之积小于 4^n。这是我们第一个*定量*的结论：它限制了质数可以排得多密。取对数后，它说的是 $\sum_{p \le n} \ln p < 1.39\, n$，而质数定理（Prime Number Theorem）指出真实的值大约是 n。我们这个 STEP 程度的论证，已经相当接近了。

我们*没有*做到的，是确切找出质数怎样分布。这由黎曼 ζ 函数（Riemann zeta function）所支配，而把这件事说得精确，就是黎曼猜想（Riemann Hypothesis），也就是数学中最著名的未解决问题。你现在所知的，已足以明白它在说甚么。

2. 怎样证明一个数不是分数？

我们找到了三种截然不同的方法。

- **取最小的那一个**（Section 1.3）。假如 $\sqrt{2}$ 是分数，就会有一个最小的正整数 k 使 $k\sqrt{2}$ 为整数；但这样一来，$(\sqrt{2}-1)k$ 就是一个更小的。同样的论证也适用于立方根。
- **把一个整数困在 0 与 1 之间**（Section 1.4）。假如 e 或 e^2 是分数，乘以一个阶乘（factorial）后，就会得出一个严格介乎 0 与 1 之间的整数。埃尔米特（Hermite）和林德曼（Lindemann）证明 e 和 π 是超越数（transcendental），背后正是这个想法。
- **找出一个不可能成立的整除关系**（Section 1.5，习题 3）。证明 $x^2 + y^2 = 3$ 上没有有理点（rational point），等同于证明当 x 是有理数时，$\sqrt{3 - x^2}$ 永远不是有理数。

我们也认识到还有多少事情仍然未知：至今没有人能证明 $e + \pi$ 是无理数（irrational）。

3. 哪些方程有整数解?

有时答案是*没有*，而 Section 1.5 展示了费马（Fermat）证明这一点的方法：任何一个解都会产生一个更小的解，如此无休无止。以一个精心挑选的质数为模（modulo）来运算，就能告诉我们该利用哪一个整除关系。

有时答案是*有，而且这里有一部制造解的机器*。在 Section 1.6 中，复数（complex numbers）把二平方和变成了一条乘法法则。在 Section 1.9 中，三次曲线（cubic curve）的一条切线给出了 $20^3 = 17^3 + 14^3 + 7^3$。

还有些时候，正确的一步是*换一个数系*。高斯整数（Gaussian integers）令二平方和变得容易。但 Section 1.7 提出了一个警告：在一个看来完全合理的数系中，数可以用两种不同的方式分解成“质数”。这个警告在 1847 年被人忽视，结果催生了理想（ideals）和近世代数。

Section 1.8 还另添了一重转折：分数看似远比整数多，却可以逐一列举出来。无穷比乍看之下有更丰富的结构。

4. 证明的工具箱

以下是你在本章用过的方法。每一种都会在大学数学课程中一再出现。

方法	你在哪里用过
归纳法（induction）与强归纳法（strong induction）	费马数（1.1）；质数之积（1.2）
良序性（well-ordering）、最小反例（minimal counterexample）	$\sqrt{2}$ 与 $\sqrt[3]{2}$（1.3）；列举分数（1.8）
无穷递降法（infinite descent）	不可能成立的方程（1.5）；互质（coprime）数对（1.8）
反证法（proof by contradiction）	贯穿全章，尤其是 1.3 和 1.4
模 n 运算	1.5；1.7
估界与夹逼（bounding and squeezing）	1.2；1.4
构造双射（bijection）	1.8
构造反例（counterexample）	非唯一分解（1.7）
转换场景	复数（1.6）；几何（1.9）

回头看看这些题目，你会留意到 STEP 的拟题人往往替你走了第一步：“考虑二项展开式（binomial expansion）…”、“考虑复数 $2 + \mathrm{i}$ 和 $10 + 11\mathrm{i}$…”。到了大学，这种帮助会逐渐消失，而找出正确的第一步就成了真正的本领。最好的准备正是你一直在做的事：每做完一道题，问问自己那个提示*为甚么*奏效，下次你就能自己想到。

下一步

本章有好几条线索会在书中其他地方延续下去。

- 以分数逼近 e 的课题会在 Section 2.9 再次出现，而黎曼 ζ 函数则见于 Section 2.4，那里有 $\zeta(2) = \pi^2/6$。
- 单位根（roots of unity）曾在拉梅（Lamé）那个失败的证明中出现，它是 Chapter 5 的主题；而在 Section 1.1 中预告过的高斯（Gauss）正十七边形作图法，就是 Section 6.2。
- Section 1.6 中那条恒等式的四平方版本，会通过四元数（quaternions）在 Section 7.8 中出现。

如果你想在上大学前走得更远

- **Harold Davenport, *The Higher Arithmetic***：初等数论入门书中最好的一本。
- **Aigner & Ziegler, *Proofs from THE BOOK***：第 1–4、8 和 17 章与本章直接相关。
- **Joseph Silverman, *A Friendly Introduction to Number Theory***：几乎涵盖本章的全部内容，而且还有更多，节奏平缓。
- **Marcus du Sautoy, *The Music of the Primes*** 和 **Simon Singh, *Fermat's Last Theorem***：数学背后的人物故事。

数论总有办法把人吸引进去。它的问题简单得第一天就可以提出，却又深奥得足以耗上一生。如果其中某些章节令你很想知道接下来会怎样，那正是本书的用意所在。

延伸练习答案

以下只是简短的答案和提示，并非完整解答：只列出结果和关键步骤，方便你核对自己的做法。请先认真做过每一道练习，然后才看答案。

1.1 质数工厂

1. 假设 $p_1, \dots, p_k$ 就是全部的质数 (prime)；那么 $p_1 \cdots p_k + 1$ 有一个不在此列之中的质因数 (prime factor)，矛盾 (contradiction)。用费马数 (Fermat number) 的证明则是构造性 (constructive) 的：它具体给出无穷多个两两互质 (pairwise coprime) 的数 $F_0, F_1, \dots$，每一个都提供一个新的质数。

2. 当 $n \geq 2$ 时，2^n 是 4 的倍数，所以 $2^{2^n} = 16^{2^{n-2}}$，而 16 的任何次幂的末位数字都是 6。因此 F_n 的末位数字是 7。

3. (a) 用归纳法 (induction)：$a_{n+1} - 1 = a_n(a_n - 1) = a_n \cdot a_1 \cdots a_{n-1}$。(b) a_m 与 a_n ($m < n$) 的公因数必整除 $a_1 \cdots a_{n-1} = a_n - 1$ 和 a_n，所以只能是 1；每一项都有一个不与其他项共有的质因数。(c) $\frac{1}{a_n} = \frac{1}{a_n - 1} - \frac{1}{a_{n+1} - 1}$；求和时各项裂项相消 (telescope)，得 $1 - \frac{1}{a_{n+1} - 1}$。

4. (a) 由 $p \mid F_n$ 得 $2^{2^n} \equiv -1$；两边平方得 $2^{2^{n+1}} \equiv 1$。(b) 凡满足 $2^e \equiv 1$ 的 e 都能被 d 整除，所以 $d \mid 2^{n+1}$；又因为 $2^{2^n} \equiv -1 \not\equiv 1$ (因 p 是奇数)，所以 $d \nmid 2^n$。故 $d = 2^{n+1}$。(c) $2^{p-1} \equiv 1$，所以 $2^{n+1} \mid p - 1$。(d) $641 = 64 \times 10 + 1$。小于 641 而形如 $64k + 1$ 的质数有 193、257、449 和 577：只需试四个候选数，欧拉 (Euler) 正是这样找到 641 的。

5. 第 n 个质数不会超过 $F_0, \dots, F_{n-1}$ 之中任何一个数的最大质因数，所以 $p_n \leq F_{n-1} = 2^{2^{n-1}} + 1$。当 $n = 5$ 时，这个上界是 65 537，而 $p_5 = 11$：结论正确，但极为粗疏。

1.2 质数分布得有多稀疏?

1. $P_{1,10} = 210$，$P_{1,20} = 9\,699\,690$，$P_{1,30} = 6\,469\,693\,230$。与 $4^{30} \approx 1.15 \times 10^{18}$ 和 $e^{30} \approx 1.07 \times 10^{13}$ 比较：这个乘积远小于 4^n，增长速度大致与 e^n 相若 (事实上 $\ln P_{1,n} \sim n$，这是质数定理 (Prime Number Theorem) 的一种形式)。

2. $\binom{2m}{k+1} / \binom{2m}{k} = \frac{2m-k}{k+1}$，此值大于 1 当且仅当 $k < m$，所以各系数先递增至中间一项，然后递减。$2m + 1$ 项之和为 4^m，其中最大的一项至少等于它们的平均值。

3. 区间 $(\sqrt{n}, n]$ 内的质数每个都大于 $\sqrt{n}$,而它们的乘积小于 4^n,所以 $(\sqrt{n})^{\pi(n)-\pi(\sqrt{n})} < 4^n$；两边取对数（logarithm）。由于 $2\ln 4 \approx 2.77$ 且 $\pi(\sqrt{n}) \le \sqrt{n}$，即得所求的上界。它具有正确的形式 $n/\ln n$，只是常数为 2.8，而非真正的 1。

4. 取质数 $2, 3, 5, 7, 13, 23, 43, 83$，其中每一个都小于前一个的两倍。对任意 $n \le 50$，此列中第一个大于 n 的质数小于前一个质数的两倍，而前一个质数不超过 n，所以它介乎 n 与 $2n$ 之间。

1.3 最小反例

1. 设 k 是使 $k\sqrt{3}$ 为整数的最小正整数，则 $k' = (\sqrt{3}-1)k$ 是一个更小的正整数，而且 $k'\sqrt{3} = 3k - k\sqrt{3}$ 是整数。这里 $\sqrt{3}-1$ 所起的作用，与 $\sqrt{2}-1$ 相同。

2. 由于 $0 < \sqrt{N} - m < 1$，$k' = (\sqrt{N}-m)k$ 是一个小于 k 的正整数，而 $k'\sqrt{N} = kN - mk\sqrt{N}$ 是整数：矛盾。

3. 当 $N = 4$ 时，$m = 2 = \sqrt{4}$，所以 $\sqrt{N} - m = 0$，而 $k' = 0$ 并非正整数。“$0 < \sqrt{N} - m$”这一步不成立。

4. 在图中，边长为 a 的正方形内有两个边长为 b 的正方形互相重叠，它们把中央一个边长为 $2b - a$ 的正方形覆盖了两次，并留下两个边长为 $a - b$ 的正方形未被覆盖。用代数表示，就是 $(2b-a)^2 - 2(a-b)^2 = 2b^2 - a^2 = 0$。

1.4 e 是无理数，e^2 也是

1. 若 $\frac{1}{e} = \frac{p}{q}$，把 $\frac{1}{e} = \sum_{k\ge 0} \frac{(-1)^k}{k!}$ 乘以 $q!$。$k \le q$ 的各项都是整数，而尾部 $\sum_{k>q}(-1)^k q!/k!$ 是各项递减的交错级数（alternating series），所以其绝对值严格介乎 0 与 $\frac{1}{q+1}$ 之间。因此 $q!/e$ 不是整数：矛盾。

2. 若 $\sin 1 = \frac{p}{q}$，取 $2n+1 \ge q$，两边乘以 $(2n+1)!$。左边是整数，级数中直至 $\frac{1}{(2n+1)!}$ 的各项也是整数；尾部是交错级数，首项的大小为 $\frac{1}{(2n+2)(2n+3)} < 1$，所以它不是整数。

3. 这个方法需要一个各项均为有理数（rational）的级数，而且其尾部缩小的速度要比分母增长的速度快，就像 $\frac{1}{k!}$ 那样；对于 π，目前并不知道有这样的级数。Niven 的证明用到 $\int_0^\pi \frac{x^n(\pi-x)^n}{n!}\sin x\,\mathrm{d}x$：假如 π 是有理数，当 n 足够大时，这个积分就会是一个小于 1 的正整数。

4. $x^2 - (e+\pi)x + e\pi$ 的根是 e 和 π。假如 $e+\pi$ 和 $e\pi$ 都是代数数（algebraic number），π 就是一个代数系数多项式的根，因而也是代数数，这与林德曼（Lindemann）定理矛盾。

1.5 无穷递降法

1. $3 \mid a$；设 $a = 3a_1$，得 $9a_1^3 + b^3 = 3c^3$，所以 $3 \mid b$，于是 $3a_1^3 + 9b_1^3 = c^3$，所以 $3 \mid c$，而 (a_1, b_1, c_1) 是一个更小的解：这就是递降（descent）。至于 $p^4 + 2q^4 = 5r^4$：四次方数模（modulo）5 只能是 0 或 1（它们的末位数字是 0、1、5 或 6），所以 $p^4 + 2q^4 \equiv 0$ 迫使 $5 \mid p$ 且 $5 \mid q$；于是 $5^4 \mid 5r^4$，所以 $5 \mid r$，同样可以递降。

2. 平方数模 4 只能是 0 或 1，所以 $x^2 + y^2$ 只能是 0、1 或 2，绝不会是 3。

3. 若有有理点 $(\frac{a}{c}, \frac{b}{c})$，便会得到 $a^2 + b^2 = 3c^2$，其中 $c \neq 0$，而第 (iii) 部分已排除这个可能。

4. 固定 x、y 时，方程是二次方程 $z^2 - 3xyz + (x^2 + y^2) = 0$，其两根之和为 $3xy$。由 $(1,1,1)$ 出发：得 $(1,1,2)$；把 $(1,2)$ 所对的那个 1 换掉，得 $3 \cdot 1 \cdot 2 - 1 = 5$，即 $(1,2,5)$；然后是 $(1,5,13)$ 和 $(2,5,29)$。

1.6 两个平方数之和

1. 平方数模 4 只能是 0 或 1，所以两个平方数之和模 4 绝不会是 3。

2. $65 = 1^2 + 8^2 = 4^2 + 7^2$ 及 $85 = 2^2 + 9^2 = 6^2 + 7^2$。利用 $65 = (2+\mathrm{i})(2-\mathrm{i})(3+2\mathrm{i})(3-2\mathrm{i})$：$(2+\mathrm{i})(3+2\mathrm{i}) = 4 + 7\mathrm{i}$ 及 $(2+\mathrm{i})(3-2\mathrm{i}) = 8 - \mathrm{i}$。同理，对于 85，$(2+\mathrm{i})(4 \pm \mathrm{i})$ 给出 $7 + 6\mathrm{i}$ 和 $9 - 2\mathrm{i}$。

3. $1105 = 4^2 + 33^2 = 9^2 + 32^2 = 12^2 + 31^2 = 23^2 + 24^2$。5、13、17 各自都可分解为一个高斯整数（Gaussian integer）与其共轭（conjugate）之积；从每一对中各选一个因数，共得 8 个乘积，它们两两互为共轭（给出相同的平方和），所以剩下 4 个。

4. 取模的平方，得 $9 = (a^2 + b^2)(c^2 + d^2)$。由于 3 不是两个平方数之和，其中一个因数必为 1。

1.7 因数分解失效之时

1. $693 = 9 \times 77 = 21 \times 33$，四个因数全是 S-质数。方法：取四个形如 $4k+3$ 的质数 p, q, r, s；那么 $(pq)(rs) = (pr)(qs)$ 就给出两种分解为 S-质数的方式。

2. E-质数就是形如 $2 \pmod 4$ 的数（它们不可能是两个偶数之积）。$60 = 2 \times 30 = 6 \times 10$。

3. 展开 $(ac - 5bd)^2 + 5(ad + bc)^2$ 即得 $N(zw) = N(z)N(w)$。$a^2 + 5b^2$ 绝不会等于 2 或 3。由于 $N(2) = 4$、$N(3) = 9$ 及 $N(1 \pm \sqrt{-5}) = 6$，若要分解为非单位（non-unit）之积，就需要一个范数（norm）为 2 或 3 的因数。

4. 设 n 是具有两种不同分解 $p_1 \cdots p_r = q_1 \cdots q_s$ 的最小的数。根据欧几里得（Euclid）

引理，p_1 整除某个 q_j，所以 $p_1 = q_j$；约去后得到一个更小而又有两种不同分解的数，矛盾。

1.8 计数有理数

1. $\frac{3}{5}$：逐级追溯其父节点，$(3,5) \leftarrow (3,2) \leftarrow (1,2) \leftarrow (1,1)$，对应的位置为 $1 \to 2 \to 5 \to 10$，所以 $f(10) = \frac{3}{5}$。$\frac{5}{3}$：$(5,3) \leftarrow (2,3) \leftarrow (2,1) \leftarrow (1,1)$，位置为 $1 \to 3 \to 6 \to 13$，所以 $f(13) = \frac{5}{3}$。

2. Newman 的法则 $x_{n+1} = \frac{1}{2\lfloor x_n \rfloor - x_n + 1}$ 给出 1、$\frac{1}{2}$、2、$\frac{1}{3}$、$\frac{3}{2}$、$\frac{2}{3}$、3、$\frac{1}{4}$、$\frac{4}{3}$、$\frac{3}{5}$，与 $f(1), \ldots, f(10)$ 吻合。

3. 第 n 位数字一律取 5，除非第 n 个数的第 n 位数字是 5，这时则取 4。新的小数与列表中每一个数都不相同，而且由于避开了 0 和 9，它只有一种小数展开式，所以确实是一个不同的数。

4. 每个正整数都可唯一地写成 2 的幂乘以一个奇数，即 $2^{m-1}(2n-1)$，所以 $(m,n) \mapsto 2^{m-1}(2n-1)$ 是一个双射（bijection）。

1.9 三次曲线的切线

1. 把 $y = t(x+1)$ 代入 $x^2 + y^2 = 1$，得 $(1+t^2)x^2 + 2t^2 x + t^2 - 1 = 0$，其根为 -1 和 $\frac{1-t^2}{1+t^2}$。取 $t = \frac{q}{p}$，该点为 $\left(\frac{p^2-q^2}{p^2+q^2}, \frac{2pq}{p^2+q^2}\right)$，由此得到三元数组 $(p^2-q^2, 2pq, p^2+q^2)$。

2. 该直线为 $x + y = 3$，而 $x^3 + (3-x)^3 = 9$ 可化简为 $9x^2 - 27x + 18 = 0$：三次项互相抵消，只剩下 $x = 1, 2$。这条直线与渐近线（asymptote）$y = -x$ 平行，第三个交点在无穷远处。

3. 切线（tangent）的斜率（gradient）为 $\frac{x^2}{y^2} = \frac{289}{400}$，它与曲线再相交于 $\left(-\frac{188479}{90391}, -\frac{36520}{90391}\right)$。去分母后，得 $188479^3 = 36520^3 + 90391^3 + 180782^3$。

4. 该直线与 $x^3 + y^3 = 1729$ 再相交于 $\left(-\frac{37}{3}, \frac{46}{3}\right)$，所以 $46^3 = 37^3 + 27 \times 1729 = 37^3 + 3^3 + 36^3$（这里用到 $27 = 3^3$ 和 $27 \times 1728 = 36^3$）。

章

2

数列、级数与收敛

无穷是数学变得有趣的地方，也是最容易出错的地方。

本章内容

2.1	“最终”是甚么意思	STEP III 2015
2.2	作为极限的黄金比例	STEP II 2013
2.3	一个发散级数与它的收敛近亲	STEP III 2014
2.4	$\pi^2/6$	STEP III 2018
2.5	韦达乘积	STEP III 2022
2.6	算术几何平均	STEP II 2024
2.7	六位还是六百位	STEP II 2023
2.8	平方再减二	STEP III 2021
2.9	逐步逼近 e 的分数	STEP II 2021
A	练习答案	第 169 页

驯服无穷

无穷！从来没有别的问题如此深刻地触动人类的心灵。

— 希尔伯特（David Hilbert）

两千五百年前，埃利亚的芝诺（Zeno of Elea）论证说运动是不可能的。要走过一个房间，你必须先走一半路程，然后走余下路程的一半，再走那余下的一半，如此类推，永无止境：这是无穷多项任务，肯定无法在有限时间内完成。人人都知道芝诺错了，因为人们确实能走过房间，但在很长的一段时间里，没有人能确切说出他为甚么错。答案是：无穷和 $\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\cdots$ 有一个有限的值，就是 1。可是，无穷多个数加起来等于某个数，这句话究竟是甚么意思？

这个问题占据数学的核心达两千年之久。十七世纪时，牛顿（Newton）和莱布尼兹（Leibniz）在它之上建立了微积分（calculus），却始终没有真正回答它；他们的后继者既得出了许多精彩的结果，也制造了一些荒谬绝伦的谬论。欧拉（Euler）是有史以来最伟大的计算家，他毫不犹豫地写下 $1-1+1-1+\cdots=\frac{1}{2}$。另一些人则就同一个无穷级数（infinite series）“证明”出互相矛盾的结论。直到十九世纪，经过柯西（Cauchy）、魏尔斯特拉斯（Weierstrass）等人的努力，数学家才有了足够精确的定义，可以分辨哪些结果是真、哪些是假。这些定义就是本章的主题。

为甚么重要

你一生中用来计算的数，几乎每一个都是某样东西的极限（limit）。你的计算机并不知道 $\sqrt{2}$、π 或 e 是甚么：它只知道一些能够愈来愈接近这些数的计算步骤，并在足够接近时停下来。明白一套步骤逼近得有多快，正是一毫秒内完成的算法与永远算不完的算法之间的分别。用来预测天气、设计飞机和为金融产品定价的电脑模拟，背后也是同样的想法。而在最纯粹的一面，对极限的研究引出了数学中一些最优美的公式：在本章中，你将会证明

$$1+\frac{1}{4}+\frac{1}{9}+\frac{1}{16}+\cdots=\frac{\pi^2}{6},$$

这个结果令欧拉一举成名，至今看来仍然近乎神奇。

你在大学会遇到甚么

几乎在每一所大学，数学分析（Analysis）都是你最先修读的课程之一，而对许多学生来说，它带来的冲击也最大。在中学数学里，你学的是计算极限；在数学分析里，你学的是

从精确的定义出发，证明关于极限的命题。你会定义数列（sequence）收敛（converge）是甚么意思，用的正是 Section 2.1 中出现的“对于每一个 $\varepsilon > 0$”和“存在 N”这些说法。你会证明每一个有界递增数列都收敛，并发现这有赖于实数的一项性质——完备性（completeness），而有理数并不具备这项性质。你会学习判别级数是否收敛的各种方法，然后运用同样的想法，为连续性、微分和积分赋予严格的意义。

往后的课程会在这个基础上向多个方向发展。数值分析（Numerical analysis）探讨近似值收敛得有多快，以及怎样令它们收敛得更快（Section 2.7、Section 2.6）。动力系统（Dynamical systems）研究反复套用同一条规则时会发生甚么事，有时会趋向一个极限，有时则产生混沌（chaos）（Section 2.2、Section 2.8）。复分析（Complex analysis）解释了为甚么 π 会在那么多无穷和与无穷乘积中出现（Section 2.4、Section 2.5）。而解析数论（analytic number theory）则利用极限来回答关于整数的问题，例如 Section 2.9。

贯穿本章的问题

三大问题

1. **一个无穷的过程有答案，是甚么意思？**我们怎样才能精确地说一个数列“趋近”某个极限，或者一个无穷和“加起来等于”某个数？
2. **怎样在不知道极限是甚么的情况下，证明某样东西收敛？**而一旦知道极限存在，又怎样把它求出来？
3. **有多快？**两个过程可以有相同的极限，逼近的速度却有天壤之别。为甚么会这样？我们又怎样量度这种速度？

第一节（Section 2.1）谈的是语言：数学家怎样精确地表达“最终”这个概念。接着的五节谈的是著名的极限：以连分数（continued fraction）表示的黄金比（golden ratio）、一个发散（divergent）级数和它那个收敛的近亲、$\pi^2/6$、韦达（Viète）关于 $2/\pi$ 的乘积，以及高斯（Gauss）的算术几何平均（arithmetic–geometric mean）。最后三节谈的是速度与结构：线性收敛与二次收敛的对比、一个要么爆炸式增长要么混沌地游走的迭代（iteration），以及一族逐步逼近 e 的积分。

处理收敛的工具

有少数几种工具，你会一再遇到，每次的场合都不一样。

§	主题	工具
2.1	“最终较小”	量词（quantifiers）：找出 m、否定命题
2.2	黄金比	单调有界子数列（monotone bounded subsequences）
2.3	调和级数（harmonic series）与 Kempner 级数	分段比较
2.4	$\pi^2/6$	多项式的根，然后用夹逼法（squeeze）
2.5	韦达乘积	裂项相消（telescoping），然后求一个简单的极限
2.6	算术几何平均（AGM）	两个数列互相夹逼；一个不变量（invariant）
2.7	逼近 $\sqrt{2}$	误差的递推关系（recurrences）：线性与二次
2.8	$x \mapsto x^2 - 2$	一个令映射变得简单的代换（substitution）
2.9	逼近 e 的分数	很小但不为零的积分

怎样阅读本章

与第 1 章一样，每一节都以一道 STEP 原题开始。请先自己尝试，然后才往下读。本章的题目往往会告诉你该用哪一种工具（“考虑……”、“你可以使用以下定理……”）；在解题的过程中，请留意同样的几种工具出现得有多频密，因为到了大学，你便要自己作出选择。

有好几道题目要求你假设极限存在，或者假设无穷和与无穷乘积收敛。遇到这种情况时，该节会解释要证明这一点还需要些甚么。意识到“这显然行得通”与“我能证明这行得通”之间的差距，是学会像分析学家那样思考的第一步。

2.1 “最终”是甚么意思

题目

STEP III 2015，第 2 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/15-s3-q2`

引子

大学第一门分析（analysis）课程里，几乎每一个命题都是同一个模样：*存在*某个东西，使得*对所有*某些东西，某个条件都成立。“数列（sequence）趋向 L”的意思是：对每一个容许误差 $\varepsilon > 0$，都*存在*一个位置 N，使得*对所有* $n \geq N$，都有 $|a_n - L| < \varepsilon$。能够纯熟运用这种句式的学生，会觉得分析引人入胜；不熟悉的学生，则会觉得摸不着头脑。

这道题目设计精妙，正好是这种说话方式的一个平易入门。它定义了一个新概念“最终小于”，然后就这个概念问了四条是非题。四条之中有一条是假的，而它为假的原因，正正是大多数大学一年级学生初次尝试*否定*（negate）含有量词（quantifier）的命题时会掉进的陷阱。

动笔之前

慢慢细读定义。$(s_n) \leq (t_n)$ 并不是说对每个 n 都有 $s_n \leq t_n$，而是说*由某一项开始*，这个不等式一直成立。最初的几项完全无关痛痒：即使你把其中一个数列开首的一百万项改掉，$(s_n) \leq (t_n)$ 是否成立也不受影响。

想一想 定义中有“存在正整数 m”这句话。要证明 $(s_n) \leq (t_n)$，你究竟要拿出甚么来？要证明它是*假*的，你又要说明甚么？

要证明它，你必须*拿出*一个合适的 m，然后说明*每一个* $n \geq m$ 都符合要求。要否证它，你必须说明*没有任何* m 行得通：无论别人提出哪一个 m，总有某个 $n \geq m$ 使得 $s_n > t_n$。请把后面这个说法记在心里，第 (ii) 部分会用到。

逐步解题

阶段 1　线性函数对二次函数

第 (i) 部分问的是 $1000n \le n^2$ 是否最终成立。由于 $n > 0$，这等同于 $1000 \le n$。所以取 $m = 1000$ 便可：对每个 $n \ge 1000$，都有 $n^2 = n \cdot n \ge 1000n$。此命题为**真**。 ■

在最初的 999 项，$1000n$ 远大于 n^2，但这并不要紧。“最终”二字不理会任何有限多项。

阶段 2　“最终小于”跟普通的“小于”相似吗？

对普通的数而言，如果 $a \le b$ 不成立，那么 $b < a$，因此当然有 $b \le a$。第 (ii) 部分问的是，我们这个新关系是否也有同样的性质。

> **想一想**　“$(s_n) \le (t_n)$ 不成立”究竟是甚么意思？把它完整地写出来，而且句首不要用“不”字。

“存在 m，使得对所有 $n \ge m$，都有 $s_n \le t_n$”这个命题为假，恰好就是

$$\text{对每个 } m,\ \text{都有某个 } n \ge m \text{ 使得 } s_n > t_n。$$

换句话说，$s_n > t_n$ 出现*无限多次*。这比“由某一项开始一直有 $s_n > t_n$”弱得多。两个数列大可以不断互换位置。

> **停一停，试一试**　试找出两个正数数列，它们永远不断互换位置，以致任何一个都不是最终小于另一个。

取 $s_n = 2 + (-1)^n$，即 $s_n = 1, 3, 1, 3, \ldots$，并对所有 n 取 $t_n = 2$。对每个 m，都有一个偶数 $n \ge m$ 使得 $s_n = 3 > 2 = t_n$，所以 $(s_n) \le (t_n)$ 不成立。但同样地，对每个 m，都有一个奇数 $n \ge m$ 使得 $t_n = 2 > 1 = s_n$，所以 $(t_n) \le (s_n)$ 也不成立。此命题为**假**。 ■

阶段 3　串连起来

> **想一想**　在第 (iii) 部分，你有两个 m 的值，每个假设各给一个，两者可能不同。要得出结论，应该用哪一个？

假设对所有 $n \ge m_1$ 都有 $s_n \le t_n$，并且对所有 $n \ge m_2$ 都有 $t_n \le u_n$。若 $n \ge \max(m_1, m_2)$，两个不等式同时成立，所以 $s_n \le t_n \le u_n$。因此取 $m = \max(m_1, m_2)$ 便可，此命题为**真**。 ■

在两个门槛之中取较大者，是你在分析里会用上数百次的一着。举例说，要证明两个收敛（convergent）数列之和也收敛，用的就是这一着。

阶段 4 二次函数对指数函数

先试几个数值：

n	1	2	3	4	5	6	7	8
n^2	1	4	9	16	25	36	49	64
2^n	2	4	8	16	32	64	128	256

可见 $n^2 \le 2^n$ 在 $n = 1, 2$ 时成立，在 $n = 3$ 时不成立，然后由 $n = 4$ 起似乎一直成立（图 2.1.1）。很自然会猜 $m = 4$。

? **想一想** 要用归纳法（induction）证明对所有 $n \ge 4$ 都有 $n^2 \le 2^n$，你需要由 $n^2 \le 2^n$ 推到 $(n+1)^2 \le 2^{n+1} = 2 \cdot 2^n$。$(n+1)^2$ 与 n^2 之间要有怎样的不等式，这一步才行得通？

只要知道 $(n+1)^2 \le 2n^2$ 就足够，因为这样便有 $(n+1)^2 \le 2n^2 \le 2 \cdot 2^n$。而 $2n^2 - (n+1)^2 = n^2 - 2n - 1 = (n-1)^2 - 2$，当 $n \ge 3$ 时为正数。于是：

- 基础步骤：$4^2 = 16 = 2^4$。
- 归纳步骤：若 $n \ge 4$ 且 $n^2 \le 2^n$，则 $(n+1)^2 \le 2n^2 \le 2^{n+1}$。

因此对所有 $n \ge 4$ 都有 $n^2 \le 2^n$，此命题为**真**，可取 $m = 4$。 ■

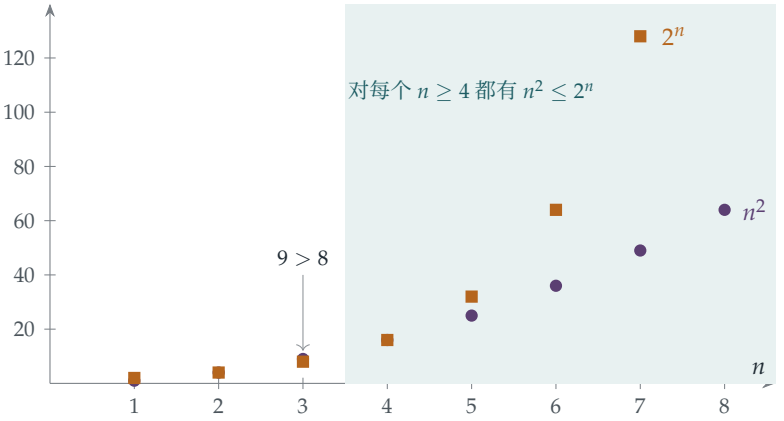


图 2.1.1 n^2（紫色圆点）对 2^n（琥珀色方块）。两个数列在最初几项交叉了两次，但由 $n = 4$ 起，2^n 永远领先。这正是 $(n^2) \le (2^n)$ 的意思，其中 $m = 4$。

回顾

- **证明“存在”，就是要拿出来。**在第 (i)、(iii) 和 (iv) 部分，证明的内容就是指明 m 是甚么，然后加以验证。

- **否定会把量词对调。**“并非（存在 m，对所有 $n \geq m$，P）”变成“对所有 m，存在 $n \geq m$，非 P”。第 (ii) 部分考的就是能否把这一点做对。
- **新的关系不一定跟旧的关系性质相同。**“最终小于”跟 $\leq$ 一样具有传递性（transitive），但跟 $\leq$ 不同的是，它并不能比较每一对数列。

背后的数学

分析的语言

在牛顿（Newton）和莱布尼兹（Leibniz）之后约 150 年间，数学家运用极限（limit）和无穷小（infinitesimal）出神入化，却说不清楚它们确切是甚么意思。贝克莱主教（Bishop Berkeley）在 1734 年嘲讽无穷小是“已逝之量的鬼魂”，他说的不无道理。补救要等到十九世纪：柯西（Augustin-Louis Cauchy）以及其后的魏尔斯特拉斯（Karl Weierstrass），以由“对每个”和“存在”构成的命题，取代了“变得无限接近”这类含糊的说法。现代的定义

$$a_n \to L \iff \text{对每个}\varepsilon > 0\text{，都存在}N\text{，使得}|a_n - L| < \varepsilon \text{ 适用于所有}n \geq N$$

跟本题的定义模样完全相同，只是前面多了一个量词。这是你在大学分析课程里最先遇到的东西。

增长的阶

第 (iv) 部分是一道著名阶梯（增长的阶，orders of growth）的其中一小级。对任何固定的 $k > 0$ 和 $c > 1$，

$$(\ln n) \leq (n^k), \qquad (n^k) \leq (c^n), \qquad (c^n) \leq (n!), \qquad (n!) \leq (n^n),$$

而事实上，每个数列不单最终小于下一个，更是最终小于下一个的*任何常数倍*。哈代（G. H. Hardy）为这类比较写了一整本书：*Orders of Infinity*（1910）。今天它已是计算机科学的日常语言，“这个算法需要 $O(n^2)$ 步”的意思，就是步数最终不超过 n^2 的某个常数倍。n^2 与 2^n 之别，就是一个会完成的程序，与一个到太阳燃尽也完成不了的程序之别。

并非全序的序

第 (ii) 部分说明“最终小于”不是*全序*（total order）：有些数列对根本无法比较。这类具有传递性但不能比较所有东西的关系，称为*预序*（preorder）或*偏序*（partial order），而且随处可见：整数的整除性、集合的包含关系，以及“是……的子群（subgroup）”都

是例子。辨别一个关系具有哪些性质、不具有哪些性质，是数学家训练的一部分。

思维技巧

处理量词

要证明“存在 m，使得对所有 $n \geq m$...”，先指明 m，再检查每一个 n。要否证它，则说明对每个 m，都有某个 $n \geq m$ 不符合要求。

当一个命题含有多个量词时，先把它完整写出来，才尝试证明或否证；否定时由左至右逐个处理：每个“对所有”变成“存在”，每个“存在”变成“对所有”，而末尾的条件则取其否定。

延伸练习

1 **样本试卷热身题。**试做 STEP III Specimen paper，第 2 题（`step.maths.org/questions/spec-s3-q2`），此题出自 1987 年首届 STEP 举行之前发出的试卷。（最后一部分的提示：利用二项展开式（binomial expansion）的下一项。）

2 **两边都成立。**证明若 $(s_n) \leq (t_n)$ 且 $(t_n) \leq (s_n)$，则对所有足够大的 n 都有 $s_n = t_n$。

3 **竞逐无穷。**试找出两个递增的正数数列，两者皆趋向无穷大，而任何一个都不是最终小于另一个。

4 **拾级而上。**证明 $(2^n) \leq (n!)$，并给出一个明确的 m。然后证明 $(n^{10}) \leq (2^n)$，并求出可行的最小 m。（第二部分用试算表会有帮助；答案大得出人意表。）

延伸阅读

- **Lara Alcock, *How to Study for a Mathematics Degree***（牛津大学出版社 Oxford University Press）。第 3 章讲的正是这种逻辑语言，以及大学新生为何觉得它困难。
- **Lara Alcock, *How to Think About Analysis***（牛津大学出版社 Oxford University Press）。一本亲切易读的指南，讲解大学一年级分析的各个定义。
- **Daniel Velleman, *How to Prove It***（剑桥大学出版社 Cambridge University Press）。有系统地介绍逻辑、量词与证明。

2.2 作为极限的黄金比例

题目

STEP II 2013，第 6 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/13-s2-q6`

引子

写下这个式子

$$1+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\cdots}}}$$

然后问它等于甚么。一个快捷的答案是：设它为 x；第一条分数线下面的部分又是 x，所以 $x=1+1/x$，由此得 $x^2=x+1$，于是 $x=\frac{1+\sqrt{5}}{2}$，即黄金比（golden ratio）φ。

但这个论证成立吗？它假设了这个无穷式子*真的有意义*，即逐层构造下去的过程最终会稳定于一个数。把同样的手法用在 $1-1+1-1+\cdots$ 上（“设它为 S；那么 $S=1-S$，所以 $S=\frac{1}{2}$”），得出的却是荒谬的结果。本题示范怎样把这件事做得严谨，所用的是 STEP 考生在事先不知道极限（limit）的情况下证明收敛（convergence）的唯一工具：*单调数列*（*monotone sequences*）。

动笔之前

先计算几项。取 $u_1=1$：

$$1,\quad 2,\quad \frac{3}{2},\quad \frac{5}{3},\quad \frac{8}{5},\quad \frac{13}{8},\quad \frac{21}{13},\quad \frac{34}{21},\quad \ldots$$

想一想 你认得这些分子和分母吗？再看看小数值：$1, 2, 1.5, 1.667, 1.6, 1.625, 1.615, \ldots$。这个数列是递增的吗？还是递减的？

它们是相邻斐波那契数（Fibonacci numbers）之比。这个数列既不递增也不递减：它呈锯齿状摆动，时而过高，时而过低。因此，题目给出的定理是关于递增数列的，不能用于整个数列。不过，奇数项 $1, 1.5, 1.6, 1.615, \ldots$ 看来在上升，而偶数项 $2, 1.667, 1.625, \ldots$ 看来在下降。这正是本题的布局。

逐步解题

阶段 1　比较相隔两项的项

第 (i) 部分比较相隔两步的 u_{n+2} 与 u_n。最自然的第一步，是利用 $(*)$ 把它们各自用前一项表示。

> **想一想**　运用 $(*)$ 两次，把 $u_{n+2}-u_n$ 用 $u_{n+1}-u_{n-1}$ 表示，再把 $u_{n+1}-u_{n-1}$ 用 u_n-u_{n-2} 表示。

由 $(*)$，

$$u_{n+2}-u_n=\frac{1}{u_{n+1}}-\frac{1}{u_{n-1}}=-\frac{u_{n+1}-u_{n-1}}{u_{n+1}u_{n-1}},$$

同理 $u_{n+1}-u_{n-1}=-\dfrac{u_n-u_{n-2}}{u_nu_{n-2}}$。把两式合起来，两个负号互相抵消：

$$u_{n+2}-u_n=\frac{u_n-u_{n-2}}{u_{n+1}u_n\cdot u_{n-1}u_{n-2}}.$$

最后，由 $(*)$ 可知 $u_{n+1}u_n=u_n+1$ 及 $u_{n-1}u_{n-2}=u_{n-2}+1$，于是得到所求的公式。（我们需要 $n\ge 3$，u_{n-2} 才存在。）∎

阶段 2　一个陷阱

> **停一停，试一试**　证明对所有 n 均有 $1\le u_n\le 2$。

对 $u_1=1$ 这是成立的。若 $1\le u_n\le 2$，则 $\frac{1}{2}\le\frac{1}{u_n}\le 1$，所以 $\frac{3}{2}\le u_{n+1}\le 2$，这当然介乎 1 与 2 之间。由归纳法（induction），对所有 n 均有 $1\le u_n\le 2$。∎

特别地，每个 u_n 都是正数，所以第 (i) 部分中的分母 $(1+u_n)(1+u_{n-2})$ 是正数。

阶段 3　两个单调子数列

> **想一想**　第 (i) 部分的公式说明，$u_{n+2}-u_n$ 是 u_n-u_{n-2} 的正数倍。由此你能看出 u_3-u_1、u_5-u_3、u_7-u_5、...的正负号有甚么关系吗?

它们的正负号全部相同！由于 $u_3-u_1=\frac{3}{2}-1>0$，由第 (i) 部分和归纳法可得 $u_5>u_3$、

$u_7 > u_5$，如此类推：奇数项是递增的。它们有上界 2，所以根据定理，它们趋向一个极限；设它为 L。

同样地，$u_4 - u_2 = \frac{5}{3} - 2 < 0$，所以偶数项是递减的。它们有下界 1。定理是就递增数列而言的，但 $-u_2, -u_4, -u_6, \ldots$ 是递增的，且有上界 -1，所以它收敛，因而偶数项也收敛；设它们的极限为 M。

阶段 4 求出极限

想一想 要求出 L，你需要一条把 u_{n+2} 与 u_n 直接联系起来的方程。你能从 $(*)$ 得出这样的方程吗?

运用 $(*)$ 两次，

$$u_{n+2} = 1 + \frac{1}{u_{n+1}} = 1 + \frac{1}{1 + 1/u_n} = 1 + \frac{u_n}{u_n + 1} = \frac{2u_n + 1}{u_n + 1}.$$

令 $n \to \infty$ 而只取奇数值，则 u_n 和 u_{n+2} 都趋向 L，所以

$$L = \frac{2L + 1}{L + 1} \quad \Longrightarrow \quad L^2 + L = 2L + 1 \quad \Longrightarrow \quad L^2 - L - 1 = 0.$$

方程的根是 $\frac{1 \pm \sqrt{5}}{2}$，而由于 $L \geq 1$，必有 $L = \frac{1+\sqrt{5}}{2} = \varphi$。对偶数值作完全相同的论证，可得 $M = \varphi$。

想一想 奇数项趋向 φ，偶数项也趋向 φ。为甚么这就表示整个数列趋向 φ?

直观地说，每一项不是奇数项就是偶数项，而两类项都在逼近 φ。严格地说：给定任意容许误差 $\varepsilon > 0$，奇数项由某个位置 N_1 起与 φ 的距离都小于 ε，偶数项则由某个位置 N_2 起如此。所以由 $\max(N_1, N_2)$ 起，*所有*项与 φ 的距离都小于 ε，而这正是收敛的意思。（这就是 Section 2.1 中“取较大的那个门槛”的做法。）

阶段 5 从别处出发

取 $u_1 = 3$ 时，数列是

$$3, \quad \frac{4}{3}, \quad \frac{7}{4}, \quad \frac{11}{7}, \quad \frac{18}{11}, \quad \frac{29}{18}, \quad \cdots$$

现在 $u_1 = 3$ 在 $[1, 2]$ 之外，所以第 (ii) 部分的归纳法无法起步。但 $u_2 = \frac{4}{3}$ 在区间之内，归纳法由此开始便行得通：对所有 $n \geq 2$ 均有 $1 \leq u_n \leq 2$。第 (i) 部分只用到由 u_{n-2} 起的各项，而且只需要它们是正数，而它们确实全是正数。这一次 $u_3 - u_1 = \frac{7}{4} - 3 < 0$ 而 $u_4 - u_2 = \frac{11}{7} - \frac{4}{3} > 0$，所以奇数项递减，偶数项递增：角色互换了，但每个子数列（subsequence）仍然是单调且有界（bounded）的。两个极限满足同一条方程，而且都是正数，所以都是 φ。结论**没有**改变。 ■

回顾

- **数列不单调时，找出单调的部分。**把数列分成奇数项和偶数项，锯齿状的数列就变成了两个规矩的数列。
- **先证明收敛，再求极限。**方程 $L = (2L+1)/(L+1)$ 只有在我们知道 L 存在之后才站得住脚。引子中那个快捷的论证所略过的，正是这一步。
- **检查该取哪个根。**极限方程有两个解；第 (ii) 部分的上下界告诉我们哪一个才对。

背后的数学

蛛网图与不动点

每个由 $u_{n+1} = f(u_n)$ 定义的数列，都可以画成一幅蛛网图（*cobweb diagram*）（图 2.2.1）。画出 $y = f(x)$ 和 $y = x$。由 x 轴上的 u_1 出发，垂直走到曲线（高度为 $u_2 = f(u_1)$），再水平走到直线 $y = x$（这样便移到 $x = u_2$），然后再垂直走到曲线，如此类推。这里 $f(x) = 1 + 1/x$ 是递减的，所以路径会向内盘旋，交替地过高和过低，趋向曲线与直线相交的那一点：不动点（*fixed point*）φ，该处 $f(\varphi) = \varphi$。

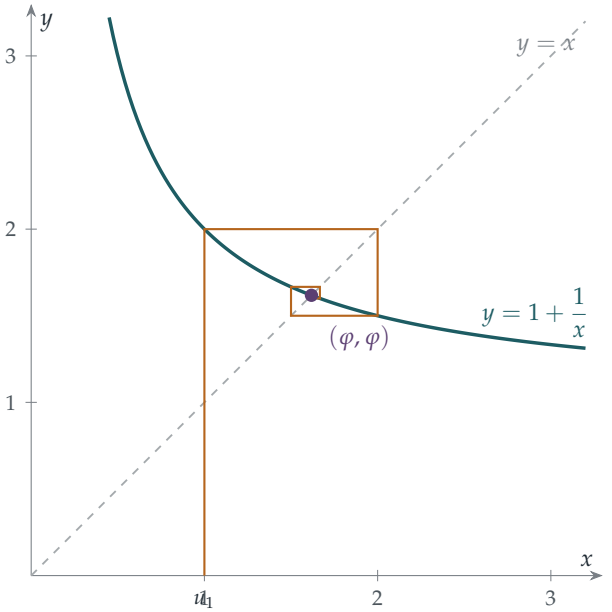


图 2.2.1　$u_{n+1} = 1 + 1/u_n$ 由 $u_1 = 1$ 出发的蛛网图。路径向上走到曲线，再横向走到直线 $y = x$，如此不断重复。由于曲线向下倾斜，路径会向内盘旋，趋向不动点 $\varphi \approx 1.618$，并且交替落在两侧：这就是奇数项与偶数项的锯齿状摆动。

有多快，为甚么

第 (i) 部分中的因子 $(1+u_n)(1+u_{n-2})$ 从何而来？在不动点附近，$u_{n+1}-\varphi = f(u_n)-f(\varphi) \approx f'(\varphi)(u_n-\varphi)$，而 $f'(\varphi) = -1/\varphi^2 \approx -0.38$。所以每走一步，误差大约乘以 -0.38：正负号反转（锯齿状摆动），大小缩减（收敛）。走两步则乘以 $1/\varphi^4 \approx 0.146$，这正是 $1/\big((1+u_n)(1+u_{n-2})\big)$ 所趋向的值。到了大学，这会发展成*压缩映射定理*（*contraction mapping theorem*）：若在不动点附近有 $|f'| \le k < 1$，则迭代会收敛至该不动点，而误差至少以几何级数的速度缩小。它是数学中最有用的定理之一，也是用来证明微分方程有解的方法（第 8 章）。在 Section 2.8 中，我们会遇到一个不动点*并不*吸引的映射，结果截然不同。

最无理的数

各项 $u_n = F_{n+1}/F_n$ 是同等大小的分数之中对 φ 最好的逼近，但所谓“最好”其实并不太好：φ 比任何其他数都更难用分数逼近。原因可从它的连分数（continued fraction）$[1;1,1,1,\dots]$ 看出：连分数中出现大的项，就表示有非常好的逼近（试想 $\pi = [3;7,15,1,292,\dots]$ 和极佳的 $\frac{355}{113}$），而 φ 的每一项自始至终都是最小的。赫维兹（Adolf Hurwitz）于 1891 年证明，对每个无理数（irrational）x，都有无穷多个分数满足 $|x-p/q| < 1/(\sqrt{5}\,q^2)$，而且 $\sqrt{5}$ 不能再改进，原因正正在于 φ。这种“最大无理性”，就是黄金角出现在向日葵种子螺旋图案中的原因。

思维技巧

单调有界必收敛

要在不知道极限的情况下证明数列收敛，可找出一个单调有界的数列，或者把数列分成若干单调有界的部分。

只有在知道极限存在之后，才可以在递推关系（recurrence）中“令 $n \to \infty$”来求出极限。在大学里，这个定理是实数完备性（*completeness*）的推论，完备性正是 $\mathbb{R}$ 有别于 $\mathbb{Q}$ 的性质，你会看到它由公理出发的证明。

延伸练习

1 藏在其中的斐波那契数。 证明 $u_n = F_{n+1}/F_n$，其中 $F_1 = F_2 = 1$ 及 $F_{n+1} = F_n + F_{n-1}$。

2 **误差缩小。**证明 $u_{n+1} - \varphi = -\dfrac{u_n - \varphi}{\varphi\, u_n}$，并由此推出对所有 n 均有 $|u_n - \varphi| \le \varphi^{-(n-1)}$。（利用 $\varphi^2 = \varphi + 1$ 及 $u_n \ge 1$。）

3 **令数列中断。**证明若 $u_1 = \frac{1-\sqrt{5}}{2}$，则数列是常数数列。求出所有使得某个 $u_n = 0$、以致数列无法继续下去的初始值 u_1。（提示：由 $u_n = 0$ 倒推回去。）

4 **$\sqrt{2}$ 的近亲。**设 $v_1 = 1$ 及 $v_{n+1} = 1 + \dfrac{1}{1 + v_n}$。仿照本题的方法，证明 $v_n \to \sqrt{2}$，并把 $\sqrt{2}$ 写成连分数。

延伸阅读

- **Mario Livio, *The Golden Ratio***（Broadway）。讲述 φ 的历史，把真正的数学与众多迷思区分开来。
- **C. D. Olds, *Continued Fractions***（Mathematical Association of America）。一本简短而非常易读的入门书，对象是能力较强的中学生。
- **J. C. Burkill, *A First Course in Mathematical Analysis***（Cambridge University Press）。严谨地处理单调数列与完备性，并附有大量习题。

2.3 一个发散级数与它的收敛近亲

题目

STEP III 2014，第 8 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/14-s3-q8`

引子

调和级数（harmonic series）$1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\cdots$ 是最著名的例子：它的各项缩小至零，但总和却无止境地增长。它增长得慢得离谱：要超过 10，需要多于 12,000 项；要超过 100，则需要大约 1.5×10^{43} 项。但它终究会超过任何一个数。

现在把分母含有数字 2 的项全部丢掉：$\frac{1}{2}$、$\frac{1}{12}$、$\frac{1}{20}$ 至 $\frac{1}{29}$、$\frac{1}{32}$ 等等统统删去。你也许会以为，对一个本来就是无穷大的和来说，删去十分之一左右的项不会有多大分别。事实却是，余下各项加起来小于 80。这个出人意表的结果由 A. J. Kempner 于 1914 年证明，而本题只用一个简单而漂亮的想法，就把你一步步带到这个结论：把和式切成一块一块。

动笔之前

想一想 第一个不等式谈的是由 $r=k^n$ 至 $r=k^{n+1}-1$ 的那一块。取 $k=2$ 和 $n=3$，这一块包含哪些项？共有多少项？

它们是 $f(8), f(9), \ldots, f(15)$，共八项。一般而言，这一块由 k^n 排到 $k^{n+1}-1$，所以共有 $k^{n+1}-k^n=k^n(k-1)$ 项。不等式所说的是：这一块的和介乎（项数）×（各项可能取的最小值）与（项数）×（最大的一项）之间。

逐步解题

阶段 1 估计一块的上下界

由于 f 是递减（decreasing）的，这一块内的每一项都满足

$$f(k^{n+1}) < f(k^{n+1}-1) \le f(r) \le f(k^n) \qquad (k^n \le r \le k^{n+1}-1).$$

把这 $k^n(k-1)$ 项加起来，便得到该不等式。 ■

阶段 2　调和级数

> **想一想**　取 $k=2$ 和 $f(r)=1/r$，由 2^n 至 $2^{n+1}-1$ 的那一块，两个界分别是甚么？

下界（lower bound）是 $2^n \cdot \frac{1}{2^{n+1}} = \frac{1}{2}$，上界（upper bound）是 $2^n \cdot \frac{1}{2^n} = 1$。所以每一块，不论排得多后，其贡献都介乎 $\frac{1}{2}$ 与 1 之间（图 2.3.1）。$n = 0, 1, \dots, N$ 的各块恰好拼合成由 $r=1$ 至 $r = 2^{N+1}-1$ 的和，所以把 $N+1$ 块加起来，便得

$$\frac{N+1}{2} \le \sum_{r=1}^{2^{N+1}-1} \frac{1}{r} \le N+1.$$

当 $N \to \infty$ 时，左边无止境地增长，所以部分和（partial sum）无界，级数发散（diverge）。 ■

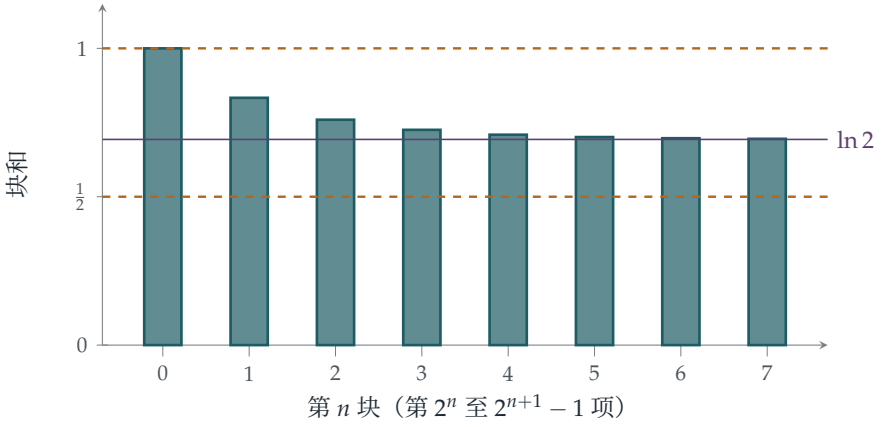


图 2.3.1　把调和级数切成分别有 $1, 2, 4, 8, \dots$ 项的各块。每一块的和都介乎 $\frac{1}{2}$ 与 1 之间（琥珀色虚线），而事实上各块的和会逐渐趋近 $\ln 2 \approx 0.693$（紫红色线）。无穷多块，每块至少值 $\frac{1}{2}$，总和不可能有限。

阶段 3　立方倒数之和

现在取 $f(r) = 1/r^3$，仍然用 $k=2$。第 n 块的上界是

$$2^n \cdot \frac{1}{(2^n)^3} = \frac{1}{4^n}.$$

$n = 0, 1, 2, \ldots$ 的各块把每个正整数恰好覆盖一次，所以

$$\sum_{r=1}^{\infty} \frac{1}{r^3} \le \sum_{n=0}^{\infty} \frac{1}{4^n} = \frac{1}{1 - \frac{1}{4}} = \frac{4}{3} = 1\tfrac{1}{3}. \qquad \blacksquare$$

（严格来说，这只证明了每个部分和都不超过 $\frac{4}{3}$；一个递增而有上界的数列必定收敛（converge），所以这个无穷和存在，而且不超过 $\frac{4}{3}$。它的真正数值是 1.2020 . . .）

阶段 4 数一数不含 2 的数

> **想一想** 如果容许在前面补零（例如把 7 写成 007），由 0 至 999 的整数之中，有多少个可以只用数字 $0, 1, 3, 4, 5, 6, 7, 8, 9$ 写出来?

三个数位每个都有 9 种选择，所以这样的数字串共有 9^3 个。它们与由 0 至 999 之中不含 2 的数一一对应，再剔除 0（它不是正数），$S(1000)$ 中便剩下 $9^3 - 1 = 728$ 个数。 $\blacksquare$

阶段 5 Kempner 级数

分块的想法现在同样适用，只是改为按位数来分块，这等同于取 $k = 10$。

> **停一停，试一试** 有多少个 d 位数不含 2? 而一个 d 位数最小可以是多少?

首位数字可以是 $1, 3, 4, \ldots, 9$ 中任何一个（八种选择：不能是 0，也不能是 2），其余 $d - 1$ 个数位各有九种选择。所以这样的数共有 $8 \times 9^{d-1}$ 个，而每一个都至少是 10^{d-1}。因此，$S(n)$ 中的 d 位数对 $\sigma(n)$ 的贡献至多是

$$8 \times 9^{d-1} \times \frac{1}{10^{d-1}} = 8\left(\frac{9}{10}\right)^{d-1}$$

。对所有位数 $d = 1, 2, 3, \ldots$ 求和，

$$\sigma(n) \le \sum_{d=1}^{\infty} 8\left(\frac{9}{10}\right)^{d-1} = \frac{8}{1 - \frac{9}{10}} = 80.$$

至于严格不等式，只须留意大部分的界都远远未达到相等（例如 $\frac{1}{3} < 1$，而我们却用 1 作为每个一位数项的上界），所以对所有 n 都有 $\sigma(n) < 80$。 $\blacksquare$

回顾

- **把各项分组。**两个和逐项处理都不容易。把它们分成大小合适的块之后，每个和都可以与一个等比级数（geometric series）比较。
- **怎样分块视乎问题而定。**2 的幂适合 $1/r$ 和 $1/r^3$；10 的幂则适合一道关于十进制数字的题目。

- **有界而递增。**两个收敛证明都建基于与 Section 2.2 相同的原理：正项的部分和是递增的，所以只需要找到一个界便足够。

背后的数学

凝聚判别法

题目中的第一个不等式正是*柯西凝聚判别法*（Cauchy's condensation test）的核心：若 f 为正且递减，则

$$\sum_{r=1}^{\infty} f(r) \text{ 收敛} \iff \sum_{n=0}^{\infty} 2^n f(2^n) \text{ 收敛。}$$

它把困难的级数变成容易的级数。应用于 $f(r)=1/r^p$，便得到 $\sum 2^n \cdot 2^{-np} = \sum (2^{1-p})^n$，这是一个等比级数，恰好在 $p>1$ 时收敛。应用于 $1/(r\ln r)$，可知连这个级数也发散；应用于 $1/(r(\ln r)^2)$，则可知这个级数收敛。这个论证可追溯至奥雷姆（Nicole Oresme），他在大约 1350 年正是用图 2.3.1 中的分块方法，证明了调和级数发散。

调和级数发散得有多慢

图 2.3.1 显示每一块的贡献大约是 $\ln 2$，所以首 N 项之和应该大约是 $\log_2 N \times \ln 2 = \ln N$。事实上

$$1+\frac{1}{2}+\cdots+\frac{1}{N} = \ln N + \gamma + (\text{某个趋向 } 0 \text{ 的量}),$$

其中 $\gamma = 0.5772\ldots$ 是欧拉－马斯刻若尼常数（Euler–Mascheroni constant）。至今没有人知道 γ 是否有理数（rational）。你会在大学一年级的分析课中用积分证明这条公式，而调和数（harmonic numbers）会在 Section 9.11 和 Section 11.12 再次出现。

大多数的数都含有 2

Kempner 级数为甚么收敛？因为在大数之中，“不含 2 的数”非常罕有：在小于 10^d 的数之中，不含 2 的所占比例至多是 $(0.9)^d$，而这个比例趋向 0。在 100 位数之中，只有不足 0.003% 能避开数字 2。所以我们删去的并不是十分之一的项，而是几乎所有的项。我们求得的界 80 相当宽松：1979 年，Robert Baillie 用一种加速方法把这个和计算到小数点后许多位，其值约为 19.257。

幂的倒数之和

第 (ii) 部分的和是 $\zeta(3)$，即我们在 Section 1.2 遇过的黎曼（Riemann）zeta 函数的一个值。在 Section 2.4 我们会求出 $\zeta(2)=\pi^2/6$ 的准确值，而欧拉（Euler）亦找到了 $\zeta(4), \zeta(6), \ldots$ 的公式。但至今没有人找到 $\zeta(3)$ 的公式。1978 年，Roger Apéry 证明了 $\zeta(3)$ 是无理数

(irrational)，震惊了整个数学界；他所用的积分与 Section 2.9 中的颇为相似。至于 $\zeta(5)$ 是否无理数，至今仍然未知。

思维技巧

分块比较

要判断一个正项级数是否收敛，可以把各项分成若干块，使每块内的项大致不变，然后把每一块与一个简单的级数比较。

比较判别法（comparison test）是大学一年级分析课的家常便饭。技巧在于选择与甚么比较：在这里，就是选大小为 2^n 或 10^d 的块，块内各项的变化不超过一个固定的倍数。

延伸练习

1 **感受一下有多慢。**利用 $\sum_{r\le N} 1/r \approx \ln N + 0.5772$，估计调和级数需要多少项才超过 10，以及需要多少项才超过 100。

2 **判别法的实际运用。**用第 (i) 部分的方法，取 $f(r) = 1/r^2$，证明 $\sum 1/r^2 \le 2$。然后把第一项分开处理，从而改进这个界。

3 **在收敛与发散之间。**利用 $k = 2$ 的分块不等式，证明 $\sum_{r\ge 2} \dfrac{1}{r \ln r}$ 发散，而 $\sum_{r\ge 2} \dfrac{1}{r(\ln r)^2}$ 收敛。

4 **排除更多数字。**较难 证明：十进制表示中既不含 2 亦不含 7 的所有正整数，其倒数（reciprocal）之和小于 35。（你看得出 35 是怎样来的吗?）

延伸阅读

- **Julian Havil, *Gamma: Exploring Euler's Constant***（Princeton University Press）。讲述调和级数、γ 以及更多相关课题，为程度较高的中学生而写。
- **Mary Hart, *Guide to Analysis***（Palgrave）。一本简明的大学一年级教科书，其中论级数与比较判别法的一章写得十分清晰。
- **Robert Baillie, "Sums of reciprocals of integers missing a given digit"**, *American Mathematical Monthly* 86 (1979)。一篇短论文（论缺少某个指定数字的整数的倒数之和），读完本节后你便有能力读懂。

2.4 $\pi^2/6$

题目

STEP III 2018，第 7 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/18-s3-q7`

引子

1650 年，Pietro Mengoli 提出了一个简单的问题：

$$1+\frac{1}{4}+\frac{1}{9}+\frac{1}{16}+\frac{1}{25}+\cdots ?$$

的值是甚么？各项缩小得够快，足以令这个和收敛（converge）（Section 2.3 已证明它不超过 2），但没有人能求出它的值。伯努利（Bernoulli）兄弟试过，却无功而还；雅各布·伯努利（Jacob Bernoulli）更公开恳求他人相助。这个问题后来以伯努利兄弟的家乡命名，称为巴塞尔问题（*Basel problem*）。

1734 年，年轻的欧拉（Leonhard Euler）解决了这个问题，答案令他一举成名：这个和等于 $\pi^2/6$。π 明明是一个关于圆的数，究竟为甚么会出现在平方倒数之和里？本题给出一个完整而严格的证明，所用的工具全部不超出高级程度（A-level）的范围：棣美弗定理（De Moivre's theorem）、多项式（polynomial）的根之和，以及夹逼（squeeze）。

动笔之前

在处理细节之前，值得先看看整体的计划。

- 第 (i) 部分构造一个多项式，它的根就是 $\cot^2\left(\frac{m\pi}{2n+1}\right)$ 这些数。
- 第 (ii) 部分利用系数（coefficient），在不知道各个根的情况下把这些根加起来。
- 第 (iii) 部分利用“当 θ 很小时 $\cot\theta \approx 1/\theta$”这个事实，把 $\cot^2$ 值之和变成 $1/m^2$ 值之和，然后夹逼。

> **想一想**　若多项式 $a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0$ 的根为 $r_1, \ldots, r_n$，那么 $r_1 + \cdots + r_n$ 以系数表示是甚么？

答案是 $-a_{n-1}/a_n$。这个事实正是第 (ii) 部分的动力所在。

逐步解题

阶段 1　棣美弗定理

> **想一想**　把 $\cot\theta + \mathrm{i}$ 写成单一分数。你留意到分子有甚么特别？

$\cot\theta + \mathrm{i} = \dfrac{\cos\theta + \mathrm{i}\sin\theta}{\sin\theta}$，同样地 $\cot\theta - \mathrm{i} = \dfrac{\cos\theta - \mathrm{i}\sin\theta}{\sin\theta}$。根据棣美弗定理，

$$(\cot\theta \pm \mathrm{i})^{2n+1} = \frac{\cos(2n+1)\theta \pm \mathrm{i}\sin(2n+1)\theta}{\sin^{2n+1}\theta}.$$

两式相减，余弦项互相抵消，正弦项则变成两倍，再除以 $2\mathrm{i}$ 便得

$$\frac{\sin(2n+1)\theta}{\sin^{2n+1}\theta}. \qquad \blacksquare$$

阶段 2　多项式

现在改用二项式定理（binomial theorem）展开左边。设 $c = \cot\theta$ 及 $N = 2n+1$。

> **想一想**　在 $(c+\mathrm{i})^N - (c-\mathrm{i})^N$ 中，两个二项展开式有哪些项互相抵消，又有哪些项变成两倍？

含 i^j 的项在第一个展开式中带 $+$ 号，在第二个展开式中则带 $(-1)^j$。因此 j 为偶数的项互相抵消，j 为奇数的项变成两倍：

$$(c+\mathrm{i})^N - (c-\mathrm{i})^N = 2\sum_{j\ \text{为奇数}} \binom{N}{j} c^{N-j}\,\mathrm{i}^j.$$

取 $j = 2k+1$，则 $\mathrm{i}^j = \mathrm{i}\,(-1)^k$ 及 $c^{N-j} = c^{2n-2k} = (c^2)^{n-k}$。除以 $2\mathrm{i}$，得

$$\frac{\sin(2n+1)\theta}{\sin^{2n+1}\theta} = \binom{2n+1}{1}x^n - \binom{2n+1}{3}x^{n-1} + \cdots + (-1)^n,$$

其中 $x = \cot^2\theta$。

> **停一停，试一试** 左边在哪些 θ 的值时等于零？利用这一点找出多项式的 n 个不同的根。

当 $\sin(2n+1)\theta = 0$ 而 $\sin\theta \neq 0$ 时，左边等于零，例如当 $\theta = \frac{m\pi}{2n+1}$，其中 $m = 1, 2, \ldots, n$。这些角全部严格介乎 0 与 $\frac{\pi}{2}$ 之间，而 cot 在这个范围内为正且严格递减，所以 $\cot^2\left(\frac{m\pi}{2n+1}\right)$ 这 n 个值各不相同。n 次多项式最多有 n 个根，所以这些就是它的全部根。 ■

阶段 3 把根加起来

根之和等于 $-$(以下一项的系数：x^{n-1})/(以下一项的系数：x^n)：

$$\sum_{m=1}^{n} \cot^2\left(\frac{m\pi}{2n+1}\right) = \frac{\binom{2n+1}{3}}{\binom{2n+1}{1}} = \frac{(2n+1)(2n)(2n-1)}{6\,(2n+1)} = \frac{n(2n-1)}{3}. \qquad \blacksquare$$

以 $n = 1$ 验算：$\cot^2(\pi/3) = \frac{1}{3}$，而 $\frac{1\times 1}{3} = \frac{1}{3}$。

阶段 4 夹逼

由 $0 < \sin\theta < \theta < \tan\theta$ 出发，取倒数（不等号方向因而倒转）再平方：

$$\cot^2\theta < \frac{1}{\theta^2} < \frac{1}{\sin^2\theta} = 1 + \cot^2\theta.$$

> **想一想** 代入 $\theta = \frac{m\pi}{2n+1}$，并对 $m = 1, \ldots, n$ 求和。外侧的两个和变成甚么？中间的和与 $\sum 1/m^2$ 有甚么关系？

中间的和是 $\displaystyle\sum_{m=1}^{n} \frac{(2n+1)^2}{m^2\pi^2}$。根据第 (ii) 部分，左边的和是 $\frac{n(2n-1)}{3}$，而右边的和是 $n + \frac{n(2n-1)}{3} = \frac{2n(n+1)}{3}$。各项同乘以 $\frac{\pi^2}{(2n+1)^2}$：

$$\frac{\pi^2}{3}\cdot\frac{n(2n-1)}{(2n+1)^2} < \sum_{m=1}^{n}\frac{1}{m^2} < \frac{\pi^2}{3}\cdot\frac{2n(n+1)}{(2n+1)^2}.$$

当 $n \to \infty$ 时，$\frac{n(2n-1)}{(2n+1)^2}$ 和 $\frac{2n(n+1)}{(2n+1)^2}$ 都趋向 $\frac{2}{4} = \frac{1}{2}$（把分子和分母同除以 n^2）。所以上下两个界都趋向 $\frac{\pi^2}{6}$，而部分和（partial sum）被夹在两者之间，亦必定趋向 $\frac{\pi^2}{6}$（图 2.4.1）。■

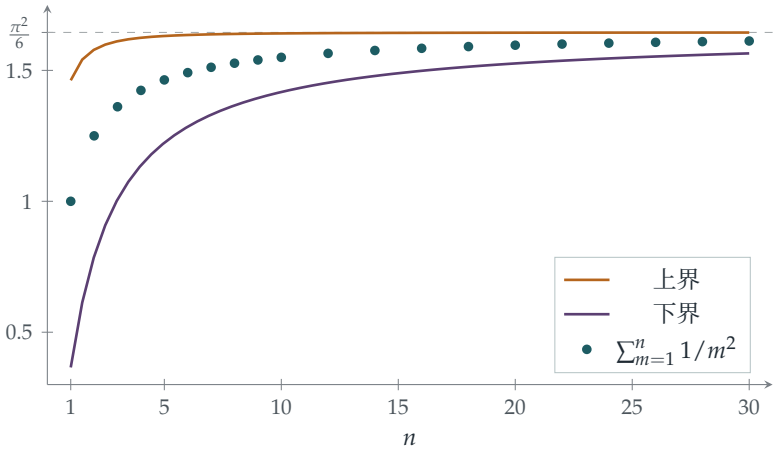


图 2.4.1　第 (iii) 部分的夹逼。$\sum 1/m^2$ 的部分和（蓝绿色点）被夹在下界（紫红色）与上界（琥珀色）之间，而两个界都趋向 $\pi^2/6 \approx 1.645$。

回顾

- **从系数取得信息。**我们始终没有把 $\cot^2(m\pi/(2n+1))$ 这些根化成任何更简单的形式，却准确地求出了它们的和。
- **先近似，再夹逼。**$\cot\theta$ 只是近似等于 $1/\theta$，但误差在两个方向上都受到精确的不等式控制。
- **由有限到无限。**除了最后一步，每一步处理的都是有限和。无穷和只在我们于不等式中令 $n \to \infty$ 时才出现，这正是处理无限的稳妥方法。

背后的数学

欧拉的大胆论证

以现代标准来看，欧拉本人在 1734 年的论证并不严格，却令人目眩神迷。一个多项式若以 $\pm a, \pm b, \pm c, \ldots$ 为根，且在 0 处取值 1，便可分解为 $(1 - x^2/a^2)(1 - x^2/b^2)\cdots$。函数 $\frac{\sin x}{x}$ 在 0 处取值 1，而它的根是 $\pm\pi, \pm 2\pi, \pm 3\pi, \ldots$，于是欧拉把它当作一个“无限多项式”，写下

$$\frac{\sin x}{x} = \left(1 - \frac{x^2}{\pi^2}\right)\left(1 - \frac{x^2}{4\pi^2}\right)\left(1 - \frac{x^2}{9\pi^2}\right)\cdots .$$

右边 x^2 的系数是 $-\frac{1}{\pi^2}\left(1 + \frac{1}{4} + \frac{1}{9} + \cdots\right)$，而由麦克劳林级数（Maclaurin series）$\frac{\sin x}{x} = 1 - \frac{x^2}{6} + \cdots$ 可知它是 $-\frac{1}{6}$。两相比较便得 $\pi^2/6$。我们这个 STEP 证明正是同一想法的

有限而严格的版本：一个多项式，它的根与 π 的倍数息息相关，再加上它的第二个系数。

平方倒数以外

欧拉其后还求出 $\sum 1/m^4 = \pi^4/90$、$\sum 1/m^6 = \pi^6/945$，而一般而言 $\sum 1/m^{2k}$ 是 π^{2k} 的有理数倍，当中涉及伯努利数（*Bernoulli numbers*）。奇数次幂则是另一回事：正如我们在 Section 2.3 所见，没有人知道 $\sum 1/m^3$ 的公式。这些和就是黎曼 ζ 函数（Riemann zeta function）的值 $\zeta(2), \zeta(4), \ldots$，而这个函数在复数自变量上的表现，支配着质数的分布（Section 1.2）。

日后在哪里再遇上

在大学里，你会至少再遇到 $\zeta(2) = \pi^2/6$ 的三个证明：一个来自傅里叶级数（*Fourier series*）（把函数写成正弦和余弦之和），一个来自二重积分（double integral），还有一个来自复分析（complex analysis）。这个数也在概率中出现：随机选出的两个整数没有公因数的机会是 $6/\pi^2 \approx 0.61$。（质数 p 同时整除两者的概率是 $1/p^2$，所以没有任何质数同时整除两者的机会是 $\prod_p (1 - 1/p^2)$，而欧拉证明了这个乘积等于 $1/\zeta(2)$。）π 在概率极限中的另一次现身，是 Section 12.23 的主题。

思维技巧

夹逼，以及根的对称函数（symmetric functions）

要求一个极限，可把该量夹在两个极限相同而较简单的式子之间。

而当你需要求一些难缠的数之和（或积）时，不妨试找一个以它们为根的多项式：系数会把这些数的对称组合白白送给你。

延伸练习

1 **1992 年的版本。**试做 STEP III 1992, 第 8 题（`step.maths.org/questions/92-s3-q8`）。它利用一个与本节相似的多项式，求出正切平方与余割平方的准确和。（提示：若你知道一个多项式的根是 r_i，那么哪个多项式的根是 $1/r_i$?）

2 **奇数的平方。**由 $\sum 1/m^2 = \pi^2/6$ 推导出 $1 + \frac{1}{9} + \frac{1}{25} + \frac{1}{49} + \cdots = \frac{\pi^2}{8}$。

3 **四次方。** 较难 利用 $\sum r_i^2 = (\sum r_i)^2 - 2\sum_{i<j} r_i r_j$ 和多项式的系数，求 $\sum_{m=1}^{n} \cot^4\!\left(\frac{m\pi}{2n+1}\right)$，并改动夹逼的论证以证明 $\sum 1/m^4 = \pi^4/90$。

4 **有多快？** 利用第 (iii) 部分的下界，证明 $\frac{\pi^2}{6} - \sum_{m=1}^{n} \frac{1}{m^2} < \frac{\pi^2}{4n}$。按这个结果，$n$ 要多大才能保证部分和与 $\pi^2/6$ 相差不超过 0.001？试与真正的误差比较，后者接近 $1/n$。

延伸阅读

- **William Dunham, *Euler: The Master of Us All***（Mathematical Association of America）。第 3 章讲述巴塞尔问题的故事，并完整交代当中的数学。
- **Aigner & Ziegler, *Proofs from THE BOOK***，第 9 章“Three times $\pi^2/6$”。当中包括本节的证明。
- **Paul Nahin, *Dr. Euler's Fabulous Formula***（Princeton University Press）。内容涵盖复数、傅里叶级数，以及类似本节的求和。

2.5 韦达乘积

题目

STEP III 2022，第 4 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2022 年 STEP 试卷（zip 文件）内的 STEP III 试卷：

`www.ocr.org.uk/Images/696587-2022-past-papers.zip`

引子

这道题的最后一行，是数学中最优美的公式之一，也是同类公式中最古老的一条。它由法国律师兼数学家韦达（François Viète）于 1593 年发表，是历史上第一次有人把 π 准确地写成一个无穷过程的结果。在韦达之前，π 只能求近似值；在他之后，它可以用一条公式来定义，只要一直算下去，就能得出每一个数位。

这条公式看来很神秘：一层套一层的 2 的平方根，即嵌套平方根（nested square roots），究竟从何而来？这道题揭示，它其实来自反复运用倍角公式（double-angle formula），而且同一套方法还顺道得出另一条名气较小的公式，是关于 $\ln 2$ 的。

动笔之前

想一想 sinh 的倍角公式是甚么？（如果不肯定，可把 $\sinh 2y$ 和 $\sinh y \cosh y$ 用指数函数写出来，再作比较。）

$\sinh 2y = 2\sinh y \cosh y$，与 $\sin 2y = 2\sin y \cos y$ 完全一样。整道题就是靠反复运用这一条恒等式（identity）来推动的。题目一直使用双曲函数（hyperbolic functions），到最后一部分我们才转用普通的三角函数。

逐步解题

阶段 1　反复减半

$n = 1$ 的情况就是取 $y = x/2$ 的倍角公式：

$$\sinh x = 2\cosh\left(\frac{x}{2}\right)\sinh\left(\frac{x}{2}\right).$$

想一想　假设公式对 n 成立。最后一个因子是 $\sinh(x/2^n)$。怎样改写它，才能得到 $n+1$ 时的公式？

对它运用倍角公式：$\sinh(x/2^n) = 2\cosh(x/2^{n+1})\sinh(x/2^{n+1})$。代入后得

$$\sinh x = 2^{n+1}\cosh\left(\frac{x}{2}\right)\cdots\cosh\left(\frac{x}{2^n}\right)\cosh\left(\frac{x}{2^{n+1}}\right)\sinh\left(\frac{x}{2^{n+1}}\right),$$

这正是 $n+1$ 时的公式。根据归纳法（induction），公式对所有 n 成立。 ■

至于随后的推论，先把两边除以 x，再把两边乘以 $\dfrac{x/2^n}{\sinh(x/2^n)}$。在右边，$2^n\sinh(x/2^n)\cdot\dfrac{x/2^n}{x\sinh(x/2^n)} = 1$，只剩下各个 cosh 项的乘积。 ■

阶段 2　极限

由级数（series）可知，当 $y \neq 0$ 时，

$$\frac{\sinh y}{y} = 1 + \frac{y^2}{3!} + \frac{y^4}{5!} + \cdots,$$

当 $y \to 0$ 时它趋向 1，所以它的倒数 $y/\sinh y$ 也趋向 1。（如果想严谨一点：尾部 $\frac{y^2}{3!} + \frac{y^4}{5!} + \cdots$ 中的每个阶乘都至少是 6，所以当 $0 < |y| < 1$ 时，尾部至多是 $\frac{1}{6}(y^2 + y^4 + \cdots) = \frac{y^2}{6(1-y^2)}$，而这个量趋向 0。）

现在固定 x，在第 (i) 部分中令 $n \to \infty$。这时 $y = x/2^n \to 0$，所以左边趋向 $\dfrac{\sinh x}{x} \times 1$。右边是部分积（partial product），因此它也趋向 $\dfrac{\sinh x}{x}$。这正是无穷乘积（infinite product）的意思。 ■

阶段 3　$\ln 2$ 的公式

取 $x = \ln 2$，则 $e^{x/2} = \sqrt{2}$，所以

$$\cosh\left(\frac{x}{2}\right) = \frac{2^{1/2} + 2^{-1/2}}{2} = \frac{2+1}{2\sqrt{2}} = \frac{3}{2\sqrt{2}}.$$

同样地，$e^{x/4} = 2^{1/4}$，所以 $\cosh(x/4) = \dfrac{2^{1/4} + 2^{-1/4}}{2}$，把分子和分母同乘以 $2^{1/4}$，便得 $\dfrac{2^{1/2}+1}{2\times 2^{1/4}}$。 ■

想一想 $\cosh(x/2^k)$ 的一般规律是甚么？把所有分母 $2\times 2^{1/2^k}$ 乘在一起，会有甚么结果？

同样的计算显示，对每个 $k\geq 1$，

$$\cosh\left(\frac{x}{2^k}\right)=\frac{1+2^{1/2^{k-1}}}{2}\times 2^{-1/2^k}.$$

（当 $k=1$ 时，这就是 $\frac{1+2}{2}\times 2^{-1/2}=\frac{3}{2\sqrt{2}}$，与前面一致。）对 $k=1,2,3,\ldots$ 连乘，2 的各个幂合并成 $2^{-(1/2+1/4+1/8+\cdots)}=2^{-1}$，而第一个分数是 $\frac{1+2}{2}=\frac{3}{2}$。所以这个无穷乘积是

$$\frac{3}{2}\times\frac{1}{2}\times\frac{1+2^{1/2}}{2}\times\frac{1+2^{1/4}}{2}\times\frac{1+2^{1/8}}{2}\cdots.$$

另一边，$\sinh(\ln 2)=\frac{2-\frac{1}{2}}{2}=\frac{3}{4}$，所以 $\dfrac{\sinh x}{x}=\dfrac{3}{4\ln 2}$。两边的因子 $\frac{3}{4}$ 互相约去，剩下

$$\frac{1}{\ln 2}=\frac{1+2^{1/2}}{2}\times\frac{1+2^{1/4}}{2}\times\frac{1+2^{1/8}}{2}\cdots.$$ ■

阶段 4 韦达公式

想一想 第 (iv) 部分含有 π，所以它必定来自普通的三角函数，而不是双曲函数。如果把 sinh 和 cosh 换成 sin 和 cos，第 (i) 和第 (ii) 部分的一切是否仍然成立？

仍然成立。归纳法只用到倍角公式，而 $\sin 2y=2\sin y\cos y$。极限只用到当 $y\to 0$ 时 $\frac{\sin y}{y}\to 1$，这可以用同样的方法由 $\sin y=y-\frac{y^3}{3!}+\cdots$ 推出。所以当 $x\neq 0$（并且 x 不是 π 的倍数，以确保过程中没有因子为零）时，

$$\frac{\sin x}{x}=\cos\left(\frac{x}{2}\right)\cos\left(\frac{x}{4}\right)\cos\left(\frac{x}{8}\right)\cdots.$$

停一停，试一试 选取 x 使左边等于 $2/\pi$，然后算出最初几个因子。半角公式（half-angle formula）$\cos\frac{\theta}{2}=\sqrt{\frac{1+\cos\theta}{2}}$（适用于 $0\leq\theta\leq\pi$）会派上用场。

取 $x=\frac{\pi}{2}$。这时 $\frac{\sin x}{x}=\frac{1}{\pi/2}=\frac{2}{\pi}$，而各因子是 $\cos\frac{\pi}{4},\cos\frac{\pi}{8},\cos\frac{\pi}{16},\ldots$。第一个是 $\cos\frac{\pi}{4}=\frac{\sqrt{2}}{2}$。如果某个因子是 $\cos\theta=\frac{c}{2}$，由半角公式可知下一个因子是

$$\cos\frac{\theta}{2}=\sqrt{\frac{1+c/2}{2}}=\frac{\sqrt{2+c}}{2}.$$

所以分子依次是 $\sqrt{2}$、$\sqrt{2+\sqrt{2}}$、$\sqrt{2+\sqrt{2+\sqrt{2}}}$、...，每一个都是把前一个加上 2 再开平方而得。这就是韦达公式。 ■

回顾

- **乘积的相消（telescoping）。**第 (i) 部分的恒等式是一个会收拢的乘积，正如某些和式可以裂项相消一样。
- **先有有限的恒等式，再取极限。**与 Section 2.4 一样，每一步对有限的 n 都是准确的，唯一需要的极限就是非常简单的 $\frac{\sin y}{y} \to 1$。
- **双曲函数与圆函数是孪生兄弟。**同一个证明对两者都适用，因为两者满足相同的倍角公式和相同的小角极限。这并非巧合：$\cosh(ix) = \cos x$，你会在 Section 5.7 看到。

背后的数学

阿基米德（Archimedes）的多边形

韦达公式有一个几何意义，可追溯至阿基米德。内接（inscribed）于半径为 1 的圆的正 2^n 边形，面积是 $A_n = 2^{n-1}\sin\frac{2\pi}{2^n}$。把边数加倍，面积就乘以

$$\frac{A_{n+1}}{A_n} = \frac{2^n \sin(\pi/2^n)}{2^{n-1}\cdot 2\sin(\pi/2^n)\cos(\pi/2^n)} = \frac{1}{\cos(\pi/2^n)}.$$

由正方形（$A_2 = 2$）开始，不断把边数加倍，面积便趋向圆的面积 π。所以 $\pi = 2 \div \left(\cos\frac{\pi}{4}\cos\frac{\pi}{8}\cos\frac{\pi}{16}\cdots\right)$，这正是韦达公式。乘积中的每个因子，都是图 2.5.1 中相邻两个多边形的面积之比。

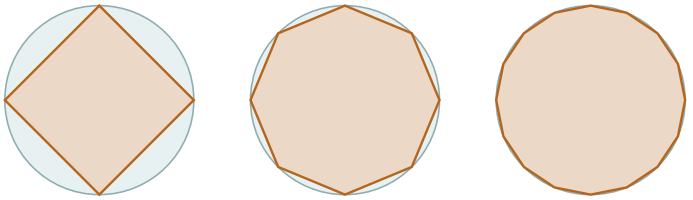


图 2.5.1　内接于半径为 1 的圆的正多边形，边数分别为 4、8 和 16。边数每加倍一次，面积就多除以韦达乘积中的一个因子，而面积不断增大并趋向 π。

无穷乘积

韦达乘积是数学史上第一个无穷乘积。其后陆续有其他例子：沃利斯（John Wallis）在 1655 年发现 $\frac{\pi}{2} = \frac{2}{1}\cdot\frac{2}{3}\cdot\frac{4}{3}\cdot\frac{4}{5}\cdot\frac{6}{5}\cdot\frac{6}{7}\cdots$，而欧拉（Euler）关于 $\frac{\sin x}{x}$ 的乘积已在 Section 2.4

出现过。到了大学，你会学到无穷乘积在甚么情况下收敛（converge）。粗略地说，当 $\prod(1+a_n)$ 中的 a_n 是细小的正数时，它的表现与 $\sum a_n$ 相似，因为 $\ln(1+a_n)\approx a_n$。这里 $1-\cos(\pi/2^n)\approx\frac{1}{2}(\pi/2^n)^2$，每次缩小 4 倍，所以这个乘积收敛得很快。

有多快?

乘了 n 个因子之后，部分积恰好等于 $\dfrac{1}{2^n\sin(\pi/2^{n+1})}$，其误差大约是 $\frac{\pi^2}{6\cdot4^{n+1}}$ 乘以 $\frac{2}{\pi}$。每多一个因子，答案的准确度大约提高 4 倍，即多得大约 0.6 个小数位。十个因子可使 $2/\pi$ 准确至六个小数位。这已经不错，但远远不及 Section 2.6 的方法那么快，那些方法每一步都令准确的位数加倍。

思维技巧

反复运用恒等式

一条把 $f(x)$ 与 $f(x/2)$ 联系起来的恒等式，可以一次又一次地运用；取极限之后，结果往往是一个无穷和或无穷乘积。

找出 n 步之后的“余项”（这里是 $2^n\sinh(x/2^n)$），然后另行求出它的极限。这个模式在分析学中随处可见，由无穷乘积以至分形（fractals）的构造都有它的踪影。

延伸练习

1 **1990 年的版本。**试做 STEP I 1990, 第 8 题（`step.maths.org/questions/90-s1-q8`）。它循一条与本节不同的途径得出韦达乘积，所用的不外乎 sin 的倍角公式。你觉得哪一种方法较为自然?

2 **另一个角。**在 $\frac{\sin x}{x}$ 的乘积中取 $x=\frac{\pi}{3}$，把 $\frac{3\sqrt{3}}{2\pi}$ 写成无穷乘积，并化简最初两个因子。

3 **要多少个因子?** 用计算机找出韦达乘积需要多少个因子，才能使 π 准确至 4 个小数位。解释为甚么每多一个因子大约多得 0.6 个数位。

4 **$\ln 3$ 的公式。**取 $x=\ln 3$ 重做第 (iii) 部分，把 $\frac{1}{\ln 3}$ 写成无穷乘积。需要多少个因子才能使 $\frac{1}{\ln 3}$ 准确至两个小数位? 为甚么这个乘积收敛得比韦达乘积慢?

延伸阅读

- **Petr Beckmann, *A History of* π**（St Martin's Press）。阿基米德、韦达、沃利斯等人的故事，写得兴味盎然。
- **Jonathan Borwein & Peter Borwein, *Pi and the AGM***（Wiley）。关于计算 π 的严谨数学，包括 Section 2.6 的方法。读完本章之后，第 1 章应该读得懂。
- **Eli Maor, *Trigonometric Delights***（Princeton University Press）。书中有一章专门讲述韦达和他的乘积。

2.6 算术几何平均

题目

STEP II 2024，第 8 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2024 年 STEP 试卷（zip 文件）内的 STEP II 试卷：

`www.ocr.org.uk/Images/725753-2024-past-papers.zip`

引子

取两个正数，例如 $\sqrt{2}$ 和 1。把它们换成两者的算术平均（arithmetic mean）和几何平均（geometric mean）。再做一次，然后再做一次。结果如下：

n	x_n（算术）	y_n（几何）
0	1.4142135623730950488	1
1	1.2071067811865475244	1.1892071150027210667
2	1.1981569480946342956	1.1981235214931201226
3	1.1981402347938772091	1.1981402346773072058
4	1.1981402347355922074	1.1981402347355922074

只需四步，两个数便有二十位小数完全相同。它们的共同极限（limit）就是*算术几何平均*（arithmetic–geometric mean，简称 AGM）$M(\sqrt{2},1) = 1.19814\ldots$

1799 年 5 月 30 日，年仅 21 岁的高斯（Carl Friedrich Gauss）在日记中写道，他发现 $1/M(\sqrt{2},1)$ 与一个常数有十一位小数相同，而这个常数出自一个完全不同的问题：一条称为双纽线（lemniscate）的曲线的弧长。他写道：“这必定会开辟一个全新的分析领域。”事实果然如此。本题要证明的，正是高斯这项发现背后的关键事实：AGM 能求出一个用任何常规方法都计算不了的积分。

动笔之前

> **想一想**　为甚么两个不相等的正数的算术平均，总是大于它们的几何平均？你能用一行证明吗？

这就是算术几何平均不等式（AM–GM inequality），而第 (i) 部分的提示正好给出这个一行证明：当 $x \neq y$ 时，$\left(\sqrt{x}-\sqrt{y}\right)^2 > 0$，展开后便是 $x+y > 2\sqrt{xy}$。第 (i) 部分的一切，都由这个不等式，以及“每个平均都介乎原来两个数之间”这个事实推出。

逐步解题

阶段 1　两个数列互相夹紧

若 $y_n < x_n$，则 $\sqrt{x_n} \neq \sqrt{y_n}$，所以 $x_n - 2\sqrt{x_n y_n} + y_n > 0$，即 $x_{n+1} > y_{n+1}$。由于 $y_0 < x_0$，用归纳法（induction）可知对所有 n 都有 $y_n < x_n$。

> **想一想**　为甚么 x_n 与 y_n 的算术平均小于 x_n，而几何平均大于 y_n？

因为 $y_n < x_n$：$x_{n+1} = \frac{x_n+y_n}{2} < \frac{x_n+x_n}{2} = x_n$，而 $y_{n+1} = \sqrt{x_n y_n} > \sqrt{y_n y_n} = y_n$。所以

$$y_0 < y_1 < y_2 < \cdots < x_2 < x_1 < x_0.$$

y_n 递增，并以 x_0 为上界（bounded above），所以根据题目所给的定理，它趋向一个极限 M。

至于两者的差距，

$$x_{n+1} - y_{n+1} < x_{n+1} - y_n = \frac{x_n+y_n}{2} - y_n = \frac{1}{2}(x_n - y_n),$$

而由第一部分可知它是正数。所以差距每次至少减半：$0 < x_n - y_n < (x_0 - y_0)/2^n$，而后者趋向 0。

最后，$x_n = y_n + (x_n - y_n)$ 是一个趋向 M 的数列与一个趋向 0 的数列之和，所以 $x_n \to M$ 亦成立。 ∎

阶段 2　代换

这个代换（substitution）看来无缘无故，但它其实经过精心设计，令根号下的两个括号都变成完全平方（或接近完全平方）。

想一想 设 $t=\frac{1}{2}\left(x-\frac{pq}{x}\right)$，计算 t^2+pq。然后计算 $t^2+\frac{1}{4}(p+q)^2$。两者之中，有没有哪一个可以漂亮地因式分解？

首先，

$$t^2+pq=\frac{1}{4}\left(x^2-2pq+\frac{p^2q^2}{x^2}\right)+pq=\frac{1}{4}\left(x+\frac{pq}{x}\right)^2.$$

其次，

$$\begin{aligned}t^2+\tfrac{1}{4}(p+q)^2&=\frac{1}{4}\left(x^2-2pq+\frac{p^2q^2}{x^2}+p^2+2pq+q^2\right)\\&=\frac{x^4+(p^2+q^2)x^2+p^2q^2}{4x^2}=\frac{(x^2+p^2)(x^2+q^2)}{4x^2}.\end{aligned}$$

另外 $\dfrac{\mathrm{d}t}{\mathrm{d}x}=\dfrac{1}{2}\left(1+\dfrac{pq}{x^2}\right)=\dfrac{x^2+pq}{2x^2}$。

停一停，试一试 把以上结果合起来。当 x 由 0 走到 ∞ 时，t 怎样变化？关于 t 的积分会变成甚么？

当 x 由 0 增加至 ∞ 时，t 由 $-\infty$ 递增至 ∞（因为它的导数是正数）。对 $x>0$，

$$\sqrt{\left(t^2+\tfrac{1}{4}(p+q)^2\right)\left(t^2+pq\right)}=\frac{\sqrt{(x^2+p^2)(x^2+q^2)}}{2x}\cdot\frac{x^2+pq}{2x},$$

所以

$$\frac{\mathrm{d}t}{\sqrt{\cdots}}=\frac{x^2+pq}{2x^2}\cdot\frac{4x^2}{(x^2+pq)\sqrt{(x^2+p^2)(x^2+q^2)}}\,\mathrm{d}x=\frac{2\,\mathrm{d}x}{\sqrt{(x^2+p^2)(x^2+q^2)}}.$$

因此关于 t 的积分等于 $2\mathrm{I}(p,q)$。

另一方面，关于 t 的积分与 I 的形式完全一样，只是 p 和 q 分别换成了 $\frac{1}{2}(p+q)=\mathrm{a}(p,q)$ 和 $\sqrt{pq}=\mathrm{g}(p,q)$，而且积分范围是整条实数线。被积函数（integrand）是 t 的偶函数（even function），所以关于 t 的积分等于 $2\mathrm{I}\big(\mathrm{a}(p,q),\mathrm{g}(p,q)\big)$。两者比较，得

$$\mathrm{I}(p,q)=\mathrm{I}\big(\mathrm{a}(p,q),\mathrm{g}(p,q)\big).\qquad\blacksquare$$

阶段 3 求积分的值

想一想 由 (x_0,y_0) 开始，反复应用这个恒等式。经过 n 步之后，$\mathrm{I}(x_0,y_0)$ 等于甚么？当 $n\to\infty$ 时又会怎样？

对每个 n，$\mathrm{I}(x_0,y_0)=\mathrm{I}(x_1,y_1)=\mathrm{I}(x_2,y_2)=\cdots=\mathrm{I}(x_n,y_n)$。当 $n\to\infty$ 时，两个变量都趋向 M，而

$$\mathrm{I}(M,M)=\int_0^\infty\frac{\mathrm{d}x}{M^2+x^2}=\left[\frac{1}{M}\arctan\frac{x}{M}\right]_0^\infty=\frac{\pi}{2M}.$$

所以

$$\mathrm{I}(x_0, y_0) = \frac{\pi}{2M}. \qquad \blacksquare$$

题目容许我们直接接受“$\mathrm{I}(x_n, y_n) \to \mathrm{I}(M, M)$”这一步，但其实可以简洁地加以证明。把 p 或 q 增大，被积函数便会变小，而 $y_n < x_n$，所以

$$\frac{\pi}{2x_n} = \mathrm{I}(x_n, x_n) < \mathrm{I}(x_n, y_n) < \mathrm{I}(y_n, y_n) = \frac{\pi}{2y_n}.$$

上下两个界都趋向 $\frac{\pi}{2M}$，而对每个 n，$\mathrm{I}(x_n, y_n)$ 都是同一个数 $\mathrm{I}(x_0, y_0)$，所以它必定等于 $\frac{\pi}{2M}$。这又是一次夹逼（squeeze），与 Section 2.4 中的做法一样。

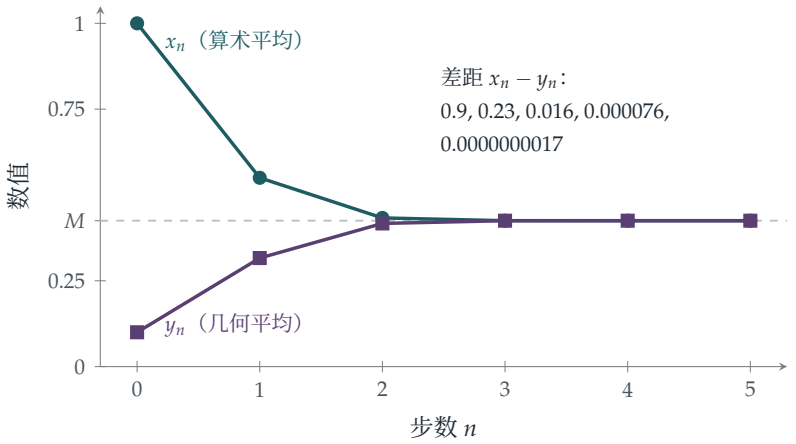


图 2.6.1　$x_0 = 1$ 与 $y_0 = 0.1$ 的算术几何平均。算术平均 x_n（蓝绿色）递减，几何平均 y_n（紫红色）递增，两者以惊人的速度逼近共同极限 $M = 0.425040709\ldots$：一旦两者接近，正确位数每一步大约增加一倍。

回顾

- **两个单调数列互相夹紧。**一个上升，一个下降，而两者的差距缩小至零。要证明两个数列有共同极限，这是最干净利落的方法。
- **不变量 (invariant)。**$\mathrm{I}(x_n, y_n)$ 这个量在每一步都保持不变。在其他一切都在变动时，找出某样固定不变的东西，是数学中最有力的想法之一。
- **巧妙的代换是设计出来的，不是猜出来的。**选用这个代换，正是为了令根号下的式子可以因式分解。

背后的数学

比题目所说的还要快

第 (i) 部分证明了差距每一步至少减半，这相当于每一步大约多得一个二进制位。引子中的表格所显示的却惊人得多：正确位数每一步大约*增加一倍*（$2, 4, 10, 21, 43, \ldots$）。事实上

$$x_{n+1} - y_{n+1} = \frac{1}{2}\bigl(\sqrt{x_n} - \sqrt{y_n}\bigr)^2 = \frac{(x_n - y_n)^2}{2\bigl(\sqrt{x_n} + \sqrt{y_n}\bigr)^2},$$

所以每个差距大约是一个常数乘以前一个差距的*平方*。这称为*二次收敛*（quadratic convergence），我们会在 Section 2.7 中仔细探讨。

椭圆积分

我们的积分 $\mathrm{I}(p, q)$ 其实是一个乔装了的*完全椭圆积分*（complete elliptic integral）。这类积分在计算椭圆的弧长时出现，而且不能用你所认识的函数表示。它最著名的应用是单摆。高级程度（A-level）课程中的周期公式 $T = 2\pi\sqrt{L/g}$ 只是小幅度摆动时的近似。由角度 θ_0 释放的单摆，其准确周期是

$$T = \frac{2\pi\sqrt{L/g}}{M\bigl(1, \cos\frac{\theta_0}{2}\bigr)},$$

这是一个可用 AGM 求值的椭圆积分。当 $\theta_0 = 90^\circ$ 时，真正的周期比简单公式所预测的长约 18%。高斯那则日记所谈的是双纽线，一条 8 字形的曲线，它的总弧长涉及 $\int_0^1 \frac{\mathrm{d}t}{\sqrt{1-t^4}}$；这个积分其实就是乔装了的 $\mathrm{I}(\sqrt{2}, 1)$，所以可以用 AGM 求值。

计算 π

1976 年，Richard Brent 与 Eugene Salamin 各自独立地把高斯的想法变成一个计算 π 的算法（algorithm），每一步都令正确位数增加一倍。二十五步便可得出数千万位。在 1980 及 1990 年代，这类算法曾用于多次刷新纪录的 π 计算；时至今日，凡是需要高精度的地方，类似的二次收敛算法已成为标准工具。试比较 Section 2.5 中的韦达（Viète）乘积，它每一步只多得约 0.6 位。

思维技巧

寻找不变量

当一个过程同时改变许多东西时，找出一个不变的量。然后在过程的终点求它的值，因为那里一切都很简单。

在这里，过程就是 AGM 的每一步，而不变量就是积分 I。在过程的终点，两个变量相等，积分变成初等积分。不变量在数学中随处可见，由力学中的能量守恒（第 10 章），以至证明某些谜题无解。

延伸练习

1 **二次收敛。**证明恒等式 $x_{n+1}-y_{n+1}=\dfrac{(x_n-y_n)^2}{2(\sqrt{x_n}+\sqrt{y_n})^2}$，并由此推出 $x_{n+1}-y_{n+1}\le\dfrac{(x_n-y_n)^2}{8y_0}$。

2 **亲手计算。**用计算机求 $M(2,1)$，准确至六位小数。你需要多少步？

3 **缩放。**证明对 $\lambda>0$ 有 $M(\lambda x,\lambda y)=\lambda M(x,y)$，以及 $\mathrm{I}(\lambda p,\lambda q)=\lambda^{-1}\mathrm{I}(p,q)$。验证这两个事实与 $\mathrm{I}(x_0,y_0)=\pi/(2M)$ 互相吻合。

4 **单摆。**利用上面的公式，求由 60° 释放的单摆的周期比小角度近似的预测值长多少，答案准确至三位有效数字。

延伸阅读

- **David Cox, “The arithmetic–geometric mean of Gauss”**, *L'Enseignement Mathématique* 30 (1984)。讲述这段历史和有关数学的经典文章；开首几节只需要微积分知识。
- **Jonathan Borwein & Peter Borwein, *Pi and the AGM*** (Wiley)。关于 AGM 与 π 的快速计算的权威著作。
- **Jonathan Borwein & Keith Devlin, *The Computer as Crucible*** (A. K. Peters)。讲述实验数学，包括像高斯那样凭着留意数值上的巧合而得出的发现。

2.7 六位还是六百位

题目

STEP II 2023，第 5 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2023 年 STEP 试卷（zip 文件）内的 STEP II 试卷：

`www.ocr.org.uk/Images/707102-2023-past-papers.zip`

引子

有两个求 $\sqrt{2}$ 近似值的简单方法，同样由 1 开始，同样只用加法、乘法和除法。十步之后，第一个方法准确至小数点后约 8 位，第二个方法则准确至超过 700 位。

两者的分别，并不在于第二个方法对付 $\sqrt{2}$ 的手法较为高明，而在于误差的形态不同：第一个方法每走一步，误差乘以一个固定的分数；第二个方法每走一步，误差就被平方一次。这道题目把这一点说得精确，同时引入数值分析（numerical analysis）的一个核心概念；每一部计算机、每一次天气预报、每一个工程模拟，背后都是这门学科。

动笔之前

先以分数形式计算两个数列（sequence）的首几项。

$$x_n:\quad 1,\ \frac{3}{2},\ \frac{7}{5},\ \frac{17}{12},\ \frac{41}{29},\ \frac{99}{70},\ \cdots \qquad y_n:\quad 1,\ \frac{3}{2},\ \frac{17}{12},\ \frac{577}{408},\ \frac{665857}{470832},\ \cdots$$

? **想一想** 你留意到这两列数有甚么特别吗？（看看哪些分数在两列中都有出现。）

每一个 y_n 都在 x 的那一列中出现，只是位置愈来愈后：$y_2 = x_3$，$y_3 = x_7$，一般而言 $y_n = x_{2^n-1}$。第二个方法其实是在同一列近似值之中，以指数速度向前跳。请记住这一点，我们会在本节末解释。

逐步解题

阶段 1　第一个方法的各项不小于 1

我们可以写成 $x_{n+1} = 1 + \dfrac{1}{x_n + 1}$。若 $x_n \geq 1$，则 $x_n + 1 > 0$，所以 $x_{n+1} > 1$。由于 $x_0 = 1$，由归纳法（induction）可知对所有 n 都有 $x_n \geq 1$。 ■

阶段 2　第一个方法的误差

? **想一想**　题目以 $x_n^2 - 2$ 而不是 $x_n - \sqrt{2}$ 来量度误差。试把 $x_{n+1}^2 - 2$ 写成以 x_n 表示的单一分数。分子有甚么特别之处？

$$x_{n+1}^2 - 2 = \frac{(x_n + 2)^2 - 2(x_n + 1)^2}{(x_n + 1)^2} = \frac{x_n^2 + 4x_n + 4 - 2x_n^2 - 4x_n - 2}{(x_n + 1)^2} = \frac{2 - x_n^2}{(x_n + 1)^2}.$$

分子正好是 $-(x_n^2 - 2)$。所以 $x_{n+1}^2 - 2$ 与 $x_n^2 - 2$ 异号（近似值交替落在 $\sqrt{2}$ 的两侧），而且由 $x_n \geq 1$ 可得 $(x_n + 1)^2 \geq 4$，因此

$$|x_{n+1}^2 - 2| = \frac{|x_n^2 - 2|}{(x_n + 1)^2} \leq \frac{1}{4}|x_n^2 - 2|. \qquad \blacksquare$$

阶段 3　十步之后

把上式连续应用十次，得 $|x_{10}^2 - 2| \leq 4^{-10}|x_0^2 - 2| = 4^{-10} = \frac{1}{1048576} < 10^{-6}$。至于正负号：$x_0^2 - 2 = -1 < 0$，而正负号逐项交替，所以对每个偶数 n 都有 $x_n^2 - 2 < 0$。于是 $x_{10}^2 < 2$，两者合起来便得 $2 - 10^{-6} \leq x_{10}^2 \leq 2$。 ■

x_n^2 的误差每一步最多乘以 $\frac{1}{4}$。这称为*线性收敛*（linear convergence）：每一步增加固定数目的准确位数（这里最少是 $\log_{10} 4 \approx 0.6$；实际上较接近 0.77）。

阶段 4　第二个方法的误差

? **想一想**　试把 $y_{n+1} - \sqrt{2}$ 写成单一分数。分子应该会令你想起完全平方。

$$y_{n+1} - \sqrt{2} = \frac{y_n^2 + 2 - 2\sqrt{2}\,y_n}{2y_n} = \frac{\left(y_n - \sqrt{2}\right)^2}{2y_n}. \qquad \blacksquare$$

若 $y_n > 0$，右方为非负数，所以 $y_{n+1} \geq \sqrt{2} > 1$。由于 $y_0 = 1$，由归纳法可知对所有 n 都有 $y_n \geq 1$。 ■（事实上，对所有 $n \geq 1$ 都有 $y_n \geq \sqrt{2}$：走完第一步之后，第二个方法永远从上方逼近 $\sqrt{2}$。）

阶段 5　把误差平方

设 $e_n = y_n - \sqrt{2}$，则当 $n \geq 1$ 时 $e_n \geq 0$，而且因为 $y_n \geq 1$，有 $e_{n+1} = \dfrac{e_n^2}{2y_n} \leq \dfrac{e_n^2}{2}$。

> **停一停，试一试** 用归纳法证明：当 $n \geq 1$ 时 $e_n \leq 2r^{2^n}$，其中 $r = \frac{\sqrt{2}-1}{2}$。请仔细检查基础情况。

基础情况：$e_1 = \dfrac{(1-\sqrt{2})^2}{2 \cdot 1} = \dfrac{(\sqrt{2}-1)^2}{2}$，而 $2r^2 = 2 \cdot \dfrac{(\sqrt{2}-1)^2}{4} = \dfrac{(\sqrt{2}-1)^2}{2}$。两者相等，所以基础情况成立（而且是等式）。

归纳步骤：若 $e_n \leq 2r^{2^n}$，则

$$e_{n+1} \leq \frac{e_n^2}{2} \leq \frac{4r^{2^{n+1}}}{2} = 2r^{2^{n+1}}. \qquad \blacksquare$$

阶段 6 六百位小数

由于 $\sqrt{2} - 1 < \frac{1}{2}$，我们有 $r < \frac{1}{4}$，所以

$$0 \leq y_{10} - \sqrt{2} \leq 2\left(\frac{1}{4}\right)^{2^{10}} = 2 \cdot 2^{-2048} = 2^{-2047}.$$

而 $2^{10} = 1024 > 10^3$，所以 $2^{-2047} < 2^{-2040} = (2^{10})^{-204} < 10^{-612}$，这当然小于 10^{-600}。 $\blacksquare$

回顾

- **找出误差的递推关系。**两部分的做法，都是把新误差写成旧误差的一个精确函数，其余一切便随之而来。
- **线性与二次之别。**“新误差 $\leq$ 常数 $\times$ 旧误差”表示每一步增加固定数目的准确位数。“新误差 $\leq$ 常数 $\times$ (旧误差)^2”则表示每一步令准确位数加倍（图 2.7.1）。
- **平方量决定正负号。**在第 (ii) 部分，走一步之后的误差是一个平方数除以一个正数，所以永远不会是负数：迭代从上方逼近。

背后的数学

牛顿法

第二个方法其实是把牛顿法（Newton's method）应用于 $f(y)=y^2-2$。牛顿法求 f 的根的做法，是反复把猜测值 y 换成 y 处的切线与横轴的交点：

$$y_{\text{新}} = y - \frac{f(y)}{f'(y)} = y - \frac{y^2-2}{2y} = \frac{y^2+2}{2y}.$$

在单根（simple root）α 附近，由泰勒展开式（Taylor expansion）可知新误差约为 $\frac{f''(\alpha)}{2f'(\alpha)} \times (\text{旧误差})^2$，这就是牛顿法二次收敛（quadratic convergence）的原因。这个特殊情况，即“取 y 与 $2/y$ 的平均值”，古巴比伦人早已知道；一块约公元前 1700 年的泥板（现称为 YBC 7289）以 60 进制给出 $\sqrt{2}$ 的值，误差小于百万分之一。

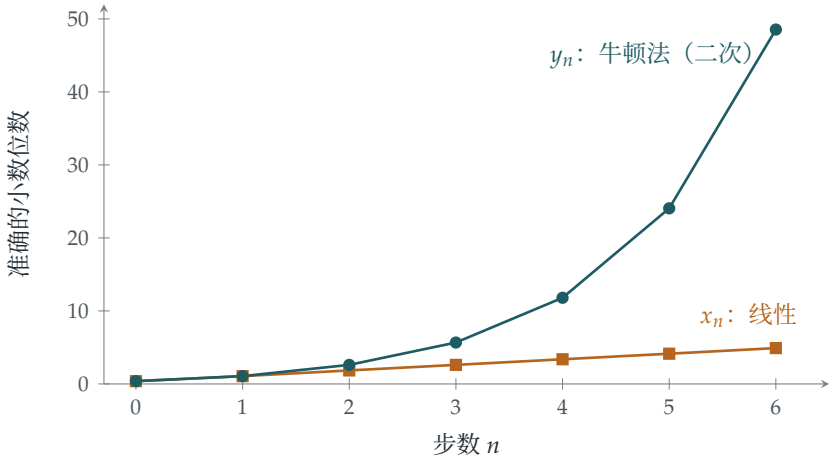


图 2.7.1　两个方法各走 n 步之后，$\sqrt{2}$ 的准确小数位数。线性的方法每一步增加约 0.77 位；牛顿法则每一步令准确位数大约加倍，到 $n=10$ 时达到约 780 位。

不动点及其收敛速度

两个方法都具有 $u_{n+1}=g(u_n)$ 的形式，而且 $g(\sqrt{2})=\sqrt{2}$，所以 $\sqrt{2}$ 是一个*不动点*（fixed point），正如 Section 2.2 的蛛网图（cobweb diagram）所示。在不动点 α 附近，误差每一步大约乘以 $g'(\alpha)$。对第一个方法而言，$g(x)=\frac{x+2}{x+1}$，而 $g'(\sqrt{2}) = -\frac{1}{(\sqrt{2}+1)^2} = -(3-2\sqrt{2}) \approx -0.17$：负号解释了正负号为何交替，0.17 则解释了每一步为何增加 0.77 位（$\log_{10}(1/0.17)\approx 0.77$）。对牛顿法而言，$g'(\alpha)=0$，所以线性项消失，误差由下一项（即平方项）决定。设计出满足 $g'(\alpha)=0$ 的迭代（iteration），正是快速数值方法的精髓

所在。

为何出现相同的分数

分数 $x_n = p_n/q_n$ 满足 $p_{n+1} = p_n + 2q_n$ 和 $q_{n+1} = p_n + q_n$，并且 $p_n^2 - 2q_n^2 = \pm 1$，正负交替。它们是佩尔方程（Pell's equation）$p^2 - 2q^2 = \pm 1$ 的解，也是连分数（continued fraction）$\sqrt{2} = [1; 2, 2, 2, \ldots]$ 的渐近分数（convergents）。它们来自 $1 + \sqrt{2}$ 的幂：$p_n + q_n\sqrt{2} = (1 + \sqrt{2})^{n+1}$。牛顿法的每一步把 $p + q\sqrt{2}$ 平方（不妨一试：$(p + q\sqrt{2})^2 = p^2 + 2q^2 + 2pq\sqrt{2}$，而 $\frac{p^2+2q^2}{2pq}$ 正是把牛顿法的公式应用于 p/q 的结果）。平方令指数加倍，这就是 $y_n = x_{2^n-1}$ 的原因。同样的“平方”模式，也是 Section 2.8 中迭代的动力来源。

思维技巧

分析误差，而不是答案

要了解一个近似方法的表现，就写下某一步的误差如何取决于上一步的误差。

这是数值分析的起点。两个方法的答案本身（$\sqrt{2}$）是一样的；不同之处，亦是实际应用中最要紧之处，在于误差如何演变。

延伸练习

1 **佩尔方程。**证明：若 $y_n = p/q$ 为最简分数，且 $p^2 - 2q^2 = 1$，则 $y_{n+1} = \dfrac{p^2 + 2q^2}{2pq}$ 亦满足这个方程。验证 $577^2 - 2 \times 408^2 = 1$。

2 **立方根。**写出应用于 $f(y) = y^3 - 2$ 的牛顿法，并证明新误差为 $\dfrac{(y_n - \alpha)^2(2y_n + \alpha)}{3y_n^2}$，其中 $\alpha = \sqrt[3]{2}$。由此推出：若 $y_n > \alpha$，则新误差最多是旧误差平方的 $\frac{1}{\alpha}$ 倍。（提示：利用 $\alpha^3 = 2$。）

3 **不用除法的除法。**证明应用于 $f(y) = \frac{1}{y} - a$ 的牛顿法给出 $y_{n+1} = y_n(2 - ay_n)$，当中无需除法。若 $e_n = 1 - ay_n$，证明 $e_{n+1} = e_n^2$。（电脑处理器正是用这类迭代来做除法的。）

4 **牛顿法失效的时候。**把牛顿法应用于 $f(y) = y^{1/3}$，它唯一的根是 0。证明无论从哪里开始，每一步都令与根的距离加倍。误差公式的哪一个假设不成立？

延伸阅读

- **Endre Süli & David Mayers, *An Introduction to Numerical Analysis***（Cambridge University Press）。标准的大学本科数值分析教科书；第 1 章讲述不动点迭代和牛顿法。
- **Nick Higham (ed.), *The Princeton Companion to Applied Mathematics***（Princeton University Press）。收录有关牛顿法和数值分析的短文，由专家执笔，面向广大读者。
- **David Fowler & Eleanor Robson, "Square root approximations in Old Babylonian mathematics: YBC 7289 in context"**（古巴比伦数学中的平方根近似值），*Historia Mathematica* 25 (1998)。

2.8 平方再减二

题目

STEP III 2021，第 8 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2021 年 STEP 试卷（zip 文件）内的 STEP III 试卷：

`www.ocr.org.uk/Images/696588-2021-past-papers.zip`

引子

“平方后减二”几乎是可以用来生成数列（sequence）的最简单规则。然而它把实数线分成两个世界。从 $[-2,2]$ 之外出发，各项会以比指数增长还要快的速度冲向无穷大。从区间之内出发，各项则永远被困在其中，四处游走，而且在一个精确的意义下完全无法预测：这是混沌（chaos）最简单的例子之一。

同一条规则，用在非常大的整数上，就是卢卡斯－莱默检验法（Lucas–Lehmer test）的核心。这是一个高效的检验法，用来判断梅森数（Mersenne number）2^p-1 是否质数（prime）；现时的最大质数纪录，以及其他已知最大的质数中的大部分，都是靠它找到的。本题探讨的是故事中“逃逸”的一面，并找出一条隐藏的公式，令这个数列完全可解。

动笔之前

先试几个初始值。取 $a=3$：$3,7,47,2207,\ldots$ 取 $a=1$：$1,-1,-1,-1,\ldots$ 取 $a=0$：$0,-2,2,2,2,\ldots$ 取 $a=0.3$：$0.3,-1.91,1.648,0.716,-1.487,0.211,\ldots$

想一想 为甚么 $[-2,2]$ 内的数永远无法逃出去？当 $-2\le x\le 2$ 时，x^2-2 最大和最小可以是多少？

若 $-2\le x\le 2$，则 $0\le x^2\le 4$，所以 $-2\le x^2-2\le 2$。区间 $[-2,2]$ 被映射到它自身之内。这就是第 (ii) 部分的一半。

逐步解题

阶段 1　逃逸

> **想一想**　第 (i) 部分出现了 4^{n-1}，暗示与 2 的距离每一步至少变成四倍。设 $x_n = 2 + d$。以 d 表示，x_{n+1} 是甚么？

$x_{n+1} = (2+d)^2 - 2 = 2 + 4d + d^2$。所以若 $d \geq 0$，新一项高出 2 的距离是 $4d + d^2 \geq 4d$。

现在用归纳法（induction）。当 $n = 1$ 时，$x_1 = a = 2 + 4^0(a-2)$。若 $x_n \geq 2 + 4^{n-1}(a-2)$，则 $d = x_n - 2 \geq 4^{n-1}(a-2) \geq 0$，所以

$$x_{n+1} - 2 \geq 4d \geq 4^n(a-2). \qquad \blacksquare$$

阶段 2　逃逸当且仅当 $|a| > 2$

> **想一想**　“当且仅当”（if and only if）表示要分别作两个证明。第 (i) 部分差不多已经给出哪一个方向？还欠甚么？

若 $|a| > 2$：当 $a > 2$ 时，由第 (i) 部分得 $x_n \geq 2 + 4^{n-1}(a-2) \to \infty$。当 $a < -2$ 时，第一步已把正负号纠正过来：$x_2 = a^2 - 2 > 4 - 2 = 2$，而由 x_2 开始的数列，就是以 $x_2 > 2$ 为初始值的同一个数列，所以它趋向无穷大。

仅当：若 $|a| \leq 2$，则 $a \in [-2, 2]$，由上面的观察可知每一项都留在 $[-2, 2]$ 之内。有界（bounded）的数列不可能趋向无穷大。 $\blacksquare$

阶段 3　一条隐藏的公式

第 (iii) 部分看来令人却步。在动手之前，值得先找出 x_n 的公式。

> **想一想**　若 $a > 2$，则存在一个数 $t > 1$ 使得 $a = t + \dfrac{1}{t}$。（为甚么？）以 t 表示，x_2 和 x_3 是甚么？

当 t 由 1 增至 ∞ 时，函数 $t + 1/t$ 由 2 递增至 ∞，所以每个 $a > 2$ 都可以写成这个形式。于是

$$x_2 = \left(t + \frac{1}{t}\right)^2 - 2 = t^2 + \frac{1}{t^2}, \qquad x_3 = \left(t^2 + \frac{1}{t^2}\right)^2 - 2 = t^4 + \frac{1}{t^4},$$

再由归纳法得

$$x_n = t^{2^{n-1}} + t^{-2^{n-1}}.$$

平方后减 2，只不过是把指数加倍。这样题目就变得容易处理了。

阶段 4 乘积裂项相消

想一想 $\left(t-\frac{1}{t}\right)\left(t+\frac{1}{t}\right)$ 是甚么？若把结果再乘以 $t^2+\frac{1}{t^2}$，会怎样？

$\left(t-\frac{1}{t}\right)\left(t+\frac{1}{t}\right)=t^2-\frac{1}{t^2}$，然后 $\left(t^2-\frac{1}{t^2}\right)\left(t^2+\frac{1}{t^2}\right)=t^4-\frac{1}{t^4}$，如此类推：每个因子 x_k 都把两个幂之差中的指数加倍。所以

$$\left(t-\frac{1}{t}\right)x_1x_2\cdots x_n=t^{2^n}-t^{-2^n}.$$

停一停，试一试 以 t 表示并计算 x_{n+1}^2-4，然后与上面的乘积比较。

$$x_{n+1}^2-4=\left(t^{2^n}+t^{-2^n}\right)^2-4=\left(t^{2^n}-t^{-2^n}\right)^2.$$

由于 $t>1$，$t^{2^n}-t^{-2^n}>0$，所以 $\sqrt{x_{n+1}^2-4}=t^{2^n}-t^{-2^n}=\left(t-\frac{1}{t}\right)x_1x_2\cdots x_n$。因此

$$\frac{\sqrt{x_{n+1}^2-4}}{x_{n+1}}=\frac{\left(t-\frac{1}{t}\right)x_1\cdots x_n}{x_{n+1}},$$

而它对每个 n 都等于 y_n，恰好就是 $A=t-\frac{1}{t}$ 的时候。最后，$\left(t-\frac{1}{t}\right)^2=\left(t+\frac{1}{t}\right)^2-4=a^2-4$，所以条件是 $A=\sqrt{a^2-4}$，即

$$a=\sqrt{A^2+4}.$$

（因为 $A>0$，这个值大于 2，符合要求。）■

阶段 5 y_n 是否收敛？

取这个 a 值时，

$$y_n=\frac{\sqrt{x_{n+1}^2-4}}{x_{n+1}}=\sqrt{1-\frac{4}{x_{n+1}^2}}.$$

由第 (ii) 部分知 $x_{n+1}\to\infty$，所以 $\frac{4}{x_{n+1}^2}\to 0$，而 $y_n\to 1$。数列收敛（converge），极限为 1。■

所以无穷乘积（infinite product）$x_1x_2x_3\cdots$ 的增长速度与 x_{n+1}/A 完全相同：我们找到了一个由急速增长的项所组成的乘积的精确增长率。

回顾

- **不变区域（invariant region）。**$[-2,2]$ 被映射到它自身之内，单凭这一点就解决了第 (ii) 部分的一半。

- **把问题线性化的代换（substitution）。**写成 $a = t + 1/t$ 之后，“平方后减 2”就变成“把指数加倍”。杂乱的非线性行为变得简单。
- **又是裂项相消（telescoping）。**与 Section 2.5 中的乘积一样，由许多因子组成的乘积之所以能折叠起来，是因为每个因子都与下一个互相配合。

背后的数学

秩序与混沌

在 $[-2, 2]$ 之内会发生甚么事？在那里 $a = t + 1/t$ 行不通（它需要 $|a| \geq 2$），但三角函数的版本却可行：设 $a = 2\cos\theta$。则 $x_2 = 4\cos^2\theta - 2 = 2\cos 2\theta$，而一般地

$$x_n = 2\cos\left(2^{n-1}\theta\right).$$

所以在区间之内，这个映射只不过是*把角度加倍*。听起来无伤大雅，但把角度加倍其实是混沌的。把 θ/π 写成二进制（binary）：加倍会把二进制小数点向右移一位，并丢掉最前面的一位数字。经过 50 步之后，x_n 取决于 θ/π 的第 51 位二进制数字，所以初始值若有 10^{-15} 的误差（一般电脑的精确度），50 步之后的答案就完全不可靠。这种对初始条件的敏感依赖（sensitive dependence on initial conditions）正是混沌的标志，见图 2.8.1。映射 $x \mapsto x^2 - 2$ 经变量代换后，等价于著名的逻辑斯谛映射（logistic map）$x \mapsto 4x(1-x)$，我们会在 Section 8.7 再遇上它。

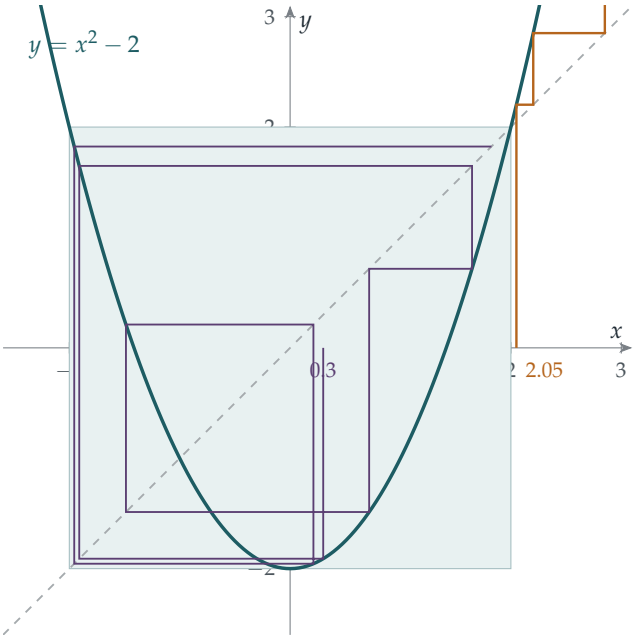


图 2.8.1 $x_{n+1} = x_n^2 - 2$ 的蛛网图（cobweb diagram）。由 0.3 出发（紫红色），轨道（orbit）永远留在阴影正方形 $[-2,2]^2$ 之内，混沌地游走。由区间外不远处的 2.05 出发（琥珀色），它就会逃逸：$2.05, 2.20, 2.85, 6.13, \ldots$

寻找最大的质数

梅森数是指 $M_p = 2^p - 1$。卢卡斯（Édouard Lucas）在 1876 年发现了一个出色的检验法，用来判断 M_p 是否质数（其中 p 为奇质数），其后莱默（Derrick Lehmer）在 1930 年把它完善：由 $s_0 = 4$ 开始，反复平方后减 2，每次都取模（modulo）M_p 化简。则

$$M_p \text{ 是质数} \iff s_{p-2} \equiv 0 \pmod{M_p}.$$

以 $M_7 = 127$ 为例：$s = 4, 14, 194 \equiv 67, 67^2 - 2 \equiv 42, 42^2 - 2 \equiv 111, 111^2 - 2 \equiv 0$，所以 127 是质数。它与本题的联系就在于那个代换：$4 = t + 1/t$，其中 $t = 2 + \sqrt{3}$，所以 $s_k = (2+\sqrt{3})^{2^k} + (2-\sqrt{3})^{2^k}$，而检验法的证明用到形如 $a + b\sqrt{3}$ 的数的算术，与 Section 1.6 的高斯整数（Gaussian integers）颇为相似。由于每一步只需要一次平方，这个检验法极为快速，这就是为甚么已知最大的质数中的大部分，包括现时的纪录保持者，都是梅森质数（并非全部都是：其他特殊形式，例如 $a^{2^n} + 1$，也有各自的快速检验法）。在本书撰写之时，纪录是 $2^{136279841} - 1$，这个数有多于 41 百万位数字，于 2024 年发现。

曼德博集合

把 -2 换成任意常数 c，并容许复数：$z \mapsto z^2 + c$。令 0 的轨道保持有界的所有 c 所组成的集合，就是曼德博集合（Mandelbrot set）。我们的映射对应 $c = -2$，即曼德博集合最

左边的一点，而轨道保持有界的初始值所组成的集合（它的*茹利亚集合*（Julia set））正好是我们在第 (ii) 部分找到的区间 $[-2,2]$。对于几乎所有其他的 c，这个集合都是分形（fractal）。

思维技巧

转换变量，令映射变得简单

若一个迭代（iteration）看来很复杂，就找一个代换，把它变成简单的东西，例如加倍或加上一个常数。

这里 $x = t + 1/t$ 和 $x = 2\cos\theta$ 把 $x \mapsto x^2 - 2$ 变成了加倍。在动力系统（dynamical systems）理论中，这称为*共轭*（conjugacy），而找到共轭往往是彻底理解一个系统的关键。

延伸练习

1 **区间之内。**证明若 $a = 2\cos\theta$，则 $x_n = 2\cos(2^{n-1}\theta)$。求所有使得 $x_3 = a$ 但 $x_2 \neq a$ 的 a 值。（提示：你需要 $4\theta \equiv \pm\theta$，模 2π。）

2 **手算卢卡斯－莱默检验法。**用这个检验法证明 $M_5 = 31$ 是质数，而 $M_{11} = 2047$ 不是。（对 M_{11} 你需要九步；用计算机会有帮助。把 2047 作因数分解，以验证你的结论。）

3 **一个具体例子。**取 $A = \frac{3}{2}$。求 a 和 t，并以数值验证公式 $y_1 = \sqrt{x_2^2 - 4}/x_2$。

4 **裂项相消的和。**较难设 $a = t + 1/t$ 且 $t > 1$，证明

$$\frac{1}{x_1 x_2 \cdots x_n} = \frac{t - t^{-1}}{t^{2^n} - t^{-2^n}}.$$

利用恒等式 $\dfrac{1}{z-1} - \dfrac{1}{z^2-1} = \dfrac{z}{z^2-1}$，其中 $z = t^{2^n}$，由此推出

$$\frac{1}{x_1} + \frac{1}{x_1 x_2} + \frac{1}{x_1 x_2 x_3} + \cdots = \frac{1}{t}.$$

当 $a = 3$ 时，以数值验证这个结果。

延伸阅读

- **Ian Stewart, *Does God Play Dice?***（Penguin）。讲述混沌的经典科普著作，当中包括逻

辑斯谛映射。

- **Steven Strogatz, *Nonlinear Dynamics and Chaos***（CRC Press）。标准的大学本科入门教材；第 10 章讲述像本题这样的一维映射。
- **Marcus du Sautoy, *The Music of the Primes***（Fourth Estate）。书中包括搜索梅森质数的故事。

2.9 逐步逼近 e 的分数

题目

STEP II 2021，第 8 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2021 年 STEP 试卷（zip 文件）内的 STEP II 试卷：

`www.ocr.org.uk/Images/696588-2021-past-papers.zip`

引子

怎样才能找到逼近 e 的好分数？你可以把级数（series）$1+1+\frac{1}{2}+\frac{1}{6}+\cdots$ 截断，但以分母的大小来衡量，这样得到的分数并不算特别好。这道题的做法高明得多：它构造出一族积分，每个积分都是形如“整数 + 整数 $\times e$”的一个很小的数。如果 $a+be$ 极小，那么 e 就非常接近 $-a/b$，于是这些积分便为我们送上一连串极佳的分数：$\frac{19}{7}$、$\frac{193}{71}$、$\frac{2721}{1001}$、…

同样的技巧，即找出既很小、又形如“你所关心的那个数的整数组合”的积分，正是数学中一些最深刻的无理性证明背后的想法，包括埃尔米特（Hermite）对 e 是超越数（transcendental）的证明，以及阿佩里（Apéry）对 $\zeta(3)$ 是无理数（irrational）的证明。

动笔之前

看看 $[0,1]$ 上的被积函数（integrand）$\frac{t^n(1-t)^n}{n!}e^t$。

> **想一想** $t(1-t)$ 在 $[0,1]$ 上的最大值是多少？由此可知，当 n 增大时 T_n 的大小会怎样变化？

$t(1-t)$ 的最大值是 $\frac{1}{4}$，在 $t=\frac{1}{2}$ 处取得。所以被积函数至多为 $\frac{e}{4^n n!}$，它缩小的速度快得惊人（图 2.9.1）。这个“小”只是故事的一半。另一半，即 T_n 具有 a_n+b_ne 这种特殊形式，则来自分部积分法（integration by parts）。

逐步解题

阶段 1　求导两次

设 $p(t)=t^n(1-t)^n$。由积法则（product rule），

$$p'(t)=nt^{n-1}(1-t)^n-nt^n(1-t)^{n-1}=nt^{n-1}(1-t)^{n-1}(1-2t).$$

> **想一想**　再求导一次。你会得到两项；把公因式 $nt^{n-2}(1-t)^{n-2}$ 抽出来。要配合题目要求的形式，你需要把 $(1-2t)^2$ 改写成以 $t(1-t)$ 表示的式子。怎样做？

用积法则对 $n\,[t(1-t)]^{n-1}(1-2t)$ 求导（留意 $\frac{\mathrm{d}}{\mathrm{d}t}[t(1-t)]=1-2t$）：

$$\begin{aligned}p''(t)&=n(n-1)[t(1-t)]^{n-2}(1-2t)^2-2n[t(1-t)]^{n-1}\\&=nt^{n-2}(1-t)^{n-2}\Big[(n-1)(1-2t)^2-2t(1-t)\Big].\end{aligned}$$

现在 $(1-2t)^2=1-4t+4t^2=1-4t(1-t)$，所以方括号内的式子是 $(n-1)-4(n-1)t(1-t)-2t(1-t)=(n-1)-2(2n-1)t(1-t)$。 ■

阶段 2　分部积分两次

> **想一想**　$p(t)$ 在 $t=0$ 和 $t=1$ 处都等于零。当 $n\ge 2$ 时，$p'(t)$ 又如何？在对 $p(t)e^t$ 作分部积分时，这一点为甚么重要？

当 $n\ge 2$ 时，$p'(t)$ 含有因式 $t^{n-1}(1-t)^{n-1}$，所以它在两个端点也等于零。分部积分两次，每一个边界项都是零：

$$\int_0^1 p\,e^t\,\mathrm{d}t=\Big[p\,e^t\Big]_0^1-\int_0^1 p'\,e^t\,\mathrm{d}t=-\Big[p'\,e^t\Big]_0^1+\int_0^1 p''\,e^t\,\mathrm{d}t=\int_0^1 p''\,e^t\,\mathrm{d}t.$$

由第 1 阶段，$p''(t)=n(n-1)\,t^{n-2}(1-t)^{n-2}-2n(2n-1)\,t^{n-1}(1-t)^{n-1}$。除以 $n!$：

$$\frac{p''(t)}{n!}=\frac{t^{n-2}(1-t)^{n-2}}{(n-2)!}-2(2n-1)\frac{t^{n-1}(1-t)^{n-1}}{(n-1)!}.$$

乘以 e^t 再积分，便得 $T_n=T_{n-2}-2(2n-1)T_{n-1}$。 ■

阶段 3　整数结构

$T_0=\int_0^1 e^t\,\mathrm{d}t=e-1$。至于 $T_1=\int_0^1(t-t^2)e^t\,\mathrm{d}t$，由分部积分得 $\int_0^1 te^t\,\mathrm{d}t=1$ 及 $\int_0^1 t^2e^t\,\mathrm{d}t=e-2$，所以 $T_1=1-(e-2)=3-e$。

想一想 $T_0 = -1+1\cdot e$ 而 $T_1 = 3+(-1)\cdot e$。为甚么每个 T_n 都必定具有 $a_n + b_n e$ 的形式，其中 a_n、b_n 是整数?

用强归纳法（strong induction）：若 $T_{n-2} = a_{n-2} + b_{n-2}e$ 及 $T_{n-1} = a_{n-1} + b_{n-1}e$，其中系数都是整数，则

$$T_n = (a_{n-2} - 2(2n-1)a_{n-1}) + (b_{n-2} - 2(2n-1)b_{n-1})e,$$

而新的系数也是整数。 ■

阶段 4 很小，因而带来信息

当 $0 < t < 1$ 时，被积函数是正数，并且 $t^n(1-t)^n \le 1$ 及 $e^t < e$。所以

$$0 < T_n < \int_0^1 \frac{e}{n!}\,\mathrm{d}t = \frac{e}{n!}.$$

停一停，试一试 把 $0 < a_n + b_n e < \dfrac{e}{n!}$ 除以 $|b_n|$。这说明 e 与 $-a_n/b_n$ 之间的距离如何?

除以 $|b_n|$（它是正整数，所以至少是 1），

$$\left| e - \left(\frac{-a_n}{b_n}\right)\right| = \frac{|a_n + b_n e|}{|b_n|} < \frac{e}{n!\,|b_n|} \le \frac{e}{n!}.$$

右边趋向 0，所以 $-a_n/b_n \to e$。 ■

以下是由递推关系（recurrence）算出的头几个分数：

n	2	3	4	5	6
$-a_n/b_n$	$\frac{19}{7}$	$\frac{193}{71}$	$\frac{2721}{1001}$	$\frac{49171}{18089}$	$\frac{1084483}{398959}$
误差	4×10^{-3}	3×10^{-5}	1×10^{-7}	3×10^{-10}	5×10^{-13}

分数 $\frac{2721}{1001}$ 准确至小数点后 6 位，而分母只有四位数。

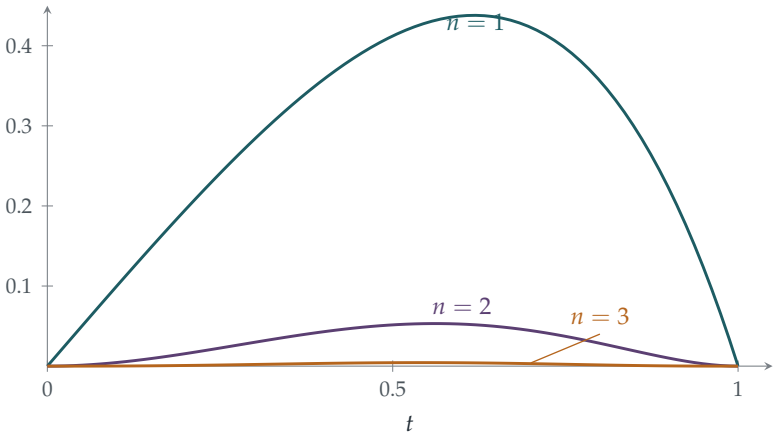


图 2.9.1　$n = 1, 2, 3$ 时的被积函数 $\frac{t^n(1-t)^n}{n!}e^t$。每一个在 $(0, 1)$ 上都是正数，所以每个 T_n 都是正数；而且它们缩小得极快，到 $n = 3$ 时曲线已几乎看不见。是正数但极小，正是这个论证所需要的。

回顾

- **分部积分是制造递推关系的机器。**由于多项式及其导数在两个端点都等于零，分部积分给出了一个没有边界项的简洁递推关系。
- **结构与大小并用。**递推关系说明 T_n 具有 $a_n + b_n e$ 的形式；上界说明它很小。单凭其中任何一项，都无法说明怎样逼近 e。
- **不需要的就不要算。**题目叫我们不必求出 a_n 和 b_n 的值，而论证也从来用不着它们；所需要的只是它们是整数。

背后的数学

e 是无理数的另一个证明

这道题离一个著名定理只差一步。假设 $e = p/q$，其中 p, q 是正整数。那么

$$q\,T_n = qa_n + b_n p$$

是一个整数。但是 $0 < q\,T_n < \dfrac{qe}{n!}$，而当 n 很大时，右边小于 1。于是 qT_n 就会是一个严格介乎 0 与 1 之间的整数，这是不可能的。因此 e 是无理数。这正是 Section 1.4 中“把一个整数困在 0 与 1 之间”的想法，只是这次由积分而不是级数来实现。正是这一类积分，配合 $t^n(1-t)^n$ 这样的多项式，在 1873 年被夏尔·埃尔米特（Charles Hermite）用来证明 e 是*超越数*：它不是任何整系数多项式的根。

连分数

表中的分数不仅是好；它们更属于*最佳之列*。e 的连分数（continued fraction）由欧拉（Euler）发现，有一个优美的规律：

$$e = [2; 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, \ldots],$$

而它的渐近分数（convergents）是 $2, 3, \frac{8}{3}, \frac{11}{4}, \frac{19}{7}, \frac{87}{32}, \frac{106}{39}, \frac{193}{71}, \ldots, \frac{2721}{1001}, \ldots$。我们的 $-a_n/b_n$ 恰好是每隔两个取一个的渐近分数。没有任何分母更小的分数能更接近 e。欧拉发现的这个规律本身就是 e 是无理数的第一个证明，因为只有有理数才有有限的连分数。

阿佩里的奇迹

1978 年，罗杰·阿佩里（Roger Apéry）宣布证明了 $\zeta(3) = 1 + \frac{1}{8} + \frac{1}{27} + \cdots$ 是无理数，这是自欧拉以来一直悬而未决的问题（见 Section 2.3）。他的证明实在太出人意表，以致在场听众起初并不相信。一年后，Frits Beukers 找到一个简短的版本，所用的积分形如

$$\int_0^1\!\!\int_0^1 \frac{-\ln(xy)}{1-xy}\,P_n(x)P_n(y)\,\mathrm{d}x\,\mathrm{d}y,$$

其中 P_n 由 $\frac{\mathrm{d}^n}{\mathrm{d}x^n}\big(x^n(1-x)^n\big)$ 构造而成。这些积分很小，是正数，并且形如“整数 + 整数 $\times\,\zeta(3)$”再除以一个受控的分母：这正是本题的策略。至于 $\zeta(5)$ 或 $e+\pi$ 是否无理数，至今仍然未知。

思维技巧

很小但不是零

如果已知某个量是若干个数的整数组合，又已知它很小但不是零，它就带有信息：关于逼近的信息，有时还有关于无理性的信息。

这个模式就是“结构来自代数，大小来自分析”。它在数论中反复出现，也是分析与代数交汇得最有成果的地方之一。

延伸练习

1 **运行递推关系。**由递推关系计算 $n = 2, 3, 4$ 时的 a_n 和 b_n，并核对表中的分数。证明 b_n 的正负号交替出现。

2 **更精确的上界。**证明 $0 < T_n < \dfrac{e}{4^n\, n!}$。这把 $-a_n/b_n$ 的误差估计改进了多少？

3 **无理性的证明。**把上面方框中的论证仔细写出来。你究竟在哪一处用到 $T_n > 0$？

4 **换一个权函数。**较难把 e^t 换成 e^{2t}，重做这道题（即 $U_n = \int_0^1 \frac{t^n(1-t)^n}{n!} e^{2t}\,\mathrm{d}t$）。求 U_n 的递推关系，证明 U_n 是 1 和 e^2 的有理系数组合，并探究这个论证是否仍能证明 e^2 是无理数。（与 Section 1.4 比较。）

延伸阅读

- **Ivan Niven, *Irrational Numbers***（Mathematical Association of America）。关于无理性证明的经典小书，当中包括埃尔米特的方法。
- **Alfred van der Poorten, “A proof that Euler missed”**, *The Mathematical Intelligencer* 1 (1979)。生动地记述阿佩里的证明及其所得到的回响。
- **Julian Havil, *The Irrationals***（Princeton University Press）。一部无理数的历史，为一般数学读者而写。

我们学到了甚么

我们由芝诺（Zeno）说起，并提出了关于无穷过程的三个问题：说无穷过程有答案，究竟是甚么意思？怎样证明它们确实有答案？它们又以多快的速度到达答案？做过九道 STEP 题目之后，让我们看看自己走了多远。

1. 说无穷过程有答案，是甚么意思？

在 Section 2.1 中，我们学会了这套语言。凡是描述数列（sequence）长远表现的命题，都具有这样的形式：“存在某一点 m，使得对所有 $n \geq m$...”。要证明它，就要把 m 找出来；要否定它，就要证明任何人提出的 m 都不管用。极限（limit）的定义只是在最前面多加一个量词（quantifier）：“对任意容许误差 $\varepsilon > 0$”。有了这个定义，芝诺悖论便迎刃而解：无论你要求的容许误差多小，$\frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^n}$ 由某一点起，与 1 的距离都在这个误差之内，而“和等于 1”的全部意思，不过如此。

这套语言也能防止我们说出毫无意义的话。在 Section 2.2 中，“设 x 等于那个无穷表达式”这种取巧的论证，为黄金比（golden ratio）给出了正确答案，但那只是因为我们能够另行证明极限存在。把同一个论证用在 $1 - 1 + 1 - \cdots$ 上，得出的答案却毫无意义。

2. 不知道极限是甚么，怎样证明收敛？

本章大部分小节所用的，都不外乎以下几种工具。

- **单调有界。**递增而有上界的数列必定收敛（converge）（Section 2.2、Section 2.6）。各项皆为正数的部分和（partial sum）亦然（Section 2.3）。到了大学，你会发现这其实是关于实数本身的一个命题：它就是实数的*完备性*（completeness）。
- **比较。**一个正项级数（series）如果比某个收敛的级数小，它就收敛；如果比某个发散（diverge）的级数大，它就发散。把各项分组成段之后，比较起来便很容易，由此可见调和级数（harmonic series）发散，而 Kempner 那个删减了部分项的版本却收敛（Section 2.3）。
- **夹逼。**把要研究的量夹在两个较简单而极限相同的量之间。我们就是这样证明 $\sum 1/m^2 = \pi^2/6$（Section 2.4），并严格地计算出算术几何平均（AGM）积分的值（Section 2.6）。

- **有限阶段的精确公式。**在 Section 2.5 和 Section 2.8 中，我们利用裂项相消（telescoping）求出 n 步之后的精确表达式，到了最后一刻才让 $n \to \infty$。

一旦知道极限存在，往往只需在递推关系（recurrence）中“让 $n \to \infty$”便可求出极限，例如黄金比的 $L = (2L+1)/(L+1)$。但次序十分重要：先证存在，后求数值。

3. 有多快？

在这方面，本章有些出人意表的发现。韦达（Viète）乘积（Section 2.5）每多乘一个因子，大约多得 0.6 个小数位。Section 2.7 中计算 $\sqrt{2}$ 的简单方法，每一步大约多得 0.8 个数位。牛顿法（Newton's method）和 AGM（Section 2.6）则每一步都令正确数位的数目*倍增*：牛顿法只需十步，便可得出 $\sqrt{2}$ 的 700 多个数位。差别在于误差递推关系的形式：一种是“新误差 $\approx$ 常数 $\times$ 旧误差”（线性），另一种是“新误差 $\approx$ 常数 $\times$（旧误差）^{2}”（二次）。而在 Section 2.8 中，我们见到了另一个极端：由仅仅大于 2 的数出发，映射（map）$x \mapsto x^2 - 2$ 会以比指数增长还要快的速度把数推开；而在 $[-2, 2]$ 之内，它会把微小的误差不断放大，直至完全无法预测。

速度对理论同样重要。Section 2.9 中的积分所产生的分数，趋近 e 的速度极快，足以证明 e 是无理数（irrational）：任何分数 p/q 都不可能被其他分数以这样快的速度逼近。

4. 分析学的工具箱

方法	你在哪里用过
找出量词所要求的对象与否定量词	“最终”（2.1）
单调有界数列	黄金比（2.2）；AGM（2.6）
比较与凝聚（condensation）	调和级数与 Kempner 级数（2.3）
夹逼	$\pi^2/6$（2.4）；AGM 积分（2.6）
根的对称函数（symmetric function）	$\cot^2$ 值之和（2.4）
裂项相消的和与积	韦达（2.5）；$x^2 - 2$（2.8）
不变量（invariant）	AGM 积分（2.6）
误差递推关系	线性收敛与二次收敛（2.7）
以变量代换简化映射	$a = t + 1/t$、$a = 2\cos\theta$（2.8）
细小但非零的整数组合	逼近 e（2.9）

留意这些工具当中，有多少附带着一条“你可以直接使用的定理”。STEP 经常直接把单调收敛定理（monotone convergence theorem）交给你，或者叫你假设极限存在。到了大学，你要由实数的公理（axiom）出发证明这些定理，而且当一个论证不动声色地假设了极限存在，你也要自己察觉得到。阅读解答的时候，不妨养成一个很好的习惯，问问自己：*究竟在哪一步取了极限？这一步有没有根据？*

下一步

本章的想法贯穿全书其余部分。

- 不动点（fixed point）和蛛网图（cobweb diagram）会在 Section 8.7 再次出现，在那里，一个混沌（chaotic）映射原来有精确解。
- 调和数（harmonic number）$1 + \frac{1}{2} + \cdots + \frac{1}{n}$ 会在 Section 9.11 的概率和 Section 11.12 的力学中出现。
- π 另一次在极限中现身，是在 Section 12.23，这次来自概率。
- 对极限的审慎态度是第 3 章的主题，该章以 Section 3.1 开始：一些看似正确、其实不然的论证。

如果你想在上大学之前再进一步

- **Lara Alcock, *How to Think About Analysis***：大学一年级的分析学是怎样的，又该如何应付。
- **David Bressoud, *A Radical Approach to Real Analysis***：循历史脉络讲授分析学，由那些迫使数学家不得不严谨起来的难题说起。
- **Tom Körner, *Calculus for the Ambitious***（Cambridge University Press）：由一位剑桥的分析学家执笔，正是为本书这样的读者而写。
- **Julian Havil, *Gamma*** 和 **William Dunham, *Euler: The Master of Us All***：本章许多级数背后的故事。

芝诺认为无穷令运动变得不可能。数学家花了两千年才回应他，而答案原来是人类发现过最强而有力的想法之一：一种能够精确地描述永无止境的过程的方法。这套语言最初的几个词语，你现在已经认识了。到了分析学，你便会学懂怎样流利地说这套语言。

延伸练习答案

以下只是简短的答案和提示，并非完整解答：这里只给出结果和关键步骤，方便你核对自己的做法。请先认真做过每一道练习，然后才翻看。

2.1 “最终”是甚么意思

1. $(1+x)^n = 1 + nx + \binom{n}{2}x^2 + \cdots$ 的各项均为非负，所以 $(1+x)^n \geq 1 + nx > nx$。设 $y = 1 + x$，则只要 $n > K/(y-1)$，便有 $y^n > n(y-1) > K$。利用下一项，$y^n > \binom{n}{2}(y-1)^2$，所以只要 $n > 1 + 2K/(y-1)^2$，便有 $y^n/n > \frac{1}{2}(n-1)(y-1)^2 > K$。

2. 若当 $n \geq m_1$ 时 $s_n \leq t_n$，而当 $n \geq m_2$ 时 $t_n \leq s_n$，则当 $n \geq \max(m_1, m_2)$ 时 $s_n = t_n$。

3. 例如：当 n 为偶数时 $s_n = n + 0.6$，当 n 为奇数时 $s_n = n$；而当 n 为奇数时 $t_n = n + 0.6$，当 n 为偶数时 $t_n = n$。两个数列（sequence）都是递增的，而且各自都有无限多次大于对方。

4. 当 $n \geq 4$ 时 $2^n \leq n!$（用归纳法（induction）：$2^{n+1} = 2 \cdot 2^n \leq (n+1)\,n!$）。至于 $n^{10} \leq 2^n$，这个不等式对由 2 至 58 的每一个 n 都不成立，而对所有 $n \geq 59$ 都成立，所以最小的 m 是 59。

2.2 作为极限的黄金比例

1. 用归纳法：$u_1 = 1 = F_2/F_1$，而若 $u_n = F_{n+1}/F_n$，则 $u_{n+1} = 1 + F_n/F_{n+1} = F_{n+2}/F_{n+1}$。

2. 由于 $\varphi - 1 = 1/\varphi$，$u_{n+1} - \varphi = \frac{1}{u_n} - \frac{1}{\varphi} = -\frac{u_n - \varphi}{\varphi u_n}$。因为 $u_n \geq 1$，所以 $|u_{n+1} - \varphi| \leq |u_n - \varphi|/\varphi$，又 $|u_1 - \varphi| < 1$，故 $|u_n - \varphi| \leq \varphi^{-(n-1)}$。

3. $\psi = \frac{1-\sqrt{5}}{2}$ 满足 $\psi^2 = \psi + 1$，所以 $1 + 1/\psi = \psi$。倒推回去，$u_{n-1} = \frac{1}{u_n - 1}$：由 0 出发，得到 $-1, -\frac{1}{2}, -\frac{2}{3}, -\frac{3}{5}, -\frac{5}{8}, \ldots$，即对某个 $k \geq 1$ 有 $u_1 = -F_{k-1}/F_k$。这些数值趋近 ψ。

4. 不动点（fixed point）满足 $v = 1 + \frac{1}{1+v}$，即 $v^2 = 2$。误差每次乘以 $-\frac{1}{(1+v_n)(1+\sqrt{2})}$，其大小不超过 $\frac{1}{2(1+\sqrt{2})}$，所以 $v_n \to \sqrt{2}$。并且 $\sqrt{2} = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cdots}}$。

2.3 一个发散级数与它的收敛近亲

1. 由 $\ln N + 0.5772 = 10$ 得 $N \approx 12\,400$；要超过 100，大约需要 $e^{99.4} \approx 1.5 \times 10^{43}$ 项。

2. 取 $k = 2$，第 n 组之和不超过 $2^n \cdot 4^{-n} = 2^{-n}$，所以总和不超过 $\sum_{n\ge 0} 2^{-n} = 2$。把 $1 + \frac14 + \frac19$ 准确加起来，再以 $\frac14 + \frac18 + \cdots = \frac12$ 作为由 $r = 4$ 起各组的上界，便得到约 1.86。（真确值是 $\frac{\pi^2}{6} \approx 1.645$。）

3. 对于 $1/(r \ln r)$，由 2^n 至 $2^{n+1} - 1$ 的一组之和至少是 $\frac{2^n}{2^{n+1}(n+1)\ln 2} = \frac{1}{2(n+1)\ln 2}$，而这些数之和的表现与调和级数（harmonic series）相同：级数发散（divergence）。对于 $1/(r(\ln r)^2)$，该组之和不超过 $\frac{1}{n^2(\ln 2)^2}$，而 $\sum \frac{1}{n^2}$ 收敛（converges）。

4. 可用的数字：首位数字有 7 个选择，其余每一位各有 8 个选择，所以符合条件的 d 位数最多有 $7 \cdot 8^{d-1}$ 个，而每个都至少是 10^{d-1}。总和不超过 $\sum_{d\ge 1} 7(0.8)^{d-1} = \frac{7}{1-0.8} = 35$。

2.4 $\pi^2/6$

1. 各个 $\tan^2 = 1/\cot^2$ 的值是倒序多项式 $\sum_j (-1)^j \binom{2n+1}{2j+1} y^j$ 的根（roots），其根之和为 $\binom{2n+1}{2n-1} = n(2n+1)$。然后由 $\operatorname{cosec}^2 = 1 + \cot^2$ 得 $n + \frac{n(2n-1)}{3} = \frac{2n(n+1)}{3}$。

2. 偶数项之和为 $\frac14 \sum \frac{1}{m^2} = \frac{\pi^2}{24}$，所以奇数项之和为 $\frac{\pi^2}{6} - \frac{\pi^2}{24} = \frac{\pi^2}{8}$。

3. 这些根满足 $e_1 = \frac{n(2n-1)}{3}$ 和 $e_2 = \binom{2n+1}{5} / \binom{2n+1}{1}$，由此得 $\sum \cot^4 = e_1^2 - 2e_2 = \frac{n(2n-1)(4n^2+10n-9)}{45}$。利用 $\cot^4\theta < \theta^{-4} < (1+\cot^2\theta)^2$ 作夹逼（squeezing），再令 $n \to \infty$，便得到 $\frac{\pi^4}{90}$。

4. 由下界得 $\sum_{m\le n} \frac{1}{m^2} > \frac{\pi^2 n(2n-1)}{3(2n+1)^2}$，所以误差小于 $\frac{\pi^2(6n+1)}{6(2n+1)^2} < \frac{\pi^2}{4n}$。这保证了当 $n \ge 2468$ 时准确度达到 0.001；事实上 $n \approx 1000$ 已经足够。

2.5 韦达乘积

1. 反复使用 $\sin\theta = 2\sin\frac{\theta}{2}\cos\frac{\theta}{2}$：乘积会逐项相消（telescopes）。令 $n \to \infty$（于是 $2^n \sin(\alpha/2^n) \to \alpha$），并取 $\alpha = \frac{\pi}{2}$，再利用 $\cos\frac{\pi}{4} = \sqrt{\frac12}$ 和 $\cos\frac{\theta}{2} = \sqrt{\frac12 + \frac12\cos\theta}$。

2. $\frac{\sin(\pi/3)}{\pi/3} = \frac{3\sqrt3}{2\pi} = \cos\frac{\pi}{6}\cos\frac{\pi}{12}\cos\frac{\pi}{24}\cdots$，其中 $\cos\frac{\pi}{6} = \frac{\sqrt3}{2}$，而 $\cos\frac{\pi}{12} = \frac{\sqrt{2+\sqrt3}}{2} = \frac{\sqrt6+\sqrt2}{4}$。

3. 取 8 个因子得 3.14157，四舍五入后是 3.1416。取 n 个因子后，误差约为 $\frac{\pi^3}{6\cdot 4^{n+1}}$，因为被略去的因子是 $\cos\frac{\pi}{2^{j+1}} \approx 1 - \frac{\pi^2}{2\cdot 4^{j+1}}$：每多一个因子，误差约缩小至原来的 4 分之一，而 $\log_{10} 4 \approx 0.6$。

4. 设 $x = \ln a$，同样的论证给出

$$\frac{a-1}{\ln a} = \prod_{j\ge 1} \frac{1 + a^{1/2^j}}{2},$$

所以 $\frac{1}{\ln 3} = \frac{1}{2}\prod_{j\geq 1}\frac{1+3^{1/2^j}}{2}$。取七个因子得 0.906，四舍五入后是 0.91（真确值是 0.9102）。这里的因子趋近 1 的方式如同 $1+\frac{\ln 3}{2^{j+1}}$，所以每多一个因子只能把误差减半（约 0.3 位数字）；而韦达（Viète）的因子趋近 1 的方式如同 $1-\frac{\pi^2}{2\cdot 4^{j+1}}$，每次把误差缩小至四分之一。

2.6 算术几何平均

1. $x_{n+1}-y_{n+1} = \frac{1}{2}(\sqrt{x_n}-\sqrt{y_n})^2 = \frac{(x_n-y_n)^2}{2(\sqrt{x_n}+\sqrt{y_n})^2}$，而 $(\sqrt{x_n}+\sqrt{y_n})^2 \geq 4y_n \geq 4y_0$。
2. $M(2,1) = 1.456791$；三步之后，x_3 与 y_3 已经有七位小数相同。
3. 迭代（iteration）的每一步都按 λ 的比例缩放，所以 $M(\lambda x, \lambda y) = \lambda M(x,y)$。在积分中代入 $x = \lambda u$，得 $\mathrm{I}(\lambda p, \lambda q) = \lambda^{-1}\mathrm{I}(p,q)$。$\mathrm{I} = \frac{\pi}{2M}$ 的两边都按 λ^{-1} 的比例缩放。
4. $M(1, \cos 30^\circ) \approx 0.9318$，所以周期变长至原来的 1.073 倍：即长了约 7.32%。

2.7 六位还是六百位

1. $(p^2+2q^2)^2 - 2(2pq)^2 = (p^2-2q^2)^2 = 1$；并且 $577^2 - 2\times 408^2 = 332\,929 - 332\,928 = 1$。
2. $y_{n+1} = \frac{2y_n^3+2}{3y_n^2}$，而 $y_{n+1}-\alpha = \frac{2y^3-3\alpha y^2+\alpha^3}{3y^2} = \frac{(y-\alpha)^2(2y+\alpha)}{3y^2}$。若 $y > \alpha$，则 $\frac{2y+\alpha}{3y^2} < \frac{3y}{3y^2} = \frac{1}{y} < \frac{1}{\alpha}$。
3. $y - \frac{1/y-a}{-1/y^2} = y + y - ay^2 = y(2-ay)$，而 $1-ay(2-ay) = (1-ay)^2$。
4. $y - \frac{y^{1/3}}{\frac{1}{3}y^{-2/3}} = y - 3y = -2y$。误差公式要求在根的附近 f′ 不等于零，而且 f″ 有界（bounded）；但这里 f′ 在 0 处是无限大。

2.8 平方再减二

1. $(2\cos u)^2 - 2 = 2\cos 2u$，所以由归纳法得 $x_n = 2\cos(2^{n-1}\theta)$。$x_3 = a$ 需要 $4\theta \equiv \pm\theta$：即 3θ 或 $5\theta \equiv 0 \pmod{2\pi}$。排除 $x_2 = a$ 的情况（不动点 $a = 2$ 和 $a = -1$）后，剩下 $\theta = \frac{2\pi}{5}, \frac{4\pi}{5}$：即 $a = \frac{\sqrt{5}-1}{2}$ 或 $a = -\frac{\sqrt{5}+1}{2}$，也就是一个 2-循环（cycle）的两个点。
2. 对于 M_5，数列是 4，14，$194 \equiv 8$，$62 \equiv 0$：是质数（prime）。对于 M_{11}，数列是 4，14，194，788，701，119，1877，240，282，$1736 \neq 0$：是合成数（composite），而事实上 $2047 = 23\times 89$。
3. 由 $A = t - \frac{1}{t} = \frac{3}{2}$ 得 $t = 2$ 和 $a = \frac{5}{2}$。于是 $x_2 = \frac{17}{4}$，$y_1 = \frac{Ax_1}{x_2} = \frac{15}{17}$，而 $\frac{\sqrt{x_2^2-4}}{x_2} = \frac{15/4}{17/4} = \frac{15}{17}$。
4. $x_k = t^{2^{k-1}} + t^{-2^{k-1}}$，而把 $x_1\cdots x_n$ 乘以 $t-t^{-1}$ 后，逐项相消得 $t^{2^n} - t^{-2^n}$。所求之

和变成 $(t-t^{-1})\sum_{n\ge1}\left(\frac{1}{t^{2^n}-1}-\frac{1}{t^{2^{n+1}}-1}\right)=\frac{t-t^{-1}}{t^2-1}=\frac{1}{t}$。当 $a=3$ 时，$t=\frac{3+\sqrt{5}}{2}$ 而 $\frac{1}{t}\approx0.3820$；级数 $\frac{1}{3}+\frac{1}{21}+\frac{1}{987}+\cdots$ 与此吻合。

2.9 逐步逼近 e 的分数

1. $T_0=e-1$，$T_1=3-e$，$T_2=7e-19$，$T_3=193-71e$，$T_4=1001e-2721$，由此得到 $\frac{19}{7}$，$\frac{193}{71}$，$\frac{2721}{1001}$。由于 $b_n=b_{n-2}-2(2n-1)b_{n-1}$，而（按归纳法）$b_{n-1}$ 与 b_{n-2} 异号，所以 b_n 与 b_{n-2} 同号。

2. $t(1-t)\le\frac{1}{4}$，所以 $T_n<\frac{e}{4^n n!}$；$-a_n/b_n$ 的误差是 $T_n/|b_n|$，现在改进了 4^n 倍。

3. 若 $e=\frac{p}{q}$，则 $qT_n=qa_n+pb_n$ 是整数。由 $T_n>0$ 可知它是正整数，所以至少是 1；但当 n 足够大时，$qT_n<\frac{qe}{n!}<1$。

4. 分部积分（integration by parts）两次（边界项为零），得 $U_n=\frac{1}{4}\big(U_{n-2}-2(2n-1)U_{n-1}\big)$，其中 $U_0=\frac{e^2-1}{2}$ 和 $U_1=\frac{1}{2}$。所以 $2^{2n+1}U_n=A_n+B_ne^2$，其中 A_n，B_n 为整数，而 $0<U_n<\frac{e^2}{4^n n!}$。若 $e^2=\frac{p}{q}$，则 $q2^{2n+1}U_n$ 是一个小于 $\frac{2qe^2}{n!}$ 的正整数，当 n 足够大时这是不可能的。所以这个论证确实证明了 e^2 是无理数（irrational）。

章

3

函数、不等式与严谨

知道一件事是对的，不等于知道它为甚么对；数学就活在这两者的差别之中。

本章内容

3.1	找出错处	STEP I 1991
3.2	倒过来的 AM–GM	STEP II 1993
3.3	凸性的威力	STEP II 2007
3.4	以面积证明不等式	STEP III 2011
3.5	π^e 还是 e^π?	STEP II 2013
3.6	斜率的界	STEP I 1988
3.7	$\pi^2 \leq 10$	STEP II 2014
3.8	由加法公式决定的函数	STEP III 2019
3.9	创造一个新函数	STEP II 2021
A	练习答案	第 237 页

把事情做对

> 在数学里，你并不是理解了甚么，你只是习惯了它们。
>
> — 冯·诺伊曼（John von Neumann）

以下是 $1 = 2$ 的一个“证明”（proof）。设 $a = b$。那么 $a^2 = ab$，所以 $a^2 - b^2 = ab - b^2$，因式分解后得 $(a+b)(a-b) = b(a-b)$。约去 $a-b$ 得 $a+b = b$，而由于 $a = b$，这就是说 $2b = b$，所以 $2 = 1$。

你大概已经看出问题所在：由于 $a = b$，“约去 $a-b$”其实就是除以零。但请留意，每一步看起来都多么合理，而错误又多么容易藏在一个例行的步骤之中。真正的数学里，真正的错误很少这样明显，有史以来最伟大的一些数学家也曾被它们骗倒。本章要谈的，正是能把这些错误揪出来的思考习惯：精确的定义、小心运用定理，以及对看来太过容易的论证抱持适度的怀疑。

为甚么严谨很重要

在中学，数学主要是求出答案。到了大学，数学主要是知道自己的答案是对的，并且能够说服别人。这个转变是中学数学与大学数学之间最大的一个变化，也可能令人不安。一件“显然”成立的事，为甚么要花一整页去证明？

答案是：“显然”并不可靠。分析（analysis）的历史上，满是看似显然、实则错误的结果：连续（continuous）函数除了可能在少数几个孤立点之外必定可导（differentiable）（安培（André-Marie Ampère）在 1806 年发表过一个“证明”；到了 1872 年，魏尔斯特拉斯（Weierstrass）找到一个处处不可导的连续函数）；连续函数的收敛（convergent）和仍然连续（柯西（Cauchy）相信这一点；阿贝尔（Abel）找到了反例（counterexample））；收敛级数（series）的各项可以重新排列而不改变其和（一般而言，不可以）。严谨（rigour）并不是吹毛求疵，而是分辨真确与似是而非的唯一可靠方法。

不等式：分析的基本工具

本章的第二个主题是不等式（inequality）。在中学数学里，大部分结果都是等式：这个等于那个。在分析里，大部分结果都是不等式：这个至多是那个。要证明一个数列收敛，你要证明误差小于任何容许的限度；要证明一个积分存在，你要证明某个和有界；要估计一个近似值有多好，你要给误差定出一个界。专业的分析学家花在证明不等式上的时间，远多于证明等式，而你将在这里遇到的经典不等式（AM–GM、詹森（Jensen）、杨

（Young）、赫尔德（Hölder）、柯西—施瓦兹（Cauchy–Schwarz））正是这一行的看家工具。

这些不等式本身也十分精彩。一条关于函数的不等式，配上一个精心挑选的例子，就能证明 $\pi^2 \leq 10$（Section 3.7）。关于对数图像形状的一个事实，就能判定 π^e 与 e^π 哪个较大（Section 3.5）。一位化学家关于量度误差的提问，引出了数学中最重要的一些多项式（Section 3.6）。

你在大学会遇到甚么

严谨是大学一年级分析（Analysis）课程的主题，也是另一门常称为数与集合（Numbers and Sets）或数学基础（Foundations）的课程的主题，你会在那里学习怎样精确地书写证明。你会遇到连续性和可导性的正式定义，以及带有精确假设（hypothesis）的定理（介值定理（intermediate value theorem）、中值定理（mean value theorem）、极值定理（extreme value theorem）），而这些假设确实缺一不可。往后的课程会把本章的不等式发展成一门门完整的学科：泛函分析（functional analysis）研究函数空间，并以 Section 3.4 和 Section 3.7 中那类积分来量度其中的函数；凸分析（convex analysis）和最优化（optimisation）由 Section 3.3 发展而来；逼近论（approximation theory）由 Section 3.6 发展而来；而像 Section 3.9 中那样的特殊函数（special functions），则在应用数学和物理学中随处可见。

贯穿本章的问题

三个大问题

1. **怎样知道一个论证是正确的？** 看似合理的论证通常会在哪些地方出错？又应该怎样检查？
2. **怎样证明一条不等式？** 经典不等式从何而来？又怎样选择用哪一条？
3. **函数究竟是甚么？** 一个函数可否由它具有的某种性质来定义，而不是由公式来定义？当旧有的函数不够用时，我们可否创造新的函数？

本章以找错开始（Section 3.1），然后逐一探讨各条经典不等式，每条都从不同的角度切入：一种不寻常的归纳法（induction）、凸性（convexity）、面积、图像的形状、多项式插值（polynomial interpolation）和试验函数（test function）。最后两节讨论由性质定义的函数：由加法定律（addition law）得出的 e^x 和 $\tanh x$，以及一个以反函数（inverse）形式定义的全新函数。

本章的数学思想

§	课题	思想
3.1	似是而非的错误论证	检查计算、逻辑和隐藏的假设
3.2	AM–GM	前推—后推归纳法
3.3	詹森不等式	凹（concave）曲线位于其弦之上
3.4	杨氏不等式	在图中比较面积
3.5	π^e 还是 e^π?	化为单一函数并研究其图像
3.6	二次函数的斜率	以已知量表示未知量
3.7	$\pi^2 \leq 10$	把试验函数代入一般的不等式
3.8	e^x 与 $\tanh x$	由函数方程（functional equation）定义的函数
3.9	W 函数	以反函数的形式创造一个函数

怎样阅读本章

和之前一样，请先自己尝试每道题目，然后才阅读解答。在这一章，请再多做一件事：读完每个解答之后，回头找出粗心的学生会做错的一两个步骤。隐藏的假设会在哪里悄悄混进来？哪条定理的哪个假设需要验证？这样带著批判眼光、主动地阅读，是你在上大学之前所能养成的最有用的一个习惯。

3.1 找出错处

题目

STEP I 1991，第 6 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/91-s1-q6`

引子

大多数考试题目都要求你写出一个论证。这一题却给你两个论证，要你把它们逐一拆解。这是一种截然不同的技巧，也是你在大学里最常用到的技巧：检查自己的作业、阅读书中的证明（proof）并判断是否相信它、批改朋友的解答、审阅论文。

这里的论证错得很有启发性。每一步看起来都像你见过上百次的东西。有些步骤含有计算上的失误，但更严重的是逻辑上的错误：即使计算无误，结论也推不出来。这类错误会一直残留到专业数学之中，而本章要谈的正是如何把它们找出来。

动笔之前

? **想一想** 当题目要求你“评论每一步”时，你在每一步应该检查甚么？动笔之前先列一张清单。

一张好的清单至少有三个问题。计算正确吗？这一步能从前面的步骤推出来吗？有没有隐藏的假设（例如函数是连续的、极限存在、驻点必定是极大点或极小点）？题目还问到结论会否无论如何都是对的。错误的论证也可以有正确的结论，所以这一点需要另行检查。

逐步解题

阶段 1 导数对吗？

想一想 用商法则（quotient rule）对 g 求导。$g'(x) = 0$ 在哪里成立？直接检查 $x = 1$ 和 $x = -1$。

由商法则，

$$g'(x) = \frac{6x^2(x^4+4) - (2x^3+3)\cdot 4x^3}{(x^4+4)^2} = \frac{-2x^6 - 12x^3 + 24x^2}{(x^4+4)^2} = \frac{-2x^2(x^4+6x-12)}{(x^4+4)^2}.$$

所以当 $x = 0$ 或 $x^4 + 6x - 12 = 0$ 时，$g'(x) = 0$。在 $x = 1$，$x^4 + 6x - 12 = -5$；在 $x = -1$，其值为 -17。两者都不是驻点（stationary point）。第一个断言是**错的**。

四次多项式 $x^4 + 6x - 12$ 在 $x = 0$ 的值是 -12，而且在两个方向上都趋向 $+\infty$，所以它有（至少）一个正根和一个负根；它恰好有两个实根，因为它的二阶导数 $12x^2$ 非负，所以它是凸（convex）的，最多只能穿过零两次。数值上，两个根是 $x \approx 1.386$ 和 $x \approx -2.247$。

阶段 2　函数值对吗？

$g(0) = \frac{3}{4}$ 是正确的，而 $g(1) = \frac{5}{5} = 1$。但 $g(-1) = \frac{-2+3}{1+4} = \frac{1}{5}$，不是 1。即使按它自己的思路，这个论证也有计算错误：它似乎假设了 g 是偶函数。

阶段 3　比较函数值能判断哪个是极大点吗？

想一想 假设一个函数恰好有两个驻点，驻点值分别是 $\frac{3}{4}$ 和 1。能否由此推出前者是极小点、后者是极大点？

不能。驻点的*性质*是局部（local）的性质：它取决于函数在该点附近的表现，而不是该点的函数值与别处的函数值孰大孰小。一个函数的局部极大值可以低于局部极小值（想想 $x + 2\sin x$），驻点也可以两者皆不是。在这里，在 $x = 0$ 附近，因式 $x^4 + 6x - 12$ 接近 -12，所以

$$g'(x) = \frac{-2x^2(x^4+6x-12)}{(x^4+4)^2} > 0 \qquad \text{对足够小的} x \neq 0.$$

导数在 0 的两侧都是正的：$x = 0$ 是一个驻*拐点*（stationary point of inflection），不是极小点。

阶段 4　驻点值能界定函数的范围吗？

停一停，试一试 即使驻点都求对了，为甚么仍不能断定 g 的每一个值都介乎最小与最大的驻点值之间？找出一个使 $g(x) < \frac{3}{4}$ 的 x 值。

驻点只描述*局部*的表现。全局（global）的表现还取决于定义域两端的情况，在这里即 $x \to \pm\infty$ 时的情况。由于当 $x \to \pm\infty$ 时 $g(x) \to 0$，而且每当 $2x^3 + 3 < 0$，即 $x <$

$-\sqrt[3]{3/2} \approx -1.14$ 时，$g(x) < 0$，所以这个函数会取负值。例如 $g(-2) = \frac{-13}{20}$。因此结论 $\frac{3}{4} \le g(x) \le 1$ 是**错的**（图 3.1.1）。

正确的说法是：g 递增至 $x \approx 1.386$ 处约为 1.083 的全局极大值，然后向 0 递减；在左边，它由 0 递减至 $x \approx -2.247$ 处约为 -0.668 的全局极小值，然后递增，经过 $x = 0$（在该处有一个水平拐点），一直升到极大值。所以对所有 x，$-0.668 \lesssim g(x) \lesssim 1.083$。

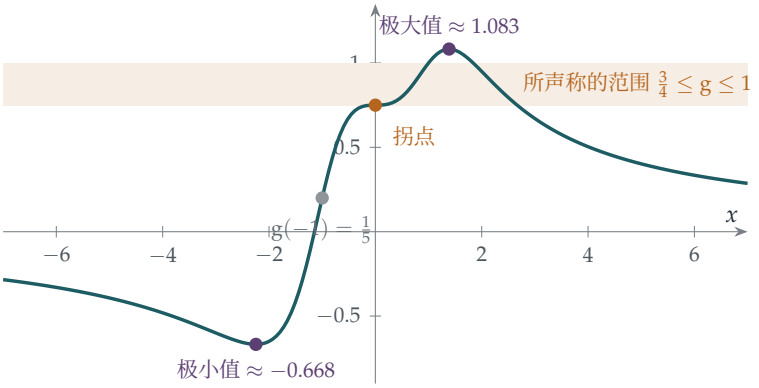


图 3.1.1 $g(x) = \frac{2x^3+3}{x^4+4}$ 的真实图像。该论证所声称的范围（琥珀色带）处处皆错：$x = -1$ 不是驻点，而且 $g(-1) = \frac{1}{5}$；$x = 0$ 是拐点，不是极小点；函数更远远超出色带之外，极大值高于 1，极小值低于 0。

阶段 5 积分

想一想 对 $-3(1-x)^{-4}$ 求导。你会得到 $(1-x)^{-3}$ 吗？

$\frac{d}{dx}\left[-3(1-x)^{-4}\right] = -3 \cdot (-4)(1-x)^{-5} \cdot (-1) = -12(1-x)^{-5}$。幂和系数都错了。正确的原函数（antiderivative）是

$$\int (1-x)^{-3}\,dx = \tfrac{1}{2}(1-x)^{-2} + C,$$

你可以求导来验证。

想一想 用正确的原函数，$\left[\frac{1}{2}(1-x)^{-2}\right]_{-1}^{3} = \frac{1}{2}\cdot\frac{1}{4} - \frac{1}{2}\cdot\frac{1}{4} = 0$。那么答案原来还是对的。真的吗？

并不是。微积分基本定理（fundamental theorem of calculus）$\int_a^b F'(x)\,dx = F(b) - F(a)$ 要求被积函数在整个区间 $[a,b]$ 上连续（continuous）。这里 $(1-x)^{-3}$ 在 $x = 1$ 处没有定义，而这一点位于 $[-1,3]$ 之内，而且在 $x = 1$ 附近函数值极其巨大。这个积分必须当作瑕积分（improper integral）来处理，在出问题的那一点拆开：

$$\int_{-1}^{1} (1-x)^{-3}\,dx = \lim_{\varepsilon\to 0^+}\left[\tfrac{1}{2}(1-x)^{-2}\right]_{-1}^{1-\varepsilon} = \lim_{\varepsilon\to 0^+}\left(\frac{1}{2\varepsilon^2} - \frac{1}{8}\right) = +\infty.$$

由 -1 到 1 的积分是无穷大（而由 1 到 3 的积分是 $-\infty$），所以 $\int_{-1}^{3}(1-x)^{-3}\,\mathrm{d}x$ **不存在**。结论“$=0$”是错的，原因不在于真正的值是另一个数，而在于根本没有值。 ■

回顾

- **三类错误。**计算错误（导数、$g(-1)$、原函数）、逻辑错误（靠比较函数值来判别驻点的性质；把驻点值当作全局的界）以及隐藏的假设（微积分基本定理中的连续性）。
- **答案对，理由错。**在第 (ii) 部分改正原函数之后，仍然得出答案 0，但论证依然无效。答案看似正确，并不能证明推理站得住脚。
- **局部与全局。**导数告诉你的是函数在一点附近的表现。要得出全局的结论，还需要更多资料：定义域两端的情况，或者函数表现异常之处的情况。

背后的数学

大数学家也会出错

分析学的奠基者也犯过一模一样的错误。柯西（Augustin-Louis Cauchy）在 1821 年的 *Cours d'analyse* 中“证明”了连续函数的收敛（convergent）和必定连续。1826 年，阿贝尔（Niels Henrik Abel）指出了一个反例（counterexample）：傅里叶级数（Fourier series）

$$\sin x - \tfrac{1}{2}\sin 2x + \tfrac{1}{3}\sin 3x - \cdots$$

对每一个 x 都收敛，而且每一项都连续，但它的和在 $x = \pm\pi$ 处出现跳跃（当 $-\pi < x < \pi$ 时，它等于 $x/2$）。要弄清楚柯西的论证究竟错在哪里，花了数十年时间，并由此产生了*一致收敛*（uniform convergence）的概念，如今每个修读分析的学生都要学。隐藏的假设，尤其是关于交换两个极限次序的假设，是整个分析学中最常见的错误来源。

带有前提的定理

这一题中有两个错误，都源于忽视了定理的前提（hypotheses）。

- *极值定理*（extreme value theorem）指出，在有界闭区间 $[a,b]$ 上的连续函数必定取得最大值和最小值。要找出它们，须检查端点、驻点，以及函数不可导的点：$|x|$ 在 $[-1,1]$ 上的最小值在 0 处取得，而这一点既非端点，亦非驻点。如果 f 在区间内部处处可导，则位于内部的最大点或最小点必定是驻点。在整条实数线上，没有端点可供检查，函数甚至可能根本没有最大值；你必须改为考察 $x \to \pm\infty$ 时的表现。
- *微积分基本定理*要求被积函数在 $[a,b]$ 上连续。如果不连续，积分可能仍然存在（试试

$\int_0^1 x^{-1/2}\,\mathrm{d}x$），也可能不存在（如本题），你必须用极限来检查。

在大学里，你遇到的每一条定理都附有一张精确的前提清单。训练的一部分，就是每次使用定理之前，都要把这些前提检查一遍。

发散积分的一个值

第 (ii) 部分还有一个耐人寻味的注脚。区间 $[-1,3]$ 关于出问题的点 $x=1$ 对称，而 $(1-x)^{-3}$ 关于该点是“奇”的。如果你挖去一个对称的缺口 $(1-\varepsilon,1+\varepsilon)$，再令 $\varepsilon\to 0$，两块无穷大的部分便会互相抵消，你确实会得到 0。这称为*柯西主值*（Cauchy principal value），在物理学和复分析中确实很有用。但它是一个刻意选取的*不同定义*，一个误打误撞得出这个值的论证，甚么也没有证明。

思维技巧

以批判的眼光阅读论证

每一步都要检查三件事：计算对不对，这一步在逻辑上是否成立，以及有甚么是没有明说却已假设了的？

然后另行检查结论：不正确的论证可以有正确的结论，而“正确”的答案也可以出自无效的论证。对自己的作业也要同样严格审视；你最容易相信的论证，正是你自己的论证。

延伸练习

1 **会说谎的代换。**试做 STEP I 1987，第 8 题（`step.maths.org/questions/87-s1-q8`）。一个代换（substitution）似乎显示某个正的积分等于一个负的积分。准确找出论证在哪里出错，然后正确地计算两个积分。

2 **是对还是错？**试做 STEP I 1998，第 3 题（`step.maths.org/questions/98-s1-q3`）：有四个断言要你证明或否证。对每一个断言，动笔之前先判断你预期是哪一种。

3 **所有马都是同一种颜色。**找出以下用归纳法（induction）写成的“证明”的漏洞：任何一匹马都与自己颜色相同；给定 $n+1$ 匹马，由归纳假设，前 n 匹颜色相同，后 n 匹也是；两组有重叠，所以全部 $n+1$ 匹颜色相同。

4 **一个确实存在的积分。**证明 $\int_{-1}^{1}|x|^{-1/2}\,\mathrm{d}x$ 存在，尽管被积函数在 0 处没有定义，并求出它的值。

延伸阅读

- **E. A. Maxwell, *Fallacies in Mathematics***（Cambridge University Press）。一本经典的谬误论证集，收录各种似是而非的错误论证，按课题编排，为高中第六学级（sixth form）学生而写。
- **Bernard Gelbaum & John Olmsted, *Counterexamples in Analysis***（Dover）。数以百计的例子，说明定理的前提为何重要。开始修读分析课程之后，不妨随手翻阅。
- **Imre Lakatos, *Proofs and Refutations***（Cambridge University Press）。一篇精彩的对话，探讨数学家如何发现并修正证明中的错误。

3.2 倒过来的 AM–GM

题目

STEP II 1993，第 8 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/93-s2-q8`

引子

算术平均—几何平均不等式（arithmetic mean–geometric mean inequality，简称 AM–GM）也许是数学中最有用的一条不等式。它说的是：对于正数，普通的平均数总是不小于“乘法意义下的平均数”，而且只有当所有数都相等时，等号（equality）才成立。它不用微积分就能解决最优化（optimisation）问题，可以用来证明其他不等式，并且在分析学（analysis）的每一个分支中都会出现。

这道题用一种奇特而优美的归纳法（induction）来证明它，这个方法由柯西（Augustin-Louis Cauchy）于 1821 年提出。它不是由 n 走到 $n+1$，而是由 2 跃到 4，再到 8、16，只涵盖 2 的幂，然后掉头往下走，把中间的数补上。它提醒我们，“归纳法”这个想法远比课本上的版本灵活。

动笔之前

网上试题库所载的这份试卷有一处误植：它写的是“$b_1 = a_2$”。那当然应该是 $b_1 = a_1$；意思是 b 这一列数就是 a 那一列数，再用它们的平均数的若干个副本补足。

? **想一想** 为甚么由 2^m 个数走到 2^{m+1} 个数，可能比由 n 走到 $n+1$ 容易？

因为 2^{m+1} 个数可以自然地分成两组,每组 2^m 个;你可以对每一半应用归纳假设(inductive hypothesis)，然后用两个数的情形把两半合起来。由 n 走到 $n+1$，就没有这样对称的结构可供利用。

逐步解题

阶段 1　两个数

$\left(\frac{a_1+a_2}{2}\right)^2 - a_1a_2 = \frac{a_1^2+2a_1a_2+a_2^2-4a_1a_2}{4} = \frac{(a_1-a_2)^2}{4} \geq 0.$ ■

阶段 2　加倍

$m = 1$ 的情形就是第 1 阶段。假设 $(*)$ 对 2^m 个数成立，现取 2^{m+1} 个正数。

> **想一想**　把这 2^{m+1} 个数分成前半和后半，设 P 和 Q 分别为两半的算术平均。归纳假设对每一半说了甚么？

对每一半应用 $(*)$，

$$a_1 \cdots a_{2^m} \leq P^{2^m} \qquad \text{及} \qquad a_{2^m+1} \cdots a_{2^{m+1}} \leq Q^{2^m}.$$

两式相乘（所有数都是正数），全部 2^{m+1} 个数的积不大于 $(PQ)^{2^m}$。现在由两个数的情形得 $PQ \leq \left(\frac{P+Q}{2}\right)^2$，而 $\frac{P+Q}{2}$ 正好是全部 2^{m+1} 个数的平均数。所以

$$a_1 \cdots a_{2^{m+1}} \leq (PQ)^{2^m} \leq \left(\frac{P+Q}{2}\right)^{2^{m+1}} = \left(\frac{a_1 + \cdots + a_{2^{m+1}}}{2^{m+1}}\right)^{2^{m+1}},$$

这就是 $m+1$ 时的 $(*)$。由归纳法可知，只要项数是 2 的幂，AM–GM 就成立。 ■

阶段 3　填补空隙

现在取任意的 n，并选取 m 使得 $2^m > n$。我们在数列 $a_1, \ldots, a_n$ 之后补上 $2^m - n$ 个它自己的平均数 A。

> **想一想**　补足后的数列 $b_1, \ldots, b_{2^m}$ 的平均数是多少？

其总和为 $(a_1 + \cdots + a_n) + (2^m - n)A = nA + (2^m - n)A = 2^m A$，所以平均数就是 A 本身。加入平均数的副本并不会改变平均数。因此，对 b 这一列数应用 $(*)$，得

$$a_1 \cdots a_n \cdot A^{2^m - n} \leq A^{2^m}.$$

> **停一停，试一试**　完成这个证明。

两边同除以 A^{2^m-n}（这是正数）：$a_1 \cdots a_n \leq A^n$。取 n 次方根，

$$(a_1 \cdots a_n)^{1/n} \leq A = \frac{a_1 + \cdots + a_n}{n}.$$ ■

回顾

- **沿另一条路径作归纳。**第 2 阶段是对 m 作归纳，涵盖 $n = 2, 4, 8, \ldots$；第 3 阶段则是由任何一个 2 的幂往下走到任何较小的 n 的方法（图 3.2.1）。
- **用平均数补足。**把数列加长而不改变其平均数，这个技巧值得记住；它能把关于任意 n 的问题，化为关于一个方便的 n 的问题。
- **等号何时成立？**沿着证明往回追溯，等号成立需要每一次使用两个数的情形时等号都成立，而这只有在所有数都相等时才会发生。

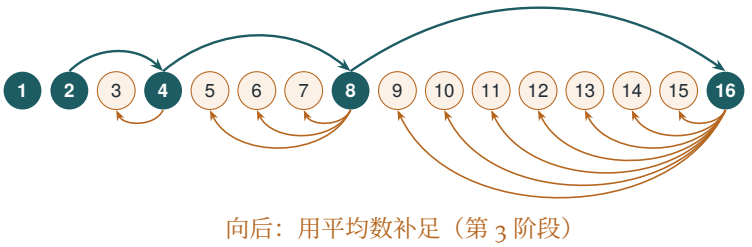


图 3.2.1 柯西的向前—向后归纳法。蓝绿色箭嘴以加倍的方式，证明 AM–GM 对 $2, 4, 8, 16, \ldots$ 个数成立；琥珀色箭嘴则把数列补足至下一个 2 的幂，从而填上其余每一个 n。

背后的数学

柯西的归纳法

柯西在他的《*Cours d'analyse*》(1821) 中发表了这个证明；我们在第 2 章谈过，正是这本书开创了分析学的严格处理。这种技巧现在称为*向前—向后归纳法* (forward–backward induction)：先对无穷多个情形组成的一个序列（这里是 2 的幂）证明命题成立，然后证明 n 的情形蕴含 $n-1$ 的情形。两者合起来，就能到达每一个 n。我们的第 3 阶段是由 2^m 直接跳到 n，但你也可以一步一步来：在 $n-1$ 个数之后补上一个它们的平均数，就能证明 n 的情形蕴含 $n-1$ 的情形。

一条不等式，多种证明

AM–GM 有数十种证明，把它们互相比较，是认识不同数学风格的好方法。

- **利用凸性（convexity）。**对数（logarithm）函数是凹（concave）函数，所以对数的平均不大于平均的对数。取指数之后，这就是 AM–GM；这也是 Section 3.3 的主题。
- **波利亚(George Pólya)的证明。**利用 $e^{x-1} \geq x$(e^{x-1} 在 $x=1$ 处的切线位于曲线下方)。设 A 为平均数，把这条不等式应用于每一个 a_i/A，然后相乘：$\prod \frac{a_i}{A} \leq e^{\sum(a_i/A-1)} = e^0 = 1$。波利亚说这个证明是他在梦中想到的。
- **利用磨光法（smoothing）。**如果其中两个数不相等，就把它们都换成两者的平均数。算术平均不变，而由两个数的情形可知，乘积会增大。重复这个步骤，所有数就会被推向平均数。

不用微积分的最优化

AM–GM 能把许多最优化问题变成一行就能解决的问题。在表面积固定的所有长方体盒子当中，哪一个体积最大？在周界固定的所有长方形当中，哪一个面积最大？（是正方形，由两个数的情形可知：$xy \leq \left(\frac{x+y}{2}\right)^2$，等号当 $x=y$ 时成立。）等号成立的条件告诉你最优值在哪里出现，而不等式则告诉你那确实是最优值，而不仅仅是一个驻点（stationary point）。可与 Section 3.1 中的告诫比较。

思维技巧

灵活的归纳法

归纳法不一定要按 $1, 2, 3, \ldots$ 的次序进行。只要一条路径能到达每一个情形，你就可以沿着它作归纳：可以加倍，可以用强归纳法（*strong induction*），也可以先在一个子数列（*subsequence*）上向前走，再向后走以填补空隙。

当由 n 到 $n+1$ 这显而易见的一步难以走通时，不妨问问哪一步会容易走，以及能否靠这一步到达每一个情形。

延伸练习

1. **后退一步。**补上一个平均数的副本，直接证明：若 AM–GM 对 n 个数成立，则它对 $n-1$ 个数也成立。
2. **一个递增数列。**把 AM–GM 应用于 $1+\frac{1}{n}, \ldots, 1+\frac{1}{n}, 1$ 这 $n+1$ 个数（其中 $1+\frac{1}{n}$ 有 n 个），证明 $\left(1+\frac{1}{n}\right)^n < \left(1+\frac{1}{n+1}\right)^{n+1}$。

3 **最好的盒子。**一个密封的长方体盒子的表面积为 S。把 AM–GM 用于三个不同的面的面积，证明其体积不大于 $(S/6)^{3/2}$，等号在盒子为正方体时成立。

4 **调和平均（harmonic mean）。**把 AM–GM 应用于 $1/a_i$ 这些数，证明调和平均 $\dfrac{n}{1/a_1 + \cdots + 1/a_n}$ 不大于几何平均。

延伸阅读

- **J. Michael Steele, *The Cauchy–Schwarz Master Class***（Cambridge University Press）。一本以问题带动、讲不等式的出色著作，正好由这个程度起步。第 2 章讲的就是 AM–GM。
- **Hardy, Littlewood & Pólya, *Inequalities***（Cambridge University Press）。经典参考书，由二十世纪三位伟大的分析学家撰写。
- **Arthur Engel, *Problem-Solving Strategies***（Springer）。第 7 章收录了大量来自数学奥林匹克的不等式题目。

3.3 凸性的威力

题目

STEP II 2007，第 7 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/07-s2-q7`

引子

在 Section 3.2 中，我们用一个巧妙的归纳法（induction）证明了 AM–GM 不等式。在这一节，它只需一行便水到渠成，因为它不过是某个更宏大的结果的特殊情形。这个更宏大的东西就是**凹性**（concavity）：向下弯的曲线所具有的形状，就像山顶一样。关于这类曲线，有一个简单的几何事实：曲线位于它的弦（chord）之上。单凭这一点，竟然可以证明范围广得惊人的各种不等式（inequality），从关于三角形的一个结果，到 AM–GM 不等式，以至概率（probability）、信息论（information theory）和经济学中的不等式。

到了大学，这个想法会发展成**凸性**（convexity），它是现代数学中最能贯通各个领域的重要主题之一。有些学科，例如凸优化（convex optimisation），整个就是建基于它之上。

动笔之前

想一想 画一条向下弯的曲线，例如 $(0, \pi)$ 上的 $y = \sin x$。在曲线上标出两点，并用直线把它们连起来。这条弦在曲线之上还是之下？

在曲线之下。对于凹曲线，每一条弦在两个端点之间都位于曲线之下。两个函数值的平均，就是弦的中点的高度；而函数在平均值处的取值，就是该中点正上方的曲线高度。因此，“弦在曲线之下”正正就是 $n = 2$ 时的琴生不等式（Jensen's inequality）：

$$\frac{f(x_1) + f(x_2)}{2} \leq f\left(\frac{x_1 + x_2}{2}\right).$$

若有 n 个点，则各点 $(x_k, f(x_k))$ 的平均就是它们的质心（centre of mass），它位于曲线下方的区域之内，所以同样在曲线之下（图 3.3.1）。

逐步解题

阶段 1　验证凹性

当 $0 < x < \pi$ 时，$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\sin x = -\sin x < 0$；当 $x > 0$ 时，$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\ln x = -\frac{1}{x^2} < 0$。 ■

阶段 2　三角形

想一想　应该用哪个函数、取哪些点？它们的平均是甚么？三角形的各个角是否都落在函数呈凹性的区间之内？

取 $\mathrm{f}(x) = \sin x$ 和三个点 A, B, C。每个角都严格介乎 0 与 π 之间，而 sin 在这个区间上是凹的，它们的平均是 $\frac{A+B+C}{3} = \frac{\pi}{3}$。由琴生不等式得

$$\frac{\sin A + \sin B + \sin C}{3} \le \sin\frac{\pi}{3} = \frac{\sqrt{3}}{2},$$

所以 $\sin A + \sin B + \sin C \le \frac{3\sqrt{3}}{2}$，而等号恰好在等边三角形（equilateral triangle）时成立。 ■

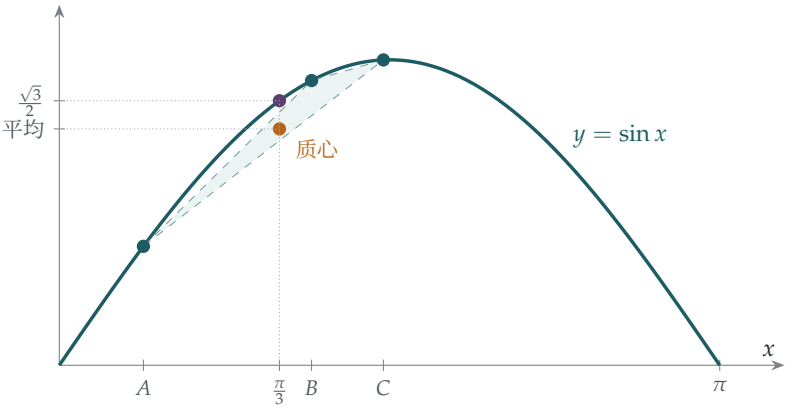


图 3.3.1　以三角形的三个角 $A = 0.4$、$B = 1.2$、$C = \pi - 1.6$ 说明 $\sin x$ 的琴生不等式。曲线上三点的质心（琥珀色）位于阴影三角形之内，因而在曲线之下。它的高度是 $\sin A$、$\sin B$、$\sin C$ 的平均，而它正上方的曲线高度是 $\sin\frac{\pi}{3}$。

阶段 3　一行证明 AM–GM

想一想　AM–GM 不等式比较的是一个积与一个和。哪个凹函数能把积变成和？

对数（logarithm）。把琴生不等式应用于 $\mathrm{f} = \ln$ 和正数 $t_1, \dots, t_n$：

$$\frac{\ln t_1 + \cdots + \ln t_n}{n} \le \ln\left(\frac{t_1 + \cdots + t_n}{n}\right).$$

左边是 $\ln\left(\sqrt[n]{t_1 \cdots t_n}\right)$。由于 ln 是递增函数，我们可以把两边的对数同时去掉，这就是 AM–GM 不等式。等号恰好在所有 t_k 都相等时成立。 ■

阶段 4 运用 AM–GM

> **想一想** 在 (a) 中，左边是四项之和，右边含有三个变量的积。你应该对多少项应用 AM–GM 不等式？那个 16 又为何有用？

对四个正数 x^4、y^4、z^4、16 应用 AM–GM 不等式：

$$\frac{x^4 + y^4 + z^4 + 16}{4} \geq \sqrt[4]{16\, x^4 y^4 z^4} = 2xyz.$$

两边乘以 4，得 $x^4 + y^4 + z^4 + 16 \geq 8xyz$。 ■

> **停一停，试一试** 现在求 $x^5 + y^5 + z^5 - 5xyz$ 的最小值。（提示：你想要的是 $x^5y^5z^5$ 的五次方根，所以需要五个数。另外两个数应该是甚么?）

对 $x^5, y^5, z^5, 1, 1$ 应用 AM–GM 不等式：

$$\frac{x^5 + y^5 + z^5 + 1 + 1}{5} \geq \sqrt[5]{x^5y^5z^5} = xyz,$$

所以 $x^5 + y^5 + z^5 - 5xyz \geq -2$。等号在 $x^5 = y^5 = z^5 = 1$ 时成立，即 $x = y = z = 1$，这时该式的值为 $3 - 5 = -2$。因此最小值是 -2。 ■

留意等号成立的条件在这里身兼两职：它说明这个界确实可以取到，所以 -2 是最小值，而不仅仅是一个下界（lower bound）。

回顾

- **一个想法，多个不等式。**关于凹曲线的一个几何事实，既给出了一个关于三角形的结果，又给出了 AM–GM 不等式，而所需的计算不过是验证一个二阶导数（second derivative）。
- **选择函数和点。**运用琴生不等式，就是要决定把它应用于哪个函数、取哪些点。对数能把积变成和，是常用之选。
- **补项以配合题目。**在 (b) 中，补上两个 1 令 AM–GM 不等式切合题目，正如在 Section 3.2 中补上平均值一样。

背后的数学

凸函数

在大学里，凸函数（convex function，即把凹函数上下倒转）的标准定义完全不涉及导数：若

$$\mathrm{f}\big(tx_1 + (1-t)x_2\big) \le t\,\mathrm{f}(x_1) + (1-t)\,\mathrm{f}(x_2) \qquad \text{对所有} x_1, x_2 \text{及} 0 \le t \le 1,$$

则称 f 为凸函数，这正正是说弦位于曲线之上。这个定义对不可微（not differentiable）的函数（例如 $|x|$）以及多元函数同样适用。丹麦工程师兼数学家琴生（Johan Jensen）在 1906 年提出了以他命名的不等式，使凸函数成为一门独立的学问。

无处不在的凸性

凸性在数学及其应用中随处可见。

- **概率。**琴生不等式指出，对于凹函数 f 和任意随机变量（random variable）X，有 $\mathrm{E}[\mathrm{f}(X)] \le \mathrm{f}(\mathrm{E}[X])$。在经济学中，若以 f 为一个凹的“效用”（utility）函数，这就解释了为何人们宁愿稳拿 £100，也不愿参与一场各有一半机会得到 £0 或 £200 的赌局，亦解释了为何会有保险。
- **信息论。**均匀分布（uniform distribution）使熵（entropy）达到最大，这个结果是数据压缩的基础，而它的证明就是 ln 的琴生不等式。
- **优化。**凸函数除了全局最小值之外，再没有其他局部极小值（local minima），所以“一直往下坡走”总能找到最佳答案。这就是凸优化问题能够有效率地求解的原因，也是人们花那么多工夫把问题改写成凸形式的原因。

在 Section 3.4 中，我们会看到凸性的另一面：函数与其反函数（inverse）之间的一种几何对偶（duality）。

思维技巧

找出一般原理

当几个结果的证明相似时，要找出它们背后的同一个原理。然后把这个原理证明一次，再多次应用。

AM–GM 不等式、关于正弦的三角形不等式，以及许多其他结果，全都是“凹曲线位于它的弦之上”的实例。认出这个原理，就能把一堆零散的技巧变成一套方法。

延伸练习

1 **定义。**STEP II 2018，第 2 题（`step.maths.org/questions/18-s2-q2`）的第 (i) 及 (ii) 部分由上面方框中以弦刻划凹性的定义出发，逐步建立三个点的琴生不等式。

2 **正弦之积。**试做 STEP II 2018，第 2 题（`step.maths.org/questions/18-s2-q2`）的第 (iii) 部分。

3 **锐角三角形。**证明 $\tan x$ 在 $(0, \frac{\pi}{2})$ 上是凸的，并由此推出在锐角三角形（acute-angled triangle）中，$\tan A + \tan B + \tan C \geq 3\sqrt{3}$。

4 **幂平均。**证明 x^2 是凸的，并由此推出 $\left(\frac{a_1+\cdots+a_n}{n}\right)^2 \leq \frac{a_1^2+\cdots+a_n^2}{n}$。试把这个结果诠释为关于一组数的方差（variance）的命题。

延伸阅读

- **J. Michael Steele, *The Cauchy–Schwarz Master Class***（Cambridge University Press）。第 6 章讲述凸性和琴生不等式。
- **Claudi Alsina & Roger Nelsen, *When Less Is More: Visualizing Basic Inequalities***（Mathematical Association of America）。以图形证明不等式，当中不少用到凸性。
- **Stephen Boyd & Lieven Vandenberghe, *Convex Optimization***（Cambridge University Press）。一部现代经典，可在网上免费阅读；开首几章说明凸性在实际应用中为何重要。

3.4 以面积证明不等式

题目

STEP III 2011，第 4 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/11-s3-q4`

引子

这里有一个不等式（inequality），证明它的图简单得可以画在一枚邮票上。画一条通过原点的递增曲线。曲线下方的面积，加上曲线左方的面积，总能盖住一个矩形。单凭这一个观察，就得出杨氏不等式（Young's inequality）$ab \le \frac{a^p}{p} + \frac{b^q}{q}$，而它又是通往赫尔德不等式（Hölder's inequality）和柯西–施瓦兹不等式（Cauchy–Schwarz inequality）的钥匙，这两个不等式是分析学中最重要的工具之一。

这幅图还蕴藏着一个更深刻的想法。它以完全对称的方式把一个函数与它的反函数（inverse function）配成一对；把这种配对加以推广，就是勒让德变换（Legendre transform），它把古典力学的两大表述方式联系起来，也把热力学的两种主要描述联系起来。

动笔之前

想一想 $\int_0^a \mathrm{f}(x)\,\mathrm{d}x$ 是曲线与 x 轴之间的面积。在同一幅图上，$\int_0^b \mathrm{f}^{-1}(y)\,\mathrm{d}y$ 代表哪一块面积?

$x = \mathrm{f}^{-1}(y)$ 的图像与 $y = \mathrm{f}(x)$ 是同一条曲线，只是横着看而已。所以 $\int_0^b \mathrm{f}^{-1}(y)\,\mathrm{d}y$ 是曲线与 y 轴之间、由高度 0 到高度 b 的面积。明白了这一点，第 (i) 部分只需要画图。

逐步解题

阶段 1 图形

想一想　画出以 $(0,0)$ 和 (a,b) 为顶点的矩形。用一种颜色涂上曲线下方由 0 到 a 的区域，再用另一种颜色涂上曲线左方由 0 到 b 的区域。这两个区域能盖住整个矩形吗？甚么时候既没有重叠，也没有多出来的部分？

这两个区域合起来盖住整个矩形 $[0,a]\times[0,b]$，通常还盖住矩形以外的一些面积（图 3.4.1）。矩形内的每一点 (x,y)，不是在曲线下方（因为 $x\le a$，所以属于第一个区域），就是在曲线左方（因为 $y\le b$，所以属于第二个区域）。因此涂色部分的总面积至少等于矩形的面积 ab。这就是 $(*)$。

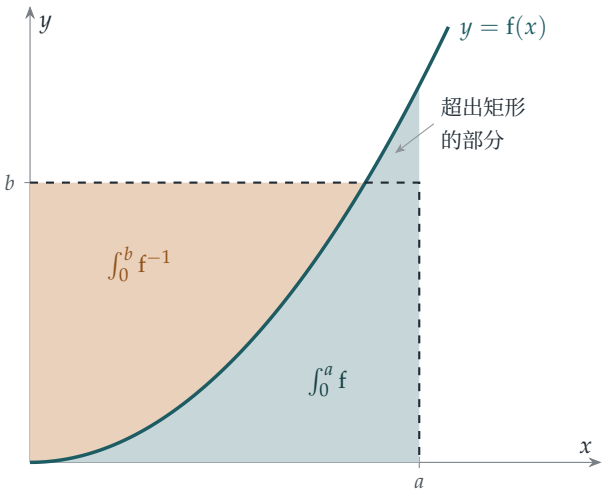


图 3.4.1　用一幅图看杨氏不等式。曲线下方由 0 到 a 的面积（蓝绿色）与曲线左方由 0 到 b 的面积（琥珀色）合起来盖住面积为 ab 的虚线矩形，而且还有剩余。恰好当顶点 (a,b) 位于曲线上时，才没有多出来的部分。

多出来的部分，就是曲线与矩形顶点之间那一小块曲边区域。它恰好在曲线通过顶点 (a,b) 时消失，也就是当

$$b=\mathrm{f}(a).$$

阶段 2　杨氏不等式

想一想　取 $\mathrm{f}(x)=x^{p-1}$，$\mathrm{f}^{-1}(y)$ 是甚么？计算两个积分，并用 $\frac{1}{p}+\frac{1}{q}=1$ 化简第二个。

由于 $p>1$，$\mathrm{f}(x)=x^{p-1}$ 连续而且递增，并且 $\mathrm{f}(0)=0$，而 $\mathrm{f}^{-1}(y)=y^{1/(p-1)}$。于是

$$\int_0^a x^{p-1}\,\mathrm{d}x=\frac{a^p}{p},\qquad \int_0^b y^{1/(p-1)}\,\mathrm{d}y=\frac{b^{1+1/(p-1)}}{1+1/(p-1)}.$$

由 $\frac{1}{p}+\frac{1}{q}=1$ 得 $q=\frac{p}{p-1}$，所以 $1+\frac{1}{p-1}=\frac{p}{p-1}=q$，第二个积分就是 $\frac{b^q}{q}$。因此 $(*)$ 变成

$$ab \le \frac{a^p}{p}+\frac{b^q}{q}. \qquad \blacksquare$$

等号应在 $b=\mathrm{f}(a)=a^{p-1}$ 时成立。验证：这时 $ab=a^p$，而 $b^q=a^{(p-1)q}=a^p$，所以右边是 $a^p\left(\frac{1}{p}+\frac{1}{q}\right)=a^p$。 $\blacksquare$

取 $p=q=2$，这就是 $ab\le\frac{a^2+b^2}{2}$，其实是 Section 3.2 中两个数的算术–几何平均不等式（AM–GM inequality）换了个样子。

阶段 3 反正弦

现在取 $[0,\frac{\pi}{2}]$ 上的 $\mathrm{f}(x)=\sin x$，它连续而且递增，并且 $\mathrm{f}(0)=0$，而在 $[0,1]$ 上 $\mathrm{f}^{-1}=\arcsin$。

> **想一想** $\int_0^a \sin x\,\mathrm{d}x$ 很容易。你会怎样求 $\arcsin y$ 的积分？

$\int_0^a \sin x\,\mathrm{d}x = 1-\cos a$。至于反正弦，把它看成 $\arcsin y\times 1$，用分部积分法（integration by parts）：

$$\int_0^b \arcsin y\,\mathrm{d}y = \Big[y\arcsin y\Big]_0^b - \int_0^b \frac{y}{\sqrt{1-y^2}}\,\mathrm{d}y = b\arcsin b+\sqrt{1-b^2}-1.$$

两式相加，由 $(*)$ 得 $ab\le b\arcsin b+\sqrt{1-b^2}-\cos a$。 $\blacksquare$

> **停一停，试一试** 现在推导关于 $\arcsin(t^{-1})$ 的不等式。应该选哪一个 a 的值？（试试最简单的那一个。）

取 $a=0$：这时 $0\le b\arcsin b+\sqrt{1-b^2}-1$，即 $b\arcsin b\ge 1-\sqrt{1-b^2}$。（这不过是说 $\int_0^b \arcsin y\,\mathrm{d}y\ge 0$。）设 $b=1/t$，其中 $t\ge 1$，于是 $0<b\le 1$，再除以 b：

$$\arcsin\left(\frac{1}{t}\right)\ge t\left(1-\sqrt{1-\frac{1}{t^2}}\right)=t-\sqrt{t^2-1}. \qquad \blacksquare$$

回顾

- **以图为证。**只要看出合适的图，第 (i) 部分就是一个完整的证明。把几何事实转化为积分之间的不等式，是一种威力强大的技巧。
- **把一般结果特殊化。**同一个一般的不等式，选取不同的 f，就得出杨氏不等式和一个关于 arcsin 的不等式。
- **最简单的选择可能正是对的选择。**最后的推导用了 $a=0$，这时不等式几乎是显然的。简单的情况一定要试试。

背后的数学

从杨氏到赫尔德

杨（William Henry Young）在 1912 年发表他的不等式。它之所以重要，在于它所引出的结果。假设 $\sum a_i^p = 1$ 及 $\sum b_i^q = 1$。对每一对数应用杨氏不等式，然后相加，

$$\sum a_i b_i \le \frac{\sum a_i^p}{p} + \frac{\sum b_i^q}{q} = \frac{1}{p} + \frac{1}{q} = 1.$$

把各个 a 和各个 b 按比例缩放，使两个和都等于 1，就得到赫尔德不等式：对正数，

$$\sum a_i b_i \le \Big(\sum a_i^p\Big)^{1/p} \Big(\sum b_i^q\Big)^{1/q}.$$

取 $p = q = 2$，这就是柯西—施瓦兹不等式 $\sum a_i b_i \le \sqrt{\sum a_i^2}\sqrt{\sum b_i^2}$，它说的是两个向量夹角的余弦最多为 1。赫尔德不等式是 L^p 空间理论的基础；在这些函数空间里，“距离”是用 $\left(\int |f|^p\right)^{1/p}$ 来量度的，你会在泛函分析（functional analysis）和概率的课程中遇到它们。

勒让德变换

图 3.4.1 对 f 和 f^{-1} 的处理完全对称。记 $F(a) = \int_0^a \mathrm{f}$ 及 $G(b) = \int_0^b \mathrm{f}^{-1}$。不等式说的是 $F(a) + G(b) \ge ab$，在曲线上等号成立，所以

$$G(b) = \max_a \big(ab - F(a)\big).$$

这个把凸函数（convex function）F 变成另一个凸函数 G 的运算，就是勒让德变换，而连续做两次就会回到 F。在力学中，它把拉格朗日量（Lagrangian，位置与速度的函数）转换成哈密顿量（Hamiltonian，位置与动量的函数）：这是牛顿（Newton）定律的两种等价表述，你会在大学二年级的课程中学到。在热力学中，它在能量、焓（enthalpy）和自由能（free energy）之间互相转换。对 $F(a) = a^p/p$，变换的结果是 $G(b) = b^q/q$，这就是杨氏不等式中两个指数由 $\frac{1}{p} + \frac{1}{q} = 1$ 联系起来的原因。

思维技巧

从图形得出不等式

积分就是面积。在一幅精心挑选的图中比较面积，可以证明一个从代数上难以看出的不等式。

留意这样的图：其中一个区域明显包含另一个区域，或者几个区域明显盖住第三个区域。然后把每块面积转化为积分。

延伸练习

1 **1987 年的版本。**试做 STEP I 1987, 第 6 题（`step.maths.org/questions/87-s1-q6`）。它用同一类图来求反函数的积分。用求导来验算你的答案。

2 **柯西–施瓦兹。**用 $p = q = 2$ 的杨氏不等式，证明对正数有 $a_1b_1 + a_2b_2 + a_3b_3 \le \sqrt{a_1^2 + a_2^2 + a_3^2}\,\sqrt{b_1^2 + b_2^2 + b_3^2}$。（提示：先处理两个平方根都等于 1 的情况。）

3 **指数版本。**在 $(*)$ 中取 $\mathrm{f}(x) = e^x - 1$，证明对 $a, b \ge 0$，

$$ab \le e^a - 1 - a + (1+b)\ln(1+b) - b.$$

等号何时成立？

4 **这个界有多好？** 较难 用计算机比较 $t = 1, 2, 10$ 时的 $\arcsin(1/t)$ 与 $t - \sqrt{t^2-1}$。证明当 t 很大时，$t - \sqrt{t^2-1} \approx \frac{1}{2t}$，而 $\arcsin(1/t) \approx \frac{1}{t}$。为甚么这个界相差了一个因子 2？（看看 $a = 0$ 时的图：多出来的部分有多少被浪费了？）

延伸阅读

- **J. Michael Steele, *The Cauchy–Schwarz Master Class***（Cambridge University Press）。第 1 章和第 9 章讲述柯西–施瓦兹不等式和赫尔德不等式，途中也谈到杨氏不等式。
- **Roger Nelsen, *Proofs Without Words***（Mathematical Association of America）。收录了杨氏不等式和许多其他的图形证明。
- **V. I. Arnold, *Mathematical Methods of Classical Mechanics***（Springer）。第 14 节正是用这幅图来解释勒让德变换。内容较深，但相关的一节很短。

3.5 π^e 还是 e^π?

题目

STEP II 2013，第 1 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/13-s2-q1`

引子

π^e 和 e^π，哪一个较大？计算机一秒钟就能告诉你：$e^\pi \approx 23.14$ 而 $\pi^e \approx 22.46$。但计算机只告诉你结果是这样，却没有告诉你为甚么，更无法一次过解答无穷多条同类的问题。3^π 比 π^3 大吗？除了显而易见的 $n = m$ 之外，还有没有整数满足 $n^m = m^n$？

原来这些问题全都关乎同一条曲线 $y = \ln x$，以及与它相交的直线。本题说明，只要运用得当，一幅草图就能解决一些单靠代数看似无从入手的大小比较，也说明 e 这个数怎样自然而然地成为分界点。

动笔之前

想一想 取对数（logarithm）后，$a^b = b^a$ 等同于 $b \ln a = a \ln b$。试把这条式重新整理，使 a 只出现在一边，b 只出现在另一边。

$b \ln a = a \ln b$ 等同于

$$\frac{\ln a}{a} = \frac{\ln b}{b}.$$

因此，每次比较 a^b 与 b^a，其实都是比较同一个函数 $h(x) = \frac{\ln x}{x}$ 在 a 和 b 两处的值。而 $\frac{\ln a}{a}$ 正是由原点（origin）连到曲线 $y = \ln x$ 上的点 $(a, \ln a)$ 的直线的斜率（gradient）。这就是本题与直线有关的原因。

逐步解题

阶段 1　通过原点的切线

想一想　若 $y = mx$ 与 $y = \ln x$ 在 $x = x_0$ 处相切，哪两条方程必须成立？

直线与曲线必须相交，即 $mx_0 = \ln x_0$，而且斜率相同，即 $m = 1/x_0$。把第二式代入第一式，得 $1 = \ln x_0$，所以 $x_0 = e$ 及 $m = 1/e$。由原点引出的切线（tangent）与曲线相切于 $(e, 1)$。

若 $y = mx$ 改为与曲线相交于 $x = a$ 和 $x = b$，则 $\ln a = ma$ 及 $\ln b = mb$，所以 $\frac{\ln a}{a} = m = \frac{\ln b}{b}$，由此得 $b \ln a = a \ln b$，因而 $a^b = b^a$。 ■

至于草图：一条通过原点、比切线稍为平坦的直线会与曲线相交两次，切点两侧各一次，所以 $a < e < b$。斜率大于 $1/e$ 的直线则完全碰不到曲线，因为 $\ln x$ 是凹（concave）函数，整条曲线都在其切线 $y = x/e$ 之下。

阶段 2　与 y 轴相交于原点上方的直线

想一想　现在直线是 $y = mx + c$，其中 $c > 0$。试以 m 和 c 表示 $\frac{\ln p}{p}$ 和 $\frac{\ln q}{q}$。

由 $\ln p = mp + c$ 及 $\ln q = mq + c$，得

$$\frac{\ln p}{p} = m + \frac{c}{p}, \qquad \frac{\ln q}{q} = m + \frac{c}{q}.$$

由于 $c > 0$ 且 $p < q$，可知 $\frac{c}{p} > \frac{c}{q}$，所以 $\frac{\ln p}{p} > \frac{\ln q}{q}$。两边乘以 $pq > 0$，得 $q \ln p > p \ln q$，因此 $p^q > q^p$。 ■

从几何角度看：若通过 $(p, \ln p)$ 和 $(q, \ln q)$ 两点的直线与 y 轴相交于原点上方，那么从原点望去，前一点的仰角比后一点的更陡。

阶段 3　e 右方的弦

想一想　设 $e \le p < q$。由 $(p, \ln p)$ 连到 $(q, \ln q)$ 的弦（chord），其斜率小于曲线在 p 处的斜率。为甚么？又为甚么这样会令它的 y 截距（intercept）为正数？

因为 $\ln x$ 是凹函数，它的斜率 $1/x$ 递减，所以由 p 到 q 的弦比 p 处的切线平坦：弦的斜率 s 满足 $s < \frac{1}{p}$。弦的 y 截距是

$$c = \ln p - sp > \ln p - \frac{1}{p} \cdot p = \ln p - 1 \ge 0,$$

当中用到 $p \ge e$。所以这条弦与 y 轴相交于原点上方。 ■

停一停，试一试 利用这个结果和第 2 阶段，判断 π^e 与 e^π 哪一个较大。

取 $p = e$ 及 $q = \pi$。连接这两点的弦的截距为正数，所以根据第 2 阶段，$p^q > q^p$，即

$$e^\pi > \pi^e.$$

阶段 4 与切线平行的弦

现在假设由 p 到 q 的弦的斜率刚好是 $\frac{1}{e}$，与切线 $y = \frac{x}{e}$ 的斜率相同。

想一想 这条弦与 $(e, 1)$ 处的切线平行。由于 $\ln x$ 是凹函数，曲线位于该切线之下。这条弦与 y 轴相交于原点的上方还是下方?

这条弦是直线 $y = \frac{x}{e} + c$，其中 c 为某个常数。它的两个端点都在曲线上，而曲线位于切线 $y = \frac{x}{e}$ 上或其下方，并且只在 $x = e$ 处与切线相接。两个端点不可能同时位于 $x = e$，所以最少有一个端点严格位于切线下方，于是整条平行的弦都在切线之下：$c < 0$。现在以 $c < 0$ 重做第 2 阶段：不等式的方向反转，$\frac{\ln p}{p} < \frac{\ln q}{q}$，因此 $q^p > p^q$。 ■

回顾

- **化为单一函数。**比较 a^b 与 b^a，变成了比较 $h(a)$ 与 $h(b)$，其中 $h(x) = \frac{\ln x}{x}$。
- **有理有据的草图。**每一处“藉草图说明”的背后，都有确切的事实支持：ln 的凹性（弦在曲线之下，切线在曲线之上），以及截距的正负。
- **e 的特殊角色。**由原点引出的切线在 $x = e$ 处与曲线相切，而这一点正是以上所有比较的分界点。

背后的数学

$\frac{\ln x}{x}$ 的图像

本题的一切，都可以在 $h(x) = \frac{\ln x}{x}$ 的图像中看到（图 3.5.1）。它的导数（derivative）是 $h'(x) = \frac{1-\ln x}{x^2}$，所以 h 在 $(0, e)$ 上递增，在 $x = e$ 处达到最大值（maximum）$\frac{1}{e}$，然后递减并趋向 0。一条高度为 m 的水平线，若高度介乎 0 与 $\frac{1}{e}$ 之间，便会与图像相交两次，交点为 $a < e < b$，且 $h(a) = h(b)$，这正正就是 $a^b = b^a$。在 e 的右方，h 递减，所以当 $e \le p < q$ 时，我们有 $h(p) > h(q)$，即 $p^q > q^p$。特别地，$e^\pi > \pi^e$；更一般而言，对每个 $x \neq e$ 都有 $e^x > x^e$：在这场较量中，没有其他数能胜过 e。

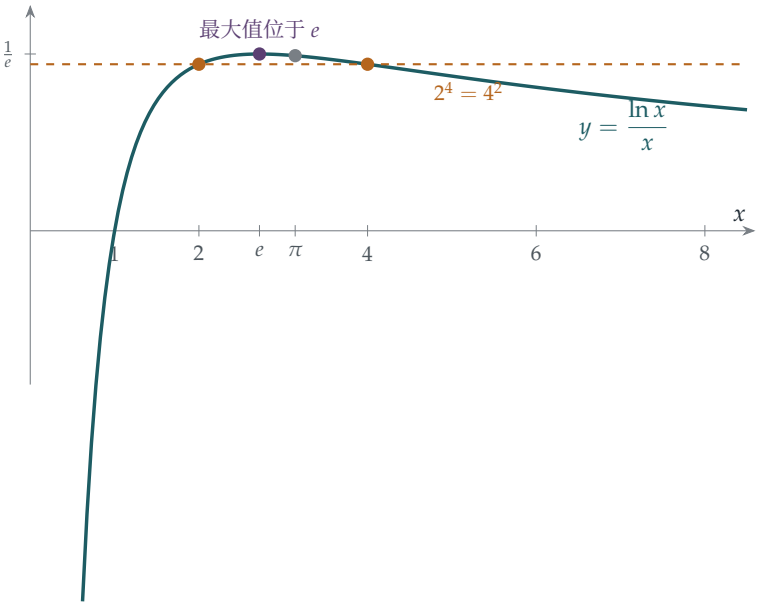


图 3.5.1　$h(x) = \frac{\ln x}{x}$ 的图像。它上升至 $x = e$ 处的最大值 $\frac{1}{e}$，然后下降。因为 $h(e) > h(\pi)$，所以 $e^\pi > \pi^e$。高度为 $\frac{\ln 2}{2}$ 的水平线（琥珀色）与图像相交于 $x = 2$ 和 $x = 4$：这是唯一一对满足 $n^m = m^n$ 的相异整数。

$x^y = y^x$ 的所有解

哪些相异的正实数对满足 $x^y = y^x$？有一个巧妙的方法可以把它们全部找出来。设 $y = tx$，其中 $t > 1$；那么 $x^{tx} = (tx)^x$ 给出 $x^t = tx$，所以 $x^{t-1} = t$，并且

$$x = t^{1/(t-1)}, \qquad y = t^{t/(t-1)}.$$

写成 $t = 1 + \frac{1}{n}$，便得 $x = \left(1 + \frac{1}{n}\right)^n$ 及 $y = \left(1 + \frac{1}{n}\right)^{n+1}$。取 $n = 1$ 时，这就是 $(2, 4)$；取 $n = 2$ 时则是 $\left(\frac{9}{4}, \frac{27}{8}\right)$，而的确有 $\left(\frac{9}{4}\right)^{27/8} = \left(\frac{27}{8}\right)^{9/4}$。当 $n \to \infty$ 时，两者都趋向 e，也就是两个分支会合之处。1728 年，丹尼尔·伯努利（Daniel Bernoulli）与哥德巴赫（Christian Goldbach）曾在书信中讨论这个问题，其后欧拉（Euler）亦有研究。

最经济的底

函数 $x^{1/x} = e^{h(x)}$ 同样在 $x = e$ 处取得最大值。这解答了一个有趣的问题：假如一个记数系统的“成本”是它的底（base）乘以所需的数位数目，那么最经济的底就是 e；若只考虑整数，则是 3（仅仅胜过 2）。有少数实验性质的电脑采用过以 3 为底的记数法，例如 1958 年苏联的 Setun。

思维技巧

重新整理以分离变量，然后研究单一函数

一条不等式或方程，若当中两个量纠缠在一起，往往可以化为关于同一个函数在两点取值的命题。

一旦找到这个函数，它的图像（在哪里递增、在哪里递减、在哪里达到顶点）就能一次过回答所有问题。在这里，a^b 与 b^a 的比较变成了 $h(a)$ 与 $h(b)$ 的比较。

延伸练习

1 **整数。**试做 STEP I 1988，第 1 题（`step.maths.org/questions/88-s1-q1`）。

2 **更多比较。**试做 STEP II 2025，第 3 题（载于 OCR 提供下载的 2025 年试卷，`www.ocr.org.uk/Images/750486-2025-past-papers.zip`）。该题利用本节的图像，解决更多 a^b 与 b^a 之间的比较，其中一个要靠 e 的数值界限才能完成。

3 **一行证明。**利用当 $x \neq 0$ 时 $e^x > 1 + x$，并取 $x = \frac{\pi}{e} - 1$，直接证明 $e^\pi > \pi^e$。

延伸阅读

- **Eli Maor, *e: The Story of a Number***（Princeton University Press）。讲述 e 的历史，包括方程 $x^y = y^x$。
- **Paul Nahin, *An Imaginary Tale***（Princeton University Press）。不只谈 $\sqrt{-1}$：书中满载正是这一类的有趣计算。
- **Michael Spivak, *Calculus***（Cambridge University Press）。第 11 章以完全严谨的方式，讲解如何运用导数了解图像的形状。

3.6 斜率的界

题目

STEP I 1988，第 7 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/88-s1-q7`

引子

假如你知道一个函数在某个区间上的值总是介乎 -1 与 1 之间，你能否说出它有多陡？对一般的函数来说，答案是不能：$\sin(1000x)$ 的值介乎 -1 与 1 之间，但它的斜率可以大至 1000。不过二次函数（quadratic）不可能这样摆动。这道题目要证明的是：一个困在方框 $|x| \leq 1$、$|y| \leq 1$ 之内的二次函数，在方框内的斜率永远不会大于 4，而且 4 已是最佳的界。

这个问题的来历出人意表。1887 年，以元素周期表闻名的化学家门捷列夫（Dmitri Mendeleev）正在研究溶液的密度如何随浓度变化，并用二次函数拟合他的数据。他需要知道数据中的误差对曲线斜率的影响最坏可以有多大，于是证明了与本题完全相同的结果。两年后，年轻的马可夫（Andrei Markov）解决了任意次数多项式（polynomial）的一般问题，而答案直接引出了数学中最重要的一些多项式。

动笔之前

想一想 $f'(x)$ 的公式只用到 f 的三个值：在 -1、0 和 1 的值。为甚么三个值就足以确定一个二次函数的导数（derivative）？

二次函数有三个系数，所以只要知道它在任意三点的值，就能把它确定下来。它的导数同样是由这些系数决定的函数，因此必定可以用这三个值表示出来。题中的公式正好告诉你怎样表示。而这恰恰是你所需要的，因为假设 $|f(x)| \leq 1$ 告诉你的是 f 的值，而结论谈的却是它的导数。

逐步解题

阶段 1　公式

$\mathrm{f}(1)=a+b+c$，$\mathrm{f}(-1)=a-b+c$，而 $\mathrm{f}(0)=c$。所以右方是

$$(a+b+c)\left(x+\tfrac{1}{2}\right)+(a-b+c)\left(x-\tfrac{1}{2}\right)-2cx.$$

x 的系数是 $(a+b+c)+(a-b+c)-2c=2a$，常数项是 $\tfrac{1}{2}(a+b+c)-\tfrac{1}{2}(a-b+c)=b$。所以右方等于 $2ax+b=\mathrm{f}'(x)$。 ■

阶段 2　上界

> **想一想**　对公式取绝对值，然后运用三角不等式（triangle inequality），再配合 $|\mathrm{f}(1)|,|\mathrm{f}(-1)|,|\mathrm{f}(0)|\le 1$。你会得到 $|\mathrm{f}'(x)|$ 的甚么上界（写成 x 的函数）？

根据三角不等式 $|u+v+w|\le|u|+|v|+|w|$，

$$|\mathrm{f}'(x)|\le|\mathrm{f}(1)|\,|x+\tfrac{1}{2}|+|\mathrm{f}(-1)|\,|x-\tfrac{1}{2}|+2|\mathrm{f}(0)|\,|x|\le|x+\tfrac{1}{2}|+|x-\tfrac{1}{2}|+2|x|.$$

> **停一停，试一试**　证明当 $|x|\le 1$ 时，$|x+\tfrac{1}{2}|+|x-\tfrac{1}{2}|+2|x|\le 4$。（按正负号分情况讨论。）

若 $|x|\le\tfrac{1}{2}$，则 $x+\tfrac{1}{2}\ge 0$ 而 $x-\tfrac{1}{2}\le 0$，所以 $|x+\tfrac{1}{2}|+|x-\tfrac{1}{2}|=(x+\tfrac{1}{2})-(x-\tfrac{1}{2})=1$，总和是 $1+2|x|\le 2$。

若 $\tfrac{1}{2}\le|x|\le 1$，两个括号的正负号都与 x 相同，所以 $|x+\tfrac{1}{2}|+|x-\tfrac{1}{2}|=2|x|$，总和是 $4|x|\le 4$。

无论哪种情况，都有 $|\mathrm{f}'(x)|\le 4$。 ■

这个上界只会在 $x=\pm1$ 处达到；在区间中间的一半，斜率最多只是 2。

阶段 3　一个极端例子

> **想一想**　要在 $x=1$ 处达到上界，第 2 阶段中的每个不等式都必须取等号。$\mathrm{f}(1)$、$\mathrm{f}(-1)$ 和 $\mathrm{f}(0)$ 必须取甚么值？正负号又是甚么？

在 $x=1$ 处，公式是 $\mathrm{f}'(1)=\tfrac{3}{2}\mathrm{f}(1)+\tfrac{1}{2}\mathrm{f}(-1)-2\mathrm{f}(0)$。要令它等于 4，我们需要 $\mathrm{f}(1)=1$、$\mathrm{f}(-1)=1$ 和 $\mathrm{f}(0)=-1$。通过 $(-1,1)$、$(0,-1)$ 和 $(1,1)$ 的二次函数是

$$\mathrm{f}(x)=2x^2-1,$$

所以 $a = 2$，$b = 0$，$c = -1$。在 $|x| \le 1$ 上，$x^2 \in [0,1]$，所以 $2x^2 - 1 \in [-1,1]$，而 $\mathrm{f}'(1) = 4$。 ■

回顾

- **用已知的量表示想求的量。**公式把导数改写成 f 的值，而假设提供的资料正正在这些值上。
- **三角不等式。**用各项绝对值之和来估计一个和的上界，是分析学（analysis）中最常见的手法之一。
- **从等号成立的条件找出极值情况。**追问每个不等式何时取等号，就直接得出例子 $2x^2 - 1$。

背后的数学

马可夫不等式与切比雪夫多项式

马可夫在 1889 年证明：若 p 是 n 次多项式，且在 $[-1,1]$ 上 $|p(x)| \le 1$，则

$$|p'(x)| \le n^2 \qquad \text{当} |x| \le 1.$$

当 $n = 2$ 时，这就是我们的上界 4。极端情况是*切比雪夫多项式*（Chebyshev polynomials）T_n，其定义为

$$T_n(\cos\theta) = \cos n\theta.$$

例如 $\cos 2\theta = 2\cos^2\theta - 1$ 给出 $T_2(x) = 2x^2 - 1$，即我们的极端二次函数；而 $\cos 3\theta = 4\cos^3\theta - 3\cos\theta$ 给出 $T_3(x) = 4x^3 - 3x$。由于 $T_n(x) = \cos n\theta$，这些多项式在 -1 与 1 之间来回振荡的次数，已是 n 次多项式所能做到的最多，而且它们在两端也是最陡的（图 3.6.1）。马可夫的弟弟 Vladimir 在 1892 年把结果推广到高阶导数。

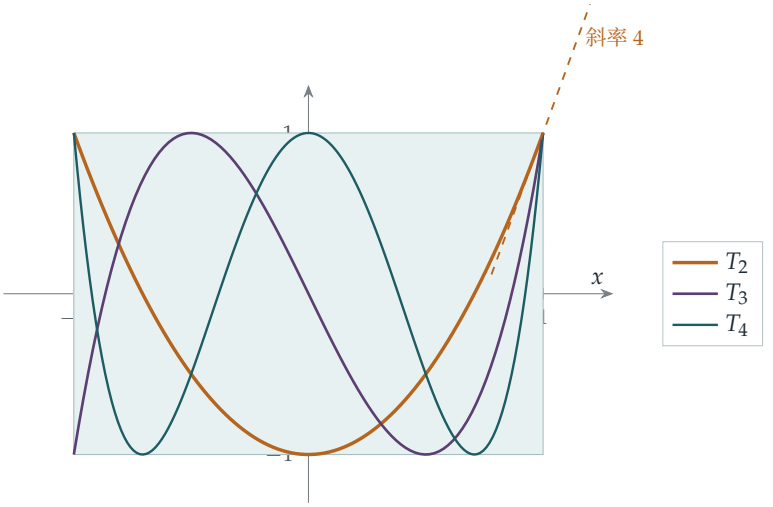


图 3.6.1　切比雪夫多项式 $T_2 = 2x^2 - 1$、$T_3 = 4x^3 - 3x$ 和 $T_4 = 8x^4 - 8x^2 + 1$ 都留在方框 $|x| \le 1$、$|y| \le 1$ 之内，并且尽可能多次触及方框的顶边和底边。在同次数的多项式中，它们在两端最陡：$T_n'(1) = n^2$。虚线是 T_2 在 $x = 1$ 处的切线，斜率为 4。

插值

题目中的公式是*拉格朗日插值*（Lagrange interpolation）的一个例子：已知函数在若干点的值，写出通过这些点而次数最低的唯一多项式，再用它来逼近该函数或其导数。这里的三个点是 $-1, 0, 1$。这类公式在数值分析（numerical analysis）中无处不在，例如用来从量度所得的数据估计导数。马可夫不等式就是随之而来的警告：函数值的小误差，在导数中最多可以被放大 n^2 倍，这正是门捷列夫当年担心的问题。

切比雪夫多项式的实际应用

切比雪夫多项式是逼近理论（approximation theory）的主力工具。在 T_n 的根处对函数插值，而不是在等距的点上插值，可以把误差控制住；若用等距的点，高次插值在区间两端附近可以错得非常离谱（龙格现象，Runge's phenomenon）。那些把函数计算至机器精度极限、或以极高准确度求解微分方程（differential equation）的软件，内部往往使用切比雪夫展开式。

思维技巧

用已知的量表示未知的量

如果假设告诉你的是某一类量，而结论谈的是另一类量，就去找一个把两者连系起来的恒等式，然后逐项估计上界。

得到一个界之后，还要问它是否最佳。沿着等号成立的情况倒推回去，通常会把你带到极端例子，而极端例子本身往往就很值得研究。

延伸练习

1 **一般的想法。**证明若 f 是二次函数，则 $\mathrm{f}(x) = \mathrm{f}(-1)\dfrac{x(x-1)}{2} + \mathrm{f}(0)(1-x^2) + \mathrm{f}(1)\dfrac{x(x+1)}{2}$，然后求导，重新得出题目中的公式。

2 **唯一性。**较难 证明在 $[-1,1]$ 上满足 $|\mathrm{f}(x)| \le 1$、且在 $[-1,1]$ 内某处有 $|\mathrm{f}'(x)| = 4$ 的二次函数，只有 $\pm(2x^2-1)$。

3 **切比雪夫多项式。**利用 $\cos 3\theta = 4\cos^3\theta - 3\cos\theta$ 验证 $T_3(x) = 4x^3 - 3x$，并验证 $T_3'(1) = 9$。证明 $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$，并求 T_4。

4 **T_n 的斜率。**把 $T_n(\cos\theta) = \cos n\theta$ 对 θ 求导，证明 $T_n'(\cos\theta) = \dfrac{n\sin n\theta}{\sin\theta}$，再令 $\theta \to 0$，推导出 $T_n'(1) = n^2$。

延伸阅读

- **Lloyd N. Trefethen,** ***Approximation Theory and Approximation Practice*** （SIAM）。一本出色的现代著作，讲述切比雪夫多项式与插值，能力强的中学生也读得懂，全书穿插电脑实验。
- **Peter Borwein & Tamás Erdélyi,** ***Polynomials and Polynomial Inequalities*** （Springer）。想了解马可夫不等式及其相关结果的来龙去脉，可读此书。程度较深。
- **George M. Phillips,** ***Interpolation and Approximation by Polynomials*** （Springer）。一本大学本科教科书，涵盖拉格朗日插值、切比雪夫多项式和马可夫型不等式。

3.7 $\pi^2 \leq 10$

题目

STEP II 2014，第 2 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/14-s2-q2`

引子

每个中学生都知道 π 比 3 大一点。但你会怎样证明它的一个上界（upper bound）呢？阿基米德（Archimedes）用的是多边形。这道题的做法却完全出人意表：取一条关于函数的一般不等式，代入一条简单的抛物线（parabola），便直接读出 $\pi^2 \leq 10$。换另一个函数，更可得出著名的近似值 $\pi \leq \frac{22}{7}$。

不等式 $(*)$ 称为 *Wirtinger* 不等式（Wirtinger's inequality）。它说的是：一个两端都固定为零的函数，不可能只“大”而不“陡”。它的背后是数学物理中最核心的思想之一：弦的振动、量子粒子的能级，以及特征值（eigenvalue）理论。

动笔之前

想一想 一个在 $[0, \pi]$ 两端都为零的函数，如果在中间取值很大，为甚么就必须很陡？条件 $f(0) = f(\pi) = 0$ 为甚么重要？

要在长度为 π 的距离之内，由 0 升到某个高度 H 再降回 0，函数上升的平均速率至少要有 $H/(\pi/2)$。因此它的导数（derivative）不可能处处都小。若没有两端的条件，就没有这种限制：常数函数可以任意大，而导数为零。不等式 $(*)$ 把这个想法精确化，以平方的积分（integral）来量度“大小”和“陡峭程度”。

逐步解题

阶段 1　正弦波

想一想　$\int_0^{\pi}\sin^2 nx\,\mathrm{d}x$ 等于甚么？$\int_0^{\pi}\cos^2 nx\,\mathrm{d}x$ 呢？（利用 $\sin^2\theta=\frac{1}{2}(1-\cos 2\theta)$。）

两者都等于 $\frac{\pi}{2}$，因为 $\int_0^{\pi}\cos 2nx\,\mathrm{d}x=0$。取 $\mathrm{f}(x)=\sin nx$，则 $\mathrm{f}'(x)=n\cos nx$，所以

$$\int_0^{\pi}\mathrm{f}^2=\frac{\pi}{2}\le\frac{n^2\pi}{2}=\int_0^{\pi}(\mathrm{f}')^2,$$

因为 $n\ge 1$。　■留意 $n=1$ 时等号成立，所以这条不等式已经紧得不能再紧。

阶段 2　端点条件不可或缺

想一想　找一个简单的函数，满足 $\mathrm{f}(0)\neq 0$，而 $(*)$ 对它不成立。然后再找一个满足 $\mathrm{f}(0)=0$ 但 $\mathrm{f}(\pi)\neq 0$ 的。

常数函数 $\mathrm{f}(x)=1$ 有 $\mathrm{f}(0)=1\neq 0$，而且 $\int_0^{\pi}1\,\mathrm{d}x=\pi>0=\int_0^{\pi}0\,\mathrm{d}x$。

至于第二个，$\mathrm{f}(x)=x$ 有 $\mathrm{f}(0)=0$ 但 $\mathrm{f}(\pi)=\pi$，而且

$$\int_0^{\pi}x^2\,\mathrm{d}x=\frac{\pi^3}{3}\approx 10.3>\pi=\int_0^{\pi}1^2\,\mathrm{d}x.$$

所以 $(*)$ 在两种情况下都不成立。　■

阶段 3　抛物线

想一想　哪个二次函数在 $x=0$ 和 $x=\pi$ 处都等于零？（若不计常数倍，就只有一个。为甚么这个倍数无关紧要？）

这个二次函数必定是 $x(\pi-x)$ 的倍数，而把 f 乘以常数 k，$(*)$ 的两边都会乘以 k^2，所以我们可以取 $\mathrm{f}(x)=x(\pi-x)$，即 $a=-1$、$b=\pi$、$c=0$。于是 $\mathrm{f}'(x)=\pi-2x$，再用代换（substitution）$u=x-\frac{\pi}{2}$（这样两个被积函数都变成偶函数），

$$\int_0^{\pi}x^2(\pi-x)^2\,\mathrm{d}x=\int_{-\pi/2}^{\pi/2}\left(\frac{\pi^2}{4}-u^2\right)^2\mathrm{d}u=\frac{\pi^5}{30},$$

$$\int_0^{\pi}(\pi-2x)^2\,\mathrm{d}x=\int_{-\pi/2}^{\pi/2}4u^2\,\mathrm{d}u=\frac{\pi^3}{3}.$$

（第一个积分：$\int_{-h}^{h}(h^2-u^2)^2\,\mathrm{d}u=2(h^5-\frac{2}{3}h^5+\frac{1}{5}h^5)=\frac{16}{15}h^5$，其中 $h=\frac{\pi}{2}$。）所以 $(*)$ 给出 $\frac{\pi^5}{30}\le\frac{\pi^3}{3}$，即

$$\pi^2\le 10.\qquad\blacksquare$$

阶段 4 三角函数的选择

想一想 对于 $f(x) = p\sin\frac{1}{2}x + q\cos\frac{1}{2}x + r$，$f(0) = 0$ 和 $f(\pi) = 0$ 对 p、q、r 施加了甚么条件?

$f(0) = q + r = 0$ 及 $f(\pi) = p + r = 0$。所以 $p = q = -r$，经缩放后，我们取 $f(x) = \sin\frac{x}{2} + \cos\frac{x}{2} - 1$。

停一停，试一试 对这个函数计算 $\int_0^\pi f^2$ 和 $\int_0^\pi (f')^2$。由此得出关于 π 的甚么不等式?

记 $s = \sin\frac{x}{2}$ 及 $c = \cos\frac{x}{2}$。则 $(s+c)^2 = 1 + 2sc = 1 + \sin x$，所以

$$f^2 = (s+c)^2 - 2(s+c) + 1 = 2 + \sin x - 2\sin\tfrac{x}{2} - 2\cos\tfrac{x}{2}.$$

利用 $\int_0^\pi \sin x = 2$、$\int_0^\pi \sin\frac{x}{2} = 2$ 及 $\int_0^\pi \cos\frac{x}{2} = 2$，

$$\int_0^\pi f^2\,dx = 2\pi + 2 - 4 - 4 = 2\pi - 6.$$

另外 $f' = \frac{1}{2}(c - s)$，所以 $(f')^2 = \frac{1}{4}(1 - \sin x)$，而 $\int_0^\pi (f')^2\,dx = \frac{1}{4}(\pi - 2)$。于是 $(*)$ 给出

$$2\pi - 6 \le \frac{\pi - 2}{4} \quad\Longrightarrow\quad 8\pi - 24 \le \pi - 2 \quad\Longrightarrow\quad \pi \le \frac{22}{7}. \qquad \blacksquare$$

阶段 5 哪个较好?

由 $\pi \le \frac{22}{7}$ 可得 $\pi^2 \le \frac{484}{49} \approx 9.878$，比 $\pi^2 \le 10$ 更好。（真实值是 $\pi^2 \approx 9.870$。）留意这里证明的是 $\frac{22}{7}$ 为 π 的一个上界；而 Section 4.3 中一个截然不同的积分，则证明它并不等于 π。

回顾

- **一般的不等式，特定的函数。**选定一个测试函数（test function），就把一个关于所有函数的命题，变成一个关于某个数的命题。
- **缩放。**$(*)$ 的两边按同一个因子缩放，所以在每一族函数中，我们都可以把那个自由常数固定下来。
- **测试函数愈好，界愈好。**测试函数愈接近极端情况 $\sin x$，不等式就愈接近等式，估计也就愈准确（图 3.7.1）。

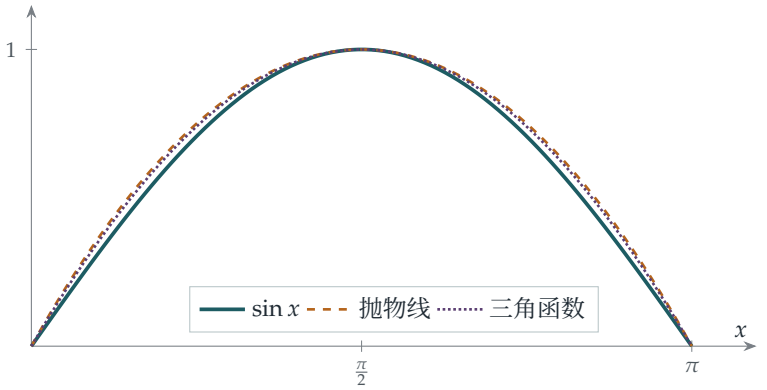


图 3.7.1　本节的三个函数，各自缩放至最大值为 1：极端情况 $\sin x$（实线），它令 $(*)$ 的等号成立；抛物线 $x(\pi - x)$（虚线），它给出 $\pi^2 \le 10$；以及 $\sin\frac{x}{2} + \cos\frac{x}{2} - 1$（点线），它给出 $\pi \le \frac{22}{7}$。两个测试函数都非常接近 $\sin x$，所以得出的界才这样好。

背后的数学

用傅里叶级数的证明

为甚么 $(*)$ 成立？到了大学你会学到，$[0, \pi]$ 上任何在两端都为零的合理函数，都可以写成傅里叶正弦级数（Fourier sine series）：

$$\mathrm{f}(x) = b_1 \sin x + b_2 \sin 2x + b_3 \sin 3x + \cdots .$$

由于当 $m \neq n$ 时 $\int_0^\pi \sin mx \sin nx \, \mathrm{d}x = 0$，平方后再积分，所有交叉项都会消失，得出

$$\int_0^\pi \mathrm{f}^2 = \frac{\pi}{2}\left(b_1^2 + b_2^2 + b_3^2 + \cdots\right), \qquad \int_0^\pi (\mathrm{f}')^2 = \frac{\pi}{2}\left(1^2 b_1^2 + 2^2 b_2^2 + 3^2 b_3^2 + \cdots\right).$$

右式的每一项都不小于左式的对应项，这就证明了 $(*)$，同时说明等号只在 f 是 $\sin x$ 的倍数时成立。题目的第 (i) 部分就是只有一项的情况。

振动的弦与量子粒子

比值 $\int (\mathrm{f}')^2 / \int \mathrm{f}^2$ 称为瑞利商（Rayleigh quotient），而 $(*)$ 说的是：对于固定在 0 和 π 的函数，它的最小可能值是 1。这个数有物理意义。对一条两端固定的弦来说，$\int (\mathrm{f}')^2$ 量度拉伸的能量，$\int \mathrm{f}^2$ 量度位移的大小，而这个比值的最小值决定了弦所能奏出的最低音。在量子力学（quantum mechanics）中，同一个最小值就是困在盒子里的粒子的最低能量：它不可能是零，因为两端固定的波必须弯曲。瑞利勋爵（Lord Rayleigh）正是用这一招

——猜一个测试函数，再计算比值——来估计振动系统的频率；由于极端情况给出最小值，任何猜测都会给出真正最低频率的一个上界，而好的猜测更会给出非常准确的上界。

变分法

找出令某个积分之比取得最小值的函数，是*变分法*（calculus of variations）的问题，当中的未知量是整个函数，而不是一个数。这里取得最小值的函数 $\sin x$ 满足微分方程（differential equation）$f'' + f = 0$，而这并非巧合：求积分的最小值，总会导出一条由取得最小值的函数所满足的微分方程。我们会在 Section 4.7 再谈这个想法。

思维技巧

测试函数

一条对某类函数中每一个都成立的不等式，可以套用在一个巧妙选取的函数上，从而得到关于某个特定数的信息。

个中艺术在于选择：接近极端情况的测试函数，会给出近乎最佳的结果。这个想法贯穿分析学、物理学和工程学，由估计特征值到有限元法（finite element method）皆然。

延伸练习

1 **四次函数。**在 $(*)$ 中取 $f(x) = x^2(\pi - x)^2$。证明它只能给出 $\pi^2 \le 12$。为甚么这个测试函数比抛物线差？（把它的形状与 $\sin x$ 比较。）

2 **两个正弦波。**对 $f(x) = \sin x + \sin 2x$ 直接验证 $(*)$，把两边都计算出来。

3 **另一个区间。**用代换证明：若 $g(0) = g(L) = 0$，则 $\displaystyle\int_0^L g^2 \le \frac{L^2}{\pi^2}\int_0^L (g')^2$。这对不同长度的弦说明了甚么？

4 **瑞利法。**较难取 $f(x) = x(\pi - x) + \lambda x^2(\pi - x)^2$，并选取 λ 使 π^2 的界尽可能好。（代数运算相当繁重，用电脑代数系统会有帮助。）

延伸阅读

- **T. W. Körner, *Fourier Analysis***（Cambridge University Press）。一本精彩而旁征博引的书，讲述傅里叶级数及其应用，作者是剑桥的分析学家。

- **Hardy, Littlewood & Pólya, *Inequalities***（Cambridge University Press）。第 7 章讨论 Wirtinger 不等式及相关的不等式。
- **Richard Feynman, *The Feynman Lectures on Physics***，第 II 卷第 19 章（“The principle of least action”，最小作用量原理）。费曼（Feynman）关于求积分最小值的一讲，通篇贯穿物理直觉。可于网上免费阅读。

3.8 由加法公式决定的函数

题目

STEP III 2019，第 2 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2019 年 STEP 试卷（zip 文件）内的 STEP III 试卷：

`www.ocr.org.uk/Images/696590-2019-past-papers.zip`

引子

你怎样知道 e^x 是甚么？在学校里，它是通过复利息引入的，或者说成是导数（derivative）等于自身的函数，又或者写成 $\lim(1+x/n)^n$。这道题给出另一个答案：e^{kx} 是唯一能把加法变成乘法的可微（differentiable）函数，即 $\mathrm{f}(x+y)=\mathrm{f}(x)\mathrm{f}(y)$。单凭这条规则，再加上一个数 $\mathrm{f}'(0)$，就足以把这个函数完全确定下来。

第二部分对一条较陌生的规则做同样的事：$\mathrm{g}(x+y)=\frac{\mathrm{g}(x)+\mathrm{g}(y)}{1+\mathrm{g}(x)\mathrm{g}(y)}$。如果你觉得这条公式似曾相识，那就对了：它正是狭义相对论（special relativity）中爱因斯坦（Einstein）的速度相加法则，只是把光速设为 1。这条规则迫使函数必然是 tanh，而这正好解释了为何任何物体都不能被加速至超越光速。

动笔之前

像 $\mathrm{f}(x+y)=\mathrm{f}(x)\mathrm{f}(y)$ 这样以函数为未知量的方程，称为*函数方程*（functional equation）。

? **想一想** 这条方程对所有 x 和 y 都成立。最值得首先尝试的是甚么？

特殊值。设 $y=0$（或 $x=y$，或 $y=-x$），往往不费甚么工夫就能得到大量资料。两个部分都是这样起步的。

逐步解题

阶段 1 在零处的值

代入 $x=y=0$：$\mathrm{f}(0)=\mathrm{f}(0)^2$，所以 $\mathrm{f}(0)=0$ 或 1。

> **想一想**　如果 $\mathrm{f}(0)=0$，会出甚么问题？

若 $\mathrm{f}(0)=0$，则代入 $y=0$ 可得对每个 x 都有 $\mathrm{f}(x)=\mathrm{f}(x)\mathrm{f}(0)=0$。于是 f 恒等于零，且 $\mathrm{f}'(0)=0$，与 $\mathrm{f}'(0)=k\neq 0$ 矛盾。所以 $\mathrm{f}(0)=1$。 ■

阶段 2　由零处的导数得出处处的导数

> **想一想**　在 $(*)$ 中，写成 $\mathrm{f}(x+h)=\mathrm{f}(x)\mathrm{f}(h)$。有甚么可以抽到极限之外？

$$\begin{aligned}\mathrm{f}'(x)&=\lim_{h\to 0}\frac{\mathrm{f}(x)\mathrm{f}(h)-\mathrm{f}(x)}{h}=\mathrm{f}(x)\lim_{h\to 0}\frac{\mathrm{f}(h)-1}{h}\\&=\mathrm{f}(x)\lim_{h\to 0}\frac{\mathrm{f}(0+h)-\mathrm{f}(0)}{h}=\mathrm{f}(x)\,\mathrm{f}'(0)=k\,\mathrm{f}(x).\end{aligned}$$ ■

函数方程让我们把 0 附近的性态搬到其他每一点去。

阶段 3　求解 $\mathrm{f}'=k\mathrm{f}$

考虑 $\mathrm{u}(x)=\mathrm{f}(x)e^{-kx}$。则 $\mathrm{u}'(x)=\big(\mathrm{f}'(x)-k\mathrm{f}(x)\big)e^{-kx}=0$，所以 u 是常数，而 $\mathrm{u}(0)=\mathrm{f}(0)=1$。因此

$$\mathrm{f}(x)=e^{kx}.$$ ■

（分离变量（separating the variables），$\int \frac{\mathrm{df}}{\mathrm{f}}=\int k\,\mathrm{d}x$，会得到相同的答案，但需要 $\mathrm{f}(x)\neq 0$；用 u 这个技巧就免却了这重顾虑。）

阶段 4　第二条规则

> **想一想**　在 g 的规则中代入 $y=0$。这告诉你 $\mathrm{g}(0)$ 是甚么？$|\mathrm{g}(x)|<1$ 在哪里派上用场？

取 $y=0$：$\mathrm{g}(x)\big(1+\mathrm{g}(x)\mathrm{g}(0)\big)=\mathrm{g}(x)+\mathrm{g}(0)$，化简后得 $\mathrm{g}(0)\big(\mathrm{g}(x)^2-1\big)=0$。由于 $|\mathrm{g}(x)|<1$，括号内的式子永不为零，所以 $\mathrm{g}(0)=0$。

> **停一停，试一试**　现在利用这条规则计算 $\mathrm{g}(x+h)-\mathrm{g}(x)$，化简，除以 h，再令 $h\to 0$。

$$\mathrm{g}(x+h)-\mathrm{g}(x)=\frac{\mathrm{g}(x)+\mathrm{g}(h)-\mathrm{g}(x)-\mathrm{g}(x)^2\mathrm{g}(h)}{1+\mathrm{g}(x)\mathrm{g}(h)}=\frac{\mathrm{g}(h)\big(1-\mathrm{g}(x)^2\big)}{1+\mathrm{g}(x)\mathrm{g}(h)}.$$

除以 h。当 $h\to 0$ 时，$\frac{\mathrm{g}(h)}{h}=\frac{\mathrm{g}(h)-\mathrm{g}(0)}{h}\to \mathrm{g}'(0)=k$，而 $\mathrm{g}(h)\to \mathrm{g}(0)=0$（可微函数必

定连续（continuous）），所以分母趋向 1。因此

$$\mathrm{g}'(x) = k\bigl(1 - \mathrm{g}(x)^2\bigr). \qquad \blacksquare$$

阶段 5 求解

> ? **想一想** 分离变量。当 $|g| < 1$ 时，哪个函数的导数是 $\frac{1}{1-g^2}$？

由于 $|\mathrm{g}| < 1$，我们可以相除：$\dfrac{\mathrm{g}'}{1-\mathrm{g}^2} = k$。函数 $\operatorname{artanh} g = \frac{1}{2}\ln\frac{1+g}{1-g}$ 的导数是 $\frac{1}{1-g^2}$，所以 $\frac{\mathrm{d}}{\mathrm{d}x}\operatorname{artanh}\mathrm{g}(x) = k$，而 $\operatorname{artanh}\mathrm{g}(x) = kx + C$。由 $\mathrm{g}(0) = 0$ 得 $C = 0$，所以

$$\mathrm{g}(x) = \tanh kx. \qquad \blacksquare$$

验算：加法公式 $\tanh(A+B) = \dfrac{\tanh A + \tanh B}{1 + \tanh A \tanh B}$ 正正就是题目给出的规则。

回顾

- **先试特殊值。**代入 $y = 0$ 便求得 $\mathrm{f}(0)$ 和 $\mathrm{g}(0)$。
- **由一点到处处。**加法定律（addition law）把 0 处的导数搬到每个 x，把函数方程化为微分方程（differential equation）。
- **唯一性（uniqueness）。**这个论证不单是验证 e^{kx} 和 $\tanh kx$ 符合要求，更证明了它们是仅有的可能。函数方程之所以能成为定义，原因正在于此。

背后的数学

相对论中的速度

在日常生活中，如果一列火车以速率 u 行驶，而你在车厢内以速率 v 向前走，你相对于地面的速率就是 $u + v$。爱因斯坦在 1905 年证明这只是一个近似。正确的法则（其中 c 为光速）是

$$u \oplus v = \frac{u + v}{1 + uv/c^2}.$$

若以 c 的分数来量度速率，这正是 g 的规则。我们的结果说明，如果把每个速度写成 $v = \tanh\phi$，那么速度的合成就是把各个 ϕ 相加。$\phi = \operatorname{artanh} v$ 这个量称为快度（rapidity）。火箭引擎每次推进都会增加固定份量的快度，但由于对每个 ϕ 都有 $\tanh\phi < 1$，速度只会愈来愈接近光速，却永远达不到（图 3.8.1）。把 $0.9c$ 与 $0.9c$ 合成，得到的不是 $1.8c$，而是 $\frac{1.8}{1.81}c \approx 0.9945c$。

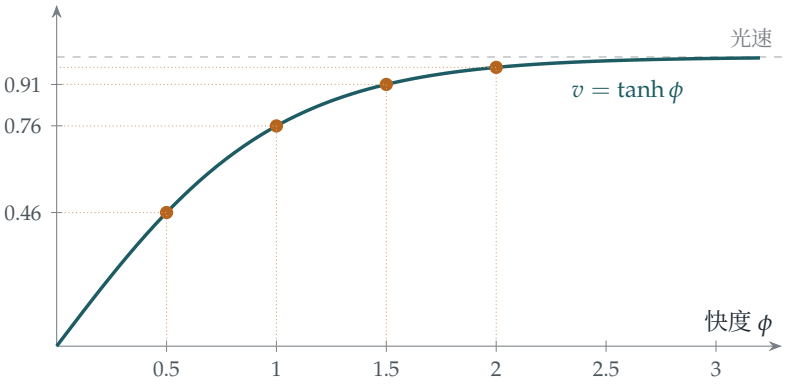


图 3.8.1　速度（以光速的分数表示）对快度的图像。快度每次增加相同的份量（引擎每次相同的推进所提供的正是这样），速度的增幅却愈来愈小：0.46、0.76、0.91、0.96、...，永远达不到 1。

柯西函数方程

最简单的函数方程是 $f(x+y) = f(x) + f(y)$，柯西（Cauchy）在 1821 年研究过它。如果 f 可微，用本题的方法可证明 $f(x) = kx$。事实上只要连续就够了。但若完全不作任何正则性（regularity）假设，就会有其他极其怪异的解，它们的图像稠密地散布在整个平面上。Georg Hamel 在 1905 年构造出这些解；这个构造需要用到*选择公理*（axiom of choice），而且不可能把其中任何一个明确地写出来。这就是题目坚持 f 可微的原因之一：这个假设并非例行公事。

同态

$f(x+y) = f(x)f(y)$ 的意思是，f 把实数的加法变成正实数的乘法，并且保持两者的结构。用大学一年级代数的语言来说，f 是由 $(\mathbb{R}, +)$ 到 $(\mathbb{R}_{>0}, \times)$ 的*群同态*（group homomorphism），而对数就是它的反函数。数学中许多最重要的函数，都是这样由它们所保持的结构来刻划的。在 Section 7.6 中，我们会看到 tanh 的几何意义：它量度的是“双曲角”（hyperbolic angle），而它的加法定律来自双曲旋转的复合。

思维技巧

用性质来定义函数

不用公式，而是用函数所遵守的规则来描述它，并证明只有一个函数遵守这条规则。

数学家往往就是这样看待 exp、sin、cos 和许多其他函数的：不把它们看成公式，而是看成函数方程或微分方程的唯一解。这解释了它们*为何*重要，而不只是怎样计算它们。

延伸练习

1 **正切的规则。**设 g 可微，$\mathrm{g}'(0) = k \neq 0$，且只要等式两边都有定义，就有 $\mathrm{g}(x+y) = \dfrac{\mathrm{g}(x)+\mathrm{g}(y)}{1-\mathrm{g}(x)\mathrm{g}(y)}$。证明 $\mathrm{g}'(x) = k\big(1+\mathrm{g}(x)^2\big)$，并由此推出在 $x=0$ 附近 $\mathrm{g}(x) = \tan kx$。

2 **柯西方程。**若 f 可微，且对所有 x、y 都有 $\mathrm{f}(x+y) = \mathrm{f}(x)+\mathrm{f}(y)$，证明 $\mathrm{f}(x) = kx$。

3 **不用导数。**较难 只假设 $\mathrm{f}(x+y) = \mathrm{f}(x)\mathrm{f}(y)$ 及 $\mathrm{f}(1) = a > 0$。证明对正整数 n 有 $\mathrm{f}(n) = a^n$，然后证明 $\mathrm{f}(1/n) = a^{1/n}$，再证明对每个正有理数 q 都有 $\mathrm{f}(q) = a^q$。（为甚么要再进一步就需要连续性?）

4 **火箭科学。**一枚火箭每次点火，都能相对于它当时所处的参考系把速率增加 $0.5c$。由静止开始，经过 2、3 和 4 次点火后，它（相对于起点）的速率是多少？请使用快度。

延伸阅读

- **Christopher Small, *Functional Equations and How to Solve Them***（Springer）。一本浅白易读的函数方程入门书，以喜欢解题的读者为对象。
- **Edwin Taylor & John Wheeler, *Spacetime Physics***（W. H. Freeman）。狭义相对论的经典入门书，全书都使用快度。
- **Tom Körner, *Calculus for the Ambitious***（Cambridge University Press）。由指数函数和三角函数的性质出发来定义它们，与本节的精神完全一致。

3.9 创造一个新函数

题目

STEP II 2021，第 4 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2021 年 STEP 试卷（zip 文件）内的 STEP II 试卷：

`www.ocr.org.uk/Images/696588-2021-past-papers.zip`

引子

你能解 $e^{-x} = 5x$ 吗？计算机上没有任何一个函数帮得上忙。移项整理也无济于事：x 同时困在指数函数的里面和外面。你可以用试误法（trial and improvement）求出数值答案，但并没有甚么“公式”。

真的没有吗？这道题踏出了大胆的一步：如果某一族方程无法用已知的函数解出来，那就*发明一个新函数*来解它们。数学家当年正是这样得到 $\sqrt{x}$（x^2 的反函数）、$\ln x$（e^x 的反函数）和 $\arcsin x$ 的。这里的新函数是 xe^x 的反函数（inverse function），现称为*兰伯特 W 函数*（Lambert W function）。有了它，从 $x^x = 10$ 到一条物理定律，范围广得出人意料的各种方程都可以精确地解出来。

动笔之前

想一想 “解”一条方程，究竟是甚么意思？$x = \sqrt{2}$ 是 $x^2 = 2$ 的解，还是只不过是那个解的名字？

两者皆是。$\sqrt{2}$ 是“平方等于 2 的那个正数”的名字；我们接受它作为答案，是因为函数 $\sqrt{\ }$ 已为人熟知、容易计算，而且性质清楚。一个答案算不算“公式”，取决于我们约定了为哪些函数命名。这道题为这份名单再添一个。

逐步解题

阶段 1　曲线 $y = xe^x$

$\frac{dy}{dx} = e^x + xe^x = (1+x)e^x$，它只在 $x = -1$ 处为零，此时 $y = -e^{-1}$。导数（derivative）在 $x < -1$ 时为负，在 $x > -1$ 时为正，所以这是一个极小值（minimum）。

> **想一想**　当 $x \to +\infty$ 和 $x \to -\infty$ 时，xe^x 会怎样？曲线在哪里与坐标轴相交？

当 $x \to \infty$ 时，$xe^x \to \infty$。当 $x \to -\infty$ 时，$xe^x \to 0$（从下方趋近），因为指数衰减胜过线性增长。曲线通过原点。所以它从负 x 轴的稍下方进来，下降至极小点 $(-1, -\frac{1}{e})$，然后上升穿过 $(0,0)$，再急速向上（图 3.9.1）。

阶段 2　反函数

要有反函数，f 必须是一对一（one-to-one）的，所以必须把它限制在一个严格递增（strictly increasing）或严格递减的区域上。包含右边部分的最大区间是 $[-1,\infty)$，所以 $a = -1$。在 $[-1,\infty)$ 上，f 由 $-\frac{1}{e}$ 递增至 ∞，所以反函数 g 定义在 $[-\frac{1}{e},\infty)$ 上，由 $g(-\frac{1}{e}) = -1$ 开始递增，经过 $g(0) = 0$，并且增长缓慢（当 x 很大时，比 $\ln x$ 还要慢）。它的图像是 $y = xe^x$ 右边部分关于直线 $y = x$ 的反射，在 $(-\frac{1}{e}, -1)$ 处有一条垂直切线。

g 的定义性质是

$$g(z)\, e^{g(z)} = z, \qquad g(z) \geq -1.$$

所以 (iii) 中每一部分的策略都一样：把方程整理成 (某个式子) $e^{(\text{同一个式子})} = (\text{一个数})$ 的形式。

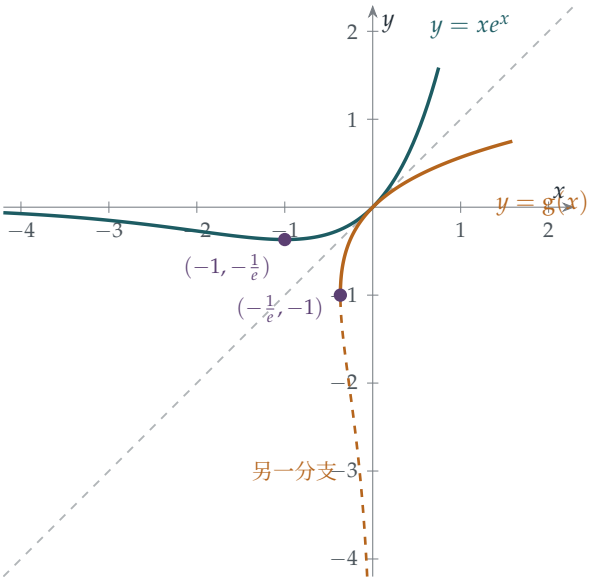


图 3.9.1 曲线 $y = xe^x$（蓝绿色）在 $(-1, -\frac{1}{e})$ 处有极小值。限制在 $x \geq -1$ 上，再关于 $y = x$ 反射，便得到反函数 g（琥珀色实线），即兰伯特 W 函数。琥珀色虚线是 $y = xe^x$ 左边部分的反射：这是第二个分支，第 (iii) 部分的第二个根要用到它。

阶段 3 四条方程

(a) $e^{-x} = 5x$。乘以 e^x 再除以 5：$xe^x = \frac{1}{5}$。所以 $x = g(\frac{1}{5}) \approx 0.169$。由于 xe^x 在 $x < 0$ 时为负，任何根都是正数，而 xe^x 在 $x > 0$ 时递增，所以这是唯一的根。

> **想一想** 在 (b) 和 (c) 中，未知数以 $x \ln x$ 的形式出现。用甚么代换（substitution）可以把 $x \ln x$ 变成 ue^u 的形式？

设 $x = e^u$，则 $\ln x = u$，而 $x \ln x = ue^u$。

(b) $2x \ln x + 1 = 0$ 变成 $ue^u = -\frac{1}{2}$。但对所有实数 u 都有 $ue^u \geq -\frac{1}{e} \approx -0.368$，而 $-\frac{1}{2} < -\frac{1}{e}$。所以**没有实根**。

(c) $3x \ln x + 1 = 0$ 变成 $ue^u = -\frac{1}{3}$，而 $-\frac{1}{e} < -\frac{1}{3} < 0$。高度为 $-\frac{1}{3}$ 的水平线与 $y = ue^u$ 相交**两次**，在极小点 $u = -1$ 的两侧各一次。右边的根是 $u = g(-\frac{1}{3})$，由此得

$$x = e^{g(-1/3)} \approx 0.538.$$

另一个根满足 $u < -1 < g(-\frac{1}{3})$，而 $x = e^u$ 随 u 递增，所以我们找到的根是两者之中**较大**的一个。（另一个约为 0.220。）

> **停一停，试一试** 现在试做 (d)，$x = 3\ln x$。（提示：两边取指数，设法凑成 (某个式子) $e^{\text{某个式子}}$ 的形式，其中“某个式子”是 x 的倍数。）

(d) $x = 3\ln x$ 即 $e^{x/3} = x$，所以 $xe^{-x/3} = 1$，再乘以 $-\frac{1}{3}$，

$$\left(-\frac{x}{3}\right) e^{-x/3} = -\frac{1}{3}.$$

所以 $-\frac{x}{3}$ 是在 (c) 中找到的 $ue^u = -\frac{1}{3}$ 的两个解之一。取由 g 给出的那一个，$x = -3\,\mathrm{g}\left(-\frac{1}{3}\right) \approx 1.857$。另一个根来自另一个解 $u < \mathrm{g}(-\frac{1}{3})$，它给出较大的 $x = -3u$。所以这一次我们找到的根是两者之中**较小**的一个。（另一个约为 4.536。） ■

阶段 4 $x^x = 10$

取对数，$x\ln x = \ln 10$。设 $x = e^u$，这就是 $ue^u = \ln 10$，所以 $u = \mathrm{g}(\ln 10)$，而

$$x = e^{\mathrm{g}(\ln 10)} \approx 2.506.$$

由于 $\mathrm{g}(z)e^{\mathrm{g}(z)} = z$，我们也可以把它写成 $x = \dfrac{\ln 10}{\mathrm{g}(\ln 10)}$。 ■

回顾

- **为缺少的函数命名。**当一条方程无法用已知函数解出时，把一个新函数定义为反函数，是正当的做法，而且往往富有启发性。
- **整理成标准形式。**(iii) 和 (iv) 的每一部分最终都归结为 $ue^u =$ 常数，用的是 $x = e^u$ 或 $u = -\frac{x}{3}$ 之类的代换。
- **用图像数根的个数。**$ue^u = c$ 有零个、一个还是两个解，只取决于 c 相对于极小值 $-\frac{1}{e}$ 的位置。

背后的数学

历史悠久、名字却很新的函数

兰伯特（Johann Heinrich Lambert）在 1758 年研究过这类方程，欧拉（Euler）在 1783 年作了进一步的研究。但这个函数一直没有标准名称，直到 1990 年代，电脑代数系统（computer algebra system）Maple 的开发者需要为它起一个名字。1996 年，Robert Corless、Gaston Gonnet、David Hare、David Jeffrey 和 Donald Knuth 发表了论文“On the Lambert W function”，汇集了它的性质和应用。这篇论文已被引用数千次，而 W

现已内置于大多数数学软件之中。题目中的函数 g 是*主分支*（principal branch）W_0；图 3.9.1 中的虚线，即给出 (c) 和 (d) 中第二个根的那条曲线，是另一个实分支 W_{-1}。

W 出现在哪里

- **维恩位移定律（Wien's displacement law）。**热的物体辐射最强的波长与其温度成反比。当中的常数来自解方程 $5(1-e^{-x})=x$，其正数解为 $x=5+W\left(-5e^{-5}\right)\approx 4.965$。这就是为甚么约 5800 K 的太阳看起来是黄白色，而温度较低的红矮星看起来是红色。
- **幂塔（power tower）。**无穷幂塔 $x^{x^{x^{\cdots}}}$ 在 $e^{-e}\le x\le e^{1/e}$ 时收敛（converge），其值为 $\dfrac{W(-\ln x)}{-\ln x}$。
- **人口与传染病模型。**在最简单的模型中，一场传染病的最终规模，以及某些增长模型中达到指定人口所需的时间，都由 W 给出。

初等函数

W 真的是新函数吗？还是可以用 exp、ln、根式和三角函数写出来？它确实是新的：一条可追溯至 1830 年代刘维尔（Joseph Liouville）的定理表明，W 不是*初等函数*（elementary function）。同一套理论还表明，一些看似平平无奇的积分，例如 $\int e^{-x^2}\,dx$，无法用初等函数表示，这就是为甚么正态分布（normal distribution）需要自己专用的特殊函数。

思维技巧

以反函数作为解

方程 $f(x)=c$ 的解是 $x=f^{-1}(c)$，前提是把 f 限制成一对一的函数。这个限制决定了你得到的是哪一个解。

许多看似不同的方程，其实是同一条方程的不同伪装；找出能揭示共同形式的代换，就是关键的一步。

延伸练习

1 **二、四，还有一个。**方程 $x^2 = 2^x$ 有显而易见的解 $x = 2$ 和 $x = 4$（比较 Section 3.5）。证明它还有一个负数解，而这个解是

$$x = -\frac{2}{\ln 2}\,\mathrm{g}\left(\frac{\ln 2}{2}\right) \approx -0.767.$$

2 **一个导数。**对 $\mathrm{g}(x)e^{\mathrm{g}(x)} = x$ 作隐函数微分（implicit differentiation），证明当 $x \neq 0$ 时，

$$\mathrm{g}'(x) = \frac{\mathrm{g}(x)}{x\big(1 + \mathrm{g}(x)\big)}.$$

在 $x = -\frac{1}{e}$ 处会怎样？

3 **一个积分。**利用代换 $x = we^w$ 证明 $\displaystyle\int_0^e \mathrm{g}(x)\,\mathrm{d}x = e - 1$。

4 **幂塔。**假设幂塔 $T = \sqrt{2}^{\sqrt{2}^{\sqrt{2}^{\cdots}}}$ 收敛，证明 $T = \sqrt{2}^{T}$，并由此推出 $T = 2$ 或 $T = 4$。究竟是哪一个？（计算数列 $t_1 = \sqrt{2}$、$t_{n+1} = \sqrt{2}^{t_n}$ 的头几项，并想想 Section 2.2 中的蛛网图（cobweb diagram）。）

延伸阅读

- **Robert Corless et al., “On the Lambert W function”**，*Advances in Computational Mathematics* 5 (1996)。为这个函数命名的论文。读完本节之后，论文的头几节应可读懂。
- **Brian Hayes, “Why W?”**，*American Scientist* 93 (2005)。一篇生动的文章，讲述这个函数的历史和用途。可在网上免费阅读。
- **Brian Conrad, “Impossibility theorems for elementary integration”**。网上可找到的讲义，解释刘维尔的理论。程度较深。

我们学到了甚么

本章以一个 $1 = 2$ 的“证明”（proof）和三个问题开始：怎样知道一个论证是正确的？怎样证明不等式（inequality）？函数（function）究竟是甚么？以下是本章九道题目教晓我们的东西。

1. 怎样知道一个论证是正确的？

Section 3.1 给了我们一份检查清单。每走一步，都要问三件事：*计算对吗？这一步在逻辑上成立吗？有甚么假设是没有明说的？*那一节出现的错误都很典型。比较两个驻点（stationary point）上的函数值，并不能说明哪一个是极大值。在无限的定义域上，驻点的函数值不能作为函数的界。微积分基本定理（fundamental theorem of calculus）要求被积函数连续（continuous）。此外，一个论证也可能只是碰巧得出“看似正确”的答案。

同样的警觉贯穿本章其余部分。Section 3.7 用反例（counterexample）说明，去掉一个假设（$f(0) = f(\pi) = 0$），不等式便不再成立。在 Section 3.8 中，函数可微（differentiable）这个假设必不可少：没有它，柯西（Cauchy）函数方程（functional equation）便有一些谁也写不出来的怪异解。而在 Section 3.6 和 Section 3.3 中，等号成立的条件告诉我们某个界是否已是最佳，这正是严谨的数学家总会问的那类问题。

2. 怎样证明不等式？

我们从几个不同方向接触了各个经典不等式，也看到它们之间的关系。

- **平方非负。**两个数的算术－几何平均不等式（AM–GM）其实就是 $(a-b)^2 \geq 0$。柯西的前推－后推归纳法（forward–backward induction）由此建立起一般情形（Section 3.2）。
- **凸性（convexity）。**凹（concave）曲线位于它的弦之上。单凭这一点便得出琴生不等式（Jensen's inequality）；配合对数，一行便得出 AM–GM；配合 $\sin x$，则得出一个关于三角形的精确不等式（Section 3.3）。
- **面积。**曲线之下的面积加上曲线旁边的面积，足以盖过一个长方形。这给出杨氏不等式（Young's inequality），再由此推出赫尔德（Hölder）不等式和柯西－施瓦兹（Cauchy–Schwarz）不等式（Section 3.4）。
- **图像的形状。**把 a^b 与 b^a 的比较改写成 $\frac{\ln x}{x}$ 在两点上的比较，便解决了 π^e 与 e^π 孰大孰小的问题（Section 3.5）。

- **插值（interpolation）与三角不等式（triangle inequality）。**把导数用函数值表示，然后逐项估计其界，便得出二次多项式的马尔可夫不等式（Markov's inequality）（Section 3.6）。
- **试验函数（test function）。**把一个一般的不等式应用于巧妙选取的函数，便得出 $\pi^2 \leq 10$ 和 $\pi \leq \frac{22}{7}$（Section 3.7）。

在这些方法之中作出选择，是一种要靠练习才能养成的技巧。粗略的指引是：见到积与和，可考虑 AM–GM；见到函数值的平均，可考虑琴生不等式；见到一个函数和它的反函数（inverse function），可考虑杨氏不等式；要比较两个形式相同的式子，则可尝试找出它们背后的那一个函数。

3. 函数究竟是甚么？

最后两节从两方面回答了这个问题。在 Section 3.8 中，e^{kx} 和 $\tanh kx$ 根本不是由公式给出的；它们是满足某些加法定律（addition law）的仅有可微函数。这是比公式更深一层的理解：它解释了为何这些函数不断出现，既见于复利息，也见于爱因斯坦（Einstein）的速度相加公式。在 Section 3.9 中，我们更进一步，创造了一个函数，即 xe^x 的反函数，因为手上已有的函数不足以解一族很自然的方程。你计算机上的每一个函数，当初都是这样诞生的新事物。

4. 严谨论证与不等式的工具箱

方法	用于何处
检查每一步和每一个假设	似是而非的错误论证（3.1）
局部性质与整体性质	驻点与无穷远处的极限（3.1）
前推－后推归纳法	AM–GM（3.2）
补上额外的项	AM–GM（3.2、3.3）
琴生不等式	三角形；AM–GM（3.3）
比较面积	杨氏不等式（3.4）
分离变量，归结为单一函数	π^e 与 e^π 的比较（3.5）
三角不等式、等号成立的情况	马尔可夫不等式（3.6）
试验函数	$\pi^2 \leq 10$（3.7）
函数方程	e^x 和 $\tanh x$（3.8）
反函数与代换	W 函数（3.9）

下一步

- 在 Section 3.7 中略窥一二的变分法（calculus of variations），是 Section 4.7 的主题。
- tanh 及其加法定律背后的几何，见于 Section 7.6。

- 第 4 章会以本章这种审慎的态度看待积分（integration）本身，由积分究竟是甚么谈起。

如果你想在入读大学前再进一步

- **J. Michael Steele, *The Cauchy–Schwarz Master Class***：不等式的最佳入门书，全书由题目构成。
- **Daniel Velleman, *How to Prove It***：有系统地教你写出正确证明的指南。
- **Michael Spivak, *Calculus***：由根基建立起来的严谨微积分；不少剑桥学生在入学前的那个暑假读这本书。
- **E. A. Maxwell, *Fallacies in Mathematics***：用来练习找出哪里出了错。

冯纽曼（Von Neumann）曾打趣说，在数学里你不是理解事物，只是习惯了它们。这话不无道理：本章那些精确的定义和严谨的论证，起初或会令人觉得陌生。但正是习惯了它们，一个懂得计算的学生才会变成一个能够确信无误的数学家。$1 = 2$ 的证明再也骗不到你；数学学位的目标，就是让每一个更隐蔽的错误都同样一眼可见。

“延伸练习”答案

以下只是简短的答案和提示，并非完整解答：这里只列出结果和关键步骤，方便你核对自己的做法。请先认真做过每道练习，然后才看答案。

3.1 找出错处

1. 当 x 由 -1 变到 1 时会经过 0，而 $t = 1/x$ 在该处没有定义：这个代换（substitution）在区间上并不连续（continuous）。正确的结果是 $\int_{-1}^{1} \frac{\mathrm{d}x}{(1+x^2)^2} = \frac{1}{2} + \frac{\pi}{4}$（用 $x = \tan\theta$），而 $\int_{-1}^{1} \frac{-t^2}{(1+t^2)^2}\,\mathrm{d}t = \frac{1}{2} - \frac{\pi}{4}$。

2. (a) 正确：两边的对数（logarithm）都是 $\ln a \ln b$。(b) 错误：当 $\theta = 0$ 时，$\cos 0 = 1 \neq \sin 1$。（事实上两者永不相等，因为 $|\sin\theta \pm \cos\theta| \leq \sqrt{2} < \frac{\pi}{2}$。）(c) 错误：非常数的多项式（polynomial）是无界的，而一个常数不可能同时与 1 和 -1 相差不超过 10^{-6}。(d) 正确：由算术—几何平均不等式（AM–GM），$x^4 + x^{-4} \geq 2$，等号在 $x = 1$ 时成立。

3. 由 $n = 1$ 推到 $n = 2$ 的一步并不成立：两匹马当中的“前 1 匹”与“后 1 匹”并不重叠。

4. $\int_{\varepsilon}^{1} x^{-1/2}\,\mathrm{d}x = 2 - 2\sqrt{\varepsilon} \to 2$，所以由对称性，该积分（作为极限）存在，其值为 4。

3.2 倒过来的 AM–GM

1. 设 $a_1, \ldots, a_{n-1}$ 的平均值为 A，几何平均（geometric mean）为 G，再加入 A 作为第 n 个数。平均值仍然是 A，所以 $(G^{n-1}A)^{1/n} \leq A$，由此得 $G \leq A$。

2. 这 $n+1$ 个数的平均值为 $1 + \frac{1}{n+1}$，几何平均为 $(1 + \frac{1}{n})^{n/(n+1)}$，而且它们并非全部相等，所以不等式是严格的；把两边同时取 $n+1$ 次方即可。

3. 各个面的面积为 ab、bc、ca，而 $S = 2(ab + bc + ca)$。由 AM–GM 得 $\frac{S}{6} \geq (abc)^{2/3} = V^{2/3}$，等号在 $ab = bc = ca$ 时成立，即正方体。

4. 对 $1/a_i$ 应用 AM–GM：$\frac{1}{n}\sum \frac{1}{a_i} \geq \frac{1}{G}$。再取倒数。

3.3 凸性的威力

1. 若 $\mathrm{f}'' < 0$，斜率递减，所以曲线位于每条弦（chord）的上方。取 $t = \frac{1}{3}$：$\mathrm{f}\big(\frac{u+v+w}{3}\big) \geq \frac{1}{3}\mathrm{f}(u) + \frac{2}{3}\mathrm{f}\big(\frac{v+w}{2}\big) \geq \frac{1}{3}\big(\mathrm{f}(u) + \mathrm{f}(v) + \mathrm{f}(w)\big)$。

2. $(\ln\sin x)'' = -\operatorname{cosec}^2 x < 0$，所以 $\ln\sin$ 在 $(0,\pi)$ 上是凹（concave）函数。由延森不等式（Jensen's inequality）得 $\frac{1}{3}\ln(\sin A \sin B \sin C) \leq \ln\sin\frac{\pi}{3}$，所以该乘积最大为 $\big(\frac{\sqrt{3}}{2}\big)^3$。

3. 在 $(0, \frac{\pi}{2})$ 上 $(\tan x)'' = 2\sec^2 x \tan x > 0$；由延森不等式得 $\tan A + \tan B + \tan C \geq 3\tan\frac{\pi}{3} = 3\sqrt{3}$。

4. $(x^2)'' = 2 > 0$；由延森不等式即得所求的不等式，它说明 $\frac{1}{n}\sum a_i^2 - \bar{a}^2$（即方差（variance））是非负的。

3.4 以面积证明不等式

1. 取 $\mathrm{f} = \sin$，定义在 $[0, \frac{\pi}{2}]$ 上，等号在 $b = \mathrm{f}(a)$ 时成立；取 $a = \arcsin t$、$b = t$：$t\arcsin t = (1 - \cos(\arcsin t)) + \int_0^t \arcsin y \,\mathrm{d}y$。把所得结果的右边求导，得 $\arcsin t + \frac{t}{\sqrt{1-t^2}} - \frac{t}{\sqrt{1-t^2}} = \arcsin t$。

2. 在杨氏不等式（Young's inequality）中取 $p = q = 2$，得 $ab \leq \frac{1}{2}(a^2 + b^2)$。若 $\sum a_i^2 = \sum b_i^2 = 1$，逐项相加得 $\sum a_i b_i \leq 1$。一般情况下，把这个结果应用于 $a_i / \sqrt{\sum a_j^2}$ 和 $b_i / \sqrt{\sum b_j^2}$。

3. $\int_0^a (e^x - 1)\,\mathrm{d}x = e^a - 1 - a$，而 $\int_0^b \ln(1+y)\,\mathrm{d}y = (1+b)\ln(1+b) - b$。等号在 $b = e^a - 1$ 时成立。

4. $t = 1$：1.571 对 1；$t = 2$：0.524 对 0.268；$t = 10$：0.1002 对 0.0501。当 t 很大时，$t - \sqrt{t^2-1} = \frac{1}{t+\sqrt{t^2-1}} \approx \frac{1}{2t}$。这个界来自 $a = 0$ 的情况，此时 $(*)$ 只是说 $0 \leq \int_0^b \arcsin y \,\mathrm{d}y$；当 b 很小时，这块面积约为 $\frac{b^2}{2}$，是 $b\arcsin b \approx b^2$ 的一半，把它舍去便损失了因子 2。

3.5 π^e 还是 e^π？

1. $n^m = m^n$ 即 $h(n) = h(m)$。h 在 $(1,e)$ 上递增，其后递减，所以两数之中有一个小于 e：它必定是 2（因为 $h(1) = 0$ 这个值不会再次取到），于是另一个是 4。唯一的一对是 $\{2,4\}$。

2. $e < 3 < \pi$，而 h 在该处递减，所以 $h(3) > h(\pi)$，即 $3^\pi > \pi^3$。$\frac{9}{4}$ 和 $\sqrt{5}$ 都小于 e，而 $\frac{9}{4} > \sqrt{5}$，所以 $h(\frac{9}{4}) > h(\sqrt{5})$，由此得第二个结果。对于 $1 < x < 2$：由于 $3 < x+2 < 4$，有 $h(x) < h(2) = h(4) < h(x+2)$，所以 $x^{x+2} < (x+2)^x$。最后，

$9^{\sqrt{2}} = 3^{2\sqrt{2}}$，而 $\sqrt{2}^{9} = 16\sqrt{2} = (2\sqrt{2})^3$。因为 $e^2 < 8$，所以 $e < 2\sqrt{2} < 3$，而 h 在该处递减，故 $h(2\sqrt{2}) > h(3)$，由此得 $(2\sqrt{2})^3 > 3^{2\sqrt{2}}$：$\sqrt{2}^{9}$ 较大。

3. $e^{\pi/e-1} > \frac{\pi}{e}$，所以 $e^{\pi/e} > \pi$；再把两边取 e 次方。

3.6 斜率的界

1. 两边都是二次多项式，并且在 $x = -1, 0, 1$ 处取值相同，所以两者相等（拉格朗日插值（Lagrange interpolation））。求导得 $\mathrm{f}(-1)(x - \frac{1}{2}) - 2\mathrm{f}(0)x + \mathrm{f}(1)(x + \frac{1}{2})$。

2. 在 $[-1, 1]$ 上 $|\mathrm{f}'(x)| \le |x - \frac{1}{2}| + 2|x| + |x + \frac{1}{2}| \le 4$，等号只在 $x = \pm 1$ 时成立，而且必须 $\mathrm{f}(1)$、$\mathrm{f}(-1)$、$-\mathrm{f}(0)$ 全部等于 1 或全部等于 -1。二次多项式由这三个值唯一确定，于是得 $\pm(2x^2 - 1)$。

3. 由 $T_3(\cos\theta) = \cos 3\theta$ 得 $4x^3 - 3x$，而 $T_3'(1) = 12 - 3 = 9$。由 $\cos(n+1)\theta + \cos(n-1)\theta = 2\cos\theta\cos n\theta$，得 $T_{n+1} = 2xT_n - T_{n-1}$，以及 $T_4 = 8x^4 - 8x^2 + 1$。

4. $-\sin\theta\, T_n'(\cos\theta) = -n\sin n\theta$。当 $\theta \to 0$ 时，$\frac{n\sin n\theta}{\sin\theta} \to n^2$。

3.7 $\pi^2 \le 10$

1. $\int_0^\pi \mathrm{f}^2 = \frac{\pi^9}{630}$，而 $\int_0^\pi (\mathrm{f}')^2 = \frac{2\pi^7}{105}$，所以由 $(*)$ 得 $\pi^2 \le 12$。与 $\sin x$ 相比，这个四次多项式在两端附近过于平坦，在中间又过于尖耸，所以它的斜率与大小之比离最优值较远。

2. $\int_0^\pi (\sin x + \sin 2x)^2 = \frac{\pi}{2} + \frac{\pi}{2} = \pi$（交叉项的积分为 0），而 $\int_0^\pi (\cos x + 2\cos 2x)^2 = \frac{\pi}{2} + 2\pi = \frac{5\pi}{2}$。

3. 设 $\mathrm{g}(x) = \mathrm{f}(\pi x/L)$。左边按 $\frac{L}{\pi}$ 的比例缩放，而 $\int(\mathrm{g}')^2$ 则按 $\frac{\pi}{L}$ 的比例缩放。对弦线而言，最低频率与 $\frac{1}{L}$ 成正比：长度减半，音高便加倍。

4. 界 $\pi^2 \le \int(\mathrm{f}')^2 / \int \mathrm{f}^2$（采用适当的单位）在 $\lambda \approx 1.133/\pi^2$ 时取得最小值，由此得 $\pi^2 \le 9.86975$，而真确值为 9.86960。

3.8 由加法公式决定的函数

1. 代入 $x = y = 0$ 得 $\mathrm{g}(0) = 0$。对 y 求导并取 $y = 0$，得 $\mathrm{g}'(x) = k(1 + \mathrm{g}(x)^2)$，所以 $\frac{\mathrm{d}}{\mathrm{d}x}\arctan\mathrm{g}(x) = k$，于是 $\mathrm{g}(x) = \tan kx$。

2. 对 y 求导得 $\mathrm{f}'(x+y) = \mathrm{f}'(y)$，所以 f' 是常数 k；又 $\mathrm{f}(0) = 0$（代入 $x = y = 0$）。

3. $\mathrm{f}(x) = \mathrm{f}(x/2)^2 \ge 0$，而且 f 永不为 0（否则它处处为 0）。用归纳法（induction）得 $\mathrm{f}(n) = a^n$；由 $\mathrm{f}(1/n)^n = \mathrm{f}(1) = a$ 及 $\mathrm{f}(1/n) > 0$ 得 $a^{1/n}$；于是 $\mathrm{f}(m/n) = \mathrm{f}(1/n)^m$。若没有连续性，函数在无理数处的值便不受限制：存在一些利用“哈默尔基（Hamel

basis）”构造出来的奇怪解，它们并不是 a^x。

4. 每次点火都使快度（rapidity）增加 $\phi = \operatorname{artanh}\frac{1}{2}$，所以点火 k 次之后，速率为 $\tanh k\phi = \frac{3^k-1}{3^k+1}$：$0.8c$、$\frac{13}{14}c \approx 0.929c$ 和 $\frac{40}{41}c \approx 0.976c$。

3.9 创造一个新函数

1. 当 $x = -1$ 时，$x^2 - 2^x = \frac{1}{2} > 0$；当 $x = 0$ 时其值为 -1：所以有一个负根。设 $x = -2u/\ln 2$，则 $x^2 = 2^x$（其中 $x < 0$，取 $x = -2^{x/2}$）化为 $ue^u = \frac{\ln 2}{2}$，所以 $u = \mathrm{g}\left(\frac{\ln 2}{2}\right)$，而 $x \approx -0.7667$。

2. $\mathrm{g}'e^{\mathrm{g}}(1+\mathrm{g}) = 1$，而 $e^{\mathrm{g}} = x/\mathrm{g}$。当 $x = -\frac{1}{e}$ 时，$\mathrm{g} = -1$，斜率为无限大（图像在该处是垂直的）。

3. 设 $x = we^w$，则 $\mathrm{d}x = (1+w)e^w\,\mathrm{d}w$，而 w 由 0 变到 1：$\int_0^1 w(1+w)e^w\,\mathrm{d}w = \left[(w^2-w+1)e^w\right]_0^1 = e-1$。

4. $T = \sqrt{2}^T$ 即 $\frac{\ln T}{T} = \frac{\ln 2}{2}$，所以 $T = 2$ 或 4。各项 $1.414, 1.633, 1.761, \ldots$ 递增，并且一直小于 2（若 $t_n < 2$，则 $t_{n+1} < \sqrt{2}^2 = 2$），所以 $T = 2$；不动点（fixed point）4 是排斥（repelling）的。

章

4

严谨的微积分

积分是分析、几何与数论交汇之处，数学中一些最优美的公式亦在此诞生。

本章内容

4.1	甚么是积分?	STEP III 1989
4.2	由一个积分建立三角学	STEP III 2017
4.3	22/7 大于 π	STEP I 1997
4.4	沃利斯、Beta 与二分之一的阶乘	STEP II 1995
4.5	大二学生之梦	STEP III 2009
4.6	没有公式的积分	STEP III 2016
4.7	变分法	STEP III 2014
4.8	切球形面包	STEP I 2001
4.9	Holditch 定理	STEP III 2011
A	练习答案	第 305 页

不只是原函数

那些没有导数的函数，实在是可悲的祸害，我怀着惊恐与厌恶避之则吉。

— 埃尔米特（Charles Hermite），1893 年的一封信

问一位中学生积分（integration）是甚么，你多半会听到“微分的逆运算”。这是个了不起的事实，称为*微积分基本定理*（fundamental theorem of calculus），它令你在中学遇到的大部分积分都变成例行公事。但它是一条关于积分的定理，而不是积分的定义，而且它把积分之所以有趣的大部分原因都遮盖了。

积分比微分（differentiation）早出现约二千年。阿基米德（Archimedes）把图形切成薄片再求和，求出了抛物线弓形的面积和球的表面积，那时还远远未有人想到导数（derivative）。而且积分能到达微分无法到达的地方。每个合理的函数都可以按机械式的法则求导，但大多数函数却*无法*用公式积分出来，这就引出一个自然的问题：甚么时候可以？不可以的时候又该怎么办？积分可以定义新的函数，证明数与数之间的不等式，计算无法用其他方式描述的面积和体积，还能给出无穷和与无穷乘积的精确值。本章谈的就是这一切。

积分究竟是甚么

本章从头说起，由定义开始（Section 4.1）。积分是许多幼长矩形之和的极限（limit）；单凭这个定义、不用原函数（antiderivative）去计算 $\int_0^a x^5\,\mathrm{d}x$，已是相当费工夫的事，而这会令你懂得珍惜基本定理。有了精确的定义，随之而来的是一个精确的问题：哪些函数是可积（integrable）的？这个问题的答案，以及它所引发的种种难题，在二十世纪初催生了勒贝格（Lebesgue）的积分理论，那正是现代概率论的基础。

用来证明的积分

本章一些最出人意表的结果，是用积分去证明一些看来与面积毫无关系的事实。某个正值函数的积分恰好等于 $\frac{22}{7}-\pi$，于是一行就证明了 $\pi<\frac{22}{7}$（Section 4.3）。$\sin t$ 的幂引出沃利斯（Wallis）关于 π 的无穷乘积（infinite product），以及 $\left(\frac{1}{2}\right)!=\frac{\sqrt{\pi}}{2}$ 这个值（Section 4.4）。x^x 的积分化成了和式 $\sum(-1)^{n+1}n^{-n}$（Section 4.5）。而仅仅一个积分 $\int_0^x \frac{\mathrm{d}u}{1+u^2}$，就足以从零开始建立整套三角学（Section 4.2）。

你在大学会遇到甚么

大学一年级的分析（Analysis）课程会从积分的定义出发，严格证明微积分基本定理。其后的课程向多个方向发展。测度论（measure theory）和勒贝格积分（Lebesgue integral）令积分与极限配合得顺畅无碍，并且是概率论的根基。向量微积分（vector calculus）在曲线、曲面和立体上积分，而格林（Green）定理和斯托克斯（Stokes）定理则把它们联系起来（Section 4.9 可让你浅尝一下）。复分析（complex analysis）取道复平面来计算实积分。变分法（calculus of variations）问的是哪个函数令某个积分尽可能小（Section 4.7），并由此通向物理学中的最小作用量原理（principle of least action）。而微分代数（differential algebra）则解释了哪些积分根本有公式可言（Section 4.6）。

贯穿本章的问题

三大问题

1. **积分是甚么？哪些函数有积分？**以和的极限作出的定义，与原函数有何关连？
2. **积分能证明甚么？**一块面积怎能告诉我们有关 π、阶乘和无穷和的事？
3. **没有公式的时候，我们可以做甚么？**我们怎样知道某个积分没有公式？即使如此，我们还能算出甚么？

本章的思想

§	课题	思想
4.1	积分的定义	下和与上和（lower and upper sums）；裂项相消（telescoping）
4.2	由 $\int \frac{\mathrm{d}u}{1+u^2}$ 建立三角学	保持积分不变的代换（substitution）
4.3	$\frac{22}{7} > \pi$	把不等式写成某个正值量的积分
4.4	沃利斯与 $\left(\frac{1}{2}\right)!$	递降公式（reduction formulae）、不变量（invariant）、夹逼（squeeze）
4.5	$\int_0^1 x^x \,\mathrm{d}x$	级数展开与逐项积分（term-by-term integration）
4.6	没有公式的积分	藉多项式代数证明不可能
4.7	变分法	在积分之内配方（completing the square）
4.8	阿基米德的帽盒定理（hat-box theorem）	化简被积函数（integrand），几何意义随之显现
4.9	霍尔迪奇（Holditch）定理	由边界积分求面积；消元（elimination）

如何阅读本章

有好几节涉及繁重但例行的计算：多项式除法、分部积分法（integration by parts）、展开乘积。请先拿起纸笔亲自做一遍，然后才阅读正文。接着找出令计算得以成功的那个思想：一个不变量、一种对称、一次相消。计算是考官要检查的东西；思想才是你会记住并再次运用的东西。

4.1 甚么是积分?

题目

STEP III 1989，第 10 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/89-s3-q10`

引子

你做积分已经很多年了。你知道 $\int_0^a x^5\,\mathrm{d}x = \frac{1}{6}a^6$，因为“积分是微分的逆运算”。但那是一条定理，不是定义。积分究竟是甚么?

老实的答案是:积分是面积,定义为许多细长矩形面积之和的极限(limit)。这道题要你单凭这个定义计算一个积分:不用原函数(antiderivative),不用微积分基本定理(fundamental theorem of calculus)，只靠把矩形加起来再取极限。这比你想像的要难，而这正好说明，令这件事变得轻而易举的微积分基本定理，是何等了不起的成果。

动笔之前

想一想 第 (iii) 部分的两个和，都是阔度为 $\frac{a}{n}$ 的矩形的面积之和。对于递增函数（increasing function）f，哪一个和低估了曲线下的面积，哪一个高估了?

第一个和取每一条长条左端的高度，即 $x = \frac{ra}{n}$，其中 $r = 0, \ldots, n-1$。对递增函数来说，那是长条上的最低点，所以这些矩形全在曲线之下：这是下和（lower sum）。第二个和取右端，即 $r = 1, \ldots, n$，所得的矩形会突出于曲线之上：这是上和（upper sum）（图 4.1.1）。定义说明：当长条愈来愈窄时，若两个和趋向同一个极限，积分便存在。

对于 $\mathrm{f}(x) = x^5$，两个和都涉及 $\sum r^5$。第 (i) 和第 (ii) 部分分别给出这个和的上界和下界。

逐步解题

阶段 1　可以裂项相消的和

想一想　设 $P(r)=r(r+1)(r+2)(r+3)(r+4)(r+5)$，即六个连续整数之积。$P(r)-P(r-1)$ 是甚么？

$P(r)$ 与 $P(r-1)=(r-1)r(r+1)(r+2)(r+3)(r+4)$ 有五个共同因子 $r(r+1)(r+2)(r+3)(r+4)$，所以

$$\begin{aligned}P(r)-P(r-1)&=r(r+1)(r+2)(r+3)(r+4)[(r+5)-(r-1)]\\&=6\,r(r+1)(r+2)(r+3)(r+4).\end{aligned}$$

由 $r=1$ 加到 n，右方各项裂项相消（telescoping），只余下 $P(n)-P(0)=P(n)$。所以

$$\sum_{r=1}^{n} r(r+1)(r+2)(r+3)(r+4)=\tfrac{1}{6}P(n)=\tfrac{1}{6}n(n+1)(n+2)(n+3)(n+4)(n+5).$$

■

由于对所有 $r\geq 1$ 都有 $r^5<r(r+1)(r+2)(r+3)(r+4)$，不等式随即成立。 ■

阶段 2　下界

停一停，试一试　把第 1 阶段的做法改用于五个向下数的连续整数之积：$r(r-1)(r-2)(r-3)(r-4)$。求它由 $r=0$ 到 $n-1$ 的和，并与 r^5 比较。

设 $Q(r)=(r-5)(r-4)(r-3)(r-2)(r-1)r$。则 $Q(r+1)-Q(r)=(r-4)(r-3)(r-2)(r-1)r[(r+1)-(r-5)]=6\,(r-4)(r-3)(r-2)(r-1)r$。由 $r=0$ 加到 $n-1$，裂项相消后得 $Q(n)-Q(0)=Q(n)$，所以

$$\sum_{r=0}^{n-1}(r-4)(r-3)(r-2)(r-1)r=\tfrac{1}{6}(n-5)(n-4)(n-3)(n-2)(n-1)n.$$

现在逐项比较。当 $r=0$ 时，两边都是 0。当 $1\leq r\leq 4$ 时，乘积含有一个零因子，而 $r^5>0$。当 $r\geq 5$ 时，每个因子都是正数，且不大于 r，其中有些更严格小于它，所以乘积小于 r^5。由于 $n>1$，和式包含 $r=1$ 一项，该处不等式严格成立，所以

$$\sum_{r=0}^{n-1} r^5>\tfrac{1}{6}(n-5)(n-4)(n-3)(n-2)(n-1)n.$$ ■

阶段 3　积分

想一想 把 $\mathrm{f}(x)=x^5$ 的下和与上和用 $\sum r^5$ 表示。分母中出现 n 的几次方?

取 $\mathrm{f}(x)=x^5$,

$$L_n=\sum_{r=0}^{n-1}\frac{a}{n}\left(\frac{ra}{n}\right)^5=\frac{a^6}{n^6}\sum_{r=0}^{n-1}r^5,\qquad U_n=\frac{a^6}{n^6}\sum_{r=1}^{n}r^5.$$

由第 (i) 和第 (ii) 部分,并因为 $L_n<U_n$,

$$\frac{a^6}{6}\cdot\frac{(n-5)(n-4)\cdots n}{n^6}\;<\;L_n\;<\;U_n\;<\;\frac{a^6}{6}\cdot\frac{n(n+1)\cdots(n+5)}{n^6}.$$

两个分数各自是六个因子之积,这些因子形如 $\frac{n-3}{n}=1-\frac{3}{n}$ 或 $\frac{n+3}{n}=1+\frac{3}{n}$,每一个都趋向 1。所以外侧两个界都趋向 $\frac{a^6}{6}$,根据夹逼定理(squeeze),L_n 和 U_n 亦然。两个极限都存在而且相等,所以按定义

$$\int_0^a x^5\,\mathrm{d}x=\tfrac{1}{6}a^6.\qquad\blacksquare$$

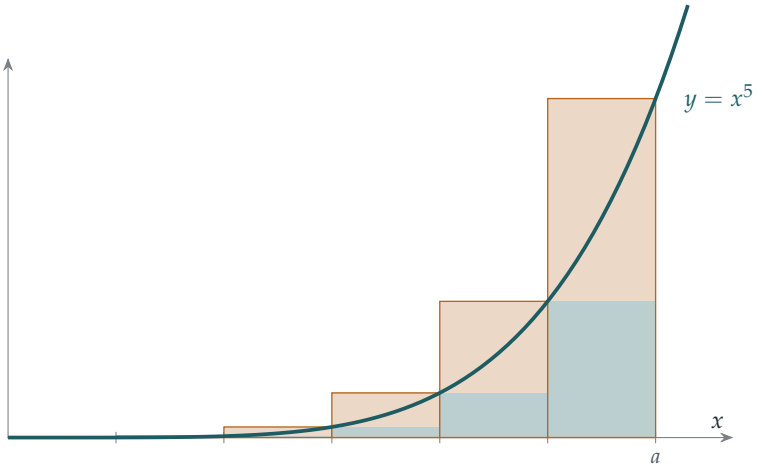


图 4.1.1 x^5 在 $[0,a]$ 上的下和与上和,分成 $n=6$ 条长条。下和是蓝绿色的面积(曲线下的矩形);上和则再加上琥珀色的部分。两者之差正好是最后、最高的那个矩形的面积 $\frac{a}{n}\mathrm{f}(a)$,当 n 增大时它趋向 0。

回顾

- **又是裂项相消。**连续整数之积的和一下子收拢起来,与第 2 章的和与积如出一辙。
- **先定界,再夹逼。**我们从没有求出 $\sum r^5$ 的准确值(它是 $\frac{1}{12}n^2(n+1)^2(2n^2+2n-1)$);有两边的界已经足够。
- **定义担起全部工作。**原函数由始至终没有出现过。一切都来自积分作为和的极限这个定义。

背后的数学

黎曼的定义

把细长条加起来求面积的想法，可追溯至阿基米德（Archimedes）：他在公元前 250 年左右，正是用这个论证的一个版本求出抛物线下的面积。牛顿（Newton）和莱布尼兹（Leibniz）把积分变成微分的逆运算，其后一个半世纪，大多数人都是这样理解积分的。1854 年，黎曼（Bernhard Riemann）提出了大学课程至今仍用作起点的定义：把 $[a,b]$ 分成许多小段，构作（阔度）$\times$（f 在该小段内某处的高度）之和；若小段不断缩小时，所有这样的和都趋向同一个极限，便说 f 可积（integrable）。本题的版本采用等阔长条和左右端点，是一个适用于递增函数的特殊情况；达布（Jean-Gaston Darboux）的“上和与下和”则把一般情况处理得干净利落。

并非每个函数都有积分

有了精确的定义，便可以问哪些函数可积。闭区间上的每个连续函数都可积，每个递增函数也可积。但试考虑狄利克雷（Dirichlet）函数：若 x 是有理数，$\mathrm{f}(x)=1$；若是无理数，则为 0。每一条长条无论多窄，都同时包含有理数和无理数，所以每个上和都是 1，每个下和都是 0。两者永不一致，因此 $\int_0^1 \mathrm{f}$ 在黎曼的意义下并不存在。1902 年，勒贝格（Henri Lebesgue）发明了一种更强大的积分，能够处理这个函数（它的勒贝格积分（Lebesgue integral）是 0），而且在取极限时表现好得多。勒贝格积分是现代概率论的基础，你会在大学二年级或三年级遇上它。

离散微积分

第 1 阶段的裂项相消，其实是乔装了的微积分。“上升阶乘”（rising factorial）$r(r+1)\cdots(r+k-1)$ 在求和之下的表现，与 x^k 在积分之下的表现完全一样：求和令因子的数目增加一个，并除以新的因子数目，正如 $\int x^k\,\mathrm{d}x=\frac{x^{k+1}}{k+1}$。和与积分、差分与导数之间的这种对应，称为*有限差分演算*（calculus of finite differences），是计算和式的有力工具。

思维技巧

从定义出发

题目要求你“用定义”证明某事时，就只用定义：不用任何依赖于待证结果的定理，也不用熟悉的捷径。

这样做出奇地难，却让你确切看到那些捷径的价值。在大学，每一科的头几个星期都是这样进行的：先有定义，然后才有定理，让你不必再从定义出发。

延伸练习

1 **把极限化为积分。**试做 STEP III 1999，第 3 题（`step.maths.org/questions/99-s3-q3`）。这道题把和的极限化为积分：在每个和式中，找出曲线下的矩形。

2 **阿基米德的抛物线。**利用 $\sum_{r=1}^{n} r^2 = \frac{1}{6}n(n+1)(2n+1)$ 和第 (iii) 部分的定义，证明 $\int_0^a x^2\,\mathrm{d}x = \frac{1}{3}a^3$。

3 **上升阶乘。**证明对每个正整数 k，都有 $\displaystyle\sum_{r=1}^{n} r(r+1)\cdots(r+k-1) = \frac{n(n+1)\cdots(n+k)}{k+1}$。

4 **准确的和。**利用恒等式 $r(r+1)(r+2)(r+3)(r+4) = r^5 + 10r^4 + 35r^3 + 50r^2 + 24r$，以及 $\sum r^k$ 在 $k \le 4$ 时的已知公式，求出 $\sum_{r=1}^{n} r^5$ 的准确值。

延伸阅读

- **David Bressoud, *A Radical Approach to Lebesgue's Theory of Integration***（Cambridge University Press）。由黎曼积分走到勒贝格积分的故事，以逼出每一步发展的问题为线索娓娓道来。
- **Michael Spivak, *Calculus***（Cambridge University Press）。第 13 章以上和与下和定义积分，与本节完全一样。
- **Graham, Knuth & Patashnik, *Concrete Mathematics***（Addison-Wesley）。第 2 章建立有限差分演算，并讨论上升阶乘和下降阶乘。

4.2 由一个积分建立三角学

题目

STEP III 2017，第 6 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/17-s3-q6`

引子

π 究竟是甚么？在中学，它是圆周与直径之比，而 sin 和 cos 则用三角形来定义。这对几何来说已经足够，却留下了一些棘手的问题。沿着曲线的“长度”到底是甚么？为甚么 $\sin x$ 的麦克劳林级数（Maclaurin series）偏偏是那个样子？我们又怎样知道加法公式（addition formulas）对所有角都成立，而不只限于画得出来的角？

这道题展示了一条截然不同的入门途径。忘掉三角形和圆吧。只从一个积分 $\mathrm{T}(x) = \int_0^x \frac{\mathrm{d}u}{1+u^2}$ 出发，它所需要的不过是 Section 4.1 中积分的定义。把 π 定义为 $2\mathrm{T}_\infty$。然后只用代换法（substitution），证明 $\mathrm{T}(\sqrt{3}) = \frac{\pi}{3}$ 和 $\mathrm{T}(\sqrt{2}-1) = \frac{\pi}{8}$，也就是你所熟悉的 $\arctan\sqrt{3} = 60^\circ$ 和 $\tan 22\frac{1}{2}^\circ = \sqrt{2}-1$。三角学就这样从无到有地重建起来。

动笔之前

你心里其实知道 $\mathrm{T}(x) = \arctan x$ 和 $\mathrm{T}_\infty = \frac{\pi}{2}$。题目不准你使用这些事实，而这正是用意所在：一切都必须由积分推导出来。不过，这点心照不宣的知识是很有用的向导。

? **想一想** 第 (i) 和第 (ii) 部分分别对应哪些三角恒等式（trigonometric identities）？（想想 $\arctan x + \arctan\frac{1}{x}$，以及 $\tan(A+B)$ 的公式。）

第 (i) 部分就是 $\arctan x + \arctan\frac{1}{x} = \frac{\pi}{2}$。第 (ii) 部分则是乔装了的加法公式 $\tan(A+B) = \frac{\tan A+\tan B}{1-\tan A\tan B}$：设 $u = \tan A$ 和 $a = \tan B$，则 $v = \tan(A+B)$。这道题不靠任何几何，单凭积分就证明了这些恒等式。

逐步解题

阶段 1　反射 $u \mapsto 1/u$

在 $\mathrm{T}(x)$ 中代入 $u = 1/w$。于是 $\mathrm{d}u = -\mathrm{d}w/w^2$，而 $\dfrac{1}{1+u^2} = \dfrac{w^2}{w^2+1}$，所以

$$\mathrm{T}(x) = \int_{\infty}^{1/x} \frac{w^2}{w^2+1} \cdot \left(-\frac{1}{w^2}\right) \mathrm{d}w = \int_{1/x}^{\infty} \frac{\mathrm{d}w}{1+w^2} = \mathrm{T}_{\infty} - \mathrm{T}(x^{-1}). \qquad \blacksquare$$

换句话说，$\frac{1}{1+u^2}$ 之下由 0 到 x 的面积，等于由 $\frac{1}{x}$ 到 ∞ 的面积（图 4.2.1）。

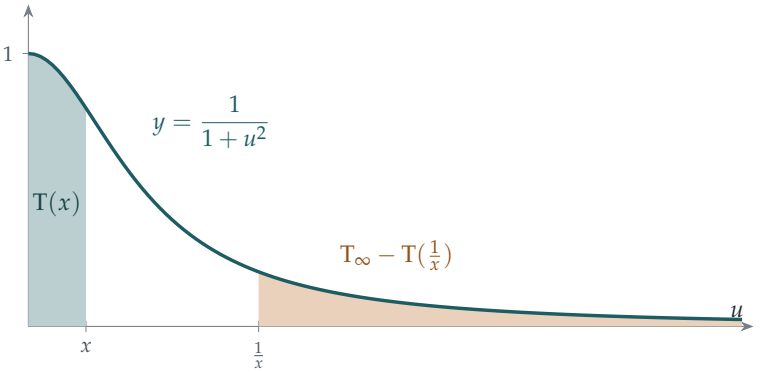


图 4.2.1　代换 $u = 1/w$ 显示 $\frac{1}{1+u^2}$ 之下的两块阴影面积相等：由 0 到 x 的面积（蓝绿色），以及由 $\frac{1}{x}$ 起无限伸延的面积（琥珀色）。图中 $x = \frac{1}{2}$。

阶段 2　加法代换

想一想　用商法则（quotient rule）对 $v = \dfrac{u+a}{1-au}$ 求导。然后把 $1+v^2$ 化成单一分式，再作比较。

$$\frac{\mathrm{d}v}{\mathrm{d}u} = \frac{(1-au)+a(u+a)}{(1-au)^2} = \frac{1+a^2}{(1-au)^2},$$
$$1+v^2 = \frac{(1-au)^2+(u+a)^2}{(1-au)^2} = \frac{(1+a^2)(1+u^2)}{(1-au)^2}.$$

两式相除，得 $\dfrac{1+v^2}{1+u^2} = \dfrac{1+a^2}{(1-au)^2} = \dfrac{\mathrm{d}v}{\mathrm{d}u}$。 $\blacksquare$

想一想　这表示 $\dfrac{\mathrm{d}u}{1+u^2} = \dfrac{\mathrm{d}v}{1+v^2}$。所以代换 $u \mapsto v$ 完全没有改变被积函数（integrand）。那么它对 $\mathrm{T}(x)$ 的积分上下限有甚么影响？

当 u 由 0 变到 x 时（其中 $x<\frac{1}{a}$，故 $1-au>0$ 始终成立，而 v 连续且递增），v 由 a 变到 $\frac{x+a}{1-ax}$。所以

$$\mathrm{T}(x)=\int_0^x \frac{\mathrm{d}u}{1+u^2}=\int_a^{(x+a)/(1-ax)} \frac{\mathrm{d}v}{1+v^2}=\mathrm{T}\left(\frac{x+a}{1-ax}\right)-\mathrm{T}(a). \qquad \blacksquare$$

阶段 3 再多两个恒等式

结合第 (i) 部分，并使用两次：

$$\mathrm{T}(x^{-1})=\mathrm{T}_\infty-\mathrm{T}(x)=\mathrm{T}_\infty-\mathrm{T}\left(\frac{x+a}{1-ax}\right)+\mathrm{T}(a)=2\mathrm{T}_\infty-\mathrm{T}\left(\frac{x+a}{1-ax}\right)-\mathrm{T}(a^{-1}). \qquad \blacksquare$$

现在更换变量名称：设 $y=1/x$ 和 $b=1/a$。条件 $x<1/a$ 变成 $1/y<b$，即 $y>1/b$，而

$$\frac{x+a}{1-ax}=\frac{1/y+1/b}{1-1/(by)}=\frac{b+y}{by-1}.$$

所以 $\mathrm{T}(y)=2\mathrm{T}_\infty-\mathrm{T}\left(\frac{y+b}{by-1}\right)-\mathrm{T}(b)$。 $\blacksquare$

阶段 4 特殊值

> **停一停，试一试** 在最后一个恒等式中选取 y 和 b，使自变量 $\frac{y+b}{by-1}$ 刚好等于 y 本身。你会得到甚么？

取 $y=b=\sqrt{3}$：这时 $y>\frac{1}{b}$，而 $\frac{\sqrt{3}+\sqrt{3}}{3-1}=\sqrt{3}$。恒等式变成 $\mathrm{T}(\sqrt{3})=2\mathrm{T}_\infty-2\mathrm{T}(\sqrt{3})$，所以

$$\mathrm{T}(\sqrt{3})=\tfrac{2}{3}\mathrm{T}_\infty. \qquad \blacksquare$$

> **想一想** 至于 $\sqrt{2}-1$，试在加法恒等式中取 $x=a=\sqrt{2}-1$。$\frac{2x}{1-x^2}$ 是多少？$\mathrm{T}(1)$ 又是多少？

取 $x=a=\sqrt{2}-1$（由于 $x^2=3-2\sqrt{2}<1$，故 $x<\frac{1}{a}$），加法恒等式给出 $2\mathrm{T}(x)=\mathrm{T}\left(\frac{2x}{1-x^2}\right)$。而 $1-x^2=2\sqrt{2}-2=2(\sqrt{2}-1)=2x$，所以 $\frac{2x}{1-x^2}=1$，于是 $2\mathrm{T}(\sqrt{2}-1)=\mathrm{T}(1)$。最后，在第 (i) 部分取 $x=1$，得 $\mathrm{T}(1)=\mathrm{T}_\infty-\mathrm{T}(1)$，所以 $\mathrm{T}(1)=\frac{1}{2}\mathrm{T}_\infty$。因此

$$\mathrm{T}(\sqrt{2}-1)=\tfrac{1}{4}\mathrm{T}_\infty. \qquad \blacksquare$$

回顾

- **保持被积函数不变的代换。**两个关键步骤所用的代换，都令 $\frac{\mathrm{d}u}{1+u^2}$ 保持不变，只有积分上下限移动了。积分的这类“对称性”（symmetry）威力强大。

- **不动点（fixed point）。**$\mathrm{T}(\sqrt{3})$ 的值，来自选取适当的数，使某个表达式等于它本身，于是恒等式变成了方程。
- **定义可以来自分析。**我们从没有用过三角形，却重新得到了 $\frac{\pi}{3}$ 和 $\frac{\pi}{8}$（其中 π 定义为 $2\mathrm{T}_\infty$）。

背后的数学

建构初等函数

在严格的分析课程中，函数 exp、sin、cos 和 ln 不能随便假设存在；它们必须被*建构*出来，其性质也必须加以证明。标准的做法有三种。你可以用幂级数（power series）来定义它们（$\sin x = x - \frac{x^3}{3!} + \cdots$）。你可以把它们定义为微分方程（differential equation）的解，就如练习中那道 1986 年样本试卷的题目（$\mathrm{f}' = \mathrm{f}$、$\mathrm{f}(0) = 1$ 定义了 exp）。你也可以像这里一样，用积分来定义反函数（inverse function）：$\ln x = \int_1^x \frac{\mathrm{d}u}{u}$ 和 $\arctan x = \int_0^x \frac{\mathrm{d}u}{1+u^2}$，然后定义 $\pi = 2\int_0^\infty \frac{\mathrm{d}u}{1+u^2}$。每种做法各有所长，而证明它们全都给出相同的函数，是一道令人满足的练习。试与 Section 3.8 的函数方程（functional equation）做法比较。

圆以外：椭圆积分

积分的做法可以推广，而三角形的做法却做不到。1718 年，Giulio Fagnano 研究*双纽线*（lemniscate）这条 8 字形曲线的弧长（arc length），由此引出 $\int_0^x \frac{\mathrm{d}u}{\sqrt{1-u^4}}$，并找到一个能把弧长加倍的代换，正如我们的代换把 T 加倍一样。欧拉（Euler）在 1751 年读到 Fagnano 的著作，找出了完整的加法公式；这最终经由阿贝尔（Abel）和雅可比（Jacobi），发展成*椭圆函数*（elliptic functions）的理论：椭圆函数就是这类积分的反函数，它们之于双纽线，正如 sin 和 cos 之于圆。我们在高斯（Gauss）关于算术几何平均（AGM）的工作中（Section 2.6）遇过同一个积分。椭圆函数如今是数论的核心；Section 1.9 的椭圆曲线（elliptic curves）就是它们在几何上的投影。

计算 π

加法恒等式也是计算 π 的那些最著名公式的基础。举例说，反复运用它可以证明 $\mathrm{T}(1) = 4\mathrm{T}(\frac{1}{5}) - \mathrm{T}(\frac{1}{239})$，这就是 1706 年的梅钦公式（Machin's formula）；配合级数 $\mathrm{T}(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots$，它算出了 π 的 100 位数字，并在其后两个半世纪一直是首选的方法。这是 Section 5.1 的主题。

思维技巧

积分的对称性

寻找令被积函数保持不变的代换。它们会化为该积分所定义的函数的恒等式。

这里 $u \mapsto 1/u$ 和 $u \mapsto \frac{u+a}{1-au}$ 都保持 $\frac{\mathrm{d}u}{1+u^2}$ 不变，并分别给出反射公式和加法公式。同一个想法用于其他积分，正是特殊函数（special functions）理论的基础。

延伸练习

1 **指数函数与余弦。**试做 STEP II Specimen paper，第 6 题（`step.maths.org/questions/spec-s2-q6`）。这道题只凭指数函数和余弦各自的微分方程，推导出它们的加法公式。（提示：对于余弦，找出两个理应为零的组合，并证明它们的平方和是常数。）

2 **反正弦。**试做 STEP III 2007，第 7 题（`step.maths.org/questions/07-s3-q7`）。这道题用处理本节积分的同一手法来处理反正弦（inverse sine），并以一个代换把两者联系起来。

3 **一条早期的梅钦型公式。**利用加法恒等式证明 $\mathrm{T}(\frac{1}{2}) + \mathrm{T}(\frac{1}{3}) = \mathrm{T}(1) = \frac{1}{2}\mathrm{T}_\infty$。

4 T_∞ **是有限的。**证明对所有 $x \geq 1$ 都有 $\mathrm{T}(x) \leq 1 + \int_1^x \frac{\mathrm{d}u}{u^2} < 2$，解释为甚么当 $0 < x < 1$ 时同样有 $\mathrm{T}(x) < 2$，并由单调收敛定理（monotone convergence theorem）（Section 2.2）推断 T_∞ 存在。

延伸阅读

- **Tom Körner,** ***Calculus for the Ambitious***（Cambridge University Press）。严格地建构初等函数，并讨论各种不同的做法。
- **Walter Rudin,** ***Principles of Mathematical Analysis***（McGraw-Hill）。即著名的“baby Rudin”；第 8 章的引言只用寥寥数页，便优雅地由幂级数建构出 exp、sin 和 cos。程度较深。
- **Victor Prasolov & Yuri Solovyev,** ***Elliptic Functions and Elliptic Integrals***（American Mathematical Society）。想知道双纽线通往何处，可读此书。程度较深。

4.3 22/7 大于 π

题目

STEP I 1997，第 6 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/97-s1-q6`

引子

一代又一代的学生都听过 π “大约是 $\frac{22}{7}$” 这个说法。这是一个非常好的近似值（approximation）：$\frac{22}{7} = 3.142857\ldots$，而 $\pi = 3.141592\ldots$。但 $\frac{22}{7}$ 究竟是偏大还是偏小？你又能否证明这一点，而不是单靠查阅 π 的各位数字？那些数字毕竟是要别人先计算出来的。

这道题给出了整个数学之中最优雅的证明之一。只需一个正函数的积分（integral），其值因而必定为正，而它竟然恰好等于 $\frac{22}{7} - \pi$。被积函数（integrand）为正，积分便为正，所以 $\frac{22}{7} > \pi$。整个证明只需寥寥数行，而且精巧得被人当作谜题印在数学 T 恤上。

动笔之前

想一想 为甚么 $\int_0^1 \frac{\text{多项式}}{1+x^2}\,\mathrm{d}x$ 会算出 π 来？这种形式的积分之中，哪一个简单的积分本身就给出 π？

$\int_0^1 \frac{\mathrm{d}x}{1+x^2} = \arctan 1 = \frac{\pi}{4}$。把一个多项式（polynomial）除以 $1+x^2$，会得到一个多项式商式（quotient）（其积分是一个有理数），再加上一个常数余式（remainder）除以 $1+x^2$（其积分是 π 的有理数倍）。题目的第一部分正是要你做这个除法。

逐步解题

阶段 1 多项式除法

先展开：$(1-x)^4 = 1 - 4x + 6x^2 - 4x^3 + x^4$，所以

$$x^4(1-x)^4 = x^8 - 4x^7 + 6x^6 - 4x^5 + x^4.$$

想一想 把它除以 x^2+1，可以用长除法（long division），也可以用比较系数（comparing coefficients）的方法。余式是甚么?

在 $(a_6x^6 + \cdots + a_0)(x^2+1) + b$ 中由最高次项开始比较系数：$a_6 = 1$；$a_5 = -4$；$a_4 + a_6 = 6$，所以 $a_4 = 5$；$a_3 + a_5 = -4$，所以 $a_3 = 0$；$a_2 + a_4 = 1$，所以 $a_2 = -4$；$a_1 + a_3 = 0$，所以 $a_1 = 0$；$a_0 + a_2 = 0$，所以 $a_0 = 4$；最后 $a_1 = 0$（x 的系数）及 $a_0 + b = 0$，所以 $b = -4$。即是说，

$$x^4(1-x)^4 = (x^6 - 4x^5 + 5x^4 - 4x^2 + 4)(x^2+1) - 4.$$

阶段 2 计算积分

除以 $1+x^2$，得

$$\frac{x^4(1-x)^4}{1+x^2} = x^6 - 4x^5 + 5x^4 - 4x^2 + 4 - \frac{4}{1+x^2},$$

所以

$$\int_0^1 \frac{x^4(1-x)^4}{1+x^2}\,\mathrm{d}x = \frac{1}{7} - \frac{4}{6} + \frac{5}{5} - \frac{4}{3} + 4 - 4\cdot\frac{\pi}{4} = \frac{1}{7} - \frac{2}{3} + 1 - \frac{4}{3} + 4 - \pi.$$

由于 $-\frac{2}{3} - \frac{4}{3} = -2$，有理数部分是 $\frac{1}{7} + 3 = \frac{22}{7}$，因此积分等于 $\frac{22}{7} - \pi$。 ■

阶段 3 估计积分的界

想一想 被积函数在 $(0,1)$ 上为正，所以立即得到 $\frac{22}{7} > \pi$。要得到 π 的下界（lower bound），我们需要这个积分的上界（upper bound）。$\frac{1}{1+x^2}$ 在 $[0,1]$ 上最大可以有多大?

在 $[0,1]$ 上，$1 \le 1+x^2 \le 2$，所以

$$\tfrac{1}{2}x^4(1-x)^4 \le \frac{x^4(1-x)^4}{1+x^2} \le x^4(1-x)^4,$$

而且除了在端点之外，不等式（inequality）都是严格成立的。

停一停，试一试 计算 $\int_0^1 x^4(1-x)^4\,\mathrm{d}x$，然后完成证明。

把展开式逐项积分，

$$\int_0^1 x^4(1-x)^4\,\mathrm{d}x = \frac{1}{9} - \frac{4}{8} + \frac{6}{7} - \frac{4}{6} + \frac{1}{5} = \frac{1}{630}.$$

（以 2520 为公分母：$280 - 1260 + 2160 - 1680 + 504 = 4$，而 $\frac{4}{2520} = \frac{1}{630}$。）所以

$$0 < \frac{22}{7} - \pi < \frac{1}{630},$$

亦即 $\frac{22}{7} > \pi > \frac{22}{7} - \frac{1}{630}$。 ■

以数值计算，$\frac{22}{7} - \frac{1}{630} = 3.14127\ldots$，所以我们已证明 $3.1412 < \pi < 3.1429$。若同时用上左边的不等式，便有 $\frac{1}{1260} < \frac{22}{7} - \pi$，这把上界改进为 $\pi < 3.14207$。

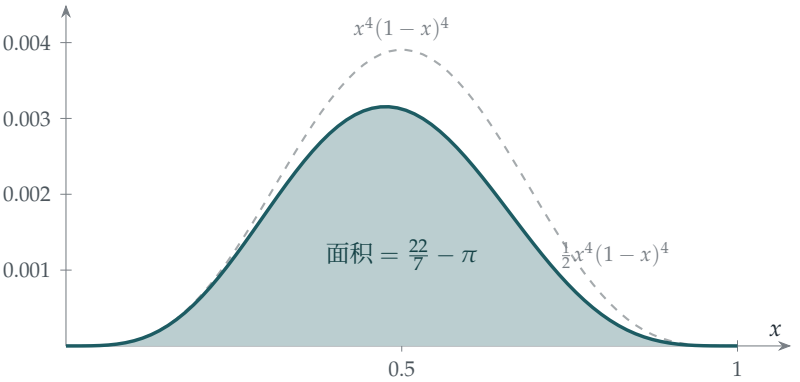


图 4.3.1 被积函数 $\frac{x^4(1-x)^4}{1+x^2}$（蓝绿色）为正，但数值极小：它的最大值约为 0.003。阴影部分的面积恰好是 $\frac{22}{7} - \pi \approx 0.00126$。被积函数介乎 $\frac{1}{2}x^4(1-x)^4$（点线）与 $x^4(1-x)^4$（虚线）之间，这两条曲线下的面积分别是 $\frac{1}{1260}$ 和 $\frac{1}{630}$。

回顾

- **以正值证明不等式。**正函数的积分为正。一旦认出积分等于 $\frac{22}{7} - \pi$，不等式便立即成立。
- **被积函数愈小，界愈精确。**因子 $x^4(1-x)^4$ 在 $[0,1]$ 上极小（最多只有 $\frac{1}{256}$），这使积分很小，所得的界也就很精确。
- **有理数加上 π 的倍数。**正如 Section 2.9 所见，一个形如“有理数 + 某常数的有理数倍”的积分，若其值很小，便给出该常数的一个良好有理近似值。

背后的数学

一个有来历的证明

阿基米德（Archimedes）约于公元前 250 年证明了 $\frac{223}{71} < \pi < \frac{22}{7}$，方法是计算圆内接和圆外切正 96 边形的周界（可与 Section 2.5 中的多边形比较）。积分证明则晚近得多。它由 D. P. Dalzell 于 1944 年发表，其后在 1968 年被选为美国大学本科生普特南数学竞赛（Putnam competition）的题目，因而广为人知。它的迷人之处在于完全不需要任何数值计算：不等式 $\frac{22}{7} > \pi$ 直接来自“正函数的积分为正”这个事实。

用同一思路得出更佳的近似值

这个方法可以推广。以 $x^8(1-x)^8$ 取代 $x^4(1-x)^4$，同样的除法给出

$$\int_0^1 \frac{x^8(1-x)^8}{1+x^2}\,\mathrm{d}x = 4\pi - \frac{188684}{15015},$$

由此证明 $\pi > \frac{47171}{15015} = 3.1415917\ldots$，准确至六位有效数字（significant figures）。2009 年，Stephen Lucas 找到了多族这类被积函数非负的积分，它们能证明关于其他著名近似值的不等式，其中包括准确至七位有效数字的 $\frac{355}{113}$。

近似值与无理性

为甚么 π 会有 $\frac{22}{7}$ 和 $\frac{355}{113}$ 这样好的分数近似值？每个无理数（irrational number）都有无穷多个分数 $\frac{p}{q}$ 与它相距不超过 $\frac{1}{q^2}$，而 π 碰巧在很早的阶段就有几个特别好的，因为它的连分数（continued fraction）$[3;7,15,1,292,\ldots]$ 含有 292 这个很大的项（见 Section 2.2）。像本题这样的积分，若令其愈来愈小，便是另一类证明的起点：这类证明指出某些数不可能被分数逼近得太好，因而是无理数。Ivan Niven 那个著名的、只有一页的 π 是无理数的证明（1947 年），用的就是 $\frac{x^n(1-x)^n}{n!}\sin(\pi x)$ 的积分，它们是 Section 2.9 中那些积分的三角函数近亲。

思维技巧

把不等式写成某个正值函数的积分

要证明 $A > B$，可找一个积分等于 $A - B$ 的正函数。

这样便把一个不等式化为一个恒等式（identity）再加上一个显然的事实。这是分析学（analysis）中最有用的技巧之一；同一思路，以平方和取代积分，就在

Section 3.2 中证明了算术平均—几何平均不等式（AM–GM）。

延伸练习

1 **核对数字。**计算 $\frac{22}{7} - \pi$ 至四位有效数字，并确认它介乎 $\frac{1}{1260}$ 与 $\frac{1}{630}$ 之间。

2 **一个 Beta 积分。**利用公式 $\int_0^1 x^n(1-x)^n\,\mathrm{d}x = \frac{(n!)^2}{(2n+1)!}$（在 Section 4.4 中证明）验证 $\int_0^1 x^4(1-x)^4\,\mathrm{d}x = \frac{1}{630}$。

3 **一个偏小的近似值。**证明

$$\int_0^1 \frac{x(1-x)^2}{1+x^2}\,\mathrm{d}x = \frac{\pi}{2} - \frac{3}{2} \quad \text{及} \quad \int_0^1 \frac{x^2(1-x)^4}{1+x^2}\,\mathrm{d}x = \pi - \frac{47}{15},$$

并由此推出 $\pi > 3$ 及 $\pi > \frac{47}{15}$。为甚么这些近似值不及 $\frac{22}{7}$ 准确?

4 **更进一步。**较难验证上文给出的 $\int_0^1 \frac{x^8(1-x)^8}{1+x^2}\,\mathrm{d}x$ 的恒等式。（除法可以交给电脑代数系统去做；重点是看出余式是 16，从而得出 4π。）

延伸阅读

- **Petr Beckmann, *A History of* π**（St Martin's Press）。详细讲述阿基米德的多边形方法，另有许多其他内容。
- **Stephen Lucas, "Approximations to π derived from integrals with nonnegative integrands"**, *American Mathematical Monthly* 116 (2009)。篇幅短小，容易阅读。
- **Ivan Niven, *Irrational Numbers***（Mathematical Association of America）。收录了 Niven 对 π 是无理数的证明。

4.4 沃利斯、Beta 与二分之一的阶乘

题目

STEP II 1995，第 4 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/95-s2-q4`

引子

$\left(\frac{1}{2}\right)!$ 是甚么？这个问题听来毫无道理：$n!$ 的意思是 $1 \times 2 \times \cdots \times n$，而你不可能把“半个数那么多”的项相乘。然而它确有一个自然的答案，就是

$$\left(\frac{1}{2}\right)! = \frac{\sqrt{\pi}}{2}.$$

π 从何而来？来自 $\sin t$ 的幂的积分（integral），这正是本题研究的对象。过程中，本题证明了关于 π 的其中一条最古老、最奇特的公式，即沃利斯（John Wallis）在 1655 年提出的无穷乘积（infinite product）

$$\frac{\pi}{2} = \frac{2}{1}\cdot\frac{2}{3}\cdot\frac{4}{3}\cdot\frac{4}{5}\cdot\frac{6}{5}\cdot\frac{6}{7}\cdots,$$

并且找出 $n!$ 的斯特灵公式（Stirling's formula）中隐藏的常数，这条公式是每一位概率学家和统计学家都会用到的。

动笔之前

想一想 u_0 和 u_1 是甚么？当 n 很大时，$\sin^n t$ 在 $[0, \frac{\pi}{2}]$ 上的图像大致是甚么样子？

$u_0 = \int_0^{\pi/2} 1\,\mathrm{d}t = \frac{\pi}{2}$，而 $u_1 = \int_0^{\pi/2} \sin t\,\mathrm{d}t = 1$。当 n 很大时，$\sin^n t$ 极小，只有在 $t = \frac{\pi}{2}$ 附近例外，因为那里 $\sin t$ 接近 1：图像变成紧贴右边界的一个狭窄尖峰（图 4.4.1）。所以 u_n 随 n 增大而缩小。本题准确地找出它缩小得有多快：与 $\sqrt{\pi/(2n)}$ 相若。

逐步解题

阶段 1　一条递推公式

> **想一想**　用分部积分法（integration by parts）计算 $\int \sin t \cdot \sin^{n-1} t \, \mathrm{d}t$：对 $\sin t$ 积分，对 $\sin^{n-1} t$ 求导。新的被积函数（integrand）含有 $\cos^2 t$；怎样才能变回 $\sin t$ 的幂？

取 $\mathrm{d}v = \sin t \, \mathrm{d}t$（故 $v = -\cos t$）及 $w = \sin^{n-1} t$（故 $w' = (n-1)\sin^{n-2} t \cos t$），

$$u_n = \Big[-\cos t \sin^{n-1} t\Big]_0^{\pi/2} + (n-1)\int_0^{\pi/2} \sin^{n-2} t \cos^2 t \, \mathrm{d}t.$$

当 $n \geq 2$ 时，边界项为零。写成 $\cos^2 t = 1 - \sin^2 t$，该积分便是 $u_{n-2} - u_n$，所以 $u_n = (n-1)(u_{n-2} - u_n)$，即

$$n u_n = (n-1) u_{n-2}. \qquad \blacksquare$$

两边同乘以 u_{n-1}，得 $n u_n u_{n-1} = (n-1) u_{n-1} u_{n-2}$。 $\blacksquare$

阶段 2　一个不变量

> **想一想**　这个关系式说明，当 n 减一时，$n u_n u_{n-1}$ 这个量保持不变。它在 $n = 1$ 时等于甚么？

反复运用，$n u_n u_{n-1} = (n-1) u_{n-1} u_{n-2} = \cdots = 1 \cdot u_1 u_0 = 1 \cdot 1 \cdot \frac{\pi}{2}$。所以对所有 $n \geq 1$，$n u_n u_{n-1} = \frac{1}{2}\pi$。 $\blacksquare$

阶段 3　夹逼

对于 $0 < t < \frac{\pi}{2}$，我们有 $0 < \sin t < 1$，所以 $\sin^n t < \sin^{n-1} t$：除端点外，$\sin^n t$ 的图像位于 $\sin^{n-1} t$ 的图像之下。积分后得 $0 < u_n < u_{n-1}$。

> **停一停，试一试**　利用 $u_n < u_{n-1}$ 和 $n u_n u_{n-1} = \frac{1}{2}\pi$ 把 $n u_n^2$ 夹住。

把 u_{n-1} 换成较小的 u_n，得 $n u_n^2 < n u_n u_{n-1} = \frac{1}{2}\pi$。把 u_n 换成较大的 u_{n-1}，得 $\frac{1}{2}\pi = n u_n u_{n-1} < n u_{n-1}^2$。所以 $n u_n^2 < \frac{1}{2}\pi < n u_{n-1}^2$。 $\blacksquare$

在右边的不等式中以 $n+1$ 代替 n：$\frac{1}{2}\pi < (n+1) u_n^2$，所以 $n u_n^2 > \frac{n}{n+1} \cdot \frac{1}{2}\pi$。合起来，

$$\left(\frac{n}{n+1}\right)\frac{1}{2}\pi < n u_n^2 < \frac{1}{2}\pi. \qquad \blacksquare$$

由于 $\frac{n}{n+1} \to 1$，由夹逼可得 $n u_n^2 \to \frac{1}{2}\pi$。 $\blacksquare$

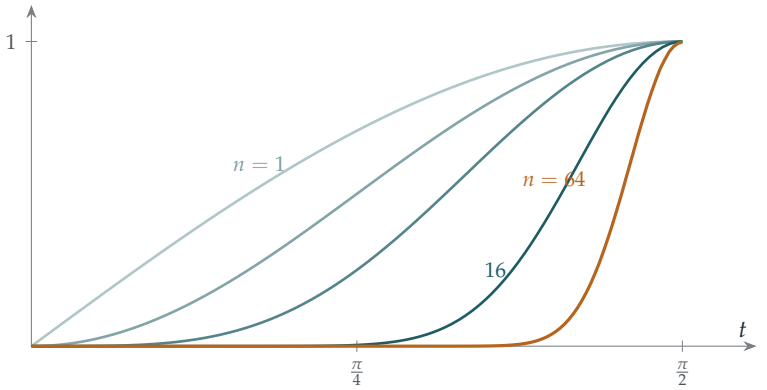


图 4.4.1 $n = 1, 2, 4, 16, 64$ 时 $\sin^n t$ 的图像。随着 n 增大，曲线下的面积 u_n 不断缩小，并集中在 $t = \frac{\pi}{2}$ 附近。本题证明 $u_n \approx \sqrt{\pi/(2n)}$；当 $n = 64$ 时，这个值约为 0.157。

回顾

- **递推公式（reduction formulae）。**分部积分法把 u_n 化为 u_{n-2}，所以所有 u_n 都由 u_0 和 u_1 决定。
- **寻找不变量（invariant）。**乘积 nu_nu_{n-1} 保持不变，正如 Section 2.6 中的积分在 AGM 步骤下保持不变一样。
- **单调性（monotonicity）加上不变量，便得出夹逼（squeeze）。**知道 $u_n < u_{n-1}$ 以及 nu_nu_{n-1} 的准确值，便把 nu_n^2 夹在两个极限相同的量之间。

背后的数学

沃利斯乘积

把递推公式一直用到 u_0 或 u_1，便得到明确的值：

$$u_{2m} = \frac{1\cdot 3\cdot 5\cdots(2m-1)}{2\cdot 4\cdot 6\cdots 2m}\cdot\frac{\pi}{2}, \qquad u_{2m+1} = \frac{2\cdot 4\cdot 6\cdots 2m}{3\cdot 5\cdot 7\cdots(2m+1)}.$$

由夹逼可知 $u_{2m}/u_{2m+1} \to 1$。把这个比写出来再整理，便得到沃利斯乘积

$$\frac{\pi}{2} = \lim_{m\to\infty}\frac{2\cdot 2}{1\cdot 3}\cdot\frac{4\cdot 4}{3\cdot 5}\cdots\frac{2m\cdot 2m}{(2m-1)(2m+1)}.$$

沃利斯在 1655 年以一种大胆的插值(interpolation)方法发现了它：他把 $\int_0^1(1-x^2)^{1/2}\,dx$ 放进一个指数为整数的积分表中，从而猜出它“应该”是甚么。这是数学史上最早的无

穷乘积之一，仅次于韦达（Viète）的乘积（Section 2.5），也是第一个只用有理数的无穷乘积。

二分之一的阶乘

沃利斯的插值想法，在 1729 年由欧拉（Euler）赋予了严格的意义。他找到一个函数，现称为*伽玛函数*（Gamma function），

$$\Gamma(s) = \int_0^\infty t^{s-1} e^{-t}\,\mathrm{d}t,$$

它满足 $\Gamma(s+1) = s\,\Gamma(s)$（用分部积分法），而且对整数 n 有 $\Gamma(n+1) = n!$。它把阶乘（factorial）平滑地延拓至所有正实数。与它密切相关的*贝塔函数*（Beta function）

$$\mathrm{B}(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}\,\mathrm{d}x = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$$

包含了 Section 4.3 和 Section 2.9 中的积分，而且经代换 $x = \sin^2 t$ 之后，也包含我们的 u_n。特别地，$\int_0^1 x^{1/2}(1-x)^{1/2}\,\mathrm{d}x = \frac{\pi}{8}$（练习 2），由此得 $\Gamma(\frac{3}{2})^2/\Gamma(3) = \frac{\pi}{8}$，即 $\left(\frac{1}{2}\right)! = \Gamma(\frac{3}{2}) = \frac{\sqrt{\pi}}{2}$。等价地，$\Gamma(\frac{1}{2}) = \sqrt{\pi}$，这与著名的高斯积分（Gaussian integral）$\int_{-\infty}^{\infty} e^{-x^2}\,\mathrm{d}x = \sqrt{\pi}$ 是同一个事实，也正是 π 出现在正态分布（normal distribution）公式中的原因。

斯特灵公式

斯特灵公式说 $n! \approx \sqrt{2\pi n}\,(n/e)^n$。其中 $(n/e)^n$ 的部分相对容易得到，但常数 $\sqrt{2\pi}$ 则较难，而求出它的标准方法正是本题：u_{2m} 的明确值涉及 $(2m)!/(m!)^2$，而极限 $nu_n^2 \to \frac{\pi}{2}$ 则把这个常数确定下来。另有一条截然不同的途径，通过概率得到同一个常数，见 Section 12.23。

思维技巧

由积分得出递推关系

分部积分法往往把带有参数 n 的积分，与参数较小的同一类积分联系起来。这个递推关系（recurrence）随后可以解出，或用来分析 $n \to \infty$ 时的表现。

留意递推关系下保持不变的量，以及对 n 的单调性；两者结合，往往不需要任何明确公式便能求出极限。

延伸练习

1 **沃利斯乘积。**利用 u_{2m} 和 u_{2m+1} 的明确值，以及 $u_{2m+1} < u_{2m} < u_{2m-1}$，证明沃利斯乘积。

2 **一个贝塔积分。**试做 STEP II 2013, 第 2 题（`step.maths.org/questions/13-s2-q2`）。这题用递推公式计算 $\int_0^1 x^n(1-x)^n\,\mathrm{d}x$，然后问当 $n=\frac{1}{2}$ 时答案有何启示：这是对二分之一的阶乘的初次一瞥。

3 **贝塔函数。**试做 STEP III 2025，第 1 题（载于 OCR 提供下载的 2025 年试卷，`www.ocr.org.uk/Images/750486-2025-past-papers.zip`）。这题建立一般的贝塔积分 $\int_0^1 x^{p-1}(1-x)^{q-1}\,\mathrm{d}x$、它的对称性和一个递推关系，并用它们计算一个在 $[0,\infty)$ 上的积分。

4 **中心二项式系数（central binomial coefficients）。**较难证明 $u_{2m}=\binom{2m}{m}\dfrac{\pi}{2^{2m+1}}$，并利用 $nu_n^2\to\frac{\pi}{2}$ 推导出当 m 很大时 $\binom{2m}{m}\approx\dfrac{4^m}{\sqrt{\pi m}}$。就 $m=10$ 验证这个结果。

延伸阅读

- **Julian Havil, *Gamma***（Princeton University Press）。内容包括伽玛函数及其历史。
- **Philip J. Davis, “Leonhard Euler’s integral: a historical profile of the Gamma function”,** *American Mathematical Monthly* 66 (1959)。一篇获奖的经典文章，读过本节之后便能看懂。
- **Emil Artin, *The Gamma Function***（Dover）。一本简短而优美的专著，适合学过一些分析之后阅读。

4.5 大二学生之梦

题目

STEP III 2009，第 8 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/09-s3-q8`

引子

每个学生迟早都会学到 $(a+b)^2 \neq a^2+b^2$。那条错误的等式有个戏称，叫做“大一新生之梦”（freshman's dream）。本题要证明的结果，看起来同样像初学者会犯的错：

$$\int_0^1 x^x \,\mathrm{d}x = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^n} = 1 - \frac{1}{2^2} + \frac{1}{3^3} - \frac{1}{4^4} + \cdots,$$

以及与它配对的 $\int_0^1 x^{-x}\,\mathrm{d}x = \sum_{n=1}^{\infty} \frac{1}{n^n}$。$x^x$ 的积分（integral）竟然变成 n^{-n} 的和：这似乎好得令人难以置信，所以它被称为大二学生之梦（sophomore's dream）。但它确实成立，由约翰·伯努利（Johann Bernoulli）于 1697 年发现。

证明用到三个你在大学会不断遇到的想法：比较指数函数与幂的极限（limit）、归约公式（reduction formula），以及把函数展开成幂级数（power series）再逐项积分（integrating term by term）。

动笔之前

想一想 函数 x^x 在 $x=0$ 处无法用你所知的任何公式定义（0^0 的意义并不明确）。当 x 从上方趋近 0 时，它的表现如何？试用计算机计算 $x = 0.1$、0.01、0.001 的情况。

$0.1^{0.1} \approx 0.794$，$0.01^{0.01} \approx 0.955$，$0.001^{0.001} \approx 0.993$。看来它趋近 1。第 (i) 部分正是证明这一点，由此可知被积函数（integrand）在 0 附近表现良好，这个积分是有意义的。x^x 在 $(0,1]$ 上的图像由接近 1 的地方出发，下降至极小值，再在 $x=1$ 处回到 1（图 4.5.1）。

逐步解题

阶段 1　幂胜过对数

想一想　当 $x \to 0^+$ 时，$\ln x \to -\infty$。哪一个代换（substitution）可以把 $x \to 0^+$ 变成 $t \to \infty$，并把 $x^m (\ln x)^n$ 变成类似 $e^{-mt} t^n$ 的式子？

设 $x = e^{-t}$。当 $x \to 0^+$ 时，$t \to \infty$，而 $\ln x = -t$，所以

$$x^m (\ln x)^n = e^{-mt} (-t)^n = (-1)^n e^{-mt} t^n \to 0.$$

用文字来说：当 $x \to 0$ 时，x 的任何正数次幂趋向零的速度，都快过 $\ln x$ 的任何次幂趋向无穷的速度。 ■

取 $m = n = 1$，得 $x \ln x \to 0$；由于指数函数是连续（continuous）的，所以 $x^x = e^{x \ln x} \to e^0 = 1$。 ■

阶段 2　归约公式

想一想　用分部积分法（integration by parts）计算 $I_{n+1} = \int_0^1 x^m (\ln x)^{n+1}\,\mathrm{d}x$。应该对哪一个因子积分？边界项（boundary term）在 $x = 0$ 处会怎样？

对 x^m 积分，对 $(\ln x)^{n+1}$ 求导：

$$I_{n+1} = \left[\frac{x^{m+1}}{m+1}(\ln x)^{n+1}\right]_0^1 - \int_0^1 \frac{x^{m+1}}{m+1} \cdot \frac{(n+1)(\ln x)^n}{x}\,\mathrm{d}x.$$

在 $x = 1$ 处，因为 $\ln 1 = 0$，边界项等于 0；当 $x \to 0$ 时，根据第 (i) 部分（取 $m+1$ 和 $n+1$），它趋向 0。余下的积分是 $\frac{n+1}{m+1} I_n$，所以

$$I_{n+1} = -\frac{n+1}{m+1} I_n. \qquad ■$$

由 $I_0 = \int_0^1 x^m\,\mathrm{d}x = \frac{1}{m+1}$ 出发，应用这条公式 n 次，得

$$I_n = (-1)^n \frac{n!}{(m+1)^n} \cdot \frac{1}{m+1} = \frac{(-1)^n\, n!}{(m+1)^{n+1}}. \qquad ■$$

阶段 3　级数

想一想　利用 e^y 的级数展开 $x^x = e^{x \ln x}$。每一项是甚么样子？m 取甚么值，才会出现第 (ii) 部分的积分？

由于对任何 y 都有 $e^y = \sum_{n=0}^{\infty} \frac{y^n}{n!}$，

$$x^x = e^{x\ln x} = \sum_{n=0}^{\infty} \frac{x^n (\ln x)^n}{n!}.$$

停一停，试一试 由 0 至 1 逐项积分，并使用第 (ii) 部分中 $m = n$ 的情况（$n = 0$ 另行处理）。你得到甚么？

$n = 0$ 的一项是 1，积分后得 1。当 $n \geq 1$ 时，在第 (ii) 部分取 $m = n$，得 $\int_0^1 x^n (\ln x)^n \, dx = \frac{(-1)^n n!}{(n+1)^{n+1}}$，所以第 n 项的贡献是 $\frac{(-1)^n}{(n+1)^{n+1}}$。把各项相加，

$$\int_0^1 x^x \, dx = 1 - \frac{1}{2^2} + \frac{1}{3^3} - \frac{1}{4^4} + \cdots . \qquad \blacksquare$$

（$n = 0$ 的一项也符合同一规律，即 $\frac{1}{1^1}$。）各项缩小得极快：取七项，便得到准确至小数点后六位的数值 0.783430…。

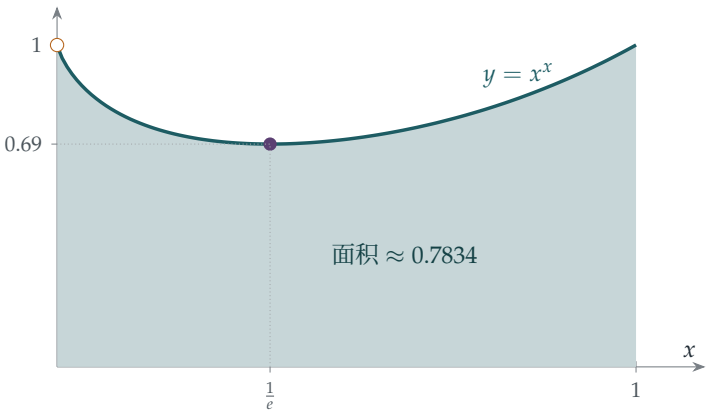


图 4.5.1 x^x 在 $(0, 1]$ 上的图像。当 $x \to 0$ 时它趋向 1（空心圆点），在 $x = \frac{1}{e}$ 处下降至极小值 $e^{-1/e} \approx 0.692$，再在 $x = 1$ 处回到 1。阴影部分的面积是 $\sum (-1)^{n+1} n^{-n} \approx 0.7834$。

回顾

- **用代换比较增长速度。**设 $x = e^{-t}$，便把在 0 处的极限变成在 ∞ 处一个熟悉的极限。
- **边界项必须检查。**归约公式有赖边界项在 $x = 0$ 处消失，这正是用到第 (i) 部分的地方。
- **先展开成级数，再积分。**把被积函数展开成幂级数并逐项积分，便把一个没有初等原函数（elementary antiderivative）的积分，变成一个收敛得很快的和。

背后的数学

交换求和与积分

第 3 阶段不动声色地把无穷和与积分交换了次序：$\int \sum = \sum \int$。这样做并非总是容许的。在大学里，你会遇到交换后不成立的例子，也会学到一些定理，说明甚么时候可以放心交换。最简单的一个定理指出，当级数一致收敛（converges uniformly）时，交换是安全的：即级数的尾部可以同时在每一个 x 处都变得很小。在这里，$|x \ln x| \le \frac{1}{e}$ 在 $(0, 1]$ 上成立，所以无论 x 是甚么，第 n 项的大小最多是 $\frac{e^{-n}}{n!}$。这些上界之和收敛，因此尾部一致地小，交换也就有了根据。这个论证称为*魏尔斯特拉斯 M 判别法*（Weierstrass M-test），是分析学（analysis）中最常用的工具之一。

为何这个“梦”会成真

这条恒等式很特别。它有赖 $\frac{(x \ln x)^n}{n!}$ 中的幂与积分 $\int_0^1 x^n (\ln x)^n \, dx = \frac{(-1)^n n!}{(n+1)^{n+1}}$ 完全吻合，当中的 $n!$ 恰好全部约去。追根究柢，这个积分其实是乔装了的伽玛函数（Gamma function）：代换 $x = e^{-t/(n+1)}$ 把它变成 $\frac{(-1)^n}{(n+1)^{n+1}} \int_0^\infty t^n e^{-t} \, dt = \frac{(-1)^n \, \Gamma(n+1)}{(n+1)^{n+1}}$，与 Section 4.4 连系起来。

名字的由来

“大二学生之梦”这个名字，是 Jonathan Borwein 与合著者在 2004 年出版的 *Experimentation in Mathematics* 一书中所创，用来与“大一新生之梦”$(a+b)^n = a^n + b^n$ 配对（有趣的是，后者在某些数系中是成立的：在模质数 p 的算术中，$(a+b)^p = a^p + b^p$，因为中间所有的二项式系数（binomial coefficient）都可被 p 整除）。Borwein 是*实验数学*（experimental mathematics）的先驱，这门学问利用电脑去发现像本题这样的恒等式，然后再以人手证明。

思维技巧

展开后逐项积分

如果被积函数有幂级数展开，而你又能对其中各项积分，便对每一项积分，再把结果加起来。

对于没有初等原函数的积分，这是一个很有效的求值方法。紧记要问：交换求和与积分是否有根据？通常是有的，但你应该知道原因。

延伸练习

1 **配对的等式。**证明 $\displaystyle\int_0^1 x^{-x}\,\mathrm{d}x = \sum_{n=1}^{\infty} \frac{1}{n^n} \approx 1.2913$。

2 **极小值。**证明 x^x 在 $(0,\infty)$ 上的极小值出现于 $x = \frac{1}{e}$，其值为 $e^{-1/e}$。

3 **需要多少项？**对于各项递减的交错级数（alternating series），截断后的误差小于第一个被略去的项。要准确至小数点后六位，需要取 $\int_0^1 x^x\,\mathrm{d}x$ 的级数的多少项？

4 **一个近亲。**用同样的方法证明 $\displaystyle\int_0^1 x^2 \ln x\,\mathrm{d}x = -\frac{1}{9}$，以及 $\displaystyle\int_0^1 \frac{\ln x}{1-x}\,\mathrm{d}x = -\sum_{n=1}^{\infty} \frac{1}{n^2}$。（根据 Section 2.4，这等于 $-\frac{\pi^2}{6}$。）

延伸阅读

- **Jonathan Borwein, David Bailey & Roland Girgensohn, *Experimentation in Mathematics***（A. K. Peters）。这个名字的出处，书中还有许多由电脑发现的恒等式。
- **William Dunham, *The Calculus Gallery***（Princeton University Press）。由牛顿（Newton）到勒贝格（Lebesgue）的分析学杰作，包括伯努利的工作。
- **Mary Hart, *Guide to Analysis***（Palgrave）。清楚讲解一致收敛，以及何时可以交换求和与积分。

4.6 没有公式的积分

题目

STEP III 2016，第 3 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/16-s3-q3`

引子

求导（differentiation）是机械式的工作：有一套法则可循，任何由常见部件砌成的函数，只要套用这些法则就能求导。积分（integration）却不然。每个学生都遇过各种技巧都攻不破的积分，并怀疑是否只是自己不够聪明。有时答案是：谁也不够聪明，因为根本没有公式。

本题就在一个特殊情况下证明这样的事。积分 $\int \frac{e^x}{x+1}\,\mathrm{d}x$ 作为面积当然存在，但它不能写成（有理函数）$\times e^x$ 的形式，而这正是用分部积分法（integration by parts）时预期会得到的形式。证明某件事不可能做到，而不是把某样东西找出来，是另一种数学，也极具数学的特色。

动笔之前

想一想 若 $\int \mathrm{f}(x)e^x\,\mathrm{d}x = \frac{\mathrm{P}}{\mathrm{Q}}e^x + C$，P 和 Q 必须满足甚么方程？（对右边求导。）

求导得 $\dfrac{\mathrm{d}}{\mathrm{d}x}\left(\dfrac{\mathrm{P}}{\mathrm{Q}}e^x\right) = \dfrac{\mathrm{P}'\mathrm{Q} - \mathrm{P}\mathrm{Q}' + \mathrm{P}\mathrm{Q}}{\mathrm{Q}^2}e^x$，而这必须等于 $\mathrm{f}(x)e^x$。所以整道题目其实围绕着以下的多项式恒等式（polynomial identity）：

$$\mathrm{P}'\mathrm{Q} - \mathrm{P}\mathrm{Q}' + \mathrm{P}\mathrm{Q} = \mathrm{f}\,\mathrm{Q}^2.$$

关于积分的问题，就此变成关于多项式的问题：它们的因式和次数。

逐步解题

阶段 1　因式 $x+1$

取 $\mathrm{f}(x)=\dfrac{x^3-2}{(x+1)^2}$，把恒等式两边同乘 $(x+1)^2$：

$$(x+1)^2(\mathrm{P}'\mathrm{Q}-\mathrm{P}\mathrm{Q}'+\mathrm{P}\mathrm{Q})=(x^3-2)\,\mathrm{Q}^2. \tag{†}$$

> ? **想一想**　在 (†) 中代入 $x=-1$。必然会出现甚么情况？

当 $x=-1$ 时，左边等于 0，右边则是 $(-1-2)\,\mathrm{Q}(-1)^2=-3\,\mathrm{Q}(-1)^2$。所以 $\mathrm{Q}(-1)=0$，而根据因式定理（factor theorem），$x+1$ 是 Q 的因式。 ■

阶段 2　次数

设 P 的次数（degree）为 p，Q 的次数为 q。

> ? **想一想**　在 $\mathrm{P}'\mathrm{Q}-\mathrm{P}\mathrm{Q}'+\mathrm{P}\mathrm{Q}$ 中，哪一项的次数最高？它会被其余各项抵消吗？

$\mathrm{P}\mathrm{Q}$ 的次数是 $p+q$，而 $\mathrm{P}'\mathrm{Q}$ 和 $\mathrm{P}\mathrm{Q}'$ 的次数最多是 $p+q-1$。所以 (†) 左边的次数恰好是 $2+p+q$，右边的次数是 $3+2q$。令两者相等，得 $p=q+1$。 ■

阶段 3　求 P

取 $\mathrm{Q}=x+1$，则 $\mathrm{Q}'=1$，而 P 的次数为 2。把 (†) 除以 $(x+1)^2$：

$$\mathrm{P}'(x+1)-\mathrm{P}+\mathrm{P}(x+1)=x^3-2, \qquad \text{即} \qquad (x+1)\mathrm{P}'+x\mathrm{P}=x^3-2.$$

试设 $\mathrm{P}=\alpha x^2+\beta x+\gamma$。左边是

$$(x+1)(2\alpha x+\beta)+\alpha x^3+\beta x^2+\gamma x=\alpha x^3+(2\alpha+\beta)x^2+(2\alpha+\beta+\gamma)x+\beta.$$

与 x^3-2 比较系数：$\alpha=1$；$2\alpha+\beta=0$，故 $\beta=-2$；于是 $\gamma=0$；常数项 $\beta=-2$ 亦吻合。所以

$$\mathrm{P}(x)=x^2-2x, \qquad \int \frac{x^3-2}{(x+1)^2}\,e^x\,\mathrm{d}x=\frac{x^2-2x}{x+1}\,e^x+C. \qquad ■$$

（有兴趣的话可以求导验算，这是练习商法则（quotient rule）的好题目。）

阶段 4　一件不可能的事

现在假设 $\int \frac{e^x}{x+1}\,\mathrm{d}x=\frac{\mathrm{P}}{\mathrm{Q}}e^x+C$，其中 P 和 Q 没有公因式（common factor）。恒等式变成

$$(x+1)\left(\mathrm{P}'\mathrm{Q}-\mathrm{P}\mathrm{Q}'+\mathrm{P}\mathrm{Q}\right)=\mathrm{Q}^2.$$

一如第 1 阶段，代入 $x=-1$ 可知 $Q(-1)=0$。写成 $Q=(x+1)^k R$，其中 $k\geq 1$ 且 $R(-1)\neq 0$（即尽量抽出 $x+1$ 这个因式）。

> **停一停，试一试** 代入 $Q=(x+1)^k R$，尽量约去 $(x+1)$ 的因式，然后代入 $x=-1$。你能得知 $P(-1)$ 的甚么性质？

这时 $Q'=k(x+1)^{k-1}R+(x+1)^k R'$，而

$$P'Q-PQ'+PQ=(x+1)^{k-1}\Big[(x+1)\big(P'R-PR'+PR\big)-kPR\Big].$$

所以恒等式变成 $(x+1)^k\big[(x+1)(P'R-PR'+PR)-kPR\big]=(x+1)^{2k}R^2$。除以 $(x+1)^k$：

$$(x+1)\big(P'R-PR'+PR\big)-kPR=(x+1)^k R^2.$$

当 $x=-1$ 时，由于 $k\geq 1$，上式给出 $-kP(-1)R(-1)=0$。因为 $k\neq 0$ 且 $R(-1)\neq 0$，必有 $P(-1)=0$。但这样 $x+1$ 就同时整除 P 和 Q，与两者没有公因式的假设矛盾。所以这样的 P 和 Q 并不存在。 ■

回顾

- **转化为代数。**“这个积分能否写成这种形式?”变成了“这个多项式恒等式有没有解?”
- **在特殊点取值。**代入 $x=-1$ 就得到关于因式的信息，前后共用了三次。
- **利用最简分数的反证法（proof by contradiction）。**正如 Section 1.3 中处理 $\sqrt{2}$ 的做法，先假设分数已约至最简，再找出一个公因式，便得到矛盾。

背后的数学

极点与对数

第 (ii) 部分背后有一幅简单的图像。被积函数（integrand）$\frac{e^x}{x+1}$ 在 $x=-1$ 处有一个**单极点**（simple pole）：在该点附近，它的表现像 $\frac{e^{-1}}{x+1}$。对 $\frac{1}{x+1}$ 积分得 $\ln|x+1|$，所以原函数（antiderivative）在 -1 附近的表现应该像 $e^{-1}\ln|x+1|$，以对数的速度趋向 $-\infty$（图 4.6.1）。但有理函数乘以 e^x 只能以幂 $(x+1)^{-k}$ 的方式趋向无穷，其导数因而有一个阶数为 $k+1\geq 2$ 的极点，绝不会是单极点。有理函数无法产生对数式的表现；第 4 阶段的代数运算，正是这个观察的严格版本。相反，在第 (i) 部分，被积函数有一个二阶极点，而这正是带有单极点的函数求导后的模样，所以答案才有可能是有理的。

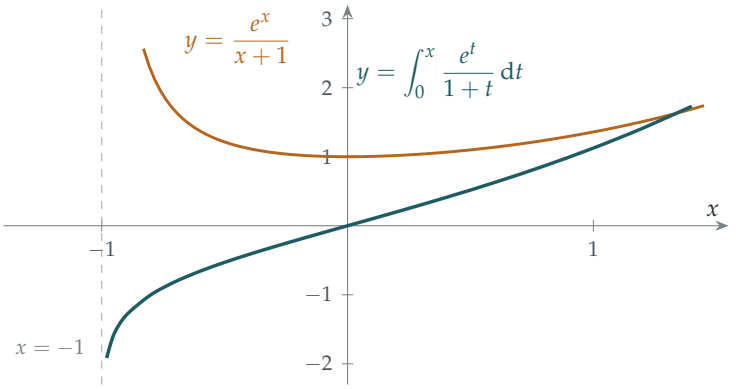


图 4.6.1　被积函数 $\frac{e^x}{x+1}$（琥珀色）在 $x = -1$ 处有一个单极点。它的原函数（蓝绿色，以数值方法计算）完全存在，但当 $x \to -1^+$ 时，它以对数的速度趋向 $-\infty$，表现像 $e^{-1}\ln(x+1)$。没有任何有理函数乘以 e^x 能有这样的表现。

刘维尔与 Risch 算法

在 1830 年代，刘维尔（Joseph Liouville）建立了一套理论，研究哪些积分可以用*初等函数*（elementary functions）表示（即由多项式、指数函数、对数函数、根式和三角函数经通常的运算构成的函数）。他的其中一个结果指出：若 f 和 g 是有理函数而 g 不是常数，则 $\int \mathrm{f}(x)e^{\mathrm{g}(x)}\,\mathrm{d}x$ 只有在等于 $\mathrm{R}(x)e^{\mathrm{g}(x)}$（另加一个常数）时才是初等的，其中 R 是某个有理函数。（对 g 的条件不可或缺：当 $\mathrm{g} = 0$ 而 $\mathrm{f}(x) = \frac{1}{x}$ 时，积分是 $\ln|x|$，它是初等函数，却不是有理函数。）所以本题排除了有理形式，其实就证明了 $\int \frac{e^x}{x+1}\,\mathrm{d}x$ 根本不是初等函数。同一定理的其他版本可证明 $\int e^{-x^2}\,\mathrm{d}x$ 和 $\int \frac{\sin x}{x}\,\mathrm{d}x$ 都不是初等的。1969 年，Robert Risch 把刘维尔的想法发展成一套算法（algorithm），可就一大类函数判定初等积分是否存在，并在存在时把它求出来。电脑代数系统（computer algebra system）内部所用的，就是这套算法的不同版本。

特殊函数

当一个积分没有初等形式时，数学家会给它起个名字，把它当作新的函数来研究，正如我们在 Section 3.9 中处理朗伯 W 函数（Lambert W function）那样。这里的积分本质上就是*指数积分*（exponential integral）$\mathrm{Ei}(x)$，它出现于热传导问题，也出现于*对数积分*（logarithmic integral）$\mathrm{li}(x) = \mathrm{Ei}(\ln x)$ 之中，后者是不超过 x 的质数（prime）个数的最佳简单近似（Section 1.2）。当 $x < 0$ 时，它是 $\int_{-\infty}^{x} \frac{e^t}{t}\,\mathrm{d}t$。当 $x > 0$ 时，这个积分会穿过 $t = 0$ 处的无穷尖峰，所以 $\mathrm{Ei}(x)$ 的定义是先挖去尖峰周围的一个小区间 $(-\varepsilon, \varepsilon)$，再令 $\varepsilon \to 0$：这就是*柯西主值*（Cauchy principal value）。

思维技巧

证明不可能

要证明某件事做不到，先假设它做得到，把这个假设转化为精确的代数式，然后推出矛盾。

困难在于转化：要为“做得到”的意思找出一个精确的陈述。在这里，它的意思是“存在没有公因式的多项式 P、Q 满足这个恒等式”。数学中许多最深刻的结果，都是这类不可能性证明。

延伸练习

1 **两个可行的例子。**证明 $\int \frac{xe^x}{(x+1)^2}\,dx = \frac{e^x}{x+1} + C$ 及 $\int \frac{(x^2+1)e^x}{(x+1)^2}\,dx = \frac{x-1}{x+1}e^x + C$。

2 **高斯函数。**证明不存在多项式 P 使得 $\int e^{-x^2}\,dx = P(x)e^{-x^2} + C$。（提示：求导后比较次数。）

3 **另一个极点。**仿照第 (ii) 部分的论证，证明 $\int \frac{e^x}{x}\,dx$ 不能写成 $\frac{P(x)}{Q(x)}e^x + C$ 的形式。

4 **何时可行？**假设 f 是多项式，且对某个多项式 P 有 $\int \frac{f(x)}{(x+1)^2}e^x\,dx = \frac{P(x)}{x+1}e^x + C$。证明 $(x+1)P' + xP = f$，并通过计算此式及其导数在 $x = -1$ 的值，证明 $f(-1) + f'(-1) = 0$。验证这个条件对 $f(x) = x^3 - 2$ 以及练习 1 的例子都成立。较难你能证明这个条件也是充分的吗？

延伸阅读

- **Brian Conrad, “Impossibility theorems for elementary integration”**（初等积分的不可能性定理）。可免费取得的讲义，配合例子讲解刘维尔定理。
- **Elena Anne Marchisotto & Gholam-Ali Zakeri, “An invitation to integration in finite terms”**（有限形式积分导引），*The College Mathematics Journal* 25 (1994)。刘维尔理论的浅白入门。
- **Manuel Bronstein, *Symbolic Integration I***（Springer）。讲述电脑代数系统实际上如何求积分。程度较深。

4.7 变分法

题目

STEP III 2014，第 4 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/14-s3-q4`

引子

中学里的最优化（optimisation）问题，大多是求最佳的*数*：例如求令面积最大的 x 值。*变分法*（calculus of variations）要求的却是最佳的*函数*。连接两点的曲线之中，珠子沿哪一条滑下最快？悬挂的链条会呈甚么形状？光线穿过透镜时走哪一条路径？在这些例子中，要优化的量都是一个积分，而未知的是一整个函数。

本题研究的是最简单而又有趣的例子。积分 $J = \int_0^1 ((y')^2 - a^2 y^2)\,\mathrm{d}x$ 在两个互相竞争的项之间取得平衡：函数愈陡，积分愈大；函数值愈大，积分愈小。当 a 较小时，陡峭程度总是占上风，所以对每个函数都有 $J \geq 0$。但 a 有一个临界值，平衡会在那里逆转，而本题正好把它准确求出：$a = \frac{\pi}{2}$。π 从何而来，就是本节要讲的故事。

动笔之前

想一想 为甚么 $I \geq 0$ 并非显而易见？试用 $y = 1 - x$，它满足 $y(1) = 0$。

I 是两个正数量之差，所以它可正可负。对于 $y = 1 - x$：$\int_0^1 1\,\mathrm{d}x = 1$ 而 $\int_0^1 (1-x)^2\,\mathrm{d}x = \frac{1}{3}$，所以 $I = \frac{2}{3} > 0$。但若函数选得更巧妙，或许可以令 $\int y^2$ 相对于 $\int (y')^2$ 变得更大。题目的窍门，是把 I 写成一个*平方*的积分，而这显然是非负的。这就是积分版本的“配方法”（completing the square）。

逐步解题

阶段 1　配方

把 I_1 中的平方展开：

$$I_1 = \int_0^1 \left((y')^2 + 2yy' \tan x + y^2 \tan^2 x\right) \mathrm{d}x,$$

所以

$$I - I_1 = -\int_0^1 \left(y^2 + y^2 \tan^2 x + 2yy' \tan x\right) \mathrm{d}x = -\int_0^1 \left(y^2 \sec^2 x + 2yy' \tan x\right) \mathrm{d}x.$$

> **想一想**　被积函数 $y^2 \sec^2 x + 2yy' \tan x$ 恰好是某个函数的导数（exact derivative）。是哪个函数？

根据乘积法则（product rule），$\dfrac{\mathrm{d}}{\mathrm{d}x}(y^2 \tan x) = 2yy' \tan x + y^2 \sec^2 x$。所以

$$I - I_1 = -\Big[y^2 \tan x\Big]_0^1 = -(y(1)^2 \tan 1 - y(0)^2 \tan 0) = 0,$$

因为 $y(1) = 0$ 而 $\tan 0 = 0$。所以 $I = I_1 \geq 0$，因为它是一个平方的积分。 ■

阶段 2　何时 $I = 0$？

> **想一想**　如果一个连续的非负函数的积分是零，你对这个函数能下甚么结论？那么若 $I = 0$，y 必须满足甚么方程？

积分为零的连续非负函数必定恒等于零（假如它在某一点取正值，它在该点附近的一个小区间上也会取正值，从而贡献一个正的量）。所以 $I = 0$ 迫使 $y' + y \tan x = 0$ 在 $[0, 1]$ 上成立。乘以 $\sec x$ 后，这就是说 $\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{y}{\cos x}\right) = \frac{y' \cos x + y \sin x}{\cos^2 x} = 0$，所以 $y = C \cos x$。于是 $y(1) = C \cos 1 = 0$，而由于 $\cos 1 \neq 0$，得 $C = 0$。所以只有当 $y \equiv 0$ 时才有 $I = 0$。 ■

阶段 3　一般情况

> **想一想**　用 $\int_0^1 (y' + ay \tan bx)^2 \,\mathrm{d}x$ 重复第 1 步。b 必须是甚么，余下的项才能恰好组成一个导数？

把这个积分记作 K。那么

$$\begin{aligned} J - K &= -\int_0^1 \left(a^2 y^2 + a^2 y^2 \tan^2 bx + 2ayy' \tan bx\right) \mathrm{d}x \\ &= -\int_0^1 \left(a^2 y^2 \sec^2 bx + 2ayy' \tan bx\right) \mathrm{d}x. \end{aligned}$$

另一方面，$\frac{\mathrm{d}}{\mathrm{d}x}(ay^2\tan bx) = 2ayy'\tan bx + ab\,y^2\sec^2 bx$。两者恰好在 $ab = a^2$ 时吻合，即 $b = a$。于是

$$J - K = -\Big[ay^2\tan ax\Big]_0^1 = -a\,y(1)^2\tan a = 0,$$

所以 $J = K \geq 0$。

> **停一停，试一试** 这个论证在哪里可能失效？（留意 $\tan ax$ 在区间 $0 \leq x \leq 1$ 上的情况。）

这个论证需要 $\tan ax$ 对 $[0,1]$ 内所有 x 都有定义而且连续；否则 K 未必存在，微积分基本定理（fundamental theorem of calculus）亦不能应用。当 x 取遍 $[0,1]$ 时，ax 取遍 $[0,a]$，而 tan 在该处连续的充要条件正是 $a < \frac{\pi}{2}$。所以这个证明在以下情况成立：

$$a < k = \frac{\pi}{2}. \qquad \blacksquare$$

阶段 4 临界情况

> **想一想** 当 $a = \frac{\pi}{2}$ 时，证明失效。在第 2 步，令 $I = 0$ 的函数须满足 $y' + y\tan x = 0$。现在对应的方程是甚么？它的解又是甚么？

对应的方程是 $y' + ay\tan ax = 0$，它的解是 $y = C\cos ax$。当 $a = \frac{\pi}{2}$ 时，$y = \cos\frac{\pi x}{2}$ 满足 $y(1) = \cos\frac{\pi}{2} = 0$。直接验证：

$$\begin{aligned} J &= \frac{\pi^2}{4}\int_0^1 \left(\sin^2\frac{\pi x}{2} - \cos^2\frac{\pi x}{2}\right)\mathrm{d}x = -\frac{\pi^2}{4}\int_0^1 \cos\pi x\,\mathrm{d}x \\ &= -\frac{\pi^2}{4}\left[\frac{\sin\pi x}{\pi}\right]_0^1 = 0. \qquad \blacksquare \end{aligned}$$

所以在临界值处，有一个非零函数令 J 恰好等于零。

回顾

- **在积分中配方。**加上再减去一个恰好是导数的式子，便把 J 化成一个平方的积分，再加上一个等于零的边界项（boundary term）。
- **查看工具在哪里失效。**这个方法只在 $\tan ax$ 在区间上保持有限时才行得通。它失效之处 $a = \frac{\pi}{2}$，正正就是临界值。
- **极端情况是某个微分方程的解。**令 $J = 0$ 的函数来自一阶微分方程（differential equation）$y' + ay\tan ax = 0$，而它同时满足 $y'' + a^2y = 0$。

背后的数学

最速降线与一门学科的诞生

1696 年，约翰·伯努利（Johann Bernoulli）向欧洲的数学家发出挑战：求一条曲线，使珠子沿着它从一点滑到另一点所需的时间最短。这就是最速降线（brachistochrone）问题。提交解答的有牛顿（Newton）（他匿名投稿，一夜之间解出；伯努利说他“从爪子认出了狮子”）、莱布尼兹（Leibniz）、洛必达（l'Hôpital）和雅各布·伯努利（Jakob Bernoulli）。答案是一条摆线（cycloid）。到了 1750 年代，欧拉（Euler）和年轻的拉格朗日（Joseph-Louis Lagrange）把这些个别的技巧变成一套通用的方法：如果 y 令 $\int F(x, y, y')\,dx$ 取平稳值，那么 y 满足**欧拉－拉格朗日方程**（Euler–Lagrange equation）

$$\frac{d}{dx}\frac{\partial F}{\partial y'} = \frac{\partial F}{\partial y}.$$

对于我们的 J，$F = (y')^2 - a^2y^2$，而欧拉－拉格朗日方程就是 $y'' + a^2y = 0$：简谐运动（simple harmonic motion）。

为何是 $\pi/2$?

当 $a < \frac{\pi}{2}$ 时，J 的最小值是 0，而且只有 $y \equiv 0$ 才能取得。当 $a = \frac{\pi}{2}$ 时，最小值仍然是 0，但多了一个函数能取得它。当 $a > \frac{\pi}{2}$ 时，同一个函数 $y = \cos\frac{\pi x}{2}$ 给出

$$J = \frac{\pi^2}{8} - \frac{a^2}{2} < 0,$$

而只要把 y 按比例放大，J 就可以负得要多大有多大：根本没有最小值（图 4.7.1）。临界值就是第一个令 $y'' + a^2y = 0$ 有非零解满足边界条件的 a（边界条件是 $y(1) = 0$，以及在自由端的“自然”条件 $y'(0) = 0$）。这是一个**特征值问题**（eigenvalue problem），而 $a^2 = \frac{\pi^2}{4}$ 是它最小的特征值。试与 Section 3.7 比较：那里把函数在长度为 π 的区间的两端都固定，得出临界值 1；这里区间长度为 1，而且只固定一端，相当于长度为 2 的区间的一半，所以临界的 a 是 $\frac{\pi}{2}$。

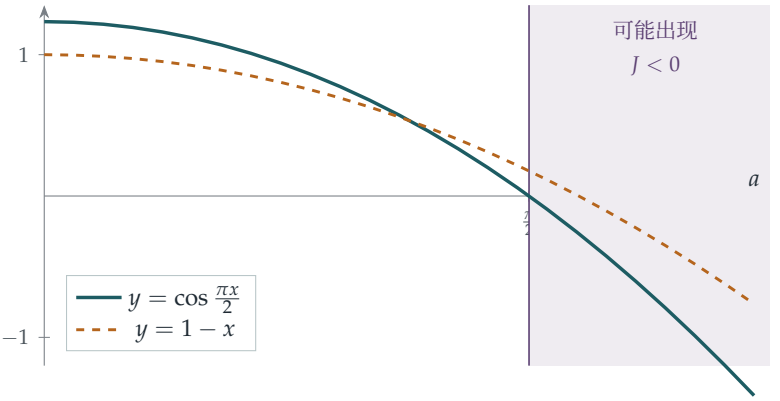


图 4.7.1 对于两个满足 $y(1) = 0$ 的试验函数，J 的值随 a 的变化。对于 $y = 1 - x$，$J = 1 - \frac{a^2}{3}$ 要到 $a > \sqrt{3}$ 才变成负数。对于 $y = \cos\frac{\pi x}{2}$，$J = \frac{\pi^2}{8} - \frac{a^2}{2}$ 在 $a > \frac{\pi}{2}$ 时随即变成负数，而且没有函数能做得更好：当 $a < \frac{\pi}{2}$ 时，对每个 y 都有 $J \geq 0$。

物理学与最小作用量原理

后来发现，变分法描述的竟是大自然本身。在十八和十九世纪，人们逐渐明白运动定律可以这样表述：力学系统的运动方式，会令一个称为作用量（action）的积分取平稳值，这就是最小作用量原理（principle of least action）。同样，光沿着需时最短的路径前进（费马原理，Fermat's principle），这解释了折射现象。到了二十世纪，同一个想法通过费曼（Richard Feynman）的“路径求和”（sum over paths），成为量子力学的基础。本题中的二次积分是最简单的例子，而它们引出的特征值问题，会在 Section 8.6 以量子振子能级的面貌再次出现。

思维技巧

先配方，再留意边界

要证明一个二次式非负，可把它写成一个平方加上一些你能控制的东西。对于积分，这些“东西”往往恰好是某个函数的导数，积分后便成为边界项。

此外，永远要问这个方法在哪里不再奏效。本题中方法失效之处并非偶然；它正是题目中最重要的那个数。

延伸练习

1 **越过临界值。**证明若 $a > \frac{\pi}{2}$，则 $y = \cos\frac{\pi x}{2}$ 给出 $J < 0$。

2 **欧拉－拉格朗日方程。**较难设 y 在满足 $y(1) = 0$ 的函数之中令 J 取最小值。对任何满足 $\eta(1) = 0$ 的函数 η 和任何数 ε，都有 $J(y + \varepsilon\eta) \geq J(y)$。把 $J(y + \varepsilon\eta)$ 按 ε 的幂展开，并由此推出 $\int_0^1 (y'\eta' - a^2 y\eta)\,\mathrm{d}x = 0$。用分部积分法（integration by parts）证明 $y'' + a^2 y = 0$ 及 $y'(0) = 0$（假设 y 足够光滑）。

3 **较长的区间。**用代换法（substitution）证明：若 $y(L) = 0$，则 $\int_0^L \left((y')^2 - a^2 y^2\right)\mathrm{d}x \geq 0$ 对所有 y 成立，当且仅当 $a \leq \frac{\pi}{2L}$；并证明只有当 $a = \frac{\pi}{2L}$ 时，才有非零的 y 令等号成立。

4 **摆线。**查阅最速降线问题，并找出摆线的参数方程（parametric equations）。证明在一条倒置的摆线上，珠子无论从哪一点由静止释放，到达最低点所需的时间都相同（这是等时（tautochrone）性质，由惠更斯（Huygens）于 1659 年发现）。

延伸阅读

- **Paul Nahin, *When Least Is Best***（Princeton University Press）。讲述最优化的历史与数学，包括最速降线，为学生而写。
- **Richard Feynman, *The Feynman Lectures on Physics***，第二卷第 19 章（“The principle of least action”，最小作用量原理）。可在网上免费阅读。
- **Bruce van Brunt, *The Calculus of Variations***（Springer）。一本大学本科教科书；头几章只需要高级程度（A-level）的微积分知识。

4.8 切球形面包

题目

STEP I 2001，第 6 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/01-s1-q6`

引子

把一个球形面包切成厚度相等的薄片。靠近中间的几片是又大又厚的圆片，只有边缘一圈窄窄的面包皮；靠近两端的几片则是小小的球冠，看起来几乎全是面包皮。两端的薄片，面包皮一定比较多吧?

其实不然。不论切在靠近中间还是靠近两端的位置，每一片的面包皮面积都完全相同。中间那圈窄带周长很大，但几乎是垂直的；两端的球冠周长很小，但倾斜得很厉害，所以面包皮铺开的阔度反而较大。这两种效应恰好互相抵消。这是阿基米德（Archimedes）的一条定理，他对这条定理珍爱有加，甚至要求在自己的墓碑上刻上一个内切于圆柱（cylinder）的球（sphere）。题目的第二部分是一则较为轻松的“规例”，但它所依据的正是同一个计算。

动笔之前

想一想 半径为 a 的球，是由哪一条曲线绕 x-轴旋转而成的？该曲线的 $\frac{\mathrm{d}y}{\mathrm{d}x}$ 是甚么？

把半圆（semicircle）$y = \sqrt{a^2 - x^2}$, $-a \le x \le a$ 旋转，便得到这个球。于是 $\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{-x}{\sqrt{a^2-x^2}} = -\frac{x}{y}$。请记住这一点：斜率（slope）在两端附近（$x \to \pm a$）很陡，在中间则很平缓，这正是引子中所描述的效应。

逐步解题

阶段 1　面包皮公式的化简

> **想一想**　对这个半圆计算 $y\sqrt{1+(y')^2}$。奇妙的事情出现了。

由于 $y' = -x/y$，

$$1 + (y')^2 = 1 + \frac{x^2}{y^2} = \frac{y^2 + x^2}{y^2} = \frac{a^2}{y^2},$$

所以 $y\sqrt{1+(y')^2} = y \cdot \frac{a}{y} = a$，竟然是一个常数！因此，任意两个平面 $x = x_1$ 与 $x = x_2$ 之间的面包皮面积为

$$2\pi \int_{x_1}^{x_2} a \, \mathrm{d}x = 2\pi a (x_2 - x_1).$$

它只取决于两个平面之间的距离，而与平面的位置无关。

阶段 2　面包皮与片数成正比

如果把面包切成厚度为 h 的薄片，每一片的面包皮面积都是 $2\pi a h$，片片相同。不论是从一端吃、从两端吃，还是从任何地方吃掉若干片，剩下的面包皮面积都是 $2\pi a h$ 乘以剩下的片数。 ■

阶段 3　“脆皮”的条件

题目的提示所描述的，是从一端吃掉若干片之后剩下的面包，我们就依照提示去做：假设从 $x = a$ 那一端开始吃，于是剩下的部分由 $x = -a$ 延伸至 $x = -a + t$。一项关于“任意片数”的规定，应该不论面包切得多薄都成立，所以我们把 t 看成可以取 0 与 $2a$ 之间的任何值。（第 4 阶段会探讨从两端吃的情况。）

> **想一想**　利用题目给出的公式计算 V。使用代换（substitution）$s = x + a$ 或许能令代数运算更简洁。

设 $s = x + a$，则 $y^2 = a^2 - x^2 = a^2 - (s-a)^2 = 2as - s^2$，所以

$$V = \pi \int_0^t (2as - s^2)\, \mathrm{d}s = \pi \left(at^2 - \frac{t^3}{3} \right), \qquad A = 2\pi a t,$$

以及

$$\frac{V}{A} = \frac{at^2 - t^3/3}{2at} = \frac{t}{2} - \frac{t^2}{6a}.$$

停一停，试一试 规定要求吃掉任意片数之后都有 $V/A < 1$，即是对所有满足 $0 < t \le 2a$ 的 t 都成立。在这个范围内，$\frac{t}{2} - \frac{t^2}{6a}$ 的最大值是多少?

对 t 求导：$\frac{1}{2} - \frac{t}{3a} = 0$ 于 $t = \frac{3a}{2}$ 时成立，这个值位于范围 $(0, 2a]$ 之内，而且是极大值（maximum）（该式是 t 的一条开口向下的抛物线）。最大值为

$$\frac{3a}{4} - \frac{9a^2/4}{6a} = \frac{3a}{4} - \frac{3a}{8} = \frac{3a}{8}.$$

脆皮面包需要对所有 t 都有 $V/A < 1$，特别是在 $t = \frac{3a}{2}$ 时也要成立，所以 $\frac{3a}{8} < 1$，即 $a < \frac{8}{3} = 2\frac{2}{3}$ 米。 ■

阶段 4 题目之外：$2\frac{2}{3}$ 足够吗?

想一想 题目只要求找出半径必须满足的条件。$a < \frac{8}{3}$ 是否同时也足够? 如果从两端吃又如何?

并不足够。从两端吃会在中间剩下一条环带，而以同样多的面包皮计算，中间的环带比末端的球冠包含更多面包。对于环带 $-\frac{w}{2} \le x \le \frac{w}{2}$，

$$V = \pi \int_{-w/2}^{w/2} (a^2 - x^2)\,\mathrm{d}x = \pi\left(a^2 w - \frac{w^3}{12}\right), \qquad A = 2\pi a w, \qquad \frac{V}{A} = \frac{a}{2} - \frac{w^2}{24a},$$

当环带很窄时，这个值接近 $\frac{a}{2}$。例如，一个半径为 2.4 米的面包，明显小于 $2\frac{2}{3}$，切成四片后吃掉两端的两片，剩下的环带有 $w = a$ 及 $\frac{V}{A} = \frac{11a}{24} = 1.1$。事实上，对于任何一组薄片都有 $\frac{V}{A} < \frac{a}{2}$，所以如果可以从任何地方吃掉薄片，面包是脆皮的当且仅当 $a \le 2$ 米。题目中的 $2\frac{2}{3}$ 是一个正确的必要条件（necessary condition），但这项规定其实比题目所透露的更为严格。

回顾

- **先化简，后积分。**面包皮的被积函数（integrand）看似复杂，结果却是常数。动手积分之前，务必先把被积函数化简。
- **仔细阅读“任意”二字。**条件必须在吃掉任意片数之后都成立，所以我们需要的是 V/A 的最大值，而不是整个面包的值（即 $\frac{a}{3}$）。
- **简单的计算，出人意表的答案。**“面包皮相等”这个性质听起来难以置信，但计算只有两行。

背后的数学

阿基米德的帽盒定理

阿基米德在《论球与圆柱》（*On the Sphere and Cylinder*）中证明了：球面上夹在两个平行平面之间的环带，其面积等于这两个平面在包裹球的圆柱上所截出的环带的面积（图 4.8.1）。我们的计算正是这条帽盒定理（hat-box theorem）：半径为 a 的圆柱每单位长度的面包皮面积为 $2\pi a$，球也是如此。考虑整个球，它的表面积（surface area）等于圆柱的曲面面积，即 $2\pi a \times 2a = 4\pi a^2$。阿基米德认为，这个结果连同“球的体积是圆柱体积的三分之二”这一事实，是他毕生最出色的成就。西塞罗（Cicero）于公元前 75 年到访叙拉古（Syracuse）时，正是凭着寻找刻在墓上的球与圆柱，找到了那座杂草丛生的坟墓。

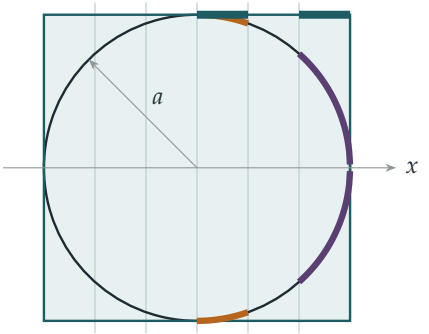


图 4.8.1　半径为 a 的球形面包置于包裹它的圆柱之内的横截面，面包被切成厚度相等的薄片。中间的薄片（琥珀色）的面包皮是一段短而接近水平的弧，离轴较远；末端的薄片（紫红色）的面包皮是一段长而陡峭的弧，离轴较近。绕 x-轴旋转后，两者的面包皮面积都恰好等于圆柱上对应环带（蓝绿色）的面积：$2\pi a$ 乘以厚度。

地球的地图

帽盒定理提供了一种绘制保持面积不变的地球地图的方法。把地球仪上的每一点沿水平方向向外投影到包裹它的圆柱上，然后把圆柱摊开。球面上的环带会映射到圆柱上面积相同的环带，而任何区域也是如此。这就是兰伯特（*Lambert*）圆柱等积投影（Lambert's cylindrical equal-area projection，1772 年）。它在两极附近会令形状严重失真，但面积是如实的，不像人们熟悉的麦卡托（Mercator）投影那样，令格陵兰看起来与非洲一样大（其实格陵兰大约只有非洲的十四分之一）。高斯（Gauss）有一条著名的定理，称为“绝妙定理”（*Theorema Egregium*，1827 年），它说明没有任何球面地图可以同时保持面积和

形状：每一幅平面的地球地图都必定有所失真。

球面上的随机点

以下是一个实际的应用。要在球面上均匀（uniformly）随机地选取一点，你或许会尝试均匀地选取纬度（latitude）和经度（longitude），但这样会令点在两极附近过于密集。帽盒定理给出了正确的做法：在 $-a$ 与 a 之间均匀地选取高度 x，并在 0 与 2π 之间均匀地选取经度。由于面积与高度成正比，这样选出的点恰好是均匀分布的。物理学和电脑图像的电脑模拟都采用这个方法。（值得一提的是，这个方法只在三维空间中有效：对于四维或更高维的球面，相应的命题并不成立。）

思维技巧

寻找隐藏的简洁

当计算得出出乎意料地简单的结果（例如一个常数）时，请停下来问一问为甚么。背后通常有几何上的原因，而找出这个原因，才是真正理解之所在。

在这里，被积函数 $y\sqrt{1+(y')^2}$ 是常数，因为在球面上，陡峭程度与离轴的距离恰好互相补偿：这就是阿基米德的定理。

延伸练习

1 **整个面包。**证明对于整个未切开的面包，$V/A = \frac{a}{3}$。为甚么题目中的条件比 $\frac{a}{3} < 1$ 更难满足？

2 **热带地区。**地球表面有多大比例位于南纬 23.4° 与北纬 23.4° 之间？（把地球视为一个球。由于面积与高度成正比，你只需要用到 $\sin 23.4°$。）

3 **其他形状。**证明对于抛物面（paraboloid）形状的面包（把 $y = \sqrt{x}$ 绕 x-轴旋转），每片的面包皮并非常数。对于哪些曲线 $y(x)$，$y\sqrt{1+(y')^2}$ 是常数？（解出相应的微分方程（differential equation）。）

4 **体积。**利用体积公式，证明球的体积是包裹它的圆柱体积的三分之二，正如阿基米德所发现的那样。

延伸阅读

- **Sherman Stein, *Archimedes: What Did He Do Besides Cry Eureka?***（Mathematical Association of America）。为现代读者讲解阿基米德的数学。
- **Timothy Gowers, *Mathematics: A Very Short Introduction***（Oxford University Press）。书中关于几何的一章介绍了曲面的几何。
- **John Snyder, *Flattening the Earth***（University of Chicago Press）。两千年来的地图投影史。

4.9 Holditch 定理

题目

STEP III 2011，第 5 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/11-s3-q5`

引子

随便取一个卵形，甚么形状都可以，只要它是凸（convex）的，然后让一根长度固定的杆在里面滑动一圈，两端始终留在曲线上。在杆上涂一个点，把杆分成长度为 a 和 b 的两段。杆绕行一周时，这个点会描出一个较小的卵形。两条曲线之间的环形区域，面积是多少？

答案令人惊讶：*无论原来的曲线是甚么形状，面积都是* πab。又长又窄的卵形、蛋形、圆角三角形：环形的面积永远相同，只取决于那个点在杆上的位置。这个结果由剑桥大学冈维尔与凯斯学院（Gonville and Caius College, Cambridge）的院士 Hamnet Holditch 牧师于 1858 年发现，是初等几何中最出人意表的定理之一。本题只用一条面积公式就把它证明出来。

动笔之前

想一想 先在最简单的情况下检验这个定理。若 $\mathcal{C}$ 是半径为 R 的圆，P 描出的是甚么曲线？环形的面积是多少？

由对称性可知，P 描出一个同心的圆，设其半径为 r。杆 AB 是大圆上一条通过 P 的弦，而*相交弦定理*（intersecting chords theorem，又称“圆幂”（power of a point））指出，对任何通过 P 的弦，都有 $PA \times PB = R^2 - r^2$。所以 $ab = R^2 - r^2$，环形的面积是 $\pi R^2 - \pi r^2 = \pi ab$。定理说的是：这个结论对每一条凸曲线都成立，不限于圆。

逐步解题

阶段 1　由坐标求面积

想一想　设 $x = r\cos\theta$ 和 $y = r\sin\theta$，全部都是 t 的函数，计算 $x\dot{y} - y\dot{x}$，其中上方的点表示 $\frac{\mathrm{d}}{\mathrm{d}t}$。

$\dot{x} = \dot{r}\cos\theta - r\sin\theta\,\dot{\theta}$，而 $\dot{y} = \dot{r}\sin\theta + r\cos\theta\,\dot{\theta}$。于是

$$x\dot{y} - y\dot{x} = r\cos\theta\,(\dot{r}\sin\theta + r\cos\theta\,\dot{\theta}) - r\sin\theta\,(\dot{r}\cos\theta - r\sin\theta\,\dot{\theta}) = r^2\dot{\theta}.$$

所以 $\frac{1}{2}\int r^2\,\mathrm{d}\theta = \frac{1}{2}\int r^2\dot{\theta}\,\mathrm{d}t = \frac{1}{2}\int (x\dot{y} - y\dot{x})\,\mathrm{d}t$。 ■

这条公式不需要极坐标（polar coordinates），只需要把笛卡儿坐标（cartesian coordinates）下的位置写成某个参数（parameter）的函数。当一点沿着一条包围 O 的闭曲线逆时针绕行一周时，公式给出的就是曲线内的面积。

阶段 2　杆的两端

杆由 A 指向 B 的方向是 $(\cos t, \sin t)$，而 P 与 A 的距离是 a，与 B 的距离是 b。所以

$$A = (x - a\cos t,\ y - a\sin t), \qquad B = (x + b\cos t,\ y + b\sin t).$$

阶段 3　OA 扫过的面积

想一想　把 $(*)$ 应用于 A。展开 $x_A\dot{y}_A - y_A\dot{x}_A$，并按 a 的幂次把各项整理好。

由 $x_A = x - a\cos t$、$y_A = y - a\sin t$，可得 $\dot{x}_A = \dot{x} + a\sin t$ 和 $\dot{y}_A = \dot{y} - a\cos t$。展开得

$$\begin{aligned} x_A\dot{y}_A - y_A\dot{x}_A &= (x - a\cos t)(\dot{y} - a\cos t) - (y - a\sin t)(\dot{x} + a\sin t) \\ &= (x\dot{y} - y\dot{x}) + a^2(\cos^2 t + \sin^2 t) - a[(x + \dot{y})\cos t + (y - \dot{x})\sin t]. \end{aligned}$$

由 0 积分至 2π 再除以二：第一项给出 $[P]$，第二项给出 $\frac{1}{2}a^2 \cdot 2\pi = \pi a^2$，第三项给出 $-af$。所以

$$[A] = [P] + \pi a^2 - af. \qquad ■$$

停一停，试一试　现在求 $[B]$。（提示：把 A 的公式中的 a 换成甚么，就得到 B?）

$B = (x + b\cos t, y + b\sin t)$ 就是把 A 的公式中的 a 换成 $-b$。所以

$$[B] = [P] + \pi b^2 + bf.$$

阶段 4 消去 f

想一想 A 和 B 都沿同一条曲线 $\mathcal{C}$ 绕行一周。由此可知 $[A]$ 和 $[B]$ 有甚么关系?

当 t 由 0 变到 2π 时，A 和 B 各自沿 $\mathcal{C}$ 逆时针绕行一周，所以 $[A] = [B] =$ 面积$(\mathcal{C})$。令两个表达式相等，

$$\pi a^2 - af = \pi b^2 + bf \quad\Longrightarrow\quad f(a+b) = \pi(a^2 - b^2) \quad\Longrightarrow\quad f = \pi(a-b).$$

由于 P 沿 $\mathcal{D}$ 绕行一周，$[P] =$ 面积$(\mathcal{D})$，于是

$$\text{面积}(\mathcal{C}) - \text{面积}(\mathcal{D}) = [A] - [P] = \pi a^2 - a\pi(a-b) = \pi ab. \qquad \blacksquare$$

那个取决于曲线形状的神秘量 f，自始至终没有计算过：它被消去了。

回顾

- **切合问题的公式。**面积公式 $(*)$ 只需要把坐标写成参数的函数，而杆的运动正是这样描述的。
- **两个表达式，一个未知量。**我们无法计算 f，但手上有两条含有它的方程，于是把它消去，就像解联立方程（simultaneous equations）一样。
- **先检验特殊情况。**动笔之前，圆的情况已经给出答案 πab，让我们知道目标是甚么。

背后的数学

格林定理与求积仪

公式 $(*)$ 是*格林定理*（Green's theorem）的特殊情况；格林定理把沿闭曲线的积分与曲线内区域上的积分联系起来。它的面积形式指出，曲线所包围的面积是 $\frac{1}{2}\oint(x\,\mathrm{d}y - y\,\mathrm{d}x)$，这个积分只需沿边界走一圈就能算出。工程师利用这一点造出了*求积仪*（planimeter），这是十九世纪发明的一种机械仪器（其中以 Jakob Amsler 在 1854 年的设计最为成功）：用指针沿着地图或图表上某个图形的边界描一圈，一个小轮就会记录所包围的面积。求积仪曾用来量度土地面积、根据示功图（indicator diagram）计算蒸汽机所作的功，还有许多其他用途，直至被数码方法取代为止。Holditch 定理与此密切相关：那根杆就像求积仪的臂。

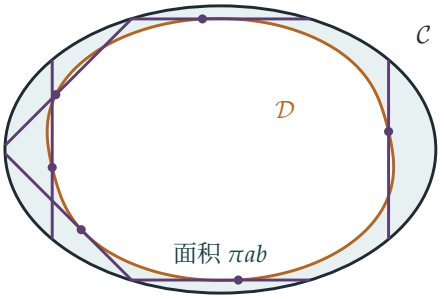


图 4.9.1　Holditch 定理的例子：椭圆的半轴（semi-axes）为 3 和 2，杆长 2.5，其中 $a = 1$、$b = 1.5$。图中画出了杆的若干位置（紫色），并标出点 P。P 所描出的曲线 $\mathcal{D}$（琥珀色）是用数值方法算出的，而两条曲线之间的阴影环形面积为 $4.7124\ldots = 1.5\pi$，与定理的预测完全吻合。

细节条款

Holditch 陈述这个定理时，对条件不大讲究，但定理其实需要一些条件。杆必须能够在两端都留在曲线上的情况下绕足一整圈，而描出的曲线 $\mathcal{D}$ 必须是只描一次的简单闭曲线（simple closed curve）；如果杆相对于 $\mathcal{C}$ 的阔度太长，$\mathcal{D}$ 可能会自相交，这时公式给出的“面积”会把某些区域计算两次，或以负号计算。题目假设 $\mathcal{D}$ 是一条包含 O 的凸曲线，从而避开了这些困难。找出定理究竟在甚么条件下成立，正好说明一个优美的想法必须经过严格化才站得住脚，这也是第 3 章的主题。

思维技巧

算不出来的，就把它消去

如果某个量太复杂而无法计算，就找两条含有它的不同方程，把它消去。

这里 f 取决于曲线的每一个细节，但它同时出现在 $[A]$ 和 $[B]$ 的表达式中，而 $[A] = [B]$ 这个事实把它消除了。答案 πab 之所以能够与曲线无关，全因取决于曲线的部分互相抵消了。

延伸练习

1 **椭圆的面积。**利用 $(*)$，取 $x = p\cos t$、$y = q\sin t$，证明半轴为 p 和 q 的椭圆面积为 πpq。

2 **圆的情况。**证明相交弦定理：若圆的两条弦相交于 P，则每条弦被分成的两段，其长度之积相等。由此推导出圆的 Holditch 定理。

3 **中点。**证明由 $f = \pi(a - b)$ 可知，当 P 是杆的中点时，$f = 0$。这时环形的面积是多少？答案以杆长 L 表示。

4 **三角形。**较难利用 $(*)$，沿边界绕行一周，求顶点为 $(0,0)$、(x_1, y_1)、(x_2, y_2) 的三角形的面积，并证明它等于 $\frac{1}{2}|x_1 y_2 - x_2 y_1|$。这就是求任意多边形面积的“鞋带公式”（shoelace formula）的起点。

延伸阅读

- **Mark Levi, *The Mathematical Mechanic***（Princeton University Press）。以杆、轮和弹簧等物理推理来证明数学结果，与本节的精神一脉相承。
- **H. M. Schey, *Div, Grad, Curl, and All That***（Norton）。向量微积分（vector calculus）的非正式入门，内容包括格林定理。
- **Hamnet Holditch, “Geometrical theorem”**，*Quarterly Journal of Pure and Applied Mathematics* 2 (1858)。原始的短文，只有短短一段，完全没有提及任何细节条款。

我们学到了甚么

本章开始时，我们提出了三个问题：积分（integral）究竟是甚么？积分可以证明甚么？当一个积分没有公式时，又该怎么办？以下是本章九道题目给出的答案。

1. 积分是甚么？哪些函数有积分？

在 Section 4.1 中，我们把积分定义为由窄长矩形组成的下和（lower sum）与上和（upper sum）的共同极限，并单凭这个定义计算出 $\int_0^a x^5\,\mathrm{d}x$。这要花一番工夫：先以裂项相消求和（telescoping sum）处理连续整数乘积之和，再从两边估计上下界，最后用夹逼（squeeze）得出结果。花了多少力气，正好说明微积分基本定理（fundamental theorem of calculus）让你不劳而获的有多少。

这个定义也告诉我们，并非每个函数都可积（integrable）。狄利克雷（Dirichlet）函数在有理数上取值 1，在无理数上取值 0，它的上和全部等于 1，下和全部等于 0。正是这类问题，促使勒贝格（Lebesgue）建立了一套更强大的积分理论，你日后在大学便会学到。

2. 积分可以证明甚么？

远远不止面积。

- **数与数之间的不等式。**一个正函数的积分等于 $\frac{22}{7}-\pi$，所以 $\frac{22}{7}>\pi$（Section 4.3）。“把 $A-B$ 写成某个正函数的积分”这个想法既简单，又非常有用。
- **无穷乘积与特殊值。**利用 $\int \sin^n t\,\mathrm{d}t$ 的归约公式（reduction formula）、一个不变量（invariant）和一次夹逼，便得到沃利斯（Wallis）乘积、斯特灵（Stirling）公式中的常数，以及 $\left(\frac{1}{2}\right)!=\frac{\sqrt{\pi}}{2}$（Section 4.4）。
- **无穷和。**把 x^x 展开成幂级数（power series）再逐项积分，便得到“二年级生之梦”（sophomore's dream）（Section 4.5）。
- **三角学的基础。**定义 $\mathrm{T}(x)=\int_0^x \frac{\mathrm{d}u}{1+u^2}$ 及 $\pi=2\mathrm{T}(\infty)$ 之后，单靠代换（substitution）便推出了加法公式，以及 $\mathrm{T}(\sqrt{3})=\frac{\pi}{3}$ 等特殊值（Section 4.2）。
- **几何。**一个化简后变成常数的被积函数（integrand），揭示了阿基米德（Archimedes）的帽盒定理（hat-box theorem）（Section 4.8）；而把面积公式写成沿边界的积分，则证明了令人惊叹的 Holditch 定理（Section 4.9）。

- **最优函数。**在积分内配方（completing the square），便准确指出 $\int\left((y')^2 - a^2y^2\right)$ 在甚么情况下必定非负，而临界值 $\frac{\pi}{2}$ 原来是一个特征值（eigenvalue）（Section 4.7）。

3. 没有公式时，我们可以怎么办？

在 Section 4.6 中，我们证明了 $\int \frac{e^x}{x+1}\,\mathrm{d}x$ 不能写成有理函数（rational function）乘以 e^x 的形式，而刘维尔（Liouville）的理论更表明它根本没有初等（elementary）公式。从极点（pole）的角度看，原因是其原函数（antiderivative）在 $x = -1$ 附近的表现必须像对数函数，而有理函数做不到这一点。然而这个积分确确实实存在，我们也可以把它算出来：可以用数值方法，可以用级数（如 Section 4.5 中处理 $\int x^x$ 那样），也可以给它起一个名字，把它当作一个新函数来研究，就像我们在 Section 3.9 中对待 W 函数那样。

4. 积分工具箱

方法	你在哪里用过
从定义出发	由矩形求 $\int_0^a x^5$（4.1）
裂项相消求和	连续整数乘积之和（4.1）
保持积分不变的代换	由 $\int \frac{\mathrm{d}u}{1+u^2}$ 建立三角学（4.2）
多项式除法	$\frac{22}{7} - \pi$（4.3）
正性与被积函数的估界	$\pi < \frac{22}{7}$（4.3）
归约公式	$\int \sin^n t$（4.4）；$\int x^m(\ln x)^n$（4.5）
不变量与夹逼	$nu_nu_{n-1} = \frac{\pi}{2}$（4.4）
逐项积分	二年级生之梦（4.5）
因式与次数论证	不存在有理原函数（4.6）
配方法	变分法（calculus of variations）（4.7）
先化简，后积分	球形面包（4.8）
以边界积分表示面积	Holditch 定理（4.9）

下一站

- Section 4.4 中的伽玛函数（Gamma function）和贝塔函数（Beta function），以及斯特灵公式中的常数 $\sqrt{2\pi}$，会在 Section 12.23 的概率问题中再度出现。
- Section 4.7 的特征值问题，到了 Section 8.6 会变成量子谐振子（quantum harmonic oscillator）。
- 梅钦（Machin）公式由 Section 4.2 的加法恒等式建立而成，是 Section 5.1 的主题。
- 第 5 章转向复数（complex numbers），它解释了为何本章有这么多积分都与 π 有关。

如果你想在上大学前走得更远

- **Paul Nahin, *Inside Interesting Integrals***（Springer）：一场巧妙积分的盛宴，并道出背后的种种窍门。
- **William Dunham, *The Calculus Gallery***：从牛顿（Newton）到勒贝格的分析学杰作。
- **Michael Spivak, *Calculus***：严谨地讲述积分的定义和微积分基本定理。
- **Sherman Stein, *Archimedes***：追溯积分的古代源头。

埃尔米特（Hermite）对没有导数的函数避之则吉，但积分从来没有这个问题：它十分宽容，连微分无法处理的函数也照单全收，而且不断带来惊喜。如果说有一个主题贯穿本章，那就是：积分不只是一种求面积的技巧。它是一种语言，用来表达和证明关于数、函数和图形的事实，而在你往后的数学生涯中，你都会一直说这种语言。

延伸练习答案

以下只是简短的答案和提示，并非完整解答：只列出结果和关键步骤，方便你核对自己的做法。请先认真做过每道练习，然后才看答案。

4.1 甚么是积分？

1. 这个和是 $y=\mathrm{f}(x)$ 之下、在 $[1,2]$ 上 n 个阔度为 $\frac{1}{n}$ 的长方形的总面积。取 $\mathrm{f}(x)=\frac{1}{x}$，得 $\frac{1}{n}\sum\frac{1}{1+m/n}=\sum\frac{1}{n+m}\to\ln 2$。第二个和是 $\frac{1}{n}\sum\frac{1}{1+(m/n)^2}\to\int_0^1\frac{\mathrm{d}x}{1+x^2}=\frac{\pi}{4}$。

2. 分成 n 条时，上和（upper sum）是 $\frac{a}{n}\sum_{r=1}^{n}\left(\frac{ra}{n}\right)^2=\frac{a^3(n+1)(2n+1)}{6n^2}$，下和（lower sum）则只须把第一个因子中的 $n+1$ 换成 $n-1$；两者都趋向 $\frac{a^3}{3}$。

3. 各项可以裂项相消（telescope）：$r(r+1)\cdots(r+k)-(r-1)r\cdots(r+k-1)=(k+1)\,r(r+1)\cdots(r+k-1)$。

4. $\sum_{r=1}^{n}r^5=\frac{n^2(n+1)^2(2n^2+2n-1)}{12}$。

4.2 由一个积分建立三角学

1. 导数（derivative）是 $-\mathrm{f}(a-x)\mathrm{f}(x)+\mathrm{f}(a-x)\mathrm{f}(x)=0$，所以 $\mathrm{f}(a-x)\mathrm{f}(x)=\mathrm{f}(a)$；代入 $a=x+y$ 即可。至于余弦，固定 y，设 $p(x)=\mathrm{c}(x+y)-\mathrm{c}(x)\mathrm{c}(y)+\mathrm{s}(x)\mathrm{s}(y)$ 及 $q(x)=\mathrm{s}(x+y)-\mathrm{s}(x)\mathrm{c}(y)-\mathrm{c}(x)\mathrm{s}(y)$。于是 $p'=-q$ 而 $q'=p$，所以 p^2+q^2 是常数，而它在 $x=0$ 时等于 0。

2. $\frac{\mathrm{d}y}{\mathrm{d}u}=(1+u^2)^{-3/2}$ 而 $\sqrt{1-y^2}=(1+u^2)^{-1/2}$，所以 $\frac{\mathrm{d}y}{\sqrt{1-y^2}}=\frac{\mathrm{d}u}{1+u^2}$。于是 $\mathrm{S}(\frac{1}{\sqrt{2}})=\mathrm{T}(1)$，而在第 (i) 部分取 $x=1$ 便得 $\mathrm{T}(1)=\frac{1}{2}\mathrm{T}_\infty$。

3. 取 $a=\frac{1}{2}$ 及 $x=\frac{1}{3}$，则 $\frac{x+a}{1-ax}=1$，所以 $\mathrm{T}(\frac{1}{3})=\mathrm{T}(1)-\mathrm{T}(\frac{1}{2})$。

4. 在 $[0,1]$ 上 $\frac{1}{1+u^2}\le 1$，在 $[1,x]$ 上则 $\le\frac{1}{u^2}$，所以当 $x\ge 1$ 时，$\mathrm{T}(x)\le 1+(1-\frac{1}{x})<2$；当 $0<x<1$ 时，$\mathrm{T}(x)<\mathrm{T}(1)\le 1$。（上界 $1+\int_1^x u^{-2}\,\mathrm{d}u$ 在 $x<1$ 时并不成立：在 $x=\frac{1}{2}$ 时它等于 0。）T 递增而且有界（bounded），所以它有极限（limit）。

4.3 $22/7$ 大于 π

1. $\frac{22}{7}-\pi\approx 0.001264$，介乎 $\frac{1}{1260}\approx 0.000794$ 与 $\frac{1}{630}\approx 0.001587$ 之间。

2. $\frac{(4!)^2}{9!} = \frac{576}{362\,880} = \frac{1}{630}$。

3. $\frac{x(1-x)^2}{1+x^2} = x - 2 + \frac{2}{1+x^2}$，积分（integrate）后得 $\frac{1}{2} - 2 + \frac{\pi}{2}$。同样地，$\frac{x^2(1-x)^4}{1+x^2} = x^4 - 4x^3 + 5x^2 - 4 + \frac{4}{1+x^2}$，积分后得 $\frac{1}{5} - 1 + \frac{5}{3} - 4 + \pi = \pi - \frac{47}{15}$。两个积分都是正数。这两个被积函数（integrand）远大于 $\frac{x^4(1-x)^4}{1+x^2}$（其最大值约为 0.003），所以误差 $\frac{\pi}{2} - \frac{3}{2} \approx 0.07$ 和 $\pi - \frac{47}{15} \approx 0.008$ 也较大。

4. 多项式除法（polynomial division）余下 $\frac{16}{1+x^2}$，积分后得 4π；整个积分是 $4\pi - \frac{188684}{15015} \approx 3.6 \times 10^{-6}$，所以 $\pi \approx 3.14159174$。

4.4 沃利斯、Beta 与二分之一的阶乘

1. $u_{2m} = \frac{1\cdot3\cdots(2m-1)}{2\cdot4\cdots2m} \cdot \frac{\pi}{2}$ 而 $u_{2m+1} = \frac{2\cdot4\cdots2m}{3\cdot5\cdots(2m+1)}$。于是 $1 < \frac{u_{2m}}{u_{2m+1}} < \frac{u_{2m-1}}{u_{2m+1}} = \frac{2m+1}{2m}$，所以这个比趋向 1，整理后即得 $\frac{\pi}{2} = \prod_{k\geq1} \frac{(2k)^2}{(2k-1)(2k+1)}$。

2. 分部积分法（integration by parts）给出 $I_n = \frac{n}{n+1} \int_0^1 x^{n+1}(1-x)^{n-1}\,\mathrm{d}x$，再写成 $x^{n+1}(1-x)^{n-1} = x^n(1-x)^{n-1} - x^n(1-x)^n$ 并利用对称性，便得 $I_n = \frac{n}{2(2n+1)} I_{n-1}$；用归纳法（induction）可得 $\frac{(n!)^2}{(2n+1)!}$。取 $x = \sin^2\theta$，得 $I_{1/2} = \int_0^{\pi/2} 2\sin^2\theta\cos^2\theta\,\mathrm{d}\theta = \frac{\pi}{8}$。公式提示 $\left(\left(\frac{1}{2}\right)!\right)^2/2! = \frac{\pi}{8}$，所以 $\left(\frac{1}{2}\right)! = \frac{\sqrt{\pi}}{2}$（即 $\Gamma\left(\frac{3}{2}\right)$）。

3. 对称性来自 $x \mapsto 1-x$；递推关系（recurrence）来自 $x^p = x^{p-1} - x^{p-1}(1-x)$；而取 $x = \frac{t}{1+t}$ 时，$1 - x = \frac{1}{1+t}$ 且 $\mathrm{d}x = \frac{\mathrm{d}t}{(1+t)^2}$。该积分是 $\mathrm{b}\left(\frac{5}{2}, \frac{7}{2}\right)$；利用递推关系和对称性化简至 $\mathrm{b}\left(\frac{1}{2}, \frac{1}{2}\right) = \pi$，便得 $\frac{3\pi}{256}$。

4. 由第 (i) 部分的显式数值，$u_{2m} = \frac{(2m)!}{(2^m m!)^2} \cdot \frac{\pi}{2} = \binom{2m}{m} \frac{\pi}{2^{2m+1}}$。于是由 $2mu_{2m}^2 \to \frac{\pi}{2}$ 得 $\binom{2m}{m}^2 \approx \frac{4^{2m}}{\pi m}$。当 $m = 10$ 时：$\binom{20}{10} = 184\,756$，而 $\frac{4^{10}}{\sqrt{10\pi}} \approx 187\,079$（高出约 1.3%）。

4.5 大二学生之梦

1. $x^{-x} = e^{-x\ln x} = \sum_{n\geq0} \frac{(-x\ln x)^n}{n!}$，而 $\int_0^1 \frac{x^n(-\ln x)^n}{n!}\,\mathrm{d}x = \frac{1}{(n+1)^{n+1}}$（代入 $x = e^{-u}$）。所以该积分是 $\sum_{n\geq1} n^{-n} \approx 1.2913$。

2. $\frac{\mathrm{d}}{\mathrm{d}x}x^x = x^x(\ln x + 1)$，在 $x = \frac{1}{e}$ 时为零；其值为 $(e^{-1})^{e^{-1}} = e^{-1/e} \approx 0.692$。

3. 各项是 $\frac{1}{n^n}$。取 7 项后停止，误差小于 $\frac{1}{8^8} \approx 6 \times 10^{-8}$，已经足够；只取 6 项的话，上界 $\frac{1}{7^7} \approx 1.2 \times 10^{-6}$ 并不足够。所以要取 7 项。

4. 用分部积分法：$\int_0^1 x^2 \ln x\,\mathrm{d}x = \left[\frac{x^3\ln x}{3}\right]_0^1 - \int_0^1 \frac{x^2}{3}\,\mathrm{d}x = -\frac{1}{9}$。一般而言 $\int_0^1 x^n \ln x\,\mathrm{d}x = -\frac{1}{(n+1)^2}$，而 $\frac{\ln x}{1-x} = \sum_{n\geq0} x^n \ln x$。

4.6 没有公式的积分

1. 求导：$\frac{\mathrm{d}}{\mathrm{d}x}\frac{e^x}{x+1}=\frac{xe^x}{(x+1)^2}$，而 $\frac{\mathrm{d}}{\mathrm{d}x}\frac{(x-1)e^x}{x+1}=e^x\left(\frac{x-1}{x+1}+\frac{2}{(x+1)^2}\right)=\frac{(x^2+1)e^x}{(x+1)^2}$。

2. $\frac{\mathrm{d}}{\mathrm{d}x}\left(\mathrm{P}e^{-x^2}\right)=(\mathrm{P}'-2x\mathrm{P})e^{-x^2}$，而 $\mathrm{P}'-2x\mathrm{P}$ 的次数是 $\deg\mathrm{P}+1\geq 1$（当 $\mathrm{P}\neq 0$），所以它不可能等于 1。

3. 假如 $\mathrm{R}=\mathrm{P}/\mathrm{Q}$（已约至最简）在 0 处有 $k\geq 1$ 阶的极点（pole），则 $\mathrm{R}'+\mathrm{R}$ 在该处会有 $k+1\geq 2$ 阶的极点，但 $\frac{1}{x}$ 的极点只是 1 阶。假如 R 在 0 处没有极点，则 $\mathrm{R}'+\mathrm{R}$ 也没有。无论哪种情况，都有 $(\mathrm{R}e^x)'\neq\frac{e^x}{x}$。

4. 对 $\frac{\mathrm{P}e^x}{x+1}$ 求导，得 $e^x\frac{(x+1)\mathrm{P}'+x\mathrm{P}}{(x+1)^2}$。在 $x=-1$ 处：$\mathrm{f}(-1)=-\mathrm{P}(-1)$，而由 $\mathrm{f}'=(x+1)\mathrm{P}''+(x+1)\mathrm{P}'+\mathrm{P}$ 得 $\mathrm{f}'(-1)=\mathrm{P}(-1)$。对于 x^3-2：$-3+3=0$；对于 x：$-1+1=0$；对于 x^2+1：$2-2=0$。这个条件也是充分的：映射 $\mathrm{P}\mapsto(x+1)\mathrm{P}'+x\mathrm{P}$ 把次数小于 d 的多项式映到次数不超过 d 的多项式，而且是单射（injective），所以它的像（image）的维数（dimension）是 d，并且包含在由满足 $\mathrm{f}(-1)+\mathrm{f}'(-1)=0$ 的 f 所组成的 d 维空间之内；因此像就是整个空间。

4.7 变分法

1. 取 $y=\cos\frac{\pi x}{2}$，得 $J=\frac{\pi^2}{4}\int_0^1\sin^2-a^2\int_0^1\cos^2=\frac{1}{2}\left(\frac{\pi^2}{4}-a^2\right)<0$。

2. $J(y+\varepsilon\eta)=J(y)+2\varepsilon\int_0^1(y'\eta'-a^2y\eta)+\varepsilon^2J(\eta)$；要在 $\varepsilon=0$ 处取得最小值，线性项必须为零。分部积分后，对每个满足 $\eta(1)=0$ 的 η 都有 $-y'(0)\eta(0)-\int_0^1(y''+a^2y)\eta=0$，由此必有 $y''+a^2y=0$ 及 $y'(0)=0$。

3. 取 $x=Lu$，得 $\int_0^L\left((y')^2-a^2y^2\right)\mathrm{d}x=\frac{1}{L}\int_0^1\left((Y')^2-(aL)^2Y^2\right)\mathrm{d}u$，所以条件是 $aL\leq\frac{\pi}{2}$。（当 $aL=\frac{\pi}{2}$ 时，正如正文所述，积分仍然不会是负数，但 $y=\cos\frac{\pi x}{2L}$ 会令它等于零；当 $aL>\frac{\pi}{2}$ 时，同一个函数会令它变成负数。）

4. 摆线（cycloid）是 $x=R(\theta-\sin\theta),y=R(1-\cos\theta)$。关于等时曲线（tautochrone），见 Section 11.4：以弧长（arc length）表示时，该运动恰好是简谐运动（simple harmonic motion）。

4.8 切球形面包

1. $V/A=\frac{\frac{4}{3}\pi a^3}{4\pi a^2}=\frac{a}{3}$。但余下一块高度为 t 的部分有 $A=2\pi at$ 及 $V=\frac{1}{3}\pi t^2(3a-t)$，所以 $V/A=\frac{t(3a-t)}{6a}$，它在 $t=\frac{3}{2}a$ 时取得最大值 $\frac{3a}{8}$：比 $\frac{a}{3}$ 还要大。

2. 这条带的高度是 $2R\sin 23.4^\circ$，而球的高度是 $2R$，所以所占比例是 $\sin 23.4^\circ\approx 0.40$。

3. 对于 $y=\sqrt{x}$，$y\sqrt{1+(y')^2}=\sqrt{x+\frac{1}{4}}$，并非常数。由 $y\sqrt{1+(y')^2}=c$ 得 $(y')^2=\frac{c^2}{y^2}-1$，其解为 $y=c$（圆柱）以及圆 $(x-x_0)^2+y^2=c^2$（球）。

4. 球的体积是 $\frac{4}{3}\pi a^3$，圆柱的体积是 $\pi a^2 \cdot 2a = 2\pi a^3$；两者之比是 $\frac{2}{3}$。

4.9 Holditch 定理

1. $\frac{1}{2}\int_0^{2\pi}(p\cos t \cdot q\cos t + q\sin t \cdot p\sin t)\,\mathrm{d}t = \frac{1}{2}\cdot 2\pi pq = \pi pq$。
2. 把两条弦（chord）的端点连起来所得的两个三角形相似（同弧所对的角相等），所以 $PA \cdot PB = PC \cdot PD$。对于半径为 R 的圆，若 P 与圆心的距离为 r，则通过 P 的直径给出 $(R-r)(R+r)$，所以 $ab = R^2 - r^2$，而圆环的面积是 $\pi R^2 - \pi r^2 = \pi ab$。
3. 若 $a = b$，则 $f = \pi(a-b) = 0$，而圆环的面积是 $\pi a^2 = \frac{\pi L^2}{4}$：即以该杆为直径的圆的面积。
4. 沿着通过 O 的两条边，$x\,\mathrm{d}y - y\,\mathrm{d}x = 0$。第三条边以 $(x_1, y_1) + s(x_2 - x_1, y_2 - y_1)$（其中 $0 \le s \le 1$）作参数表示，在这条边上它等于 $(x_1 y_2 - x_2 y_1)\,\mathrm{d}s$。所以面积是 $\frac{1}{2}|x_1 y_2 - x_2 y_1|$。

章

5

复数

实数域中两个眞理之间的最短路径，要取道复数域。

本章内容

5.1	梅钦公式	STEP III 1987
5.2	单位根筛选法	STEP I 1990
5.3	因式分解 $z^{2n}+1$	STEP III 2013
5.4	正多边形中的距离	STEP III 2016
5.5	名不副实的有理角	STEP III 2022
5.6	费马点	STEP III 2000
5.7	作为映射的复指数函数	STEP II 1997
5.8	圆盘中隐藏的距离	STEP III 2013
A	练习答案	第 367 页

虚数，实在的答案

实数域中两个眞理之间的最短路径，要取道复数域。

— 阿达马（Jacques Hadamard）

1545 年，卡尔达诺（Gerolamo Cardano）发表了一条解三次方程（cubic equation）的公式，却发现了一件令人不安的事。有些三次方程明明有三个不折不扣的实根，公式在计算途中却要求取负数的平方根。这些平方根到最后会互相抵消，留下实数答案，但在计算的中途，却出现了一些按当时任何标准都不存在的量。卡尔达诺形容这种运算“既精妙，又无用”。其后两个世纪，数学家用起这些“虚”数（imaginary number）时总是惴惴不安，只当它是一种碰巧行得通的窍门。

要到十九世纪初，Caspar Wessel、Jean-Robert Argand 和高斯（Gauss）把复数（complex number）看成平面上的点，复数才得到认可。自此之后，人们发现复数并非窍门，而是一种启示。复数解释了实数世界的现象为何如此表现，而且不断为有关实数的结果提供最短的路径，正如阿达马所言。

复数为何重要

每一条多项式方程在复数中都有解：这就是*代数基本定理*（fundamental theorem of algebra）。因此，在一个确切的意义下，代数到了复数才算完备。复数也是描述旋转和振荡的天然语言。乘以一个模（modulus）为 1 的复数，就是把平面旋转，所以凡是涉及角度、波动或周期现象的问题，用复数处理都会较为简单。电机工程师用复数分析交流电；物理学家没有复数就写不出量子力学；由手提电话到磁力共振（MRI）扫描器，信号处理全靠复数傅里叶变换（Fourier transform）运作。而在纯数学中，复变函数原来有着惊人的刚性和美感：只要知道一个复函数在一小片区域上的取值，它在各处的取值就全部确定了。

你在大学会遇到甚么

你从第一个星期起就会用到复数，但真正的一大步是*复分析*（complex analysis），通常在二年级修读。它的核心对象是*全纯*（holomorphic）函数，即可以求导的复函数，而这套理论处处是惊喜。每个全纯函数都自动地无限次可导，并且等于它的泰勒级数（Taylor series）。沿闭曲线的积分，只需察看函数趋向无穷的那些点便可求出（柯西留数定理，Cauchy's residue theorem），让你能够计算其他方法都束手无策的实积分。这套理论再

往前走，就通往保角映射（conformal map）（Section 5.7）、黎曼曲面（Riemann surface），以及支配质数的黎曼 zeta 函数（Riemann zeta function）（Section 1.2）。与此同时，单位根（roots of unity）的代数（Section 5.2、Section 5.3）通往伽罗瓦理论（Galois theory），而复几何则通往 Section 5.8 的非欧几何（non-Euclidean geometry）。

贯穿本章的问题

三个大问题

1. **复数怎样帮助解决实数问题？** 为甚么绕道复平面（complex plane），竟能得出关于 π、二项式系数（binomial coefficient）或余弦的事实？
2. **单位根有甚么特别之处？** 为甚么 $z^n = 1$ 的解不断在求和、乘积、多边形和角度之中出现？
3. **复函数是甚么模样的？** 输入和输出都是平面上的点的函数，我们怎样把它画出来？它们又揭示了怎样的几何？

本章由辐角（argument）与乘法出发（梅钦公式），经过单位根的代数（筛选、因式分解、多边形的距离和无理角），走到用复数处理的几何（费马点），最后到达作为映射的复函数（指数函数，以及单位圆盘中一个隐藏的距离）。

本章的主要想法

§	课题	想法
5.1	梅钦（Machin）公式	复数相乘时，辐角相加
5.2	每隔三项的二项式系数	对单位根取平均，可按余数（residue）筛选
5.3	余弦的乘积	按各根分解因式，然后代入
5.4	正多边形的弦	距离就是模；$z^n - 1 = \prod(z - \omega^k)$
5.5	$\cos^{-1}\frac{1}{3}$ 是无理的	棣美弗（De Moivre）定理、奇偶性（parity）与最小反例（minimal counterexample）
5.6	费马点（Fermat point）	旋转就是乘法；选取好的坐标
5.7	把 e^w 看成映射	把复函数画出来；多对一映射（many-to-one map）
5.8	一个隐藏的距离	把不等式看成双曲几何（hyperbolic geometry）中的三角不等式（triangle inequality）

怎样阅读本章

动手画图。本章几乎每一个结果都有几何意义，随手画一幅复平面的草图（单位根在哪里、某个乘法把平面转向哪一边、网格线落在何处），往往在动用任何代数之前，就已经

告诉你该怎样做。然后要小心地做代数运算，因为在正负号和 2π 的倍数上，图形可能会误导你。

5.1 梅钦公式

题目

STEP III 1987，第 3 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/87-s3-q3`

引子

1706 年，伦敦数学家梅钦（John Machin）徒手把 π 计算到小数点后 100 位，在此之前从未有人接近过这个纪录。他的秘诀就是第 (ii) 部分的公式：这条公式把 $\frac{\pi}{4}$ 用几个角表示出来，而这些角的正切都是很小的分数。分数愈小，级数（series）$\tan^{-1} x = x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots$ 收敛（converge）得愈快。梅钦公式及其各种变体，在其后两个半世纪一直是计算 π 的标准方法，直到电脑和 Section 2.6 的方法出现才被取代。

这条公式看起来像是三角恒等式（identity）变出来的奇迹。这道题目却说明，它其实是关于复数（complex number）乘法的一个事实：复数相乘时，它们的辐角（argument）相加。只要把几个实部、虚部都是整数的复数乘一次，整条公式就得证。

动笔之前

? **想一想** $5 - \mathrm{i}$ 的辐角是甚么？$1 + \mathrm{i}$ 的呢？复数相乘时，辐角会怎样变化？

$\arg(5 - \mathrm{i}) = -\tan^{-1}\frac{1}{5}$，是一个很小的负角，而 $\arg(1 + \mathrm{i}) = \frac{\pi}{4}$。复数相乘时，模（modulus）相乘，辐角相加（可相差 2π 的倍数）。因此

$$\arg((5 - \mathrm{i})^4(1 + \mathrm{i})) = \frac{\pi}{4} - 4\tan^{-1}\frac{1}{5} \quad \text{（可能再加上若干个}2\pi\text{）}.$$

如果我们能算出这个乘积，再用另一个方法求出它的辐角，就得到一条恒等式。

逐步解题

阶段 1　模的上下界

想一想　求上界（upper bound）时，把两边平方。求下界（lower bound）时，把 $|x|\cos\alpha + |y|\sin\alpha$ 看成两个向量的点积（dot product）。

上界：$(|x|+|y|)^2 = x^2+y^2+2|x||y| \geq x^2+y^2 = |z|^2$，而且两边都非负，所以 $|z| \leq |x|+|y|$。

下界：$|x|\cos\alpha + |y|\sin\alpha$ 是长度为 $|z|$ 的向量 $(|x|,|y|)$ 与单位向量（unit vector）$(\cos\alpha,\sin\alpha)$ 的点积。点积不会超过两个向量长度之积，所以它不超过 $|z|$。（代数证法：$|z|^2 - (|x|\cos\alpha+|y|\sin\alpha)^2 = (|x|\sin\alpha - |y|\cos\alpha)^2 \geq 0$。） ■

阶段 2　关键的乘法

想一想　先计算 $(5-\mathrm{i})^2$，再把结果平方，然后乘以 $1+\mathrm{i}$。

$(5-\mathrm{i})^2 = 25-10\mathrm{i}-1 = 24-10\mathrm{i}$。于是 $(24-10\mathrm{i})^2 = 576-480\mathrm{i}-100 = 476-480\mathrm{i}$。最后

$$(476-480\mathrm{i})(1+\mathrm{i}) = 476+476\mathrm{i}-480\mathrm{i}+480 = 956-4\mathrm{i}.$$

停一停，试一试　现在比较辐角。$956-4\mathrm{i}$ 的辐角是甚么？你凭甚么肯定没有多出 2π 的倍数？

$956-4\mathrm{i} = 4(239-\mathrm{i})$，其辐角为 $-\tan^{-1}\frac{1}{239}$。因此

$$\frac{\pi}{4} - 4\tan^{-1}\frac{1}{5} = -\tan^{-1}\frac{1}{239} + 2k\pi$$

其中 k 是某个整数。但 $0 < \tan^{-1}\frac{1}{5} < \frac{1}{5}$，所以左边介乎 $\frac{\pi}{4}-\frac{4}{5}$ 与 $\frac{\pi}{4}$ 之间，即大约介乎 -0.02 与 0.79 之间；而右边除去 $2k\pi$ 后介乎 -0.005 与 0 之间。由于 2π 远大于两者的差距，故 $k=0$。移项后得

$$\frac{\pi}{4} = 4\tan^{-1}\frac{1}{5} - \tan^{-1}\frac{1}{239}. \qquad ■$$

图 5.1.1 显示了整个过程：由 45° 出发，每乘一次 $5-\mathrm{i}$ 就顺时针旋转约 11.3°，乘了四次之后，刚好越过实轴，只过了 0.24°。

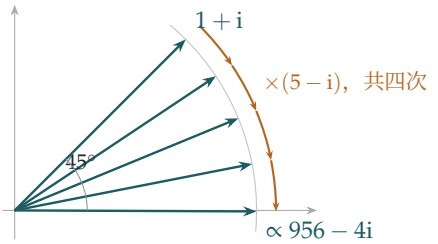


图 5.1.1 把 $1+\mathrm{i}$ 乘以 $5-\mathrm{i}$ 四次。每乘一次就顺时针旋转 $\tan^{-1}\frac{1}{5}\approx 11.31°$（同时也会拉长，这里略去不计）。四次旋转把 45° 带到 $-0.24°$，即 $956-4\mathrm{i}$ 的辐角，也就是 $-\tan^{-1}\frac{1}{239}$。

阶段 3 第二条公式

想一想 依样画葫芦，这次应该计算哪一个乘积?

考虑 $(4+\mathrm{i})^3(20+\mathrm{i})(1985+\mathrm{i})$，它的辐角是 $3\tan^{-1}\frac{1}{4}+\tan^{-1}\frac{1}{20}+\tan^{-1}\frac{1}{1985}$。现在 $(4+\mathrm{i})^2=15+8\mathrm{i}$，所以 $(4+\mathrm{i})^3=(15+8\mathrm{i})(4+\mathrm{i})=52+47\mathrm{i}$。于是 $(52+47\mathrm{i})(20+\mathrm{i})=1040+52\mathrm{i}+940\mathrm{i}-47=993+992\mathrm{i}$，而

$$(993+992\mathrm{i})(1985+\mathrm{i})=(1971105-992)+(993+1969120)\mathrm{i}=1970113+1970113\mathrm{i}.$$

实部（real part）等于虚部（imaginary part），所以辐角是 $\frac{\pi}{4}$。三个角都是很小的正角（它们的和小于 $3\cdot\frac{1}{4}+\frac{1}{20}+\frac{1}{1985}<1<2\pi$），所以毋须担心 2π 的倍数，公式由此得证。■

回顾

- **相乘时辐角相加。**单凭这一个事实，高斯整数（Gaussian integers）的乘积就变成了角与角之间的恒等式。
- **留意 2π 的倍数。**辐角只能确定到相差 2π 的倍数，所以每条恒等式都要靠估计大小来确定那个倍数。
- **整数包办一切。**所有运算都只涉及整数；关于 π 的恒等式是精确地得出的，没有用到任何近似。

背后的数学

徒手计算 π

为甚么梅钦公式这样适合用来计算？级数 $\tan^{-1} x = x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots$ 在 $|x| \le 1$ 时收敛，但在 $x = 1$ 时慢得令人难受：要用 $1 - \frac{1}{3} + \frac{1}{5} - \cdots$ 求得 $\frac{\pi}{4}$ 的三位小数，大约需要一千项。取 $x = \frac{1}{5}$ 时，每一项约为前一项的 25 分之一，即每项大约带来 1.4 位小数；取 $x = \frac{1}{239}$ 时，每项约 4.8 位。梅钦的 100 位小数，只需要第一个级数约 70 项和第二个级数约 20 项。1873 年，William Shanks 用一条梅钦型公式算出 707 位，单是运算就花了多年时间；到了 1944 年，人们发现他在第 528 位出了错，其后的数字全部不对。

再谈高斯整数

形如 $a + b\mathrm{i}$ 而 a、b 为整数的数，就是 Section 1.6 的高斯整数。$5 - \mathrm{i}$ 之所以这样好用，原因在于一个因数分解（factorisation）：$239^2 + 1 = 57122 = 2 \times 13^4$，在高斯整数中写成 $(239 + \mathrm{i})(239 - \mathrm{i}) = (1 + \mathrm{i})(1 - \mathrm{i})\,(3 + 2\mathrm{i})^4(3 - 2\mathrm{i})^4$，而 $5 - \mathrm{i} = (1 - \mathrm{i})(3 + 2\mathrm{i})$ 正是由同一批高斯质数（Gaussian primes）组成的。寻找梅钦型公式，其实就是把形如 $n^2 + 1$ 的数作因数分解的问题。1899 年，Carl Størmer 正是用这个想法证明了：形如 $\frac{\pi}{4} = a\tan^{-1}\frac{1}{m} + b\tan^{-1}\frac{1}{n}$ 而 a、b、m、n 为整数的公式只有四条，即梅钦的一条，以及欧拉（Euler）、Hermann 和 Hutton 分别发现的三条。

旋转

乘以一个模为 1 的复数，就是把平面旋转（rotation）。正因如此，凡是涉及旋转和角的事情，复数都是最自然的语言：由本章稍后的几何，到电机工程中的交流电，以至量子力学中的波函数，莫不如此。在 Section 4.2 中，$\tan^{-1}$ 的加法公式（addition formula）来自一个巧妙的代换（substitution）；在这里，它来自复数的乘法。两者其实是同一个事实，只是从不同的角度去看。

思维技巧

把角转化为复数乘法

角的和之间的恒等式，就是乘积的辐角之间的恒等式。把乘积精确地算出来，再读出它的辐角。

复数的运算机械而精确，而处理反三角函数（inverse trigonometric functions）则容易出错。把后者转化为前者，往往是最快捷的证明途径。

延伸练习

1 **一个近亲。**试做 STEP II 2000，第 4 题（`step.maths.org/questions/00-s2-q4`）。这道题用与本节相同的方法，证明另一条关于 $\frac{\pi}{4}$ 的反正切公式：把复数相乘，然后读出辐角。

2 **欧拉公式。**考虑 $(2+\mathrm{i})(3+\mathrm{i})$，证明 $\tan^{-1}\frac{1}{2}+\tan^{-1}\frac{1}{3}=\frac{\pi}{4}$。

3 **Hutton 公式。**证明 $\frac{\pi}{4}=2\tan^{-1}\frac{1}{3}+\tan^{-1}\frac{1}{7}$。

4 **要多少项？**估计一下，要把 π 计算到小数点后 20 位，梅钦公式中的两个级数各需要多少项。然后用计算机实际算出最初几位。

延伸阅读

- **Jonathan Borwein & Peter Borwein, *Pi and the AGM***（Wiley）。书中载有梅钦型公式的历史。
- **Jörg Arndt & Christoph Haenel, *Pi Unleashed***（Springer）。生动地介绍古今各种计算 π 的方法。
- **Tristan Needham, *Visual Complex Analysis***（Oxford University Press）。要从几何角度理解复数，这是最好的一本书。第 1 章讲的就是乘法即旋转加缩放。

5.2 单位根筛选法

题目

STEP I 1990，第 2 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/90-s1-q2`

引子

帕斯卡三角形（Pascal's triangle）同一行的二项式系数（binomial coefficients）加起来等于 2^n：只要在 $(1+x)^n$ 中代入 $x=1$ 即可。若改为代入 $x=-1$，便得到交错和，其值为 0；把两式相加，可知位于偶数位置的系数之和为 2^{n-1}，刚好是一半。所以偶数位置和奇数位置的项把总和平分。

那么每隔两项取一个、即每第三个系数又如何？没有任何实数 x 能够挑出位置 $0, 3, 6, \ldots$。但有一个复数可以。三次单位根（cube roots of unity）就像一个滤波器：对 1、ω 和 ω^2 取平均，幂 ω^k 在 k 是 3 的倍数时给出 1，否则给出 0。这个想法称为*单位根滤波器*（roots-of-unity filter），是*离散傅里叶变换*（discrete Fourier transform）的小型版本，而后者是有史以来最重要的算法之一。

动笔之前

想一想 ω 是 $z^3-1=0$ 的一个根。试把 z^3-1 因式分解。为甚么由此可得 $1+\omega+\omega^2=0$？

$z^3-1=(z-1)(z^2+z+1)$。由于 $\omega^3=1$ 但 $\omega\neq 1$，ω 必定是第二个因式的根：$\omega^2+\omega+1=0$。从几何上看，1、ω 和 ω^2 是一个以原点为中心的等边三角形的顶点，它们的和就是该三角形的质心，即 0。

逐步解题

阶段 1　模与辐角

由 $1+\omega+\omega^2=0$，得 $1+\omega^2=-\omega$。因为 $\omega=e^{2\pi\mathrm{i}/3}$，所以 $-\omega=e^{\mathrm{i}\pi}e^{2\pi\mathrm{i}/3}=e^{5\pi\mathrm{i}/3}=e^{-\pi\mathrm{i}/3}$。故 $1+\omega^2$ 的模（modulus）为 1，辐角（argument）为 $-\frac{\pi}{3}$。同理，$1+\omega=-\omega^2=e^{\pi\mathrm{i}/3}$，其模为 1，辐角为 $\frac{\pi}{3}$。

阶段 2　滤波器

> **想一想**　当 k 是 3 的倍数时，$1+\omega^k+\omega^{2k}$ 等于甚么？不是倍数时又如何？

若 $3\mid k$，则 $\omega^k=1$，和为 3。若否，ω^k 是 ω 或 ω^2，而无论哪一种情况，都有 $1+\omega^k+\omega^{2k}=1+\omega+\omega^2=0$。因此，表达式 $\frac{1}{3}(1+\omega^k+\omega^{2k})$ 在 $3\mid k$ 时等于 1，否则等于 0：它是一个侦测 3 的倍数的探测器。

> **停一停，试一试**　把 $(1+\omega^r)^n$ 对 $r=1,2,3$ 相加，每一项都用二项式定理（binomial theorem）展开。你得到甚么？

由于 $\omega^3=1$，

$$(1+\omega)^n+(1+\omega^2)^n+(1+1)^n=\sum_{k=0}^{n}\binom{n}{k}(\omega^k+\omega^{2k}+1)=3\sum_{3\mid k}\binom{n}{k}.$$

另一方面，由阶段 1，

$$(1+\omega)^n+(1+\omega^2)^n+2^n=e^{n\pi\mathrm{i}/3}+e^{-n\pi\mathrm{i}/3}+2^n=2\cos\frac{n\pi}{3}+2^n.$$

两边除以 3 便得到所求公式。 ■

阶段 3　两个计算

对于 $n=25$：$\frac{25\pi}{3}=8\pi+\frac{\pi}{3}$，所以 $\cos\frac{25\pi}{3}=\frac{1}{2}$，于是

$$\binom{25}{0}+\binom{25}{3}+\cdots+\binom{25}{24}=\frac{33554432+1}{3}=11184811.$$

> **想一想**　第二个和需要的是满足 $k\equiv 2\pmod 3$ 的位置。怎样修改滤波器，使它侦测的是 $k\equiv 2$ 而不是 $k\equiv 0$？

把探测器平移：$\frac{1}{3}(1+\omega^{k-2}+\omega^{2(k-2)})$ 恰好在 $k\equiv 2\pmod 3$ 时等于 1。由于 $\omega^{-2}=\omega$

及 $\omega^{-4}=\omega^2$，由此可得

$$\sum_{k\equiv 2}\binom{n}{k}=\frac{1}{3}\Big(2^n+\omega(1+\omega)^n+\omega^2(1+\omega^2)^n\Big).$$

取 $n=24$，则 $(1+\omega)^{24}=e^{8\pi\mathrm{i}}=1$，同样 $(1+\omega^2)^{24}=1$，所以括号内的值为 $2^{24}+\omega+\omega^2=2^{24}-1$。由 $2^{24}=16777216$，

$$\binom{24}{2}+\binom{24}{5}+\cdots+\binom{24}{23}=\frac{16777215}{3}=5592405.$$ ■

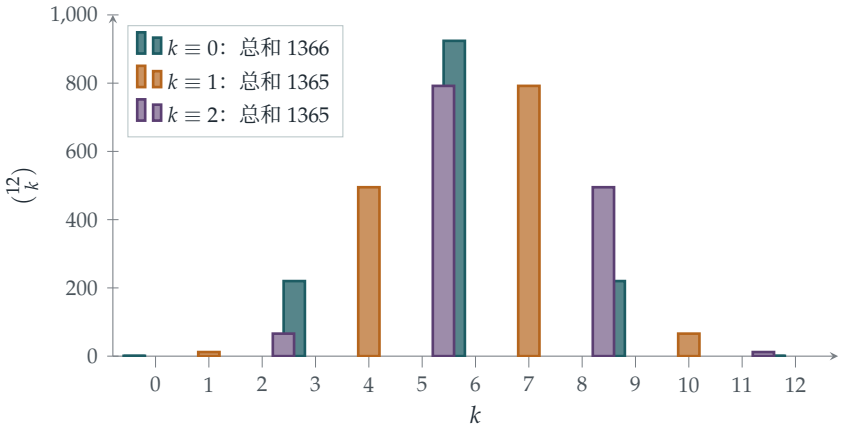


图 5.2.1 帕斯卡三角形的第 12 行，按 k 除以 3 的余数着色。三个类别的总和分别是 1366、1365 和 1365：各自几乎刚好是 $2^{12}=4096$ 的三分之一。单位根滤波器准确地预测了它们与 $\frac{4096}{3}$ 之间的微小差异：$+\frac{2}{3}$、$-\frac{1}{3}$ 和 $-\frac{1}{3}$。

回顾

- **对单位根取平均，可按余数筛选。**$\frac{1}{3}\sum_r \omega^{rk}$ 等于 1 还是 0，视乎 3 是否整除 k。把多项式在各单位根处的值相加，便能挑出每第三个系数。
- **平移滤波器。**取平均之前先乘以 ω^{-rj}，便能挑出满足 $k\equiv j$ 的系数。
- **用复数求实数答案。**答案是整数，题目问的也是整数，但解题的路径却穿过了复平面。

背后的数学

离散傅里叶变换

把 3 换成任意的 N，并把 ω 换成 $e^{2\pi i/N}$。对于一列数 $a_0, a_1, \ldots, a_{N-1}$，以下各数

$$\hat{a}_r = \sum_{k=0}^{N-1} a_k \, \omega^{rk}$$

构成它的*离散傅里叶变换*，而正是滤波恒等式 $\sum_r \omega^{rk} = N$ 或 0，让你可以从 $\hat{a}_r$ 还原出 a_k。这个变换把一个信号分解成它的各个频率成分。1965 年，James Cooley 和 John Tukey 发表了一种方法，只需大约 $N \log_2 N$ 步而非 N^2 步便能把它算出，这就是*快速傅里叶变换*（fast Fourier transform，简称 FFT）；高斯（Gauss）早在 1805 年已发现本质上相同的想法，但从未发表。FFT 应用于音讯和图像压缩（MP3 和 JPEG）、流动电话信号、医学扫描仪，以及极大数字的乘法，并且经常获列为二十世纪最重要的算法之一。

生成函数

这个滤波器只是一个更宏大想法的其中一例：数列 $a_0, a_1, a_2, \ldots$ 可以包装成一个多项式或幂级数（power series）$A(x) = \sum a_k x^k$，称为它的*生成函数*（generating function），而对数列进行的运算就变成对函数进行的运算。在 1 处求值，就是把数列求和；在单位根处求值，就是挑出各个剩余类（residue classes）；求导则是把各项以 k 加权。这样一来，许多计数问题都化为代数问题，而这正是第 9 章的核心技巧。同一个滤波器，配合五次单位根应用于 $(1+x)^n$，便会与正五边形和黄金比（golden ratio）扯上关系；我们会在 Section 6.1 继续探讨单位根的代数。

思维技巧

利用对称性来筛选

要从一个和之中挑出满足某个条件的项，可找一个函数，使它在你想要的项上取值 1、在其余的项上取值 0，而这个函数是由对称平均构造出来的。

对单位根取平均就是经典的例子。同一个原理——“对一个对称群取平均，以分离出你想要的部分”——在代数中随处可见，由单位根滤波器以至群表示论（representation theory of groups）皆然。

延伸练习

1 **偶数位置。**利用二次单位根（1 和 -1）的滤波器，证明 $\binom{n}{0}+\binom{n}{2}+\binom{n}{4}+\cdots=2^{n-1}$。

2 **另一个余数。**求 $\binom{n}{1}+\binom{n}{4}+\binom{n}{7}+\cdots$ 的公式，并验证三条公式加起来等于 2^n。

3 **每第四个。**利用 i，证明 $\binom{n}{0}+\binom{n}{4}+\binom{n}{8}+\cdots=\frac{1}{4}\left(2^n+2^{n/2+1}\cos\frac{n\pi}{4}\right)$。

4 **整除性。**$\{1,2,\ldots,12\}$ 有多少个子集的元素个数可被 3 整除？有多少个子集的元素之和可被 3 整除？（第二问的提示：考虑 $\prod_{j=1}^{12}(1+x^j)$ 在三次单位根处的值。）

延伸阅读

- **Herbert Wilf, *generatingfunctionology***（A. K. Peters；可于网上免费阅览）。一本引人入胜的生成函数入门书；单位根滤波器在书中很早便出现。
- **Arthur Benjamin & Jennifer Quinn, *Proofs That Really Count***（Mathematical Association of America）。以计数方法证明二项式恒等式。
- **T. W. Körner, *Fourier Analysis***（Cambridge University Press）。内容包括离散傅里叶变换和 FFT。

5.3 因式分解 $z^{2n}+1$

题目

STEP III 2013，第 4 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/13-s3-q4`

引子

$\cos\frac{\pi}{8}\cos\frac{3\pi}{8}\cos\frac{5\pi}{8}\cos\frac{7\pi}{8}$ 等于多少？每个因子都是涉及多重根号的难看数字，但它们的乘积却恰好是 $\frac{1}{8}$。等间隔角度的余弦之积，总是这样化简成简洁的数，而这道题正好解释了个中原因。窍门在于看出这些角度是某个多项式（polynomial）的根（roots）的辐角（argument），把该多项式因式分解（factorise），然后代入（substitute）一个巧妙选取的值。

在复数（complex numbers）范围内，每个多项式都可以完全分解成线性因式（linear factors）。在实数范围内，非实根以共轭对（conjugate pairs）的形式出现，每一对合并成一个实二次因式（real quadratic factor）。这两项事实在本题中都展现出来，而 $z^{2n}+1$ 明明完全没有实根，却可以写成实二次式之积，靠的正是第二项事实。

动笔之前

? **想一想** $z^{2n}=-1$ 的解在复数平面（complex plane）上的甚么位置？试就 $n=2$ 和 $n=4$ 把它们画出来。

它们的模（modulus）全都是 1，而辐角 θ 满足 $2n\theta=\pi+2m\pi$，所以 θ 是 $\frac{\pi}{2n}$ 的奇数倍。它们在单位圆（unit circle）上构成一个正 $2n$ 边形，并且旋转到没有任何顶点落在实轴上（图 5.3.1）。它们两两配成共轭对 $e^{\pm i\theta}$，即关于实轴互为镜像。

逐步解题

阶段 1　共轭对给出实二次式

$(z-e^{\mathrm{i}\theta})(z-e^{-\mathrm{i}\theta}) = z^2-(e^{\mathrm{i}\theta}+e^{-\mathrm{i}\theta})z+1 = z^2-2z\cos\theta+1.$ ■

阶段 2　因式分解

满足 $-\pi<\theta\le\pi$ 的 -1 的 $2n$ 次方根是 $e^{\pm\mathrm{i}\theta_k}$，其中

$$\theta_k=\frac{(2k-1)\pi}{2n},\qquad k=1,2,\dots,n.$$

这是 $2n$ 个互不相同的数，全都是 $2n$ 次多项式 $z^{2n}+1$ 的根，所以它们就是该多项式的全部根，于是

$$z^{2n}+1=\prod_{k=1}^{n}(z-e^{\mathrm{i}\theta_k})(z-e^{-\mathrm{i}\theta_k})=\prod_{k=1}^{n}(z^2-2z\cos\theta_k+1).$$ ■

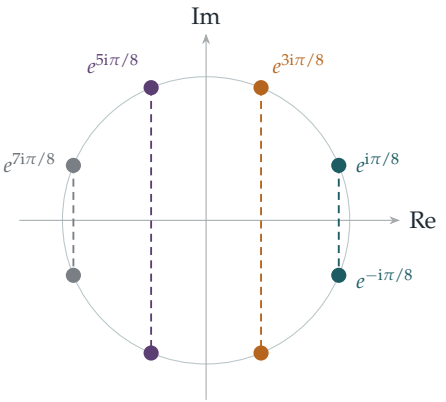


图 5.3.1　$z^8=-1$ 的八个解（即 $n=4$ 的情况）。它们配成四个共轭对，以虚线相连；每一对给出 z^8+1 的一个实二次因式 $z^2-2z\cos\theta_k+1$。

阶段 3　代入 $z=\mathrm{i}$

想一想　$\mathrm{i}^2-2\mathrm{i}\cos\theta+1$ 等于甚么？当 n 是偶数时，$\mathrm{i}^{2n}+1$ 又等于甚么？

每个因式变成

$$\mathrm{i}^2-2\mathrm{i}\cos\theta_k+1=-2\mathrm{i}\cos\theta_k,$$

所以右边是 $(-2\mathrm{i})^n\prod_k\cos\theta_k$。左边是 $\mathrm{i}^{2n}+1=(-1)^n+1$，当 n 是偶数时等于 2。而对于偶数 n，

$$(-2\mathrm{i})^n=2^n(-\mathrm{i})^n=2^n\big((-\mathrm{i})^2\big)^{n/2}=2^n(-1)^{n/2}.$$

因此

$$\prod_{k=1}^{n}\cos\theta_k=\frac{2}{2^n(-1)^{n/2}}=(-1)^{\frac12 n}2^{1-n}. \qquad \blacksquare$$

当 $n=4$ 时，这正是引子中的乘积：$(-1)^2 2^{-3}=\frac18$。

阶段 4 当 n 是奇数时

想一想 对于奇数 n，其中一个角 θ_k 等于 $\frac{\pi}{2}$。是哪一个？为甚么这会令代入 $z=\mathrm{i}$ 这一步在现有形式下毫无用处？

当 $k=\frac{n+1}{2}$ 时，$\theta_k=\frac{n\pi}{2n}=\frac{\pi}{2}$，所以该因式是 z^2+1，它在 $z=\mathrm{i}$ 时等于零；两边都是 0，我们甚么也得不到。解决办法是先把这个因式除去，而题目给出的恒等式正是这样做的。

停一停，试一试 把因式分解式的两边除以 $1+z^2$，代入 $z=\mathrm{i}$，然后化简。你需要把余下的余弦两两配对。

除以因式 z^2+1（即 $k=\frac{n+1}{2}$ 的那一项），得

$$1-z^2+z^4-\cdots+z^{2n-2}=\prod_{k\neq\frac{n+1}{2}}\left(z^2-2z\cos\theta_k+1\right).$$

当 $z=\mathrm{i}$ 时，左边每一项都是 $(-z^2)^j=1^j=1$，而共有 n 项，所以左边等于 n。右边是 $(-2\mathrm{i})^{n-1}\prod_{k\neq\frac{n+1}{2}}\cos\theta_k$。由于 $n-1$ 是偶数，$(-2\mathrm{i})^{n-1}=2^{n-1}(-1)^{(n-1)/2}$。

现在把各角配对：$\theta_{n+1-k}=\pi-\theta_k$，所以 $\cos\theta_{n+1-k}=-\cos\theta_k$。余下的 $n-1$ 个因子组成 $\frac{n-1}{2}$ 个这样的对，其中 $k=1,\dots,\frac{n-1}{2}$，而每一对贡献 $-\cos^2\theta_k$。因此

$$n=2^{n-1}(-1)^{(n-1)/2}\cdot(-1)^{(n-1)/2}\prod_{k=1}^{(n-1)/2}\cos^2\theta_k=2^{n-1}\prod_{k=1}^{(n-1)/2}\cos^2\theta_k,$$

从而得到 $\prod\cos^2\theta_k=n2^{1-n}$。最后一个因子，即 $k=\frac{n-1}{2}$ 的那个，是 $\cos^2\frac{(n-2)\pi}{2n}$，与题目所写的一致。 $\blacksquare$

以 $n=3$ 验算：$\cos^2\frac{\pi}{6}=\frac34=3\cdot2^{-2}$。

回顾

- **知道根，就知道多项式。**一个多项式由它的根决定（只差一个常数倍），所以列出 $z^{2n}+1$ 的根，就得到它的因式分解。
- **代入一个巧妙的值。**取 $z=\mathrm{i}$，每个二次因式都变成余弦的倍数。
- **当代入后得到 $0=0$，就先做除法。**对于奇数 n，有一个因式在 $z=\mathrm{i}$ 时等于零，所以我们在代入之前先把它除去。

背后的数学

代数基本定理

$z^{2n}+1$ 可以在 $\mathbb{C}$ 上完全分解成线性因式，这是*代数基本定理*（fundamental theorem of algebra）的一个特殊情况：每个复系数的非常数多项式都有一个复根，因而可以完全分解成线性因式。高斯（Gauss）在 1799 年的博士论文中给出了第一个获广泛接受的证明，其后又多次重新探讨这条定理，提出新的证明。由此可推出，每个实多项式都可以分解成实线性因式和实二次因式，因为非实根以共轭对的形式出现，正如第 1 步所见。这条定理名为“代数”，实质却属于分析：每个证明都要用到实数的某种性质，例如连续性（continuity），而你会在复分析（complex analysis）的入门课程中见到其中一个最巧妙的证明。

正弦与余弦之积

这个方法可以给出一整族恒等式。以 z^n-1 取代 $z^{2n}+1$，就得到

$$\prod_{k=1}^{n-1}\sin\frac{k\pi}{n}=\frac{n}{2^{n-1}},$$

它的几何形式在 Section 5.4 中有证明。这类恒等式会在振动系统的分析中出现（以弹簧相连的 n 个质点的频率涉及 $\sin\frac{k\pi}{2(n+1)}$），也会在计数问题中出现（某些图的生成树（spanning trees）数目，以及长方形的多米诺骨牌铺砌（domino tilings）数目，都可以用余弦之积表示）。Kasteleyn 在 1961 年提出的 $m\times n$ 长方形多米诺骨牌铺砌公式，是一个双重乘积，其中各项是含 $\cos^2\frac{j\pi}{m+1}$ 和 $\cos^2\frac{k\pi}{n+1}$ 的式子。

思维技巧

从根到乘积

要计算等间隔角度的三角函数值之积，先看出它们来自某个简单多项式的根，再作因式分解，然后代入一个方便的值。

选取甚么值，是一门艺术：本题中 $z=\mathrm{i}$ 把每个二次因式变成一个余弦。其他选择（例如 $z=1$ 或 $z=-1$）会给出其他恒等式。

延伸练习

1 **一个七边形的乘积。**利用第 (ii) 部分，取 $n=7$，证明 $\cos\frac{\pi}{14}\cos\frac{3\pi}{14}\cos\frac{5\pi}{14}=\frac{\sqrt{7}}{8}$。然后证明 $\cos\frac{\pi}{7}\cos\frac{2\pi}{7}\cos\frac{3\pi}{7}=\frac{1}{8}$。（提示：利用 $\cos\theta=\sin(\frac{\pi}{2}-\theta)$ 把各角联系起来，或者反复使用恒等式 $\sin 2x=2\sin x\cos x$。）

2 **另一个代入。**在因式分解式中代入 $z=1$，证明

$$\prod_{k=1}^{n}\sin\frac{(2k-1)\pi}{4n}=2^{\frac{1}{2}-n}.$$

3 **分解 z^6+1。**把 z^6+1 写成三个实二次式之积，并把它们乘开以验算答案。

4 **正弦。**以类似方法分解 $z^{2n}-1$，再除以 z^2-1，证明

$$\prod_{k=1}^{n-1}\sin\frac{k\pi}{2n}=\frac{\sqrt{n}}{2^{n-1}}.$$

延伸阅读

- **Tristan Needham, *Visual Complex Analysis***（Oxford University Press）。第 2 章讲述单位根（roots of unity）与多项式，配有精美的图画。
- **Benjamin Fine & Gerhard Rosenberger, *The Fundamental Theorem of Algebra***（Springer）。收录这条定理的多个证明，分别来自分析、代数和拓扑，为大学本科生而写。
- **Paul Nahin, *An Imaginary Tale***（Princeton University Press）。一部关于 $\sqrt{-1}$ 的历史，包括单位根及其应用。

5.4 正多边形中的距离

题目

STEP III 2016，第 7 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/16-s3-q7`

引子

在半径为 1 的圆内画一个正七边形（regular heptagon）。从其中一个顶点出发，画出连往其余各顶点的六条弦（chord），量度它们的长度，再把这六个长度乘起来。答案恰好是 7。若换成正 100 边形，从一个顶点出发的 99 条弦，长度之积恰好是 100。

这么简洁的结果，为甚么会成立？各条弦的长度本身并不整齐（$2\sin\frac{\pi}{7}$、$2\sin\frac{2\pi}{7}$ 等等）。这道题目说明，这个结果其实是一个简单代数恒等式在几何上的投影，这个恒等式就是 z^n-1 按单位根（roots of unity）所作的因式分解（factorisation）；题目同时说明，平面上的距离就是复数（complex number）的模（modulus）。

动笔之前

> **想一想** 把多边形的顶点放在 n 次单位根上，即 $X_k=\omega^k$。若 P 是点 z，$|PX_k|$ 用 z 表示是甚么？那么所有距离之积又是甚么？

$|PX_k|=|z-\omega^k|$，而所有距离之积是 $|(z-1)(z-\omega)\cdots(z-\omega^{n-1})|$，按题目第一部分，这等于 $|z^n-1|$。n 个距离之积变成了单独一个模。余下的工夫，全在于巧妙地选取 z。

逐步解题

阶段 1 因式分解

每个 ω^k 都满足 $(\omega^k)^n = (\omega^n)^k = 1$，所以 $1, \omega, \ldots, \omega^{n-1}$ 都是 $z^n - 1$ 的根。它们互不相同（它们的辐角（argument）$\frac{2\pi k}{n}$ 各不相同，且都在 0 与 2π 之间），所以它们就是这个首项系数（leading coefficient）为 1 的 n 次多项式的全部 n 个根，于是 $z^n - 1 = \prod_{k=0}^{n-1}(z - \omega^k)$。
■

阶段 2 垂直平分线上的一点

把图形旋转，使得 $X_k = \omega^k$。与 $X_0 = 1$ 和 $X_1 = \omega$ 等距的点，组成 X_0X_1 的垂直平分线（perpendicular bisector），即通过 O 而倾角为 $\frac{\pi}{n}$ 的直线。

想一想 所以 P 就是 $z = te^{\mathrm{i}\pi/n}$，其中 t 是某个实数（可正可负），而 $|OP| = |t|$。z^n 是甚么？

$z^n = t^n e^{\mathrm{i}\pi} = -t^n$。所以

$$|PX_0| \times \cdots \times |PX_{n-1}| = |z^n - 1| = |-t^n - 1| = |t^n + 1|.$$

若 n 是偶数，$t^n = |t|^n \geq 0$，所以上式等于 $|OP|^n + 1$。 ■

停一停，试一试 n 是奇数时又如何？分开考虑 $t > 0$ 和 $t < 0$ 两种情况，并说明每种情况在几何上代表甚么。

若 n 是奇数，t^n 与 t 同号。

- 若 $t \geq 0$，即 P 与 X_0X_1 的中点位于 O 的同一侧，则所求之积是 $|OP|^n + 1$。
- 若 $t < 0$，即 P 位于 O 的另一侧（朝向与边 X_0X_1 相对的顶点），则 $t^n = -|OP|^n$，所求之积是 $\left|1 - |OP|^n\right|$。

当 n 是奇数时，通过 O 而垂直于 X_0X_1 的直线会穿过另一侧的一个顶点；在第二种情况下，当 P 就是该顶点（$|OP| = 1$）时，所求之积等于零，这是理所当然的，因为此时其中一个距离是零。

阶段 3 从一个顶点出发的弦

想一想 现在 $P = X_0$，其中一个距离 $|X_0X_0|$ 是零，所以 $|z^n - 1| = 0$ 派不上用场。哪一个多项式以 $\omega, \omega^2, \ldots, \omega^{n-1}$ 为根，却不以 1 为根？

把 $z^n - 1$ 除以 $z - 1$，

$$(z - \omega)(z - \omega^2)\cdots(z - \omega^{n-1}) = \frac{z^n - 1}{z - 1} = 1 + z + z^2 + \cdots + z^{n-1}.$$

右边是一个多项式，所以这个等式对所有 z 都成立，包括 $z = 1$，此时它的值是 n。在

$z = 1$ 处取模，

$$|X_0X_1| \times |X_0X_2| \times \cdots \times |X_0X_{n-1}| = |1-\omega|\,|1-\omega^2| \cdots |1-\omega^{n-1}| = n. \qquad \blacksquare$$

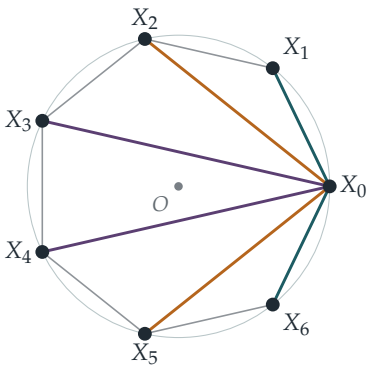


图 5.4.1 内接于半径为 1 的圆的正七边形。从 X_0 出发的六条弦，长度为 $2\sin\frac{k\pi}{7}$，约为 0.868（蓝绿色）、1.564（琥珀色）和 1.950（紫红色），每个长度各出现两次。它们的积恰好是 7。

回顾

- **距离就是模。**$|PX_k| = |z-\omega^k|$，所以距离之积就是积的模，而这个积有闭合形式（closed form）。
- **利用对称性选取坐标。**把图形旋转，使顶点成为单位根，再利用通过 O 的对称轴，z^n 便容易计算。
- **先消去零因子，再代入。**一如 Section 5.3，先除去在所关注的点上等于零的因子，代入才行得通。

背后的数学

科茨定理

第 (i) 部分是科茨（Roger Cotes）的一条定理。科茨是剑桥一位才华横溢的年轻数学家，曾编订牛顿（Newton）《自然哲学的数学原理》（*Principia*）的第二版，1716 年去世，年仅三十三岁（牛顿曾说："假如他还在世，我们或许已经学到一点东西了。"）。科茨发现，若 P 位于由 O 出发、通过正 n 边形一个顶点的射线上，则 P 到各顶点的距离之积是 $\big||OP|^n - 1\big|$；若 P 位于通过某条边中点的射线上，则这个积是 $|OP|^n + 1$。他没有用到复数，就凭这个结果把 $x^n \pm 1$ 分解成实二次因式，而这正是他计算有理函数（rational

function）积分时所需要的。这条定理在他去世后，于 1722 年才发表。

正弦之积

由于 $|1-\omega^k| = |e^{-\mathrm{i}\pi k/n} - e^{\mathrm{i}\pi k/n}| = 2\sin\frac{k\pi}{n}$，第 (ii) 部分说的就是

$$\prod_{k=1}^{n-1} \sin\frac{k\pi}{n} = \frac{n}{2^{n-1}}.$$

这与 Section 5.3 中的余弦之积互相配对。另一个相关的积给出 z^n-1 的判别式（discriminant）：所有成对的相异 n 次单位根之间距离的平方，乘起来等于 n^n。判别式量度一个多项式的根彼此相隔多远，在多项式和数域（number field）的代数理论中占有核心地位，我们会在第 6 章讨论。

静电学与等距分布

这个结果还有一个物理上的解读。把 n 个相等的电荷放在一个圆上；令势能（potential energy）最小的排列（采用二维静电学的对数势）就是正多边形，而能量所涉及的正是 $\sum \ln|X_jX_k|$，即距离之积的对数。“分布得尽量疏散”的点，在数值分析（numerical analysis）中十分重要（见 Section 3.6，当中最佳的插值点是由正多边形经 $x=\cos\theta$ 得来的），在随机矩阵（random matrix）理论中亦然。

思维技巧

用复数把几何化为代数

平面上的点是复数，距离是模，正多边形是单位根。距离之积就变成多项式的模。

经过这样的转换，像 z^n-1 的因式分解这类代数恒等式，便可以用来证明一些很难直接看出的几何事实。

延伸练习

1. **检查小的情况。**就正方形（$n=4$）和等边三角形（$n=3$）直接验证第 (ii) 部分。
2. **通过顶点。**证明：若 P 位于由 O 出发、通过 X_0 的射线上，则 $|PX_0|\times\cdots\times|PX_{n-1}| = \big||OP|^n-1\big|$。

3 **平均距离。**利用第 (ii) 部分和 Section 3.2 的算术平均—几何平均不等式（AM–GM inequality），证明从 X_0 出发的 $n-1$ 条弦，平均长度至少是 $n^{1/(n-1)}$。就 $n=6$ 求出实际的平均值 $\frac{1}{n-1}\sum_{k} 2\sin\frac{k\pi}{n}$，并加以比较。

4 **所有的弦。**较难证明：内接于单位圆的正 n 边形，全部 $\binom{n}{2}$ 条弦的长度之积是 $n^{n/2}$。

延伸阅读

- **Tristan Needham, *Visual Complex Analysis***（Oxford University Press）。用复数处理几何，插图精美。
- **Ross Honsberger, *Mathematical Gems***（Mathematical Association of America）。收录了好几道关于正多边形的精巧问题。
- **Ronald Gowing, *Roger Cotes: Natural Philosopher***（Cambridge University Press）。科茨的传记，并介绍他的数学工作。

5.5 名不副实的有理角

题目

STEP III 2022，第 8 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2022 年 STEP 试卷（zip 文件）内的 STEP III 试卷：

`www.ocr.org.uk/Images/696587-2022-past-papers.zip`

引子

中学几何里遇到的角，度数通常是整数：30°、45°、60°。它们的余弦是 $\frac{\sqrt{3}}{2}$、$\frac{1}{2}$ 这类数。反过来又如何？如果余弦是个简单的分数，例如 $\frac{1}{3}$，那么这个角的度数会不会也是个漂亮的数？

$\cos^{-1}\frac{1}{3} \approx 70.528779\ldots^\circ$ 这个角在科学中随处可见。它是正四面体（regular tetrahedron）两个面之间的夹角，而它的补角 109.47° 则是甲烷分子中两条化学键之间的夹角。本题要证明这个角的度数不是有理数（rational number）：不是 $70\frac{1}{2}$，不是 $\frac{1411}{20}$，也不是任何分数。它的小数展开永不循环。这个看似冷僻的事实，解决了大卫·希尔伯特（David Hilbert）在 1900 年提出的著名问题之一。

动笔之前

想一想 假设 θ 是有理数，设 $\theta = \frac{p}{q}$ 度。试找出一个正整数 k，使得 $\sin(k\theta) = 0$。

取 $k = 180q$：这时 $k\theta = 180p$ 度，是半圈的整数倍，所以 $\sin(k\theta) = 0$。因此，度数为有理数的角，总有某个倍数会落在 180° 的倍数上。要证明 θ 是无理数（irrational number），只需证明没有任何倍数 $k\theta$ 的正弦为零。第 (ii) 部分的做法是：假如这样的 k 存在，就取其中最小的一个，然后推出矛盾（contradiction）。这又是 Section 1.3 的良序原理（well-ordering principle）。

逐步解题

阶段 1　倍角公式

想一想　写出 $\cos k\theta + \mathrm{i}\sin k\theta = (\cos\theta + \mathrm{i}\sin\theta)^k$，并提出因子 $\cos^k\theta$。括号内剩下甚么？

$(\cos\theta + \mathrm{i}\sin\theta)^k = \cos^k\theta\,(1 + \mathrm{i}\tan\theta)^k = \cos^k\theta \sum_{j=0}^{k} \binom{k}{j} \mathrm{i}^j \tan^j\theta$。虚部（imaginary part）来自奇数 $j = 2m+1$ 的各项，这时 $\mathrm{i}^j = \mathrm{i}(-1)^m$：

$$\begin{aligned}\sin k\theta &= \cos^k\theta \sum_{m} (-1)^m \binom{k}{2m+1} \tan^{2m+1}\theta \\ &= \sin\theta\cos^{k-1}\theta \sum_{m} (-1)^m \binom{k}{2m+1} \tan^{2m}\theta.\end{aligned}$$

由于 $\tan^2\theta = \sec^2\theta - 1$，这就是所求的公式。■实部（real part）来自偶数 $j = 2m$ 的各项，由此得

$$\cos k\theta = \cos^k\theta \left(1 - \binom{k}{2}(\sec^2\theta - 1) + \binom{k}{4}(\sec^2\theta - 1)^2 - \cdots\right).$$

阶段 2　奇偶性

现在 $\sec\theta = a$ 是奇整数，所以 $\sec^2\theta - 1 = a^2 - 1$ 是偶整数。而且 θ 严格介乎 0° 与 90° 之间，所以 $\sin\theta \neq 0$ 且 $\cos\theta \neq 0$。

想一想　如果 $\sin(k\theta) = 0$，括号 $k - \binom{k}{3}(a^2-1) + \cdots$ 必须等于零。第一项之后的每一项都是 $a^2 - 1$ 的倍数。由此可知 k 有甚么性质？

括号等于 k 加上若干个偶数 $a^2 - 1$ 的倍数之和。假如 k 是奇数，括号就是奇数，而奇整数不可能等于 0。所以 k 必定是偶数。

于是 $\sin(k\theta) = 2\sin(\tfrac{1}{2}k\theta)\cos(\tfrac{1}{2}k\theta) = 0$。但 $\tfrac{1}{2}k$ 是比 k 小的正整数，根据 k 的最小性，$\sin(\tfrac{1}{2}k\theta) \neq 0$。因此 $\cos(\tfrac{1}{2}k\theta) = 0$。 ■

阶段 3　矛盾

停一停，试一试　取 $m = \tfrac{1}{2}k$，套用余弦公式。为甚么 $\cos(m\theta)$ 永远不可能是零？

根据第 1 步，$\cos(m\theta) = a^{-m}\left(1 - \binom{m}{2}(a^2-1) + \binom{m}{4}(a^2-1)^2 - \cdots\right)$。括号等于 1 加上若干个偶数 $a^2 - 1$ 的倍数，所以它是奇数，自然不等于零。因此 $\cos(m\theta) \neq 0$，与第 2 步矛盾。

所以不存在使 $\sin(k\theta) = 0$ 的正整数 k。（假如有这样的正整数，其中必有最小的一个，而我们刚刚证明了这个最小的不可能存在。）根据动笔之前的观察，θ 的度数不可能是有理数。 ■

阶段 4 余切版本

? **想一想** 如果 $\cot\phi = \frac{1}{b}$，$\sec^2\phi - 1$ 是多少？它是偶数吗？

$\tan\phi = b$，所以 $\sec^2\phi - 1 = \tan^2\phi = b^2$；由于 b 是偶数，这个数是偶数（其实更是 4 的倍数）。此外 $0° < \phi < 90°$，所以 $\sin\phi$ 和 $\cos\phi$ 都不等于零。完全相同的论证现在同样适用：如果 k 是使 $\sin(k\phi) = 0$ 的最小正整数，括号 $k - \binom{k}{3}b^2 + \cdots$ 必须等于零，于是 k 必为偶数；然后 $\cos(\frac{1}{2}k\phi) = 0$；但余弦的括号 $1 - \binom{m}{2}b^2 + \cdots$ 是奇数。这就产生矛盾，所以 ϕ 是无理数。 ■

例如，$\cot^{-1}\frac{1}{2} = \tan^{-1}2 \approx 63.43°$ 是无理数。

回顾

- **有理角的倍数具周期性。**这样重新表述之后，无理性就变成一个关于所有 k 的 $\sin(k\theta)$ 的命题。
- **奇偶性论证。**以模 2 计算，可知某些整数是奇数，因而不等于零。方法简单，却一锤定音。
- **最小反例。**假设有最小的 k，再证明必定有更小的一个（这里是 $\frac{1}{2}k$），正是第 1 章的良序论证。

背后的数学

无理旋转

如果 θ 的度数是有理数，不断重复旋转 θ，最终会回到起点。如果 θ 是无理数，就永远不会回到起点，而且会出现奇妙的现象：各点 $k\theta$（以模 360° 计算）会均匀地散布在整个圆周上，与每一个角都可以任意接近。这就是等分布定理（equidistribution theorem），由赫尔曼．外尔（Hermann Weyl）、瓦茨瓦夫．谢尔宾斯基（Wacław Sierpiński）和 Piers Bohl 在 1910 年前后各自独立证明。图 5.5.1 显示 $\cos^{-1}\frac{1}{3}$ 的首 40 个倍数：它们已经逐渐布满圆周，而且永远不会重复。

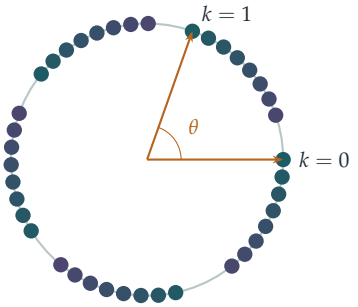


图 5.5.1　位于角 $k\theta$ 的各点，其中 $k = 0, 1, \ldots, 39$，而 $\theta = \cos^{-1}\frac{1}{3} \approx 70.53°$（颜色随 k 增加由蓝绿色渐变为紫红色）。由于 θ 是无理数，没有任何一点会重复出现，而各点均匀地散布在整个圆周上。

希尔伯特第三问题

1900 年，大卫．希尔伯特为新世纪列出了二十三个问题。第三个问题是：如果两个多面体（polyhedron）体积相同，是否总可以把其中一个切成有限多块多面体，再重新拼合成另一个？（对平面上的多边形而言，答案是肯定的：这就是 Wallace–Bolyai–Gerwien 定理。）不出一年，希尔伯特的学生马克斯．德恩（Max Dehn）证明了答案是*否定*的：正四面体不能切开再重新拼合成体积相同的立方体。德恩的证明为每个多面体赋予一个由其棱长和二面角（dihedral angle）构成的量，现称为*德恩不变量*（Dehn invariant），它在切割和重新拼合之下保持不变；立方体的德恩不变量是零，而正四面体的德恩不变量不等于零，原因正是二面角 $\cos^{-1}\frac{1}{3}$ 不是 $180°$ 的有理数倍。这恰恰就是本题所证明的事实。

尼云定理

这个结论可以推到多远？1956 年，伊凡．尼云（Ivan Niven）证明：如果 θ 的度数是有理数，而且 $\cos\theta$ 是有理数，那么 $\cos\theta$ 只能是 0、$\pm\frac{1}{2}$、± 1 其中之一。所以余弦为有理数的“漂亮”角，就只有 $0°$、$60°$、$90°$、$120°$、$180°$ 以及与它们相关的角。证明用到这个事实：$2\cos\theta$ 是某个整数系数首一多项式（monic polynomial）的根，即*代数整数*（algebraic integer），也就是 Section 1.3 末尾提到的概念。

思维技巧

以奇偶性证明不等于零

要证明一个整数算式永不为零，可证明它恒为奇数：把它以模 2 化简。

这个简单的想法与最小反例（minimal counterexample）结合，就证明了一个深刻的事实。以一个小的数为模作化简，从而排除某个方程成立的可能，是数论中最常用的手法之一，Section 1.5 便是一例。

延伸练习

1 **正四面体。**证明正四面体两个面之间的夹角是 $\cos^{-1}\frac{1}{3}$，并证明甲烷分子中两条化学键（由正四面体的中心连到两个顶点）之间的夹角是 $\cos^{-1}\left(-\frac{1}{3}\right)$。

2 **3–4–5 三角形。**证明 3–4–5 三角形最小的角的度数是无理数。(提示：这个角是 $\cot^{-1}\frac{4}{3}$；可改写上述论证，或考虑高斯整数（Gaussian integer）$4+3\mathrm{i}$ 的各次幂，并证明它们全都不是实数。)

3 **为甚么 a 要是奇数？**当 $a=2$ 时，论证在哪里出了问题？$\cos^{-1}\frac{1}{2}$ 是有理数吗？

4 **尼云定理的简单情况。**找出介乎 0° 与 90° 之间、度数为整数而 $\cos\theta$ 为有理数的所有角 θ，并验证你所列的结果与尼云定理吻合。

延伸阅读

- **Ivan Niven, *Irrational Numbers***（Mathematical Association of America）。书中载有关于三角函数有理数值的尼云定理。
- **Martin Aigner & Günter Ziegler, *Proofs from THE BOOK***，论希尔伯特第三问题的一章。对德恩的解法有精彩的阐述。
- **Benjamin Yandell, *The Honors Class: Hilbert's Problems and Their Solvers***（A. K. Peters）。讲述全部二十三个问题背后的故事。

5.6 费马点

题目

STEP III 2000，第 3 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/00-s3-q3`

引子

三条村庄想兴建一个把它们全部连接起来的道路网络，而且路程愈短愈好。交汇处应该设在哪里？大约在 1640 年，费马（Pierre de Fermat）提出了这样一个问题：在三角形内找出使 $PA + PB + PC$ 最小的点 P。托里拆利（Evangelista Torricelli）解决了这个问题：从这一点望向三条边，所见的角都是 120°。

本题利用复数（complex numbers）来证明这一点的一个漂亮作图法。在三角形的每条边上向外作一个等边三角形（equilateral triangle），再把每个新顶点与原三角形中相对的顶点连起来。这三条直线相交于同一点，即*费马点*（Fermat point），而且三条线段的长度都相同，这个长度正是道路总长的最小值。有了复数，再加上一条关于 $e^{i\pi/3}$ 的恒等式，整件事只需几行代数运算。

动笔之前

想一想 A、B 和 C 在哪里？它们的辐角（argument）是多少？OA、OB 和 OC 彼此之间的夹角是多少？

$A = p$ 位于正实轴上（辐角为 0°）。$B = q\alpha^2 = qe^{2\pi i/3}$ 的辐角是 120°。$C = -r\alpha = re^{i\pi}e^{i\pi/3}$ 的辐角是 240°。因此原点是三角形内的一点，从这一点望向三个顶点，方向两两相隔 120°，距离分别是 p、q 和 r。按照这个设置，原点就是费马点；本题要证明的，是等边三角形作图法确实能找到它。

逐步解题

阶段 1　关键恒等式

$\alpha^3 = e^{i\pi} = -1$，所以 α 是 $z^3+1 = (z+1)(z^2-z+1)$ 的一个根。由于 $\alpha \neq -1$，故 $\alpha^2 - \alpha + 1 = 0$，即 $1+\alpha^2 = \alpha$。 ■

有两个推论稍后会用得着：$1-\alpha = -\alpha^2$，以及 $\alpha - \alpha^2 = 1$。此外，$1-\alpha = \frac{1}{2} - \frac{\sqrt{3}}{2}\mathrm{i} = e^{-i\pi/3}$，即 α 的共轭复数（complex conjugate）。

阶段 2　把等边三角形看成旋转

> **想一想**　三角形 CAN 是等边三角形，顶点按顺时针方向标记。假如你站在 C，怎样从 A 走到 N？试以 $A-C$ 表示 $N-C$。

按顺时针方向依次经过 C、A、N，可知点 N 是把 A 绕 C 顺时针旋转（rotation）60° 而得。乘以 $e^{-i\pi/3} = 1-\alpha$ 相当于顺时针旋转 60°，所以

$$\begin{aligned} N &= C + (1-\alpha)(A-C) = -r\alpha + (1-\alpha)(p+r\alpha) \\ &= (1-\alpha)p + r(\alpha - \alpha^2 - \alpha) = (1-\alpha)p - \alpha^2 r. \end{aligned}$$ ■

（顺时针而非逆时针旋转，会使 N 落在 CA 的另一侧，与原点分处两边，即在三角形之外。）

> **停一停，试一试**　用同样的方法求出 L 和 M，并利用第 1 阶段的恒等式化简。

同理，$L = A + (1-\alpha)(B-A)$ 及 $M = B + (1-\alpha)(C-B)$：

$$\begin{aligned} L &= p + (1-\alpha)(q\alpha^2 - p) = \alpha p + (\alpha^2 - \alpha^3)q = \alpha p + (\alpha^2+1)q = \alpha(p+q), \\ M &= q\alpha^2 + (1-\alpha)(-r\alpha - q\alpha^2) = q\alpha^3 + r(\alpha^2 - \alpha) = -q - r. \end{aligned}$$

再利用 $1-\alpha = -\alpha^2$，可得 $N = -\alpha^2 p - \alpha^2 r = -\alpha^2(p+r)$。

阶段 3　三条通过原点的直线

> **想一想**　比较 $L = \alpha(p+q)$ 与 $C = -r\alpha$。你留意到甚么？对 M 与 A，以及 N 与 B，也作同样的比较。

L 是 α 的正数倍，而 C 是 α 的负数倍。所以 L 和 C 分别位于 O 的两侧，并且在同一条通过 O 的直线上；线段 LC 因而通过原点，长度为 $|L-C| = |\alpha|(p+q+r) = p+q+r$。

同样地，$M = -(q+r)$ 和 $A = p$ 都在实轴上，分别位于 O 的两侧，所以 MA 通过 O，

长度为 $p+q+r$。而 $N=-\alpha^2(p+r)$ 和 $B=q\alpha^2$ 都在通过 O、方向为 α^2 的直线上，并且分处两侧，所以 NB 通过 O，长度为 $p+q+r$。 ■

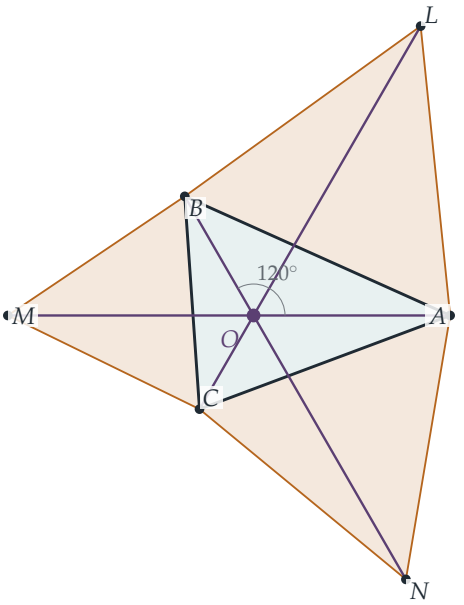


图 5.6.1 费马点，其中 $p=2$，$q=1.4$，$r=1.1$。在 ABC 的各边上向外作等边三角形（琥珀色）。直线 LC、MA 和 NB（紫红色）全部通过原点 O，而 ABC 的各边在该点所对的角均为 120°；每条线段的长度都是 $p+q+r=4.5$。

回顾

- **旋转即乘法。**作等边三角形,就是把一个顶点绕另一个顶点旋转 60°,也就是乘以 $e^{\pm i\pi/3}$。
- **巧选坐标。**把费马点放在原点，并让各顶点沿 1、α^2、$-\alpha$ 这三个方向排列，每条算式都随即大幅简化。
- **一条恒等式，反复使用。**每一步化简都来自 $\alpha^2-\alpha+1=0$ 和 $\alpha^3=-1$。

背后的数学

为甚么费马点使距离最小

为甚么 $PA+PB+PC$ 会在各边所对的角均为 120° 的那一点取得最小值？以下是一个精彩的论证。把整幅图绕 B 旋转 60°，把 A 送到 L'（即 AB 上等边三角形的顶点），把 P 送到 P'。三角形 BPP' 是等边三角形，所以 $PB=PP'$，而且 $PA=P'L'$。于是

$PA + PB + PC = L'P' + P'P + PC$，这是一条由 L' 到 C 的路径的长度。它至少等于 $|L'C|$，而等号恰好在路径成一直线时成立，这种情况出现于 P 处的各个角均为 120° 之时。所以总长度的最小值是 $|LC| = p + q + r$，与题目所得的结果一致。（这个论证成立的前提是三角形的每个角都小于 120°；否则最小值会在钝角的顶点取得。）

肥皂膜与斯坦纳树

把一个装有三根小钉（代表 A、B、C）的框架浸入肥皂水中：肥皂膜会令其表面张力能量减至最小，形成三面薄壁，在费马点以 120° 相交。若有更多的点，最短的连接网络是一棵*斯坦纳树*（Steiner tree），其中有额外的交汇点，每个交汇点都有三条边以 120° 相交。一般而言，求斯坦纳树在计算上十分困难（这是一个 NP 困难（NP-hard）问题），但它们在实际应用中相当重要，例如设计道路网络、管道网络以及电脑晶片上的线路。

拿破仑定理

同一个图形中还藏着另一件瑰宝：在任意三角形的三边上所作的三个等边三角形，它们的*中心*构成一个等边三角形。传统上这个结果归功于拿破仑（Napoleon Bonaparte）（他是一位热衷数学的业余爱好者），不过并无证据显示这是他发现的。这个定理同样可以用复数简短地证明，用到的正是 $1 + \alpha^2 = \alpha$。

思维技巧

选择切合问题的坐标

把最重要的点放在原点，并按照问题的对称性来安排各个方向。选得好，代数运算几乎可以变得轻而易举。

在这里，原点就是费马点，而三个顶点分别位于 -1 的三个立方根的方向上（或相差一个正负号）。若改用其他坐标，同样的计算会繁复得多。

延伸练习

1. **检查字母次序。**证明若各等边三角形的顶点按逆时针方向标记（即向内作），则三条直线一般不会通过 O。

2. **拿破仑定理。**证明等边三角形 ABL 的中心是 $\frac{1}{3}(A + B + L)$。利用上文的表达式，证明三个中心构成一个等边三角形。

3 **最小值。**取 $p = 2$, $q = 1.4$, $r = 1.1$，分别就 P 位于每个顶点以及位于形心（centroid）$\frac{1}{3}(A + B + C)$ 的情况计算 $PA + PB + PC$，并验证所得的值全部大于 4.5。

4 **等边三角形的判别法。**证明 z_1、z_2、z_3 构成等边三角形（不论方向），当且仅当 $z_1^2 + z_2^2 + z_3^2 = z_1z_2 + z_2z_3 + z_3z_1$。

延伸阅读

- **Richard Courant & Herbert Robbins, *What Is Mathematics?***（Oxford University Press）。第 7 章讨论费马点、斯坦纳树和肥皂膜。
- **Liang-shin Hahn, *Complex Numbers and Geometry***（Mathematical Association of America）。以复数证明多个几何定理，程度正好与本节相若。
- **Tristan Needham, *Visual Complex Analysis***（Oxford University Press）。第 1 章包括拿破仑定理及其他几何结果。

5.7 作为映射的复指数函数

题目

STEP II 1997，第 5 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/97-s2-q5`

引子

你所认识的指数函数（exponential function）是一条曲线 $y = e^x$。但复变量的函数却无法画成曲线：它的输入是平面上的一点，输出也是平面上的一点。要想像这样的函数，应把它看成一个映射（map），也就是一个变换（transformation）：它把一个平面上的每一点搬到另一个平面上的某一点，沿途把图形拉伸、弯曲。

题目中的公式是 $z = e^u(\cos v + \mathrm{i}\sin v) = e^{u+\mathrm{i}v} = e^w$。作为映射来看，复指数函数（complex exponential）做了一件奇妙的事：它把平面上的一条水平带状区域卷成整个平面（除去原点），把铅垂线绕成圆，把水平线铺成射线。又因为它是绕圈卷起来的，所以它不是一对一（one-to-one）的：每一点都有无穷多个点映到它那里。这正是复对数（complex logarithm）“多值”（multi-valued）的原因，也是复分析（complex analysis）最早带来的惊奇之一。

动笔之前

想一想 $z = e^u(\cos v + \mathrm{i}\sin v)$ 的模（modulus）和辐角（argument）用 u 和 v 表示是甚么？

$|z| = e^u$，而 $\arg z = v$（可相差 2π 的倍数）。所以 w 的实部 u 以指数方式控制与原点的距离，而虚部 v 则控制方向。单凭这一点观察，题目几乎每一部分都能解答。

逐步解题

阶段 1　解 $e^w = \mathrm{i}e$

想一想　$\mathrm{i}e$ 的模和辐角是甚么?

$|\mathrm{i}e| = e$，而 $\arg(\mathrm{i}e) = \frac{\pi}{2}$。所以我们需要 $e^u = e$，得 $u = 1$，并且 $v = \frac{\pi}{2} + 2k\pi$，其中 k 为任意整数。方程的解为

$$w = 1 + \mathrm{i}\left(\tfrac{\pi}{2} + 2k\pi\right), \qquad k \in \mathbb{Z}.$$

解有无穷多个，沿铅垂方向每隔 2π 有一个。

阶段 2　格线的像

在铅垂线 $u = c$ 上，$|z| = e^c$ 固定不变，而 v 则在变动：它的像（image）是以原点为圆心、半径为 e^c 的圆，当 v 取遍所有实数时，这个圆被绕了无穷多圈。

在水平线 $v = d$ 上，$\arg z = d$ 固定不变，而 $|z| = e^u$ 取遍所有正值：它的像是由原点出发、方向角为 d 的开射线（半直线），不包括原点本身。直线 $v = d$ 与 $v = d + 2\pi$ 给出同一条射线。

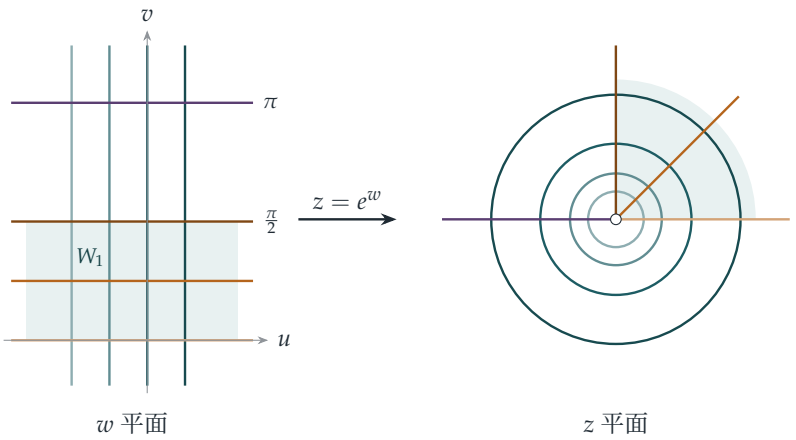


图 5.7.1　把复指数函数看成映射。w 平面上的铅垂线 $u = c$（蓝绿色）映到半径为 e^c 的圆；水平线 $v = d$（琥珀色、紫红色）映到方向角为 d 的射线。阴影带状区域 $W_1 = \{0 < v < \frac{\pi}{2}\}$ 恰好映满第一象限。留意两幅图中的格线都相交成直角：这个映射保持角度不变。原点（空心圆圈）永远不会被映到。

阶段 3　一对一的部分

想一想 哪些点 w 会映入第一象限 $x>0$、$y>0$？为甚么这个集合太大，不能与该象限构成一一对应（one-to-one correspondence）？

z 位于开的第一象限，当且仅当 $\arg z$ 严格介乎 0 与 $\frac{\pi}{2}$ 之间（模 2π），而模可取任意值。所以 w 必须满足 $0<v<\frac{\pi}{2}$ 或 $2\pi<v<\frac{5\pi}{2}$ 或 ...：共有无穷多条水平带状区域，每一条都映满整个象限。任取其中一条，它都与该象限一一对应，因为在一条带状区域之内，u 和 v 由 $|z|$ 和 $\arg z$ 唯一确定。所以我们可以取

$$W_1=\{u+\mathrm{i}v:0<v<\tfrac{\pi}{2}\},\qquad W_2=\{u+\mathrm{i}v:2\pi<v<\tfrac{5\pi}{2}\}.$$

停一停，试一试 现在试在 w 平面上找出两个集合，使它们各自与去心圆盘（punctured disc）$0<|z|<1$ 一一对应。

$|z|<1$ 即 $e^u<1$，亦即 $u<0$；而每一个辐角都必须恰好被覆盖一次，所以 v 的取值范围应是一个长度恰为 2π 的区间，包括其中一个端点而不包括另一个。例如

$$W_3=\{u+\mathrm{i}v:u<0,\ 0\le v<2\pi\},\qquad W_4=\{u+\mathrm{i}v:u<0,\ -\pi<v\le\pi\}.$$

这些都是半带状区域，一条边包括在内，另一条边不包括在内。（若两条边都包括在内，正实轴上的点便会被映到两次；若两条边都不包括，这些点便会被遗漏。）■

回顾

- **模和辐角把变量分开。** $|e^w|=e^{\operatorname{Re}w}$，而 $\arg e^w=\operatorname{Im}w$，所以 w 的实部和虚部分别独立地控制距离和方向。
- **把复函数想像成映射。** 追踪格线被映到哪里，便能清楚看见一个无法画成图像的函数的面貌。
- **周期性破坏单射性（injectivity）。** $e^{w+2\pi\mathrm{i}}=e^w$，所以这个映射是多对一的，必须限制在一条高度为 2π 的带状区域上，才有反函数。

背后的数学

复对数

在实数线上，ln 是 exp 的反函数。在复平面上，对每一个 $z\neq 0$，方程 $e^w=z$ 都有无穷多个解 w，彼此相差 $2\pi\mathrm{i}$ 的倍数，所以“$\log z$”并不是单独一个数。要得到一个真正的函数，就必须选定一个分支（branch），也就是选定一条像 W_3 或 W_4 那样的带状区域：主

分支（principal branch）取 $-\pi < \arg z \leq \pi$。每一个分支都必然在某处有跳跃，跳跃发生在由原点出发的一条射线上，这条射线称为**分支割线**（branch cut）。黎曼（Bernhard Riemann）在 1851 年的博士论文中提出了解决办法：不再坚持对数函数要定义在一个平面上。它其实定义在一道无穷的螺旋楼梯上，即一个**黎曼曲面**（Riemann surface）；在这个曲面上，对数函数是单值而且光滑的。时至今日，黎曼曲面在几何、数论和弦论中都占有核心地位。

角度保持不变

再看一次图 5.7.1。铅垂和水平的格线相交成直角，它们变成的圆和射线也是如此。这并非巧合：凡是复函数的导数不为零之处，该函数都会保持曲线之间的夹角不变。这样的映射称为**保角**（conformal）映射。在局部范围内，保角映射不过是一次旋转加上一次缩放，而这正是乘以复数 $f'(z)$ 所起的作用。保角映射可用来解决流体流动、静电学和热传导的问题，方法是把一个复杂的区域变换成一个答案已知的简单区域。

麦卡托地图

这里有一个意想不到的发现，把本节与 Section 4.8 连系起来。麦卡托（Gerardus Mercator）在 1569 年绘制的地图，设计的目的是令罗盘方位恒定的航线成为一条直线，这就要求地图保持角度不变。如果你先把地球以球极投影（stereographic projection）的方式（从北极出发）投影到一个平面上，然后取复对数，所得的结果恰好就是麦卡托投影：经线变成铅垂线，纬线变成水平线，正如射线和圆在 log 之下变成水平线和铅垂线一样。把本节的指数映射倒过来运行，就是航海图背后的数学。

思维技巧

把函数想像成变换

当一个函数无法画成图像时，就把它所做的事画出来：取一组由简单曲线组成的格网，看看它们被映到哪里。

对于复函数，水平线和铅垂线的像，或者圆和射线的像，往往已能说明一切。这种着眼于映射如何改变图形的思考方式，贯穿整个几何与分析。

延伸练习

1 **解方程。**求所有满足 $e^w = -1$ 的复数 w，以及所有满足 $e^w = 1 + \mathrm{i}$ 的 w。

2 **正方形的像。**求顶点为 0、1、$1 + \mathrm{i}$ 和 i 的正方形在 $z = e^w$ 之下的像，并画出草图。

3 **直角。**直接证明圆 $|z| = e^c$ 与射线 $\arg z = d$ 相交成直角，并解释为甚么这与直线 $u = c$ 和 $v = d$ 互相垂直是一致的。

4 **平方。**设 c 和 d 为非零实数。求直线 $x = c$ 和 $y = d$ 在映射 $z \mapsto z^2$ 之下的像。证明它们都是抛物线，并且凡相交之处都相交成直角。（若 $c = 0$ 或 $d = 0$，情况会怎样?）

延伸阅读

- **Tristan Needham, *Visual Complex Analysis***（Oxford University Press）。第 2 章和第 4 章以精彩的插图讲解指数函数、对数函数和保角映射。
- **Hilary Priestley, *Introduction to Complex Analysis***（Oxford University Press）。一本清晰易懂的大学本科入门教材。
- **Ian Stewart & David Tall, *Complex Analysis***（Cambridge University Press）。另一本出色的本科入门书，着重几何观点。

5.8 圆盘中隐藏的距离

题目

STEP III 2013，第 6 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/13-s3-q6`

引子

乍看之下，这只是一道运算模（modulus）和共轭（conjugate）的练习，最后得出一条模样古怪的不等式。为甚么会有人关心 $\frac{|z-w|}{|1-zw^*|}$ 呢?

因为它是一个距离。它不是 z 与 w 之间的普通距离，而是另一种几何中的距离：在这种几何里，单位圆盘（unit disc）本身就是一个无限大的完整宇宙，直线是圆弧，三角形的内角和小于 180°。这就是*双曲几何*（*hyperbolic geometry*）。两千年来，人们一直试图证明欧几里得（Euclid）的平行公设（parallel postulate），到了十九世纪才发现这种几何；而本题的不等式，原来正是它的三角不等式（triangle inequality），只是乔装了一番。更令人意外的是，不等式的右边正是狭义相对论（special relativity）中的速度相加公式，我们在 Section 3.8 已经见过。

动笔之前

? **想一想** 怎样从图形看出 $|z-w| \le |z|+|w|$?（把 z 和 $-w$ 想成向量。）

$z-w = z+(-w)$。先把 z 画成由原点出发的箭嘴，再把 $-w$ 画成由 z 的尖端出发的箭嘴。由原点指向终点的箭嘴就是 $z-w$，三支箭嘴组成一个三角形，边长分别为 $|z|$、$|w|$ 和 $|z-w|$。三角形任何一边的长度都不超过另外两边之和。

另外要记得，对任何复数 a 都有 $|a|^2 = aa^*$。这样就可以把题目中每一个模都化为代数运算。

逐步解题

阶段 1　展开各个模

想一想　展开 $|z-w|^2=(z-w)(z^*-w^*)$，再与 $(|z|+|w|)^2$ 比较。

$|z-w|^2=zz^*-zw^*-z^*w+ww^*=|z|^2+|w|^2-zw^*-z^*w$。另一方面，$(|z|+|w|)^2-E=|z|^2+2|z||w|+|w|^2-zw^*-z^*w-2|zw|$，而由于 $|z||w|=|zw|$，两者相等。 ■

$E=(zw^*)+(zw^*)^*+2|zw|=2\,\mathrm{Re}(zw^*)+2|zw|$，这是实数。而且根据三角不等式，$E=(|z|+|w|)^2-|z-w|^2\geq 0$。 ■

阶段 2　第二条恒等式

$|1-zw^*|^2=(1-zw^*)(1-z^*w)=1-zw^*-z^*w+|z|^2|w|^2$，而 $(1+|zw|)^2-E=1+2|zw|+|zw|^2-zw^*-z^*w-2|zw|$，两者相等。 ■

阶段 3　不等式

想一想　设 $A=(|z|+|w|)^2$ 和 $B=(1+|zw|)^2$。把不等式两边平方后，它就是 $\frac{A-E}{B-E}\leq\frac{A}{B}$。这在甚么时候成立?

所求不等式的两边都是非负的，所以它等价于平方后的不等式 $\frac{A-E}{B-E}\leq\frac{A}{B}$。假设 $B-E=|1-zw^*|^2>0$，两边乘开：

$$(A-E)B\leq A(B-E)\iff -EB\leq -EA\iff E(B-A)\geq 0.$$

由于 $E\geq 0$，只要 $B\geq A$，即 $1+|zw|\geq|z|+|w|$，上式便成立。

停一停，试一试　证明 $1+|z||w|-|z|-|w|$ 可以因式分解。它在甚么时候非负?

$1+|z||w|-|z|-|w|=(1-|z|)(1-|w|)$。若 $|z|>1$ 且 $|w|>1$，两个因式都是负数，乘积为正。所以 $B\geq A$，不等式成立。 ■

若改为 $|z|<1$ 且 $|w|<1$，两个因式都是正数，乘积同样为正，同一个论证依然适用：**是的**，不等式在这个情况下也成立。（在这两个情况下都有 $|zw^*|\neq 1$，所以 $1-zw^*\neq 0$，左边有定义。）不等式可能不成立的，是混合的情况：一点在单位圆内，另一点在单位圆外。

回顾

- $|a|^2 = aa^*$。每一条恒等式都是由展开共轭的乘积而来。这是用代数方法处理模的标准做法。
- **归结为一个量的正负。**平方并交叉相乘之后，整条不等式只取决于 $(1-|z|)(1-|w|)$ 的正负。
- **检查其他情况。**最后一问考虑 $|z|, |w| < 1$；同一个因式分解立即给出答案。

背后的数学

非欧几里得几何

两千年来，数学家一直试图由欧几里得的其他公理，证明他的第五公设（过直线外一点，恰好有一条直线与该直线平行）。在 1820 和 1830 年代，罗巴切夫斯基（Nikolai Lobachevsky）、鲍耶（János Bolyai）和（私下研究的）高斯（Gauss）意识到这是办不到的：存在一种完全相容的几何，其中有无限多条平行线。1882 年，庞加莱（Henri Poincaré）在单位圆盘内找到了这种几何的一个优美模型。点就是开圆盘内的点；“直线”是直径，以及与边界圆成直角相交的圆弧；而 z 与 w 之间的距离是

$$d(z,w) = 2\tanh^{-1}\rho(z,w), \qquad \text{其中} \qquad \rho(z,w) = \frac{|z-w|}{|1-zw^*|}.$$

当一点趋近边界圆时，它与圆心的距离趋向无限大：这个圆盘是一整个无限的平面，只是通过一块令影像变形的透镜来观看。

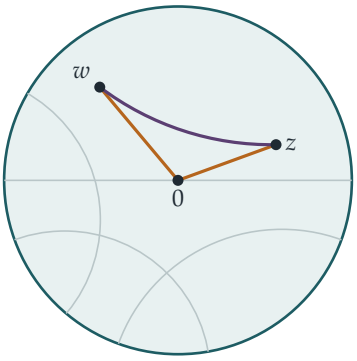


图 5.8.1 庞加莱圆盘（Poincaré disc）。双曲“直线”是直径，或与边界成直角相交的圆弧（灰色）。以 0、z 和 w 为顶点的双曲三角形，有两条边沿着直径（琥珀色），一条边沿着圆弧（紫红色）。题目中的不等式说明，以双曲距离计算，紫红色的边不会比另外两条边加起来更长。

乔装了的三角不等式

对于 $|z|, |w| < 1$，题目所证明的其实是以下结果。设 $|z| = \tanh a$ 和 $|w| = \tanh b$，则由 0 到 z 和 w 的双曲距离分别是 $2a$ 和 $2b$。根据 Section 3.8 中 tanh 的加法公式，

$$\frac{|z| + |w|}{1 + |zw|} = \frac{\tanh a + \tanh b}{1 + \tanh a \tanh b} = \tanh(a + b).$$

所以该不等式就是 $\tanh\big(\frac{1}{2}d(z,w)\big) \le \tanh(a+b)$，即

$$d(z,w) \le 2a + 2b = d(z,0) + d(0,w).$$

这就是双曲距离的三角不等式（图 5.8.1）。相对论的速度相加公式与双曲几何，骨子里是同一回事：狭义相对论中所有可能速度组成的集合，配上适当的距离概念，就是一个双曲空间。我们会在 Section 7.6 再谈这一点。

圆盘的对称

为甚么 ρ 是合适的量？对于 $|a| < 1$，映射 $z \mapsto \dfrac{z-a}{1-a^*z}$ 把单位圆盘映到自身，保持双曲距离不变，并把 a 移到圆心。它们是平移在双曲几何中的对应物，而 $\rho(z,w)$ 正是把 w 移到圆心时的 $|$映射后的$z|$。这些莫比乌斯变换（*Möbius transformation*）是复分析（complex analysis）的核心课题。艾雪（M. C. Escher）的木刻版画《圆极限》（*Circle Limit*），灵感来自数学家考克斯特（H. S. M. Coxeter）的一幅图；画中以全等的鱼和天使镶嵌（tiling）庞加莱圆盘，它们愈近边缘便愈小，因为双曲距离在那里会无限增大。

思维技巧

问一问算式有甚么意义

当某个代数式不断出现时，问一问它是否有某种意义：一个距离、一个角、一个概率、一种对称。

在这里，一个古怪的模之比原来是另一种几何中的距离，而一条古怪的不等式原来是它的三角不等式。认出当中的意义，就能解释结果为何成立，以及它通往何处。

延伸练习

1 **数值验算。**取 $z = \frac{1}{2}$ 和 $w = -\frac{1}{2}$。计算不等式的两边。然后计算双曲距离 $d(z,0)$、$d(0,w)$ 和 $d(z,w)$，并验证三角不等式在这里取等号。为甚么？

2 **混合的情况。**求满足 $|z| < 1 < |w|$ 的 z、w，使不等式不成立。

3 **等距映射（isometry）。**设 $f(z) = \dfrac{z-a}{1-a^*z}$，其中 $|a| < 1$。证明当 $|z| = 1$ 时必有 $|f(z)| = 1$，并且 $f(a) = 0$。证明 $\rho(f(z), f(w)) = \rho(z, w)$。

4 **无限近在咫尺。**证明由 0 到 r（其中 $0 < r < 1$）的双曲距离是 $\ln \frac{1+r}{1-r}$，并且当 $r \to 1$ 时它趋向无限大。一点要多接近边界，才会与圆心相距双曲距离 10?

延伸阅读

- **Caroline Series, “Hyperbolic geometry” lecture notes**（华威大学 University of Warwick，可免费取得）。一份清晰的入门课程讲义，由庞加莱圆盘讲起。
- **Jeremy Gray, *Ideas of Space***（牛津大学出版社 Oxford University Press）。由欧几里得到爱因斯坦（Einstein）的非欧几里得几何史。
- **Tristan Needham, *Visual Complex Analysis***（牛津大学出版社 Oxford University Press）。第 6 章讲述非欧几里得几何和莫比乌斯变换，并配有精美的插图。

我们学到了甚么

本章开首，我们谈到卡尔达诺（Cardano）那个令人不安的发现：要得到实数答案，中途可能非用虚数不可。我们也提出了三个问题：复数（complex number）怎样帮助我们解决实数问题？单位根（roots of unity）有何特别之处？复函数（complex function）是甚么模样？以下是本章八道题目所揭示的答案。

1. 复数怎样帮助我们解决实数问题？

一次又一次，关于实数的事实，取道复平面（complex plane）便有简短的证明。

- **角变成乘积。**梅钦（Machin）的 π 公式，说的其实是 $(5-\mathrm{i})^4(1+\mathrm{i})$ 是 $239-\mathrm{i}$ 的实数倍，只需整数运算便可验证（Section 5.1）。
- **计数变成取平均。**每隔三项取一个二项式系数（binomial coefficient）所得的和，来自在三次单位根处计算 $(1+x)^n$ 的值（Section 5.2）。
- **三角函数的乘积变成多项式的值。**等间隔角的余弦之积，来自把 $z^{2n}+1$ 因式分解，再代入 $z=\mathrm{i}$（Section 5.3）。
- **几何变成代数。**距离变成模（modulus），所以由正 n 边形的一个顶点引出的各条弦，其长度之积是 $|1+1+\cdots+1|=n$（Section 5.4）；旋转变成乘法，所以费马点（Fermat point）的作图化简为恒等式 $1+\alpha^2=\alpha$（Section 5.6）。
- **由棣美弗定理推出无理性。**展开 $(\cos\theta+\mathrm{i}\sin\theta)^k$，再从奇偶性（parity）着手论证，便知 $\cos^{-1}\frac{1}{3}$ 的度数不是有理数；这正是德恩（Dehn）解决希尔伯特（Hilbert）第三问题的关键事实（Section 5.5）。

2. 单位根有何特别之处？

n 次单位根是正 n 边形的顶点，是最简单的多项式 z^n-1 的根，并且在乘法之下构成一个群（group）。这三个事实合起来，解释了它们为何无处不在。由于它们均匀地分布在一个圆上，所以总和为零；Section 5.2 的筛选法由此而来，推而广之就是离散傅里叶变换（discrete Fourier transform）。由于它们是 z^n-1 的根，所以遍取它们的乘积，就是简单多项式的值（Section 5.3、Section 5.4）。又由于它们对乘法封闭，单位根的幂会循环往复，而像 $\cos^{-1}\frac{1}{3}$ 这样的角，欠缺的正是这一点（Section 5.5）。到了第 6 章，单位根会成为代数的核心，并通往伽罗瓦理论（Galois theory）。

3. 复函数是甚么模样?

在 Section 5.7 中，我们把 e^w 想像成一个把带状区域卷成整个平面的映射：它把铅垂线变成圆，把水平线变成射线，并且无穷无尽地一圈圈绕下去，所以对数是多值的。我们看到它保持角度不变，是共形（conformal）映射的第一个例子，也是麦卡托（Mercator）海图背后的原理。在 Section 5.8 中，一个古怪的不等式原来就是双曲几何（hyperbolic geometry）的三角不等式（triangle inequality）；在双曲几何里，单位圆盘是一个无限的平面，直线是圆弧，而距离则按相对论的速度相加公式合成。

4. 复数工具箱

方法	用于何处
相乘时辐角（argument）相加	梅钦公式（5.1）
检查 2π 的倍数	梅钦公式（5.1）
对单位根取平均	每隔三项的系数（5.2）
按根因式分解，然后代入	余弦乘积（5.3）；弦（5.4）
约去一个等于零的因式	5.3；5.4
棣美弗定理（De Moivre's theorem）	倍角公式（5.5）
奇偶性与最小反例（minimal counterexample）	无理角（5.5）
以乘法表示旋转	等边三角形（5.6）
映射网格	指数映射（5.7）
$\lvert a\rvert^2 = aa^*$	庞加莱（Poincaré）圆盘不等式（5.8）

下一步

- 单位根是第 6 章的主角，当中包括 Section 6.1 的分圆多项式（cyclotomic polynomial），以及 Section 6.2 中高斯（Gauss）的正 17 边形作图。
- 双曲几何、tanh 与相对论之间的联系，在 Section 7.6 中有进一步的探讨。
- 在梅钦公式中出现过的高斯整数（Gaussian integer），是 Section 1.6 的主题。

如果你想在入读大学之前再进一步

- **Tristan Needham, *Visual Complex Analysis***（《复分析：可视化方法》）：有史以来最优美的复数著作，图画丰富。由第 1 章读起。
- **Paul Nahin, *An Imaginary Tale***（《虚数的故事》）：$\sqrt{-1}$ 的历史，由卡尔达诺讲到现代物理。

- **Liang-shin Hahn, *Complex Numbers and Geometry***（《复数与几何》）：用复数处理欧几里得几何，程度与本章完全相同。
- **Hilary Priestley, *Introduction to Complex Analysis***（《复分析导论》）：让你初步认识大学的复分析课程。

卡尔达诺认为他的虚数量既精微又无用。五个世纪之后，在数学或物理之中，已经很难找到一个没有它们踪影的角落。本章的启示，正是阿达马（Hadamard）的那句话：当一个关于实数的问题很难时，不妨试试走出实数线。复平面更加广阔，许多看似复杂的事情，在那里原来都很简单。

延伸练习答案

以下只是简短的答案和提示，并非完整解答：只列出结果和关键步骤，方便你核对自己的做法。请先认真做过每道练习，然后才看答案。

5.1 梅钦公式

1. $(5-\mathrm{i})^2(1+\mathrm{i}) = (24-10\mathrm{i})(1+\mathrm{i}) = 34+14\mathrm{i}$，其辐角（argument）为 $\tan^{-1}\frac{7}{17}$。它的辐角同时也是 $-2\tan^{-1}\frac{1}{5}+\frac{\pi}{4}$。
2. $(2+\mathrm{i})(3+\mathrm{i}) = 5+5\mathrm{i}$，辐角为 $\frac{\pi}{4}$。
3. $(3+\mathrm{i})^2(7+\mathrm{i}) = (8+6\mathrm{i})(7+\mathrm{i}) = 50+50\mathrm{i}$。
4. 这些级数（series）都是交错级数，所以误差小于第一个被略去的项。对于 $16\tan^{-1}\frac{1}{5}$，需要被略去的项 $\frac{16}{(2N+1)5^{2N+1}}$ 小于 5×10^{-21}，即要取 $N=14$ 项；对于 $4\tan^{-1}\frac{1}{239}$，取 4 项便足够。用计算机可得 $16\left(\frac{1}{5}-\frac{1}{3\cdot5^3}+\frac{1}{5\cdot5^5}\right)-\frac{4}{239}\approx3.14162$。

5.2 单位根筛选法

1. $(1+1)^n+(1-1)^n = 2\left(\binom{n}{0}+\binom{n}{2}+\cdots\right)$，所以当 $n\ge1$ 时，该和为 2^{n-1}。
2. 先分别乘以 ω^{-1} 和 ω^{-2}，然后取平均：$\frac{1}{3}\left(2^n+\omega^{-1}(1+\omega)^n+\omega^{-2}(1+\omega^2)^n\right)$。由于 $1+\omega=\mathrm{e}^{\mathrm{i}\pi/3}$，上式等于 $\frac{1}{3}\left(2^n+2\cos\frac{(n-2)\pi}{3}\right)$。第三个剩余类（residue）给出 $\frac{1}{3}\left(2^n+2\cos\frac{(n+2)\pi}{3}\right)$，而三个余弦 $\cos x+\cos(x-\frac{2\pi}{3})+\cos(x+\frac{2\pi}{3})$ 之和为 0。
3. 对 $j=0,1,2,3$ 取 $(1+\mathrm{i}^j)^n$ 的平均：该和为 $\frac{1}{4}\left(2^n+(1+\mathrm{i})^n+0+(1-\mathrm{i})^n\right)$，而 $(1\pm\mathrm{i})^n = 2^{n/2}\mathrm{e}^{\pm\mathrm{i}n\pi/4}$。
4. 按元素个数计：$\frac{1}{3}(2^{12}+2\cos4\pi)=1366$。按元素之和计：$\prod(1+\omega^j)$ 中，2、$1+\omega$ 和 $1+\omega^2$ 这三种因子各出现四次，而 $(1+\omega)(1+\omega^2)=1$，所以该乘积等于 16（在 ω^2 处的乘积亦然）。所求的数目为 $\frac{1}{3}(4096+16+16)=1376$。

5.3 因式分解 $z^{2n}+1$

1. 在第 (ii) 部分取 $n=7$，得 $\cos^2\frac{\pi}{14}\cos^2\frac{3\pi}{14}\cos^2\frac{5\pi}{14}=\frac{7}{64}$，而三个余弦均为正数。至于第二个结果，设 $c=\cos\frac{\pi}{7}\cos\frac{2\pi}{7}\cos\frac{4\pi}{7}$。则 $8c\sin\frac{\pi}{7}=\sin\frac{8\pi}{7}=-\sin\frac{\pi}{7}$，所以 $c=-\frac{1}{8}$；另外 $\cos\frac{4\pi}{7}=-\cos\frac{3\pi}{7}$。

2. 在 $z=1$ 处，$2=\prod\left(2-2\cos\frac{(2k-1)\pi}{2n}\right)=\prod 4\sin^2\frac{(2k-1)\pi}{4n}$，所以各正弦（全为正数）的乘积为 $\sqrt{2/4^n}=2^{\frac{1}{2}-n}$。

3. $z^6+1=(z^2-\sqrt{3}z+1)(z^2+1)(z^2+\sqrt{3}z+1)$；首尾两个因式相乘得 z^4-z^2+1。

4. $z^{2n}=1$ 的根是 ±1 以及共轭（conjugate）对 $\mathrm{e}^{\pm\mathrm{i}k\pi/n}$，其中 $1\le k\le n-1$，所以 $\frac{z^{2n}-1}{z^2-1}=\prod_{k=1}^{n-1}\left(z^2-2z\cos\frac{k\pi}{n}+1\right)$。令 $z\to1$：左边趋向 n，右边则为 $\prod 4\sin^2\frac{k\pi}{2n}$。

5.4 正多边形中的距离

1. 正方形：$\sqrt{2}\times2\times\sqrt{2}=4$。三角形：$\sqrt{3}\times\sqrt{3}=3$。

2. 设 $X_k=\omega^k$，并把 $P=r$ 放在正实轴上。则 $\prod|r-\omega^k|=|r^n-1|$。

3. 由算术几何平均不等式（AM–GM），平均值不小于几何平均（geometric mean）$n^{1/(n-1)}$。当 $n=6$ 时，各弦（chord）的长度为 1、$\sqrt{3}$、2、$\sqrt{3}$、1，平均值为 $\frac{4+2\sqrt{3}}{5}\approx1.493$，而 $6^{1/5}\approx1.431$。

4. 由第 (ii) 部分，从每个顶点出发的 $n-1$ 条弦的乘积为 n。对全部 n 个顶点相乘得 n^n，其中每条弦都计算了两次，所以 $\binom{n}{2}$ 条弦的乘积为 $n^{n/2}$。

5.5 名不副实的有理角

1. 取顶点 $(1,1,1)$、$(1,-1,-1)$、$(-1,1,-1)$、$(-1,-1,1)$，中心为 O。两个键向量的点积（dot product）为 -1，长度均为 $\sqrt{3}$，所以键角（bond angle）为 $\cos^{-1}(-\frac{1}{3})$。一个面的向外法向量（outward normal），是指向对面顶点的向量的负向量，所以两个向外法向量之间的夹角同样是 $\cos^{-1}(-\frac{1}{3})$，而两个面之间的夹角是它的补角（supplement），即 $\cos^{-1}\frac{1}{3}$。

2. 设 $(4+3\mathrm{i})^k=a_k+\mathrm{i}b_k$。由于 $4+3\mathrm{i}$ 及其共轭是 $t^2-8t+25$ 的根，故 $b_{k+1}=8b_k-25b_{k-1}\equiv3b_k\pmod5$，而 $b_1=3$，所以 $b_k\equiv3^k\not\equiv0\pmod5$。因此没有任何一个幂是实数。假如该角的度数是有理数（rational number），它的某个倍数便会是 180° 的倍数，相应的幂就会是实数。

3. 当 $a=2$ 时，$a^2-1=3$ 是奇数，所以奇偶性（parity）论证不成立。事实上 $\cos^{-1}\frac{1}{2}=60^\circ$ 的确是有理数；取 $k=3$ 时，正弦的括号部分为 $3-\binom{3}{3}\cdot3=0$。

4. 0°、60° 和 90°，余弦值分别为 1、$\frac{1}{2}$ 和 0。尼云定理（Niven's theorem）指出：在 180° 的有理数倍处，cos 的有理数值只有 0、$\pm\frac{1}{2}$ 和 ±1。

5.6 费马点

1. 向内作时，各顶点为 $A+(1-\bar{\alpha})(B-A)$，余此类推。在 $p=2$、$q=1.4$、$r=1.1$ 的例子中，通过 C 和新的 L 的直线并不经过 O（检验条件 $\mathrm{Im}(\bar{C}L')=0$ 不成立），

另外两条直线也是如此。

2. 等边三角形（equilateral triangle）的中心就是它的形心（centroid）。设 $L=\alpha(p+q)$、$M=-(q+r)$ 和 $N=-\alpha^2(p+r)$，则 ABL、BCM、CAN 的中心 G_1、G_2、G_3 满足 $G_3-G_1=\alpha(G_2-G_1)$（利用 $\alpha^2=\alpha-1$ 展开），这说明 $G_1G_2G_3$ 是等边三角形。在该例子中，它的边长为 $\frac{p+q+r}{\sqrt{3}}\approx 2.598$。

3. $PA+PB+PC$ 在 A 处约为 5.68，在 B 处约为 5.13，在 C 处约为 4.89，在形心处约为 4.54；全部都大于 4.5。

4. 等边（两种定向（orientation）之一）即是 $z_1+\omega z_2+\omega^2 z_3=0$ 或 $z_1+\omega^2 z_2+\omega z_3=0$，其中 $\omega=\mathrm{e}^{2\pi\mathrm{i}/3}$。利用 $\omega+\omega^2=-1$，两式左边的乘积为 $z_1^2+z_2^2+z_3^2-z_1z_2-z_2z_3-z_3z_1$。

5.7 作为映射的复指数函数

1. 当 $w=(2k+1)\pi\mathrm{i}$ 时 $\mathrm{e}^w=-1$；当 $w=\frac{1}{2}\ln 2+\mathrm{i}\left(\frac{\pi}{4}+2k\pi\right)$ 时 $\mathrm{e}^w=1+\mathrm{i}$，其中 k 为整数。

2. 区域 $1\le|z|\le\mathrm{e}$、$0\le\arg z\le 1$（一弧度）：圆环（annulus）的一部分。底边映到实轴上的 $[1,\mathrm{e}]$，右边映到半径为 e 的一段圆弧，顶边映到角度为 1 的一段径向线段，左边映到单位圆的一段圆弧。

3. 射线沿半径方向；圆的切线则与半径垂直。这与直线 $u=c$ 和 $v=d$ 的情况吻合，因为 e^w 保持角度不变（它的导数永不为 0）。

4. $(x+\mathrm{i}y)^2=x^2-y^2+2\mathrm{i}xy$。直线 $x=c$ 映到 $u=c^2-\frac{v^2}{4c^2}$，而 $y=d$ 映到 $u=\frac{v^2}{4d^2}-d^2$：分别是开口向左和向右的抛物线（parabola），两者的焦点（focus）都在 0。令两式相等，得 $v^2=4c^2d^2$，所以它们相交两次，交点为 $u=c^2-d^2$、$v=\pm 2cd$。在 $v=2cd$ 处（即两直线交点 $c+\mathrm{i}d$ 的像），斜率 $\frac{\mathrm{d}u}{\mathrm{d}v}$ 分别为 $-\frac{d}{c}$ 和 $\frac{c}{d}$；在 $v=-2cd$ 处则为 $\frac{d}{c}$ 和 $-\frac{c}{d}$。两种情况下乘积都是 -1。（第二个交点既是第一条直线上 $c-\mathrm{i}d$ 的像，也是第二条直线上 $-c+\mathrm{i}d$ 的像：平方把 z 和 $-z$ 映到同一点。）若 $c=0$，$x=0$ 的像是射线 $u\le 0$、$v=0$；若 $d=0$，$y=0$ 的像是射线 $u\ge 0$、$v=0$：这些都不是抛物线。只要两直线的交点不在 0，射线仍会与另一个像成直角相交；但若 $c=d=0$，两个像都落在实轴上：在 0 处这个映射把角度加倍（它的导数 $2z$ 在该处为零），所以两轴之间的直角变成了平角。

5.8 圆盘中隐藏的距离

1. 两边都等于 $\frac{1}{5/4}=\frac{4}{5}$。于是 $d(z,0)=d(0,w)=2\tanh^{-1}\frac{1}{2}=\ln 3$，而 $d(z,w)=2\tanh^{-1}\frac{4}{5}=\ln 9$，所以等号成立：0 位于由 z 到 w 的双曲线段（hyperbolic line segment，这里是一条直径）上。

2. 由第 (i) 和第 (ii) 部分，当 $E > 0$ 时，该不等式等价于 $|z| + |w| \le 1 + |zw|$，即 $(1 - |z|)(1 - |w|) \ge 0$。在混合的情况下这并不成立，例如取 $z = \frac{1}{2}$、$w = 3$：左边是 $\frac{5/2}{1/2} = 5$，右边是 $\frac{7/2}{5/2} = \frac{7}{5}$。

3. 若 $|z| = 1$，则 $|1 - a^* z| = |\bar{z} - a^*| = |z - a|$。直接计算，$f(z) - f(w) = \frac{(1-|a|^2)(z-w)}{(1-a^*z)(1-a^*w)}$ 及 $1 - f(z)f(w)^* = \frac{(1-|a|^2)(1-zw^*)}{(1-a^*z)(1-aw^*)}$；两式相除再取模（modulus），得 $\rho(f(z), f(w)) = \rho(z, w)$。

4. $d(0, r) = 2\tanh^{-1} r = \ln \frac{1+r}{1-r}$，当 $r \to 1$ 时趋向 ∞。距离为 10 需要 $r = \tanh 5 \approx 0.99991$：离边界只有约 9×10^{-5}。

章

6

代数与结构

代数中每一个巧妙的技巧背后都有一个结构，而理解这个结构，比掌握多少技巧都更有价值。

本章内容

6.1	以单位根为根的多项式	STEP III 2010
6.2	高斯的正十七边形	STEP II 1991
6.3	以 $\sqrt{2}+\sqrt{3}+\sqrt{5}$ 为根的多项式	STEP II 2023
6.4	解四次方程	STEP III 2017
6.5	从公理出发的微分法	STEP II 1992
6.6	子群的并集	STEP II 1988
6.7	四元数群	STEP III 1993
6.8	由奇异矩阵组成的群	STEP III 1996
A	练习答案	第 429 页

从方程到结构

数学是给不同事物起相同名字的艺术。

— 庞加莱（Henri Poincaré）

在历史上的大部分时间里，代数就是解方程。四千年前，巴比伦人已懂得解二次方程；十六世纪的意大利数学家解决了三次和四次方程；其后近三百年间，五次方程却令一切尝试无功而还。到了 1820 至 1830 年代，问题终于有了答案，而这个答案彻底改变了代数。阿贝尔（Niels Henrik Abel）证明了五次方程不存在一般的求根公式，伽罗瓦（Évariste Galois）则确切说明了哪些方程可以求解，他研究的不是方程本身，而是*方程各根的对称性*。为此，他必须创造一种全新的数学对象：群（group）。

从计算答案转向研究结构，这个转变正是本章的主题。现代代数问的不再是“解是甚么?”，而是“这是哪一类对象？它容许哪些运算?”这些问题的答案，既解释了旧有技巧为何奏效，也解决了单靠技巧永远无法解决的问题。

结构为何重要

以下三个古老的问题，两千年来令一切尝试都无功而还：只用直尺和圆规，作出体积为已知立方体两倍的立方体；三等分任意角；作出正 7 边形。三者都不可能做到，而在十九世纪找到的证明，靠的并不是巧妙的作图，而是关于结构的论证：所讨论的是尺规所能到达的数所组成的域（field）的*次数*（degree）（Section 6.3）。同一套理论也解释了为何正 17 边形*可以*作出（Section 6.2），以及为何四次方程可解而五次方程不可解（Section 6.4）。

结构还可以搬到别处去。晶体的对称、多项式各根的置换（permutation）、三维空间的旋转、扭计骰的各种转法，以及椭圆曲线（elliptic curve）上的点，全都满足群的公理（axiom）。只要从公理出发把一条定理证明一次，它便在以上所有情况中成立。这种简约，即庞加莱所说的给“不同事物起相同名字”，正是抽象代数（abstract algebra）如此强大的原因。

你在大学会遇到甚么

在一年级，你会遇到*群*（带有一个运算并满足三条公理的集合），通常还有*环*（ring）和*域*（带有加法和乘法的集合）。你会证明拉格朗日定理（Lagrange's theorem），即子群（subgroup）的大小整除群的大小，并认识置换、对称群（symmetry group）和矩阵群（matrix group）。到了二、三年级，这些内容会引向*伽罗瓦理论*（Galois theory），它把域

和群联系起来，回答有关方程和作图的经典问题；引向表示论（representation theory），它通过矩阵来研究群；还会引向代数数论（algebraic number theory），Section 1.7 中的想法在那里会得到严格的基础。本章最后三节讲的是群。群已不在 STEP 的课程范围之内（已于 2019 年删去），但这些题目本身自成一体，每一节都会给出所需的定义。

贯穿本章的问题

三个大问题

1. **甚么可以由甚么构造出来？**从有理数出发，通过开平方根或解方程，可以到达哪些数？又怎样证明某个数无法到达？
2. **方程的公式为何奏效？**二次、三次和四次方程都有求根公式，背后隐藏的原因是甚么？到了五次，又有甚么改变了？
3. **抽象化带来甚么？**如果我们只凭对象所遵守的规则来描述它们，会有甚么结果？

本章的想法

§	课题	想法
6.1	分圆多项式（cyclotomic polynomials）	按阶（order）把单位根（roots of unity）分类
6.2	正 17 边形	高斯周期（Gauss periods）：和与积给出二次方程
6.3	最小多项式（minimal polynomials）	消去根式；次数相乘
6.4	解四次方程	对称函数（symmetric functions）；三次预解式（resolvent cubic）
6.5	从公理出发的微分法	由规则决定的运算
6.6	子群的并集（union）	从群公理出发作论证
6.7	四元数群（quaternion group）	用矩阵表示运算
6.8	由奇异矩阵（singular matrices）组成的群	按字面理解定义

如何阅读本章

这些题目当中，有不少可以靠蛮力完成：把所有式子展开、比较系数、逐一检查每种情况。如有需要，不妨先这样做。然后回头想一想，是甚么令蛮力行得通，背后有没有结构上的原因。结构上的原因通常较为简短，总是更具启发性，而当蛮力再也行不通时，你需

要的正是它。

6.1 以单位根为根的多项式

题目

STEP III 2010，第 3 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/10-s3-q3`

引子

n 次单位根（roots of unity），即 $z^n = 1$ 的解，位于正 n 边形的各个顶点上。但它们并非全都一样。例如在 12 次单位根之中，1 本身已是一次单位根，-1 是二次单位根，i 是四次单位根，而十二个之中只有四个，即本原（primitive）的那些，才需要走足十二步方能回到 1。

按所需的步数把单位根分类，再由每一类构造一个多项式，便得到分圆多项式（cyclotomic polynomials，此名源自希腊文，意即“分割圆周”）。它们是 $z^n - 1$ 的基本构件，正如质数是整数的基本构件一样；高斯（Gauss）的正 17 边形作图法（Section 6.2）、单位根筛选法（Section 5.2）以及现代数论中的不少内容，背后都有它们的身影。

动笔之前

想一想 每个 n 次单位根都满足 $z^n = 1$。但使它等于 1 的最小正整数次幂，可能小于 n。对于四个 4 次单位根 1、i、-1、$-\mathrm{i}$，这个最小的幂次分别是多少？

$1^1 = 1$，$(-1)^2 = 1$，而 $\mathrm{i}^4 = (-\mathrm{i})^4 = 1$，且没有更小的幂次可行。所以最小的幂次依次是 1、4、2、4。这个最小的幂次称为该单位根的阶（order），而本原 n 次单位根正是阶为 n 的那些单位根。整道题的关键事实是：每个单位根都恰好有一个阶。

逐步解题

阶段 1　最初的几个

本原 4 次单位根是 i 和 $-\mathrm{i}$，所以 $\mathrm{C}_4(x) = (x-\mathrm{i})(x+\mathrm{i}) = x^2+1$。 ■

> **想一想**　哪些单位根的阶是 1、2、3、5 和 6？（画出 6 次单位根，并按阶把它们分类。）

阶为 1：只有 1，所以 $\mathrm{C}_1(x) = x-1$。阶为 2：只有 -1，所以 $\mathrm{C}_2(x) = x+1$。

阶为 3：除 1 以外的三次单位根，即 $\frac{x^3-1}{x-1}$ 的根。所以 $\mathrm{C}_3(x) = x^2+x+1$。

阶为 5：除 1 以外的五次单位根，它们是 $\frac{x^5-1}{x-1}$ 的根，所以 $\mathrm{C}_5(x) = x^4+x^3+x^2+x+1$。

阶为 6：在六次单位根 $e^{k\pi\mathrm{i}/3}$，$k = 0,\dots,5$ 之中，$k = 0$、3 的阶分别是 1、2，而 $k = 2$、4 的阶是 3。余下的是 $e^{\pm\pi\mathrm{i}/3}$，所以

$$\mathrm{C}_6(x) = (x-e^{\mathrm{i}\pi/3})(x-e^{-\mathrm{i}\pi/3}) = x^2 - 2x\cos\tfrac{\pi}{3} + 1 = x^2 - x + 1. \qquad \blacksquare$$

阶段 2　哪个 n 给出 x^4+1？

$x^4+1=0$ 的根满足 $x^4=-1$，所以 $x^8=1$ 但 $x^4\neq 1$。它们的阶整除 8 但不整除 4，所以阶是 8。反过来，阶为 8 的单位根满足 $x^8=1$，$x^4\neq 1$，所以 $x^4=-1$。本原 8 次单位根有四个，x^4+1 的根也有四个，所以 $n=8$。 ■

阶段 3　质数

若 p 是质数，则 p 次单位根的阶整除 p，所以阶是 1 或 p。因此除 1 以外的每个 p 次单位根都是本原的，而

$$\mathrm{C}_p(x) = \frac{x^p-1}{x-1} = x^{p-1}+x^{p-2}+\cdots+x+1. \qquad \blacksquare$$

阶段 4　没有分圆多项式是另外两个分圆多项式之积

> **停一停，试一试**　假设 $\mathrm{C}_q = \mathrm{C}_r\mathrm{C}_s$。取 C_r 的一个根。它的阶是多少？这条恒等式对它的阶有何说法？由此必然得出甚么？

C_r 的每个根（至少有一个，因为 $e^{2\pi\mathrm{i}/r}$ 是本原的）都是 $\mathrm{C}_r\mathrm{C}_s = \mathrm{C}_q$ 的根，所以它的阶是 q。但它的阶同时也是 r，而一个单位根只有一个阶，所以 $r=q$。于是 $\mathrm{C}_q = \mathrm{C}_q\,\mathrm{C}_s$，又因为 C_q 不是零多项式，所以 $\mathrm{C}_s = 1$。这是不可能的，因为 C_s 至少有一个根 $e^{2\pi\mathrm{i}/s}$。所以这样的 q、r、s 并不存在。 ■

回顾

- **按阶分类。**每个单位根都有唯一的阶，按阶分组便把 $z^n - 1$ 的根分成互不相交的类。
- **用根来定义多项式。**每个 C_n 都是借着确切指明哪些数是它的根而确定下来的。
- **唯一性带来不可能性。**第 (iv) 部分只需要一个想法：一个根不可能有两个不同的阶。

背后的数学

分解 $x^n - 1$

每个 n 次单位根都有某个整除 n 的阶 d；而当 $d \mid n$ 时，每个阶为 d 的 d 次单位根都是 n 次单位根。所以 $x^n - 1$ 的根按阶分成若干类，由此得到

$$x^n - 1 = \prod_{d \mid n} C_d(x).$$

例如 $x^6 - 1 = C_1C_2C_3C_6 = (x-1)(x+1)(x^2+x+1)(x^2-x+1)$。图 6.1.1 展示了按这种方式分类的十二个 12 次单位根。本原 n 次单位根的个数，亦即 C_n 的次数，等于 1 至 n 之间与 n 没有公因数的 k 的个数，也就是欧拉函数（Euler's totient function）$\varphi(n)$。比较两边的次数，便得到恒等式 $\sum_{d \mid n} \varphi(d) = n$。

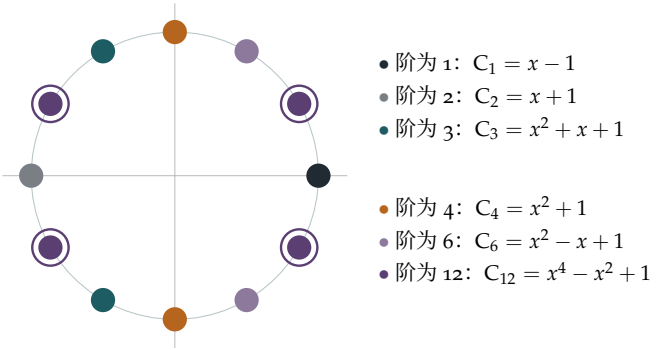


图 6.1.1 十二个 12 次单位根，按阶着色。四个本原 12 次单位根（以圆圈标示）是 C_{12} 的根。各类合起来给出 $x^{12} - 1 = C_1C_2C_3C_4C_6C_{12}$，次数为 $1+1+2+2+2+4 = 12$。

整数系数，以及一个意外

单看定义并不明显，但每个分圆多项式的系数都是整数。（利用 x^n-1 的因式分解和归纳法（induction）：C_n 等于 x^n-1 除以若干个较早的分圆多项式之积，而以首一（monic）整系数多项式对整系数多项式作长除法，所得系数仍是整数。）在很长一段时间里，所有已知例子的系数都只有 0、1 和 -1。第一个例外是 C_{105}，它含有 $-2x^7$ 这一项，由 Migotti 于 1883 年发现；$105=3\times5\times7$ 是具有三个相异奇质因数的最小的数。事实上，分圆多项式的系数可以要多大有多大。

不可约性

第 (iv) 部分证明了没有分圆多项式是另外两个分圆多项式之积。其实更强的结论也成立：C_n 根本不能分解成次数较低的有理系数多项式之积。高斯对 n 为质数的情况证明了这一点（它可由艾森斯坦判别法（Eisenstein's criterion）推出，你会在大学的第一门代数课中遇到），而克罗内克（Kronecker）和戴德金（Dedekind）则证明了一般 n 的情况。正是这种*不可约性*（irreducibility），使分圆多项式成为与单位根有关的多项式之中的“质数”，它也是单位根的伽罗瓦理论（Galois theory）的起点。

思维技巧

先分类，后构造

用一个定义明确的不变量（invariant），把问题中的对象分成若干类，然后逐类建立结构。

这里的不变量是单位根的阶。由于每个根恰好有一个阶，各类互不重叠，于是 x^n-1 可分解成遍及各类的乘积。同样的模式（由唯一的不变量得出不相交的分解）也是质因数分解和有限阿贝尔群（finite abelian groups）分类的基础。

延伸练习

1. **检查因式分解。**验证 $(x-1)(x+1)(x^2+x+1)(x^2-x+1)=x^6-1$，并验证 $x^{12}-1$ 的因式分解如图 6.1.1 所示。
2. **质数的两倍。**证明若 p 是奇质数，则 $C_{2p}(x)=C_p(-x)$。就 $p=3$ 验证这个结果。
3. **欧拉函数。**证明对于质数 p，有 $\varphi(p^k)=p^k-p^{k-1}$，并且 $C_{p^k}(x)=C_p\left(x^{p^{k-1}}\right)$。利用这个结果求 C_9。

4 较长的一题。 较难利用 $x^{15}-1=\mathrm{C}_1\mathrm{C}_3\mathrm{C}_5\mathrm{C}_{15}$，证明 $\mathrm{C}_{15}(x)=x^8-x^7+x^5-x^4+x^3-x+1$。

延伸阅读

- **Ian Stewart, *Galois Theory***（CRC Press）。大学本科的标准入门书；分圆多项式和正多边形在书中很早便出现。
- **Harold Davenport, *The Higher Arithmetic***（Cambridge University Press）。内容包括单位根及其算术性质。
- **Marcus du Sautoy, *Finding Moonshine***（Fourth Estate）。一本讲述对称与群论的科普读物，当中包括伽罗瓦的故事。

6.2 高斯的正十七边形

题目

STEP II 1991，第 4 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/91-s2-q4`

引子

1796 年 3 月 30 日，哥廷根（Göttingen）一名十八岁的学生在他的数学日记写下了第一则记录，这本日记他此后写了一辈子。他发现了怎样用尺规作图（ruler and compasses construction）作出正 17 边形。古希腊人早已懂得作出边数为 3、4、5 和 15 的正多边形（regular polygon），以及边数为这些数目反复加倍的正多边形，但其后两千年间，再没有人添上任何一个新的。高斯（Carl Friedrich Gauss）后来说，正是这项发现令他决心成为数学家，而不是语文学家；相传他还曾要求在自己的墓碑上刻一个正 17 边形。

这个作图法的核心，就是本题的后半部分：四个特定余弦之和等于 $\frac{\sqrt{17}-1}{4}$，这个数只涉及一个平方根。由平方根构成的长度都可以用尺规作出，而高斯再用三个同类的步骤，就能作出 $\cos\frac{2\pi}{17}$ 本身。

动笔之前

? **想一想** 为甚么一道关于正 17 边形的题目，第一部分却是关于 $\cos\frac{2\pi}{5}$ 的?

那是热身，把同一个想法用在较小的多边形上。对正五边形（pentagon）而言，和 $\cos\phi + \cos 2\phi$ 满足一条有理系数的二次方程（quadratic），所以解一条二次方程就能求出它。对正 17 边形而言，四个余弦之和 z 原来也满足一条二次方程。两种情况的窍门都是把余弦相乘，再用 $\cos A\cos B = \frac{1}{2}(\cos(A+B)+\cos(A-B))$ 把乘积变回同一批角的余弦之和。

逐步解题

阶段 1　正五边形

想一想　展开 $y^2=(\cos\phi+\cos 2\phi)^2$，把每个乘积化为和，再利用 $5\phi=2\pi$ 化简 3ϕ 和 4ϕ 这类角。

$y^2=\cos^2\phi+2\cos\phi\cos 2\phi+\cos^2 2\phi=\frac12(1+\cos 2\phi)+(\cos 3\phi+\cos\phi)+\frac12(1+\cos 4\phi)$.

由于 $5\phi=2\pi$，有 $\cos 3\phi=\cos(2\pi-2\phi)=\cos 2\phi$ 和 $\cos 4\phi=\cos\phi$。所以

$$y^2=1+\tfrac12(\cos\phi+\cos 2\phi)+(\cos\phi+\cos 2\phi)=1+\tfrac32 y,$$

即 $2y^2=3y+2$。　■方程 $2y^2-3y-2=(2y+1)(y-2)=0$ 的根是 $y=2$ 和 $y=-\frac12$。但 $y=2$ 需要两个余弦都等于 1，而事实并非如此，所以 $y=-\frac12$。　■

阶段 2　十七次单位根之和为零

$\sum_{k=0}^{16}\cos k\theta$ 是 $\sum_{k=0}^{16}\omega^k$ 的实部(real part),其中 $\omega=e^{\mathrm{i}\theta}$ 是一个 17 次本原单位根(primitive root of unity)。这个等比级数（geometric series）等于 $\frac{\omega^{17}-1}{\omega-1}=0$。　■

所以 $\sum_{k=1}^{16}\cos k\theta=-1$。由于 $\cos(17-k)\theta=\cos k\theta$，各项可以两两配对，于是

$$\sum_{k=1}^{8}\cos k\theta=-\tfrac12.$$

阶段 3　把余弦分成两组

想一想　设 $z=\cos\theta+\cos 2\theta+\cos 4\theta+\cos 8\theta$，并设 z' 为其余四个余弦之和 $\cos 3\theta+\cos 5\theta+\cos 6\theta+\cos 7\theta$。$z+z'$ 是多少?

由第 2 阶段，$z+z'=-\frac12$。如果还能求出 zz'，那么 z 和 z' 就是某条二次方程的两个根。

停一停，试一试　计算 zz'。把它展开成 16 个乘积 $\cos a\theta\cos b\theta$，把每个乘积化为两个余弦之和，然后数一数 $\cos\theta,\cos 2\theta,\ldots,\cos 8\theta$ 各出现多少次。（利用 $\cos k\theta=\cos(17-k)\theta$ 和 $\cos(-k\theta)=\cos k\theta$ 化简各角。）

每个乘积给出 $\cos a\theta\cos b\theta=\frac12(\cos(a+b)\theta+\cos(a-b)\theta)$，所以 zz' 是 32 个余弦之和的 $\frac12$ 倍。逐一考虑 $a\in\{1,2,4,8\}$ 和 $b\in\{3,5,6,7\}$ 的 16 对组合：例如 $a=1$、$b=3$ 给出 $\cos 4\theta$ 和 $\cos 2\theta$；$a=8$、$b=7$ 给出 $\cos 15\theta=\cos 2\theta$ 和 $\cos\theta$。对全部十六对都这样

做，$\cos\theta,\ldots,\cos 8\theta$ 每一个都恰好出现四次。因此

$$zz' = \tfrac{1}{2}\times 4\times\sum_{k=1}^{8}\cos k\theta = 2\times\left(-\tfrac{1}{2}\right) = -1.$$

阶段 4　二次方程

所以 z 和 z' 是 $t^2+\frac{1}{2}t-1=0$ 的两个根，即 $4t^2+2t-4=0$，其根为 $t=\frac{-1\pm\sqrt{17}}{4}$。哪一个是 z？按数值计算，$\cos\theta\approx 0.932$、$\cos 2\theta\approx 0.739$、$\cos 4\theta\approx 0.092$ 和 $\cos 8\theta\approx -0.983$，所以 $z\approx 0.78>0$。因此

$$z=\frac{-1+\sqrt{17}}{4}=-\frac{1-\sqrt{17}}{4}.$$ ■

回顾

- **积化和。**余弦的积化和公式把 y^2 和 zz' 变成同一批余弦之和，于是代数运算得以自成一体。
- **和与积给出二次方程。**知道 $z+z'$ 和 zz'，就等于知道以 z 和 z' 为根的二次方程。
- **分组要分得对。**把八个余弦分成 $\{1,2,4,8\}$ 和 $\{3,5,6,7\}$，令每个指标在 zz' 中出现的次数相同。随意分组是行不通的。

背后的数学

高斯周期

为甚么是 $\{1,2,4,8\}$？以 17 为模（modulo），3 的幂 $3^0,3^1,3^2,\ldots$ 会走遍所有非零剩余（residue）：

$$1,3,9,10,13,5,15,11,16,14,8,7,4,12,2,6.$$

每隔一个取一个，即 $3^0,3^2,3^4,\ldots$，得出 1、9、13、15、16、8、4、2，以 17 为模就是 ±1、±2、±4、±8：正好是 z 中的角。另一半给出 $\pm3,\pm5,\pm6,\pm7$，即 z' 中的角。高斯把 ω^k 在这类剩余类上的和称为**周期**（periods）。由于 $17-1=16=2^4$，这些剩余可以对分四次，而在每一阶段，周期都满足一些二次方程，其系数涉及上一阶段的周期。四条二次方程，也就是四个平方根，就从有理数一路通往 $\cos\frac{2\pi}{17}$：

$$\cos\tfrac{2\pi}{17}=\tfrac{1}{16}\Big(-1+\sqrt{17}+\sqrt{34-2\sqrt{17}}\\ +2\sqrt{17+3\sqrt{17}-\sqrt{34-2\sqrt{17}}-2\sqrt{34+2\sqrt{17}}}\Big).$$

每一步都只涉及平方根，而平方根可以用尺规作出（利用相似三角形或半圆），所以正 17 边形是可作图（constructible）的。

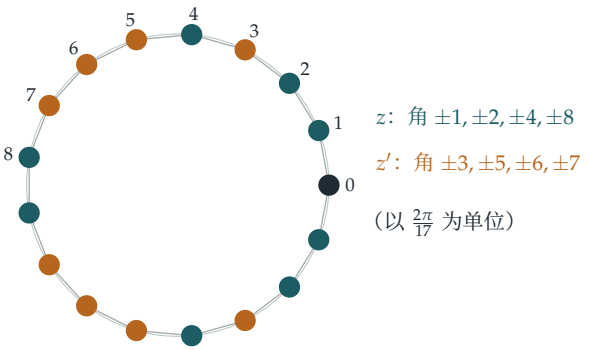


图 6.2.1　正 17 边形的顶点，按高斯的两个周期分成两组。蓝绿色顶点位于角 $\pm 1, \pm 2, \pm 4, \pm 8$（以 $\frac{2\pi}{17}$ 为单位），它们的 x 坐标之和是 $2z = \frac{-1+\sqrt{17}}{2}$；琥珀色顶点则给出 $2z' = \frac{-1-\sqrt{17}}{2}$。

哪些正多边形可以作图？

高斯断言，而 Pierre Wantzel 在 1837 年证明：正 n 边形可以用尺规作出，当且仅当 n 是 2 的幂乘以若干个互不相同的*费马质数*（Fermat primes）之积，费马质数即形如 $2^{2^m}+1$ 的质数。已知的费马质数有 3、5、17、257 和 65537，它们是 Section 1.1 中的费马数。（费马（Fermat）猜想所有这类数都是质数，但欧拉（Euler）证明了 $2^{32}+1$ 可被 641 整除，而此后再没有人找到其他费马质数。）所以正 257 边形和正 65537 边形都可以作图；十九世纪时，Johann Hermes 花了大约十年时间推算出正 65537 边形的作图法，他的手稿现存于哥廷根。正 7 边形和正 11 边形不可作图，因为 7 和 11 不是费马质数（$7-1=6$ 和 $11-1=10$ 都不是 2 的幂）；正 9 边形也不可作图，因为 $9=3^2$ 把费马质数 3 重复了。Wantzel 的证明用到 Section 6.3 中关于域次数（field degrees）的理论。

思维技巧

计算和与积来求出二次方程

要求出一个量，可以为它配上一个自然的伙伴，计算两者的和与积，再解由此得出的二次方程。

个中的艺术在于选对伙伴，使和与积都算得出来。这里的伙伴来自以 17 为模的剩余的对称性，而把这个想法重复四次，就得出高斯的作图法。

延伸练习

1 **正七边形（heptagon）。**试做 STEP II 1995，第 6 题（`step.maths.org/questions/95-s2-q6`）。这道题把高斯将单位根分组的想法用于七次单位根，并求出余弦之和与正弦之和的准确值。判断正弦之和的正负号是当中最微妙的一步。

2 **正五边形。**证明 $\cos\frac{2\pi}{5} = \frac{\sqrt{5}-1}{4}$，并由此描述一个正五边形的尺规作图法。

3 **另一个周期。**用数值验证 $z' = \cos 3\theta + \cos 5\theta + \cos 6\theta + \cos 7\theta = \frac{-1-\sqrt{17}}{4}$。

4 **生成元（generators）。**证明以 17 为模时，3 的幂给出所有非零剩余，但 2 的幂却不然。2 的幂给出哪些剩余?

延伸阅读

- **Ian Stewart, *Galois Theory***（CRC Press）。当中有一章讲述尺规作图和正 17 边形。
- **Richard Courant & Herbert Robbins, *What Is Mathematics?***（Oxford University Press）。第 3 章以浅白优美的笔触介绍作图问题，包括正 17 边形。
- **Daniel Kehlmann, *Measuring the World***（Quercus）。一本关于高斯和洪堡（Alexander von Humboldt）的诙谐小说，可作轻松读物。

6.3 以 $\sqrt{2}+\sqrt{3}+\sqrt{5}$ 为根的多项式

题目

STEP II 2023，第 4 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2023 年 STEP 试卷（zip 文件）内的 STEP II 试卷：

`www.ocr.org.uk/Images/707102-2023-past-papers.zip`

引子

$\sqrt{2}$ 是 x^2-2 的根，而这是一个整数系数多项式（polynomial）。同样，$\sqrt{3}$ 是 x^2-3 的根。那么 $\sqrt{2}+\sqrt{3}$ 是否也是某个整数系数多项式的根？$\sqrt{2}+\sqrt{3}+\sqrt{5}$ 又如何？$\sqrt[3]{2}+\sqrt[3]{3}$ 呢？

凡是整数系数多项式的根，都称为*代数数*（algebraic number）。这道题通过亲手消元（elimination），说明代数数之和仍是代数数。计算所得的次数 4、8、6、9 并非偶然：它们量度每个数有多“复杂”，而这种量度，即*域扩张的次数*（degree of a field extension），正是最终解决古希腊倍立方与三等分角问题的关键。

动笔之前

想一想 要找一个以 $\sqrt{2}$ 为根的多项式，你会把 $x=\sqrt{2}$ 平方，从而消去平方根。若有两个不同的平方根，怎样才能逐一消去？

先把其中一个平方根单独移到一边再平方，它便会消失；然后把另一个平方根单独移到一边，再平方一次。每次平方都可能令次数加倍。整道题的策略就是如此：逐一消去根式（radical）。

逐步解题

阶段 1　两个平方根

若 $(x-\sqrt{2})^2=3$，则 $x^2-2\sqrt{2}\,x+2=3$，所以 $x^2-1=2\sqrt{2}\,x$。平方便可消去余下的平方根：$(x^2-1)^2=8x^2$，即

$$x^4-10x^2+1=0. \qquad \blacksquare$$

而 $x=\sqrt{2}+\sqrt{3}$ 满足 $(x-\sqrt{2})^2=(\sqrt{3})^2=3$，所以 $\mathrm{f}(\sqrt{2}+\sqrt{3})=0$。 $\blacksquare$

阶段 2　三个平方根

想一想　设 $x=\sqrt{2}+\sqrt{3}+\sqrt{5}$，则 $x-\sqrt{5}=\sqrt{2}+\sqrt{3}$ 是 f 的根。把 $\mathrm{f}(x-\sqrt{5})=0$ 写成 $P(x)+\sqrt{5}\,Q(x)=0$ 的形式。之后怎样消去 $\sqrt{5}$?

利用 $(x-\sqrt{5})^2=x^2+5-2\sqrt{5}x$ 和 $(x-\sqrt{5})^4=\left(x^2+5-2\sqrt{5}x\right)^2=x^4+30x^2+25-4\sqrt{5}\,x(x^2+5)$ 展开：

$$\mathrm{f}(x-\sqrt{5})=\left(x^4+30x^2+25\right)-4\sqrt{5}\,x(x^2+5)-10\left(x^2+5\right)+20\sqrt{5}\,x+1.$$

合并同类项，得 $P(x)=x^4+20x^2-24$ 和 $Q(x)=-4x^3-20x+20x=-4x^3$。所以 $x^4+20x^2-24=4\sqrt{5}\,x^3$。平方得

$$\left(x^4+20x^2-24\right)^2=80x^6,$$

展开后得

$$\mathrm{g}(x)=x^8-40x^6+352x^4-960x^2+576. \qquad \blacksquare$$

（验算：$(x^4+20x^2-24)^2=x^8+40x^6+352x^4-960x^2+576$，其中用到 $400-48=352$；再减去 $80x^6$ 便得 g。）

阶段 3　一个三次方程与一个平方根

想一想　若 $x=a+\sqrt{2}$，则 $x-\sqrt{2}$ 是 t^3-3t+1 的根。代入后，把含有 $\sqrt{2}$ 的项与不含它的项分开，然后平方。

$(x-\sqrt{2})^3=x^3-3\sqrt{2}\,x^2+6x-2\sqrt{2}$，所以

$$(x-\sqrt{2})^3-3(x-\sqrt{2})+1=\left(x^3+3x+1\right)-\sqrt{2}\left(3x^2-1\right)=0.$$

因此 $(x^3+3x+1)^2=2(3x^2-1)^2$。展开得 $x^6+6x^4+2x^3+9x^2+6x+1=18x^4-12x^2+2$，所以

$$\mathrm{h}(x)=x^6-12x^4+2x^3+21x^2+6x-1. \qquad \blacksquare$$

同样的论证对 $b+\sqrt{2}$ 和 $c+\sqrt{2}$ 也成立，因为我们只用到 a 是 t^3-3t+1 的根这一事实。（h 的另外三个根是 $a-\sqrt{2}$、$b-\sqrt{2}$ 和 $c-\sqrt{2}$，它们是在平方时引入的。）

阶段 4　两个立方根

停一停，试一试 设 $x=\sqrt[3]{2}+\sqrt[3]{3}$。把两边立方。你能否把结果用 x 和单独一个立方根表示?

利用 $(u+v)^3=u^3+v^3+3uv(u+v)$，取 $u=\sqrt[3]{2}$，$v=\sqrt[3]{3}$：

$$x^3=2+3+3\sqrt[3]{6}\,x, \qquad \text{所以} \qquad x^3-5=3\sqrt[3]{6}\,x.$$

再立方一次，便消去最后一个立方根：$(x^3-5)^3=27\cdot 6\,x^3=162x^3$。展开 $(x^3-5)^3=x^9-15x^6+75x^3-125$，得

$$\mathrm{k}(x)=x^9-15x^6-87x^3-125. \qquad \blacksquare$$

回顾

- **逐一消去根式。**把一个根式单独移到一边，取适当的次方，然后重复这个步骤。
- **把有理部分与无理部分分开。**把式子写成 $P+\sqrt{5}\,Q$，再利用 $P^2=5Q^2$，是关键的一步。
- **平方会带来额外的根。**每次平方都会引入“共轭”（conjugate）根，例如除了 $a+\sqrt{2}$ 之外还有 $a-\sqrt{2}$。最后得到的多项式以它们全部为根。

背后的数学

域及其次数

所有形如 $p+q\sqrt{2}$（其中 p、q 为有理数）的数所组成的集合 $\mathbb{Q}(\sqrt{2})$，对加、减、乘、除（除数不为零）都封闭：它是一个域（field）。它是 $\mathbb{Q}$ 上的二维“向量空间”（vector space），以 $1,\sqrt{2}$ 为基底（basis），所以我们说它在 $\mathbb{Q}$ 上的次数（degree）是 2。添加 $\sqrt{3}$ 便得到 $\mathbb{Q}(\sqrt{2},\sqrt{3})$，其基底为 $1,\sqrt{2},\sqrt{3},\sqrt{6}$，次数为 4；再添加 $\sqrt{5}$，次数便是 8（图 6.3.1）。塔定律（tower law）指出，沿着一座域塔往上，次数会相乘。由一个数生成的域，其次数等于该数的最小多项式（minimal polynomial）的次数，即该数所满足的有理系数多项式之中次数最低的一个。可以证明，本题的多项式 f 和 g 都是最小多项式，所以 $\sqrt{2}+\sqrt{3}$ 的次数是 4，而 $\sqrt{2}+\sqrt{3}+\sqrt{5}$ 的次数是 8：事实上，单凭 $\sqrt{2}+\sqrt{3}+\sqrt{5}$ 这一个数，就足以生成整个 8 次域。

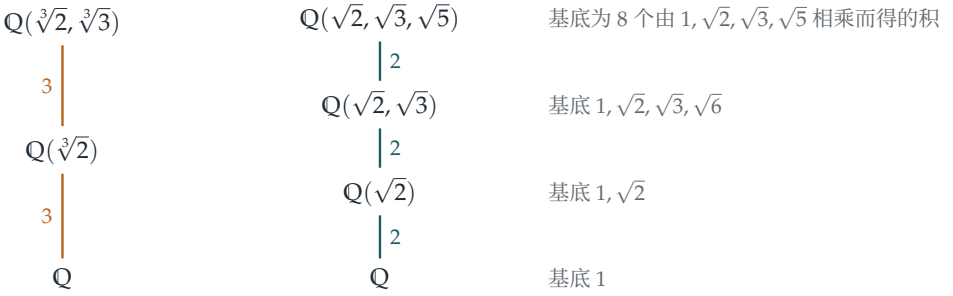


图 6.3.1　域塔。右：每添加一个平方根，次数便加倍，得出 $2\times2\times2=8$，即 g 的次数。左：每添加一个立方根，次数便变为三倍，得出 $3\times3=9$，即 k 的次数。沿着域塔往上，次数会相乘；这条*塔定律*正是证明某些尺规作图不可能完成的关键。

倍立方

尺规作图（ruler-and-compass construction）所能作出的长度，只限于由有理数出发、反复解一次和二次方程而得到的数。每一步都给出一个次数为 1 或 2 的域扩张，所以根据塔定律，每个可作图（constructible）的长度都位于一个次数为 2 的幂的域之内，而它本身的次数也是 2 的幂。但 $\sqrt[3]{2}$ 的最小多项式是 x^3-2，次数为 3。由于 3 不是 2 的幂，$\sqrt[3]{2}$ 不可作图，因此约于公元前 430 年提出的提洛岛（Delos）问题，即用尺规把立方体的体积加倍，是不可能做到的。旺策尔（Pierre Wantzel）在 1837 年证明了这一点，同时证明了一般的角不可能三等分，并完成了可作图正多边形的分类（Section 6.2）。两千年来的种种尝试，就此被一个关于次数的论证画上句号。

代数数构成一个域

本题的方法通过例子说明，代数数之和仍是代数数。一般情况则要用线性代数（linear algebra）来证明：若 α 和 β 的次数分别为 m 和 n，则所有 $\alpha^i\beta^j$（其中 $i<m$，$j<n$）这些数所张成的空间，维数至多为 mn，于是 $mn+1$ 个幂 $1,\alpha+\beta,(\alpha+\beta)^2,\ldots$ 必定线性相关（linearly dependent），由此得出一个次数至多为 mn 的多项式。我们得到的次数 $4=2\times2$、$6=3\times2$ 和 $9=3\times3$ 与这个规律完全吻合。*不是*代数数的数称为*超越数*（transcendental number）；e 和 π 都是例子（分别由埃尔米特（Hermite）于 1873 年及林德曼（Lindemann）于 1882 年证明），而林德曼关于 π 是超越数的定理，也证明了化圆为方同样不可能。

思维技巧

先消去，再量度

要找出一个含根式的式子所满足的多项式，可把根式逐个单独移到一边，再取适当的次方，逐一消去。

然后问一问：次数说明了甚么？最小多项式的次数是该数的一个不变量（invariant），而比较次数是证明某事不可能的最有力方法之一。

延伸练习

1 **本原元（primitive element）。**设 $y=\sqrt{2}+\sqrt{3}$。证明 $y^3=11\sqrt{2}+9\sqrt{3}$，并由此推出 $\sqrt{2}=\frac{1}{2}(y^3-9y)$。因此每个形如 $p+q\sqrt{2}+r\sqrt{3}+s\sqrt{6}$ 的数都可以写成 y 的多项式。

2 **混合根式。**求一个次数为 6 的整数系数多项式，使它以 $\sqrt{2}+\sqrt[3]{3}$ 为根。

3 **不可约性（irreducibility）。**证明 x^4-10x^2+1 没有有理根，而且不能写成两个整数系数二次式的积。（提示：试用 $(x^2+ax+b)(x^2-ax+c)$。）由此推出 $\sqrt{2}+\sqrt{3}$ 的次数是 4。

4 **三等分角。**证明 $\cos 3\theta=4\cos^3\theta-3\cos\theta$，并由此推出 $\cos 20^\circ$ 是 $8x^3-6x-1=0$ 的根。假设这个三次方程没有有理根（请自行验证），解释为何不能用尺规把 60° 的角三等分。

延伸阅读

- **Ian Stewart, *Galois Theory***（CRC Press）。讲述域扩张、塔定律及古典作图问题，为大学本科生而写。
- **Richard Courant & Herbert Robbins, *What Is Mathematics?***（Oxford University Press）。第 3 章证明古希腊几大问题不可能解决，毋须任何预备知识。
- **Robin Hartshorne, *Geometry: Euclid and Beyond***（Springer）。以近世代数处理尺规作图。

6.4 解四次方程

题目

STEP III 2017，第 3 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/17-s3-q3`

引子

每个学生都学过二次方程（quadratic equation）的求根公式。三次方程（cubic）和四次方程（quartic）其实也有公式；这些公式在十六世纪的意大利被发现，其间充满秘密比试，以及为争夺发现权而起的激烈争执。卡尔达诺（Gerolamo Cardano）在 1545 年出版的《大术》（*Ars Magna*）中把两者一并发表：三次方程的解法来自塔尔塔利亚（Niccolò Tartaglia），而塔尔塔利亚曾要他发誓保密；四次方程的解法则来自他自己的学生费拉里（Lodovico Ferrari）。此后近三百年，数学家一直苦苦寻找五次方程（quintic）的公式，却始终徒劳无功。

这道题目展示的，正是解开四次方程的那个窍门，而这个窍门背后大有深意。从四个根出发，按三种可能的方式把根两两配对，得出三个新的量。这三个量满足一个三次方程，而三次方程是你解得到的；有了它们，原来的四个根便随之而来。为甚么这个方法行得通？为甚么四个根可以化约为三个量，五个根却不能如法炮制？正是这个问题，引领伽罗瓦（Galois）创立了群论（group theory）。

动笔之前

想一想 把四件物件 $\alpha, \beta, \gamma, \delta$ 分成两对，共有多少种分法？

三种：$\{\alpha\beta, \gamma\delta\}$、$\{\alpha\gamma, \beta\delta\}$ 和 $\{\alpha\delta, \beta\gamma\}$（图 6.4.1）。题目中的三个量 $\alpha\beta + \gamma\delta$ 等等，正好对应这三种配对。把四个根任意重新排列，只会令这三种配对彼此对调，所以这三个量的任何对称式（symmetric expression），都是四个根的对称式，因而可以用 p、q、r、s 表示。这就是为甚么它们满足一个系数只涉及 p、q、r、s 的三次方程。

逐步解题

阶段 1　欠缺的系数

想一想　在三次方程 $y^3 + Ay^2 + By + C = 0$ 中，根之和用 A 表示是甚么？而 $(\alpha\beta + \gamma\delta) + (\alpha\gamma + \beta\delta) + (\alpha\delta + \beta\gamma)$ 用四次方程的系数表示又是甚么？

三次方程的根之和是 $-A$。那三个量之和，就是 $\alpha, \beta, \gamma, \delta$ 中两两相乘所得全部六个积之和；对四次多项式 $x^4 + px^3 + qx^2 + rx + s$ 而言，这个和是 q。所以 $A = -q$。

阶段 2　具体的三次方程

代入 $p = 0$、$q = 3$、$r = -6$、$s = 10$，三次方程是

$$y^3 - 3y^2 + (0 - 40)y + (120 - 0 - 36) = y^3 - 3y^2 - 40y + 84 = 0.$$

想一想　试试 84 的细小因数。

$y = 2$ 给出 $8 - 12 - 80 + 84 = 0$。作除法，得 $y^3 - 3y^2 - 40y + 84 = (y - 2)(y^2 - y - 42) = (y - 2)(y - 7)(y + 6)$。最大的根是 7，所以 $\alpha\beta + \gamma\delta = 7$。 ■

阶段 3　根的配对

$(\alpha + \beta)(\gamma + \delta) = \alpha\gamma + \alpha\delta + \beta\gamma + \beta\delta$，这是全部六个两两之积的和减去 $\alpha\beta + \gamma\delta$，即 $q - 7 = 3 - 7 = -4$。

此外 $\alpha\beta \cdot \gamma\delta = \alpha\beta\gamma\delta = s = 10$。所以 $\alpha\beta$ 和 $\gamma\delta$ 之和为 7，积为 10：它们是 $t^2 - 7t + 10 = (t - 2)(t - 5)$ 的根。由于 $\alpha\beta > \gamma\delta$，得 $\alpha\beta = 5$ 及 $\gamma\delta = 2$。 ■

阶段 4　解四次方程

停一停，试一试　求 $\alpha + \beta$ 和 $\gamma + \delta$。你已知道它们的和（由 p 得出）和它们的积（由第 3 步得出）；要判断哪个是哪个，就用 r。

$\alpha + \beta + \gamma + \delta = -p = 0$ 而 $(\alpha + \beta)(\gamma + \delta) = -4$，所以 $\{\alpha + \beta, \gamma + \delta\} = \{2, -2\}$。系数 r 给出四个三根之积的和：$\alpha\beta\gamma + \alpha\beta\delta + \alpha\gamma\delta + \beta\gamma\delta = -r = 6$。这可以写成 $\alpha\beta(\gamma + \delta) + \gamma\delta(\alpha + \beta) = 5(\gamma + \delta) + 2(\alpha + \beta)$。若 $\alpha + \beta = 2$，此式等于 $-10 + 4 = -6$；若 $\alpha + \beta = -2$，则等于 $10 - 4 = 6$。所以 $\alpha + \beta = -2$ 及 $\gamma + \delta = 2$。

于是 α、β 是 $t^2 + 2t + 5 = 0$ 的根，即 $-1 \pm 2\mathrm{i}$；而 γ、δ 是 $t^2 - 2t + 2 = 0$ 的根，即 $1 \pm \mathrm{i}$。

四个根是

$$x = -1 \pm 2\mathrm{i}, \qquad x = 1 \pm \mathrm{i}. \qquad \blacksquare$$

验算：$(x^2 + 2x + 5)(x^2 - 2x + 2) = x^4 + 3x^2 - 6x + 10$。

回顾

- **对称函数（symmetric functions）。**每一步都用上了根与系数的关系：根之和与根之积。
- **降低次数。**一个四次方程的问题变成了一个三次方程（即*预解三次方程*（resolvent cubic）），然后再变成两个二次方程。
- **用尽每一项资料。**p 和 q 给出了各个和与积；s 把 $\alpha\beta$ 和 $\gamma\delta$ 分开；r 则决定了哪一对根配哪一个和。

背后的数学

这个窍门为何行得通：根的置换

1770 年，拉格朗日（Joseph-Louis Lagrange）追问三次和四次方程的解法为何行得通，并在根的*置换*（permutation）中找到了答案。四个根共有 24 种重新排列的方式。在所有这些排列之下，三个量 $y_1 = \alpha\beta + \gamma\delta$、$y_2 = \alpha\gamma + \beta\delta$、$y_3 = \alpha\delta + \beta\gamma$ 只不过是彼此对调：四个根的每一种重新排列，都引致三种配对的一种重新排列（图 6.4.1）。所以它们满足一个 3 次多项式，其系数是各根的对称式。用现代的语言来说，从对称群（symmetric group）S_4（全部 24 个置换）到 S_3（三种配对的 6 个置换）有一个满射的群同态（group homomorphism），这就是四次方程可以化约为三次方程的原因。

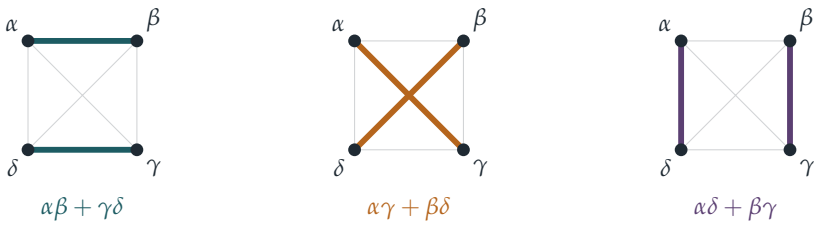


图 6.4.1　把四个根分成两对的三种方式，以及预解三次方程中与之对应的根。$\alpha, \beta, \gamma, \delta$ 的任何重新排列，都只会令这三幅图彼此置换，这就是为甚么那三个量满足一个三次方程，而其系数可以用四次方程的系数表示。

五次方程

对于五个根，拉格朗日尝试寻找一个类似的量，使它在根的置换之下只取寥寥几个值，却找不到有用的。1799 年鲁菲尼（Paolo Ruffini）提出证明，1824 年阿贝尔（Niels Henrik Abel）把证明补全，两人证明了：只用四则运算和根式（radicals），并不存在五次方程的一般求根公式。数年之后，伽罗瓦（Évariste Galois）确切地说明了哪些多项式可以用根式求解：就是那些*根的对称群*能以某种特定方式分解成较简单部分的多项式。对称群 S_4 可以这样分解（经由 S_3 和更小的群），但 S_5 不可以，因为它包含一个大小为 60 的“单”（simple）部分，无法再进一步分解。1832 年，伽罗瓦在决斗中丧生，年仅二十岁；决斗前一晚，他才匆匆把自己的理论写下来。这套理论如今是数学学位课程中的瑰宝之一，通常在三年级学到。

解三次方程

三次方程本身也可以用类似的窍门求解：把一个缺二次项的三次式（depressed cubic）$x^3 - 3px + q$ 写成两个完全立方的组合，从而化约为一个二次方程。这就是练习中那道 2011 年题目所用的方法，它与卡尔达诺公式（Cardano's formula）等价。当三次方程有三个实根时，这条公式在运算途中仍然需要用到复数（complex numbers），正是这个现象迫使数学家认真看待复数（见第 5 章的引言）。

思维技巧

> **对称性化繁为简**
>
> *如果由根构造出来的一些量，在根被置换时只是彼此置换，那么它们满足一个次数较低的多项式，而其系数是你可以计算出来的。*
>
> 这是伽罗瓦理论（Galois theory）的种子：通过理解方程各根的对称性来理解方程本身。

延伸练习

1. **验证三次方程。**验证 $y_1y_2 + y_2y_3 + y_3y_1 = pr - 4s$，其中 $y_1 = \alpha\beta + \gamma\delta$，余类推。（展开并利用 $\sum\alpha = -p$、$\sum\alpha\beta\gamma = -r$、$\alpha\beta\gamma\delta = s$。）

2. **把三次式写成立方之和。**试做 STEP III 2011，第 3 题（`step.maths.org/questions/11-s3-q3`）。这道题把三次式写成两个立方的组合来求解，与本节的路线截然不同。请用数值方法核对它的答案。

3 **另一个四次方程。**用本节的方法解 $x^4 - 4x^3 + 7x^2 - 6x + 2 = 0$。（预解三次方程有一个根是细小的整数。）

4 **数算置换。**在 $\alpha, \beta, \gamma, \delta$ 的 24 个置换中，找出所有把配对 $\{\alpha\beta, \gamma\delta\}$ 变回自身的置换。证明这样的置换共有 8 个。（它们构成 S_4 的一个子群（subgroup）。）

延伸阅读

- **Mario Livio, *The Equation That Couldn't Be Solved***（Simon & Schuster）。五次方程的故事与群论的诞生，适合一般读者。
- **Ian Stewart, *Why Beauty Is Truth: A History of Symmetry***（Basic Books）。从巴比伦人的方程讲到伽罗瓦，再到其后的发展。
- **Ian Stewart, *Galois Theory***（CRC Press）。大学本科教科书，其历史导论所涵盖的正是本节的内容。

6.5 从公理出发的微分法

题目

STEP II 1992，第 4 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/92-s2-q4`

引子

甚么是微分法（differentiation）？在中学，它是用极限（limit）$\lim_{h\to 0}\frac{f(x+h)-f(x)}{h}$ 来定义的，然后你再学各种法则：和的导数（derivative）等于导数的和、积法则（product rule），诸如此类。这道题把整件事倒转过来。请把极限完全抛开。假设你对某个运算所知的，就只有它符合和法则、积法则和常数的法则，并且把 x 变成 1。题目要证明的是：这样的运算必定就是微分，至少对多项式（polynomial）而言是如此。

这是代数学家看微积分的角度：一个运算的特征在于它所遵从的法则，而不在于某条公式。遵从积法则的运算称为导子（derivation），它们在代数、几何和物理中随处可见。出人意料的是，训练现代机器学习系统的软件，内里用的也是同一个想法。

动笔之前

想一想 法则 (i) 就是积法则。积法则怎可能告诉你关于常数的任何事情？（试把常数看成一个乘积。）

数 1 等于 1×1。把积法则应用于 1×1，便得出 $\Diamond 1$ 与它自身之间的一个关系，从而迫使它等于零。一般而言，处理这类公理（axiom）的窍门，是把它们应用于精心挑选的简单输入。

逐步解题

阶段 1　常数

想一想　取 $\mathrm{f} = \mathrm{g} = 1$，应用法则 (i)。然后用法则 (iv) 处理其他任何常数。

$\Diamond 1 = \Diamond(1 \cdot 1) = 1 \cdot \Diamond 1 + 1 \cdot \Diamond 1 = 2\Diamond 1$，所以 $\Diamond 1 = 0$。对任意常数 λ，由法则 (iv) 得 $\Diamond \lambda = \Diamond(\lambda \cdot 1) = \lambda \Diamond 1 = 0$。 ■

阶段 2　幂

利用法则 (iii)，$\Diamond x^2 = \Diamond(x \cdot x) = x\Diamond x + x\Diamond x = 2x$。于是

$$\Diamond x^3 = \Diamond(x^2 \cdot x) = x\Diamond x^2 + x^2 \Diamond x = x \cdot 2x + x^2 = 3x^2.$$

停一停，试一试　用归纳法(induction)证明：对每个正整数 n，均有 $\Diamond x^n = n x^{n-1}$。

$n = 1$ 的情况就是法则 (iii)。若 $\Diamond x^n = n x^{n-1}$，则

$$\Diamond x^{n+1} = \Diamond(x^n \cdot x) = x\Diamond x^n + x^n \Diamond x = x \cdot n x^{n-1} + x^n = (n+1)x^n. \qquad \blacksquare$$

阶段 3　所有多项式

一般的多项式是 $\mathrm{h}(x) = a_0 + a_1 x + \cdots + a_m x^m$。由法则 (ii)（重复应用）和法则 (iv)，

$$\begin{aligned} \Diamond \mathrm{h}(x) &= \Diamond a_0 + a_1 \Diamond x + a_2 \Diamond x^2 + \cdots + a_m \Diamond x^m \\ &= a_1 + 2a_2 x + \cdots + m a_m x^{m-1} = \mathrm{h}'(x). \end{aligned} \qquad \blacksquare$$

回顾

- **公理决定运算。**四条法则，完全没有提及极限，就把 $\Diamond$ 在多项式上的作用完全确定下来。
- **生成元（generator）。**每个多项式都是由常数和 x 经加法和乘法构造出来的。只要知道 $\Diamond$ 对这些基本构件的作用，以及它与这些构造运算如何互动，就能确定它在所有多项式上的作用。
- **沿着构造过程作归纳。**证明的步骤紧随多项式的构造方式：先以重复相乘得出幂，然后再求和。

背后的数学

导子

在一个代数系统上，若运算 D 具有可加性（additive），并且满足积法则 $D(fg) = fD(g) + gD(f)$，便称为*导子*。这道题说明，在多项式上，一个把常数变成零的导子完全由 $D(x)$ 决定：事实上 $D(h) = h'(x)\,D(x)$，所以唯一的自由度就是多项式 $D(x)$ 的选择。法则 (iii) 规定了 $D(x) = 1$。导子遍布数学各个领域：在几何中，曲面上的*向量场*（vector field）可以定义为该曲面上的函数的一个导子（它沿着向量场的方向对函数求导）；在代数中，一个结构的导子量度它的“无穷小对称”（infinitesimal symmetries）；在物理中，量子力学的交换子（commutator）$[A, B] = AB - BA$ 对每个变元而言都是一个导子。

没有极限的微积分

在极限完全没有意义的场合，代数式的导数依然行得通。在有限域（finite field）上，例如模质数 p 的算术，并没有极限可言，但多项式的形式导数（formal derivative）仍由同样的法则定义，而且仍然能够侦测重根（repeated root）：多项式 f 有重根，当且仅当 f 与 f' 有公因式。举例说，我们就是这样证明当 $p \nmid n$ 时 $x^n - 1$ 没有重根的；要在有限域上定义 Section 6.1 的分圆多项式（cyclotomic polynomial），就需要这个事实。编码理论和密码学也用到同样的想法。

自动微分

以下是一个现代应用。创造一个数 ε，使得 $\varepsilon^2 = 0$ 但 $\varepsilon \neq 0$，然后用*对偶数*（dual number）$a + b\varepsilon$ 来做算术。这时 $(a + b\varepsilon)(c + d\varepsilon) = ac + (ad + bc)\varepsilon$：$\varepsilon$ 的部分恰好遵从积法则。因此对任何多项式，

$$f(x + \varepsilon) = f(x) + f'(x)\,\varepsilon,$$

而在对偶数 $x + \varepsilon$ 处计算 f，就能同时求出函数值和导数，结果完全准确，既不用极限，也没有因微小差值而产生的舍入误差（图 6.5.1）。这称为*前向模式自动微分*（forward-mode automatic differentiation）。它的近亲“反向模式”（reverse mode），就是用来训练拥有数十亿个参数的神经网络（neural network）的算法（*反向传播*，backpropagation）：这类系统每学习一次，都是在把积法则和链式法则（chain rule）当作导子，一步一步地应用。

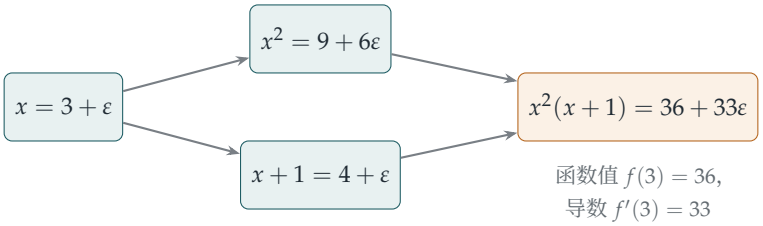


图 6.5.1　用对偶数（$\varepsilon^2 = 0$）对 $f(x) = x^2(x+1)$ 在 $x = 3$ 处作自动微分。每一步都遵从题目中的法则：$(3+\varepsilon)^2 = 9 + 6\varepsilon$，以及 $(9+6\varepsilon)(4+\varepsilon) = 36 + (9+24)\varepsilon$。$\varepsilon$ 的系数就是 $f'(3) = 3\cdot 9 + 2\cdot 3 = 33$，计算结果完全准确。

思维技巧

以性质刻划对象

用一个运算所遵从的法则来描述它，然后证明这些法则唯一地确定了它。

这就是公理化方法（axiomatic method），是大学代数的核心。它的威力在于：凡是由公理证明出来的结果，都适用于每一个满足这些公理的运算，不论那是多项式的微分、曲面上函数的微分，还是某种尚未发明的东西。

延伸练习

1. **另一个导子。**假设 $\Diamond$ 满足法则 (i)、(ii) 和 (iv)，但 $\Diamond x = x$ 而不是 1。证明 $\Diamond \mathrm{h}(x) = x\,\mathrm{h}'(x)$。$\Diamond$ 会把 x^n 变成甚么？

2. **商法则（quotient rule）。**把 $\Diamond$ 推广到有理函数（rational function）f/g，并假设法则 (i) 仍然成立。把法则 (i) 应用于 $\mathrm{g}\cdot\frac{\mathrm{f}}{\mathrm{g}} = \mathrm{f}$，从而推导出商法则。

3. **重根。**证明若 $(x-a)^2$ 整除多项式 f，则 $x-a$ 整除 f′。利用这个结果证明 $x^n - 1$ 没有重根。

4. **对偶数。**设 $f(x) = (x^2+1)^3$，用对偶数计算 $f(2)$ 和 $f'(2)$，并用链式法则验算。

延伸阅读

- **Peter Cameron, *Introduction to Algebra***（Oxford University Press）。大学一年级程度的环、多项式和域，并包括形式导数。

- **Andreas Griewank & Andrea Walther, *Evaluating Derivatives***（SIAM）。自动微分的标准参考书。程度较深。
- **Michael Nielsen, *Neural Networks and Deep Learning***（网上免费阅读）。第 2 章清楚讲解反向传播。

6.6 子群的并集

题目

STEP II 1988，第 9 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/88-s2-q9`

引子

这是全书最短的题目之一，也是最具大学数学特色的题目之一。它谈的是*群*（group），即描述对称的数学语言，问的问题听起来像个谜题：一个群能否被它里面两个较小的群完全覆盖？答案是不能，而证明只有三行。不过三个较小的群却可以做到，这倒是出人意表。

群论（group theory）在 2019 年以前一直属于 STEP 的课程范围，如今则是大学一年级的必修内容。这道题只需要用到定义，下文会一一列出；它正好让你初尝群论学者的思考方式：一个简短而巧妙的论证，除了公理（axiom）之外甚么都不用。

动笔之前

*群*是一个集合 G，连同一个运算（写成乘法 ab），满足：

- 运算符合*结合律*（associative）：$(ab)c = a(bc)$；
- 存在*单位元*（identity）e，对每个 a 都有 $ea = ae = a$；
- 每个 a 都有*逆元*（inverse）a^{-1}，使得 $aa^{-1} = a^{-1}a = e$。

例子：整数（正数、负数和零）在加法下构成群；非零实数在乘法下构成群；正方形的旋转在复合（composition）下构成群。*子群*（subgroup）是 G 的一个子集 K，它在同一运算下本身就是一个群：它包含 e，而且若 $a, b \in K$，则 $ab \in K$ 及 $a^{-1} \in K$。

? **想一想** 若 K 是子群，$a \in K$ 但 $b \notin K$，那么 ab 可以属于 K 吗？

不可以。若 $ab \in K$，则 $b = a^{-1}(ab)$ 是 K 中两个元素的乘积，因而属于 K。这个小小的观察就是整个证明。

逐步解题

阶段 1　两个子群无法覆盖

用反证法（proof by contradiction）：假设 $G = G_1 \cup G_2$，但 G_1 和 G_2 都不是整个 G。

> **想一想**　由于 $G_1 \neq G$，必有一个元素不在 G_1 之中。它必定在哪里？G_2 的情况也一样。现在把这两个元素相乘。

取 $a \in G$ 使得 $a \notin G_1$；由于每个元素都属于 G_1 或 G_2，故 $a \in G_2$。取 $b \in G$ 使得 $b \notin G_2$；于是 $b \in G_1$。考虑乘积 ab。它属于 G_1 或 G_2。

- 若 $ab \in G_1$：由于 $b \in G_1$，$a = (ab)b^{-1} \in G_1$。矛盾。
- 若 $ab \in G_2$：由于 $a \in G_2$，$b = a^{-1}(ab) \in G_2$。矛盾。

无论哪种情况都得出矛盾，所以 $G_1 = G$ 或 $G_2 = G$。 ■

留意我们从未用过 G 是有限群这个条件。这个结果对所有群都成立。

阶段 2　三个子群却可以

> **停一停，试一试**　找出一个群，它是三个子群的并集（union），而这三个子群都不是整个群。（提示：试试一个每个元素都是自己逆元的小群。）

取克莱因四元群（Klein four-group）$V = \{e, a, b, c\}$，其运算规则为 $a^2 = b^2 = c^2 = e$ 以及 $ab = ba = c$、$bc = cb = a$、$ca = ac = b$。（具体来说：一个非正方形的长方形的四个对称，即保持不动、旋转 180°，以及沿两条对称轴之一作反射。又或者是由位元 0、1 组成的数对 (x, y)，在模 2 加法下构成的群。）它的三个子群

$$H_1 = \{e, a\}, \qquad H_2 = \{e, b\}, \qquad H_3 = \{e, c\}$$

各有两个元素，没有一个是整个群，而它们合起来包含了 V 的每一个元素（图 6.6.1）。 ■

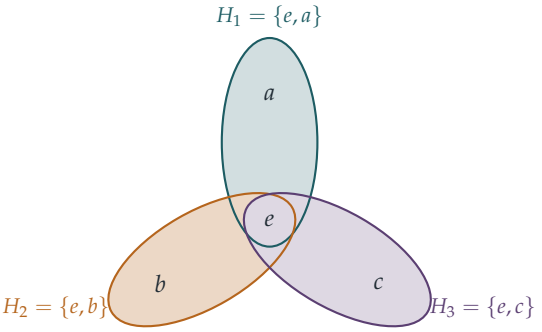


图 6.6.1 克莱因四元群 $\{e, a, b, c\}$ 是它三个 2 阶子群的并集，这三个子群只在单位元 e 处重叠。根据题目的第一部分，没有任何群可以只是两个真子群（proper subgroup）的并集。

回顾

- **选出见证元素，再把它们结合。**证明中在每个子群之外各选一个元素，然后把它们相乘。这种“各取一个再结合”的手法在代数中极为常见。
- **封闭性威力强大。**证明中只用到子群的一个性质：它们在乘法和取逆元之下封闭（closed）。
- **检查你真正需要的假设。**题目给了有限这个条件，却没有用上。察觉到某个假设并非必要，是理解一个证明的一部分。

背后的数学

对称的语言

群在十九世纪同时从几个方向出现：伽罗瓦（Galois）对多项式的根的置换（permutation）的研究（Section 6.4）、柯西（Cauchy）关于置换的工作，以及几何学和晶体学中对对称的研究。上文那个抽象定义，即凡满足三条公理的东西都是群，到该世纪末已经确立。它的威力在于一般性：由公理证明的定理，例如本题中的那一个，同时适用于由数、矩阵、对称和置换构成的各种群。今天群无处不在：在粒子物理学中（标准模型建基于对称群）、在化学中（分子的对称性决定光谱）、在密码学中（椭圆曲线群，Section 1.9），以及在扭计骰（Rubik's cube）的解法之中。

要用多少个子群?

这道题引出一个自然的问题：最少要多少个真子群，它们的并集才是整个群？根据本题，这个数目永远不会是 2。1926 年，Gaetano Scorza 证明了一个群是三个真子群的并集，当且仅当它以克莱因四元群为“商群”（quotient）。1994 年，J. H. E. Cohn 研究了一般的问题，并提出了若干猜想；1997 年，M. J. Tomkinson 证明了没有任何群的这个最小数目是七，尽管由 3 到 6 的每个数都会出现。这类问题叙述简单，却出奇地微妙，正是群论的典型。

向量空间

类似的论证可以证明，实数上的向量空间（vector space）（例如平面或三维空间）永远不可能是有限多个真子空间（proper subspace）的并集：平面不是有限多条通过原点的直线的并集。但在有限域（finite field）上却可以：两个元素的域上的平面，其实又是克莱因四元群，它正是其三条直线的并集。

思维技巧

从公理出发论证

在抽象代数中，你对所处理的对象往往一无所知，只知道它们满足哪些公理。证明必须只用这些公理。

这迫使论证简短而干净利落，就像本题的论证一样，也使它们放诸四海皆准。这需要一些时间去适应，而它正是大学一年级代数课中最独特的体验之一。

延伸练习

1. **验证这个群。**写出克莱因四元群的乘法表，并验证各条群公理。证明它与乘法下的群 $\{1, i, -1, -i\}$ 并不相同。（提示：哪些元素是自己的逆元?）

2. **循环群。**若一个群的每个元素都是某一个元素 g 的幂，则称它为*循环群*（cyclic group）。证明循环群永远不是真子群的并集，不论用多少个。（提示：哪个子群包含 g?）

3. **整数。**直接证明 $(\mathbb{Z}, +)$ 不是子群 $2\mathbb{Z}$（偶数）和 $3\mathbb{Z}$（3 的倍数）的并集。它是否某两个真子群的并集?

4. **平面上的直线。**较难 证明平面 $\mathbb{R}^2$ 不是有限多条通过原点的直线的并集。（提示：每条直线除原点外，最多只包含抛物线 $y = x^2$ 上的一个点。）

延伸阅读

- **M. A. Armstrong, *Groups and Symmetry***（Springer）。一本插图精美的入门教材，以几何例子带动全书。
- **Nathan Carter, *Visual Group Theory***（Mathematical Association of America）。用图像讲解群，最适合在入读大学前阅读。
- **Peter Neumann, Gabrielle Stoy & Edward Thompson, *Groups and Geometry***（Oxford University Press）。牛津大学一年级的经典教材。

6.7 四元数群

题目

STEP III 1993，第 5 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/93-s3-q5`

引子

1843 年 10 月 16 日，哈密顿（William Rowan Hamilton）与妻子沿着都柏林的皇家运河散步时，苦思多年的问题忽然灵光一闪，有了答案。他一直想仿照复数把一对对的数相乘那样，把三个一组的数相乘，却屡试屡败。他突然看出，答案需要的是四个数，而且乘法的次序不可调换：$ij \neq ji$。他把这些法则刻在布鲁穆桥（Brougham Bridge）的石头上：

$$\mathrm{i}^2 = \mathrm{j}^2 = \mathrm{k}^2 = \mathrm{ijk} = -1.$$

这就是*四元数*（quaternions）。本题用复数对把四元数构造出来，并在其中找到一个只有八个元素、乘法带点扭曲的小群。时至今日，每一款电脑游戏、每一部飞行模拟器、每一套太空船导航系统，都用四元数来描述三维空间中的旋转。

动笔之前

群的各项定义见 Section 6.6。这里再补充两个用语：群中元素 g 的*阶*（order），是指使 $g^n = e$ 成立的最小正整数 n；若两个群之间存在一个保持运算的一一对应，则称这两个群*同构*（isomorphic），也就是说，它们是“换了标签的同一个群”。

想一想 把数对 (z_1, z_2) 看成代表一个“数” $z_1 + z_2\mathrm{j}$，其中 j 是一个新符号，满足 $\mathrm{j}^2 = -1$，并且对任意复数 w 有 $\mathrm{j}w = w^*\mathrm{j}$。试用这些法则计算 $(z_1 + z_2\mathrm{j})(w_1 + w_2\mathrm{j})$。得出的是否就是 $\circ$ 的公式？

$(z_1 + z_2\mathrm{j})(w_1 + w_2\mathrm{j}) = z_1w_1 + z_1w_2\mathrm{j} + z_2\mathrm{j}w_1 + z_2\mathrm{j}w_2\mathrm{j} = z_1w_1 + z_1w_2\mathrm{j} + z_2w_1^*\mathrm{j} + z_2w_2^*\mathrm{j}^2$，即 $(z_1w_1 - z_2w_2^*) + (z_1w_2 + z_2w_1^*)\mathrm{j}$。答案是肯定的：运算 $\circ$ 其实就是乔装了的四元数乘

法。

逐步解题

阶段 1　借助矩阵证明结合律

直接验证 $(a\circ b)\circ c = a\circ(b\circ c)$，计算十分冗长。其实有更巧妙的方法。

> **想一想**　把 (z_1,z_2) 对应到复矩阵 $M=\begin{pmatrix} z_1 & z_2 \\ -z_2^{*} & z_1^{*}\end{pmatrix}$。矩阵乘积 $M(z_1,z_2)\,M(w_1,w_2)$ 的第一行是甚么？

乘积的第一行是 $(z_1w_1 - z_2w_2^{*},\ z_1w_2 + z_2w_1^{*})$，正好是 $(z_1,z_2)\circ(w_1,w_2)$；第二行是 $(-(z_1w_2 + z_2w_1^{*})^{*},\ (z_1w_1 - z_2w_2^{*})^{*})$，所以乘积仍具有同样的特殊形式。因此 $M(a\circ b) = M(a)M(b)$。矩阵乘法满足结合律（associativity），而 M 是一对一的（从第一行便可读出 z_1、z_2），所以 $\circ$ 满足结合律。 ■

阶段 2　不满足交换律

$(0,1)\circ(\mathrm{i},0) = (0\cdot\mathrm{i} - 1\cdot 0,\ 0\cdot 0 + 1\cdot\mathrm{i}^{*}) = (0,-\mathrm{i})$，但 $(\mathrm{i},0)\circ(0,1) = (0,\mathrm{i})$。所以 $\circ$ 不满足交换律（commutativity）。

四个特殊乘积可由定义直接得出：

$$(z,0)\circ(w,0) = (zw,0),\qquad (z,0)\circ(0,w) = (0,zw),$$
$$(0,z)\circ(w,0) = (0,zw^{*}),\qquad (0,z)\circ(0,w) = (-zw^{*},0).$$

阶段 3　八个元素的群

用四元数的记法，$A=1$，$C=\mathrm{i}$，$B=\mathrm{j}$，$D=\mathrm{ij}=\mathrm{k}$，而 E,G,F,H 则分别是它们的相反数。S_1 的每个元素都具有 $(u,0)$ 或 $(0,u)$ 的形式，其中 $u\in\{1,\mathrm{i},-1,-\mathrm{i}\}$。

> **想一想**　利用四个特殊乘积，证明 S_1 是封闭的。单位元是甚么？

由四个特殊乘积可知，两个形如 $(u,0)$ 或 $(0,u)$ 的元素相乘，结果仍是这两种形式之一，其第一个或第二个分量为 uv、uv^{*} 或 $-uv^{*}$。由于 $\{1,\mathrm{i},-1,-\mathrm{i}\}$ 在乘法、取共轭及取相反数之下都封闭（closed），乘积仍在 S_1 之内。单位元（identity）是 $A=(1,0)$：$(1,0)\circ(w_1,w_2) = (w_1,w_2)$，而且 $(w_1,w_2)\circ(1,0) = (w_1,w_2)$。

> **停一停，试一试**　求出每个元素的阶和逆元。

A 的阶为 1。$E \circ E = (-1,0) \circ (-1,0) = (1,0)$，所以 E 的阶为 2。对于 $C = (\mathrm{i},0)$：$C \circ C = (-1,0) = E$，所以 C 的阶为 4，而 $C^{-1} = C^3 = E \circ C = G$。对于 $B = (0,1)$：$B \circ B = (-1 \cdot 1^*, 0) = (-1,0) = E$，所以 B 的阶为 4，而 $B^{-1} = F$。同理，$D \circ D = (-\mathrm{i} \cdot \mathrm{i}^*, 0) = (-\mathrm{i} \cdot (-\mathrm{i}), 0) = (-1,0) = E$，所以 D 的阶为 4，逆元（inverse）为 H。至于 G、F、H，它们分别是 C、B、D 的逆元，阶同样为 4。总结如下：

元素	A	B	C	D	E	F	G	H
逆元	A	F	G	H	E	B	C	D
阶	1	4	4	4	2	4	4	4

既有封闭性和结合律（第 1 步），又有单位元和逆元，S_1 是一个群。■这个群称为四元数群（quaternion group）Q_8。

阶段 4 一个子群

$\{A, B, E, F\} = \{1, \mathrm{j}, -1, -\mathrm{j}\}$ 由 B 的各次幂组成：$B^1 = B$，$B^2 = E$，$B^3 = F$，$B^4 = A$。所以它是封闭的，并且包含各元素的逆元：这是一个子群（subgroup），而且是阶为 4 的循环群（cyclic group）。矩阵 $J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ 满足 $J^2 = -I$，$J^3 = -J$，$J^4 = I$，所以它同样生成一个阶为 4 的循环群。对应 $B^k \leftrightarrow J^k$ 是一个同构。■（事实上，J 正是 $M(0,1)$，即 B 所对应的矩阵。）

回顾

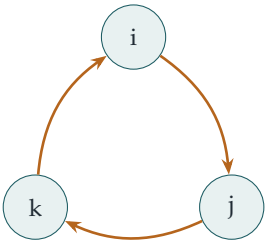
- **用矩阵来表示。**结合律直接验证起来十分费劲，这里却可从矩阵乘法继承过来。
- **找出公式背后的结构。**古怪的运算 $\circ$ 原来就是四元数乘法，看穿了这一点，一切都变得容易。
- **阶和逆元刻划了一个群。**单看各元素的阶，已可见 Q_8 与别不同：一个阶为 2 的元素，六个阶为 4 的元素。

背后的数学

复数之外

实数构成一个一维的数系，在其中可以做加法、乘法和除法；复数则构成一个二维的数系。哈密顿的四元数是四维的，除法依然可行，但乘法不再满足交换律（图 6.7.1）。1878 年，弗罗贝尼乌斯（Ferdinand Frobenius）证明了 $\mathbb{R}$、$\mathbb{C}$ 和 $\mathbb{H}$（取自 Hamilton 的首

字母）这三者，是实数域上仅有的、乘法满足结合律的有限维“可除代数”（division algebras）。如果放弃结合律，但保留一个可乘的大小，即 $|xy| = |x|\,|y|$，一如实数、复数和四元数那样，那就恰好还多一个：八维的八元数（octonions），由 John Graves 和凯莱（Arthur Cayley）在 1840 年代发现。赫维兹（Adolf Hurwitz）于 1898 年证明，这类数系就只有这四个。（若不设关于大小的条件，还有其他更古怪的非结合可除代数，但 1958 年一条由博特（Raoul Bott）、米尔诺（John Milnor）和 Michel Kervaire 得出的深刻定理指出，任何有限维实可除代数的维数必为 1、2、4 或 8。）



顺着箭头：$\mathrm{ij} = \mathrm{k}$，$\mathrm{jk} = \mathrm{i}$，$\mathrm{ki} = \mathrm{j}$

逆着箭头：$\mathrm{ji} = -\mathrm{k}$，$\mathrm{kj} = -\mathrm{i}$，$\mathrm{ik} = -\mathrm{j}$

以及 $\mathrm{i}^2 = \mathrm{j}^2 = \mathrm{k}^2 = -1$

图 6.7.1　哈密顿的乘法法则。把 i、j、k 其中两个顺着箭头方向相乘，得到第三个；逆着箭头方向相乘，则得到第三个的相反数。$\pm 1, \pm\mathrm{i}, \pm\mathrm{j}, \pm\mathrm{k}$ 这八个元素构成本题的四元数群 Q_8。

三维空间中的旋转

模（modulus）为 1 的四元数描述三维空间的旋转：绕单位轴 $\mathbf{u}$ 转过角度 θ 的旋转，对应于 $q = \cos\frac{\theta}{2} + \sin\frac{\theta}{2}(u_1\mathrm{i} + u_2\mathrm{j} + u_3\mathrm{k})$，它作用于向量 $\mathbf{v}$（写成 $v_1\mathrm{i} + v_2\mathrm{j} + v_3\mathrm{k}$）的方式是 $\mathbf{v} \mapsto q\mathbf{v}q^{-1}$。旋转的复合就是四元数的相乘。这比使用矩阵或角度更有效率，数值上也更稳定，而且避开了“万向节锁”（gimbal lock）：那是失去一个自由度的现象，当年阿波罗（Apollo）太空人使用以陀螺仪为本的导航系统时，就不得不设法绕过这个问题。半角 $\frac{\theta}{2}$ 也透露出一点深刻的端倪：旋转 360° 得到的是 $q = -1$，而不是 1，物理学家在电子的行为（自旋，spin）中正好看到这个事实。

四平方和

四元数的模是可乘的，$|pq| = |p||q|$，与复数的模一样。把它展开写出，便得到欧拉（Euler）的*四平方恒等式*（four-square identity）：两个四平方和之积，本身也是一个四平方和。这正是拉格朗日（Lagrange）定理的关键一步，该定理指出每个正整数都是四个平方数之和；我们会在 Section 7.8 证明这条定理，它是 Section 1.6 中二平方和结果的四维近亲。

思维技巧

把抽象的对象具体地表示出来

如果一个抽象的运算难以推理，就找一个具体的模型（例如矩阵），让它在其中变成熟悉的运算。

这样，结合律之类的性质便不费吹灰之力。这个想法，即把群表示为矩阵来研究，后来发展成一门庞大而优美的学科：*表示论*（representation theory）。

延伸练习

1 **哈密顿的法则。**利用数对，验证 $C \circ B = D$ 及 $B \circ C = H$，即 $\mathrm{ij} = \mathrm{k}$ 及 $\mathrm{ji} = -\mathrm{k}$。

2 **只有一个阶为 2 的元素。**证明 $E = -1$ 是 Q_8 中唯一阶为 2 的元素。由此推断 Q_8 与正方形的对称群（它有五个阶为 2 的元素）并不同构。

3 **四平方和。**证明 $\det M(z_1, z_2) = |z_1|^2 + |z_2|^2$。利用 $\det(M_1 M_2) = \det M_1 \det M_2$，推断两个四平方和之积是一个四平方和。

4 **子群。**找出 Q_8 的所有子群。证明除了 $\{A\}$ 之外，每个子群都包含 E，并证明 Q_8 是它三个阶为 4 的循环子群的并集（试与 Section 6.6 比较）。

延伸阅读

- **John Conway & Derek Smith, *On Quaternions and Octonions***（A. K. Peters）。讲述四元数和八元数的几何与算术。程度较深，但引人入胜。
- **John Baez, "The octonions"**, *Bulletin of the American Mathematical Society* 39 (2002)；网上亦可免费阅览。一篇精彩的综述；引言部分现在就读得懂。
- **Andrew Hanson, *Visualizing Quaternions***（Morgan Kaufmann）。讲述以四元数处理旋转，并介绍其在电脑图像方面的应用。

6.8 由奇异矩阵组成的群

题目

STEP III 1996，第 6 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/96-s3-q6`

引子

每个学生都学过：行列式（determinant）为零的矩阵没有逆矩阵。那么，一个由矩阵组成的群（group），既然其中每个元素都有逆元（inverse），就肯定不可能包含任何奇异矩阵（singular matrix）吧?

其实可以。关键在于“逆”这个字。在群里，$\mathbf{A}$ 的逆元是指与 $\mathbf{A}$ 结合后得出*该群的*单位元（identity element）的那个元素，而没有任何规定说群的单位元必须是单位矩阵（identity matrix）$\mathbf{I}$。本题构造出一个群，其中每个矩阵都是奇异的，而单位元本身就是一个行列式为零的矩阵。这一课教我们要按字面去读定义，也教我们分清抽象结构与我们最初认识它时所见的那个熟悉例子。

动笔之前

群的定义见 Section 6.6。

想一想 在第 (i) 部分，单位矩阵 $\mathbf{I}$ 并不是 $\begin{pmatrix} a & a \\ a & a \end{pmatrix}$ 这种形式。那么 S 的单位元会是甚么?

它必须是 S 中的某个矩阵 $\mathbf{E}$，使得对 S 中每个 $\mathbf{X}$ 都有 $\mathbf{EX} = \mathbf{XE} = \mathbf{X}$。这个条件只涉及 $\mathbf{E}$ 怎样作用于这种特殊形式的矩阵，而不是所有矩阵。解方程就可以把它找出来。

逐步解题

阶段 1　封闭性（closure）

以 $\mathbf{M}_a$ 表示四个项全都等于 a 的矩阵。那么

$$\mathbf{M}_a\mathbf{M}_b = \begin{pmatrix} a & a \\ a & a \end{pmatrix}\begin{pmatrix} b & b \\ b & b \end{pmatrix} = \begin{pmatrix} 2ab & 2ab \\ 2ab & 2ab \end{pmatrix} = \mathbf{M}_{2ab},$$

而当 $a, b \neq 0$ 时，$2ab \neq 0$。所以 S 是封闭的。 ■

阶段 2　群公理（group axioms）

> **想一想**　矩阵乘法满足结合律。至于单位元，解 $\mathbf{M}_e\mathbf{M}_a = \mathbf{M}_a$。至于逆元，解 $\mathbf{M}_a\mathbf{M}_b = \mathbf{M}_e$。

结合律（*associativity*）由矩阵乘法继承而来。*单位元*：$\mathbf{M}_e\mathbf{M}_a = \mathbf{M}_{2ea} = \mathbf{M}_a$ 对所有 a 成立，当且仅当 $e = \tfrac{1}{2}$；由于公式是对称的，$\mathbf{M}_a\mathbf{M}_{1/2} = \mathbf{M}_a$ 同样成立。所以

$$\mathbf{E} = \begin{pmatrix} \tfrac{1}{2} & \tfrac{1}{2} \\ \tfrac{1}{2} & \tfrac{1}{2} \end{pmatrix}$$

是 S 的单位元。*逆元*：当 $b = \tfrac{1}{4a}$ 时，$\mathbf{M}_a\mathbf{M}_b = \mathbf{M}_{2ab} = \mathbf{M}_{1/2}$，而这个值不为零，所以 $\mathbf{M}_{1/(4a)}$ 是 $\mathbf{M}_a$ 在 S 中的逆元。因此 S 是一个群。 ■每个元素的行列式都是 $a^2 - a^2 = 0$。

阶段 3　奇异性会蔓延

> **停一停，试一试**　假设 $\mathbf{A} \in G$ 是奇异的，并设 $\mathbf{B}$ 为 G 的任意元素。把 $\mathbf{B}$ 写成一个含有 $\mathbf{A}$ 的乘积，然后取行列式。

以 $\mathbf{A}^{-1}$ 表示 $\mathbf{A}$ *在群* G *中*的逆元（不一定是逆矩阵）。那么 $\mathbf{C} = \mathbf{A}^{-1}\mathbf{B}$ 属于 G，而且

$$\mathbf{AC} = \mathbf{A}\mathbf{A}^{-1}\mathbf{B} = \mathbf{EB} = \mathbf{B}.$$

取行列式，得 $\det\mathbf{B} = \det\mathbf{A}\,\det\mathbf{C} = 0$。所以 G 的每个元素都是奇异的。 ■特别地，如果一个矩阵群中有一个元素作为矩阵是可逆（invertible）的，那么所有元素都可逆；一个矩阵群要么全部可逆，要么全部奇异。

阶段 4　一个 3×3 的例子

想一想 你能把第 (i) 部分推广到 3×3 矩阵吗？或者找一个更简单的例子，用一个“无视”三个坐标中其中两个的矩阵？

取

$$G = \left\{ \begin{pmatrix} a & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} : a \neq 0 \right\}.$$

角上的项分别为 a 和 b 的两个矩阵，乘积就是角上的项为 ab 的矩阵，所以 G 是封闭的；乘法满足结合律；单位元是角上的项为 1 的矩阵；角上的项为 a 的矩阵，其逆元是角上的项为 $\frac{1}{a}$ 的矩阵。所以 G 是一个群，而且每个元素的行列式都是 0。 ■（另一个例子：九个项全都等于 $a \neq 0$ 的 3×3 矩阵，其单位元的所有项都是 $\frac{1}{3}$。）

回顾

- **按字面去读定义。**群的单位元就是*在该群之内*起单位元作用的那个元素；它不一定是我们熟悉的那个单位矩阵。
- **解方程求出结构。**S 的单位元和逆元都是靠解方程找出来的，而不是靠猜。
- **行列式具有可乘性。**单凭这一个性质，就说明了奇异性会由一个元素蔓延到所有元素。

背后的数学

幂等元与隐藏的可逆性

任何矩阵群的单位元 $\mathbf{E}$ 都满足 $\mathbf{E}^2 = \mathbf{E}$：它是幂等（*idempotent*）的。从几何上看，幂等矩阵就是一个投影（*projection*）：它把每个向量映射到某个子空间（subspace）上，而原本已在该子空间内的向量则保持不变。在第 (i) 部分，$\mathbf{E}$ 把平面投影到直线 $y = x$ 上，而每个矩阵 $\mathbf{M}_a$ 都是先投影到这条直线上，然后沿着它以因子 $2a$ 拉伸（图 6.8.1）。在这条直线上，这个群的作用是可逆的拉伸；它只是无视了与直线垂直的方向。所以 S 其实就是非零实数的乘法群（通过 $\mathbf{M}_a \mapsto 2a$），只不过它栖身于平面内的一条直线上。这个结论普遍成立：每个由矩阵组成的群，经过换基（change of basis）之后，都是一个由可逆矩阵组成的群，作用于其单位元的像上，其余位置则以零填充。

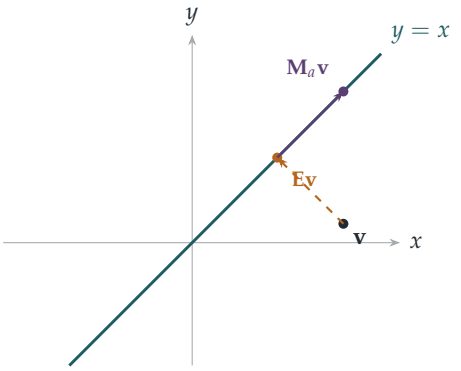


图 6.8.1　第 (i) 部分的矩阵所起的作用。单位元 $\mathbf{E}$ 把每个向量 $\mathbf{v}$ 投影到直线 $y = x$ 上（琥珀色虚线）。每个 $\mathbf{M}_a$ 都先作同样的投影，然后沿着直线以因子 $2a$ 拉伸（紫红色；图中 $2a = \frac{16}{9}$）。在这条直线本身之上，这个群的作用就是普通的可逆缩放。

抽象与具体

人们最早研究的群都是具体的：置换（permutation）、对称、可逆矩阵。抽象的定义（一个集合连同一个满足三条公理的运算）是从这些例子中提炼出来的。本题显示了抽象观点的价值：S 与非零实数的乘法群是*同构*（*isomorphic*）的，即同一个群换了一身衣服，尽管一个由奇异矩阵组成，另一个由数组成。辨认出两个外表不同的结构何时“相同”，是代数最核心的技巧之一，而数学家正是藉同构这个概念把这种“相同”说得精确。

思维技巧

质疑对熟悉对象的隐藏假设

当一个熟悉的对象出现在抽象定义中时，要查清楚定义实际上要求它具备哪些我们熟悉的性质。

“矩阵群的单位元是 $\mathbf{I}$”听起来理所当然，却不是定义的一部分；放下这个假设，就得出了一个真正崭新的例子。许多新的数学结构，正是这样靠质疑假设而发现的。

延伸练习

1 **一个同构。**证明映射 $\mathbf{M}_a \mapsto 2a$ 是由 S 到非零实数的一一对应（one-to-one correspondence），并且把矩阵乘法变成数的乘法。

2 **不是群。**证明所有形如 $\begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix}$ 且 $a \neq 0$ 的矩阵所组成的集合在乘法下封闭，但不是一个群。（提示：试找单位元。有没有左单位元？有没有右单位元?）

3 **幂等元。**求所有满足 $\mathbf{E}^2 = \mathbf{E}$ 的 2×2 实矩阵 $\mathbf{E}$，并从几何角度描述它们。

4 **所有项相等。**证明所有项都等于 $a \neq 0$ 的 $n \times n$ 矩阵在乘法下构成一个群，并求它的单位元。

延伸阅读

- **M. A. Armstrong, *Groups and Symmetry***（Springer）。讲述矩阵群与同构，并附有大量几何例子。
- **Gilbert Strang, *Introduction to Linear Algebra***（Wellesley-Cambridge）。对投影与幂等矩阵有清晰的讲解。
- **Joseph Gallian, *Contemporary Abstract Algebra***（CRC Press）。一本平易近人的入门教材，载有许多不寻常的群的例子。

我们学到了甚么

本章开首，我们谈到代数的重心由解方程转移到研究结构，并提出了三个问题：甚么可以由甚么构造出来？方程的公式为何行得通？抽象化带来了甚么好处？以下是本章八道题目所揭示的答案。

1. 甚么可以由甚么构造出来？

在 Section 6.3 中，我们逐一消去根式（radical），找出 $\sqrt{2}+\sqrt{3}$、$\sqrt{2}+\sqrt{3}+\sqrt{5}$ 和 $\sqrt[3]{2}+\sqrt[3]{3}$ 所满足的多项式，次数分别为 4、8 和 9。这些次数是不变量（invariant）：它们就是该数所生成的域（field）的次数。由于次数沿着域塔（tower of fields）逐层相乘，凡是可以用平方根得到的数，其次数必为 2 的幂。单凭这一点，便足以证明倍立方不可能做到（$\sqrt[3]{2}$ 的次数是 3），也证明 60° 角不可能三等分。

反过来，Section 6.2 展示了某些东西可以怎样构造出来。把 17 次单位根（roots of unity）分拆成高斯（Gauss）周期（period）之后，和 $\cos\frac{2\pi}{17}+\cos\frac{4\pi}{17}+\cos\frac{8\pi}{17}+\cos\frac{16\pi}{17}$ 满足一个二次方程，其根为 $\frac{-1+\sqrt{17}}{4}$；再解三个二次方程便可得到 $\cos\frac{2\pi}{17}$，从而作出正 17 边形。这两节的背后，都是 Section 6.1 的分圆多项式（cyclotomic polynomial）：它们是 x^n-1 的基本构件，把单位根按阶（order）分类便可得到。

2. 方程的公式为何行得通？

Section 6.4 揭示了四次方程（quartic）的秘密：由四个根组成三个量 $\alpha\beta+\gamma\delta$、$\alpha\gamma+\beta\delta$、$\alpha\delta+\beta\gamma$，每个量对应把四个根两两配对的一种方法。无论怎样重新排列这些根，都只会把这三个量互相调换，所以它们满足一个系数已知的三次方程。先解这个三次方程，再解两个二次方程，便得出四个根。拉格朗日（Lagrange）看出，这个方法之所以行得通，在于四个物件的置换（permutation）如何作用于那三种配对；伽罗瓦（Galois）则看出，这个方法对五个根失效，是因为五个物件的置换群（group）之中，有一部分无法再分解。这些公式，其实是群结构的影子。

3. 抽象化带来了甚么好处？

最后四节都以物件所遵守的规则来描述物件。

- 在 Section 6.5 中，四条规则（可加性、乘积法则、常数、$\Diamond x=1$）迫使一个运算必然是求导（differentiation），完全不涉及极限。同样的规则，正是现代软件中自动微分

（automatic differentiation）的基础。

- 在 Section 6.6 中，由群公理（axiom）出发的一个三行论证，一举对所有群证明了：没有任何群是两个真子群（proper subgroup）的并集。
- 在 Section 6.7 中，一个定义在复数对上的古怪运算，原来就是哈密顿（Hamilton）的四元数（quaternion）乘法；用矩阵来表示它，结合律（associativity）便自动成立。
- 在 Section 6.8 中，按字面理解群的定义，便得出一个由奇异矩阵（singular matrix）组成的群，它的单位元（identity）并不是单位矩阵，而它“其实”是非零实数群的化身。

在每一个例子中，抽象的观点都解释了一些单靠计算会被掩盖的东西。

4. 代数工具箱

方法	在哪里用过
按唯一的不变量分类	按阶把单位根分类（6.1）
由和与积得出二次方程	高斯周期（6.2）；四次方程（6.4）
积化和差公式	正 17 边形（6.2）
消去根式	最小多项式（minimal polynomial）（6.3）
次数与塔定律（tower law）	作图的不可能性（6.3）
根的对称函数（symmetric function）	三次预解式（resolvent cubic）（6.4）
由公理确定一个运算	导子（derivation）（6.5）
由群公理出发论证	子群的并集（6.6）
矩阵表示（matrix representation）	四元数（6.7）
同构（isomorphism）	奇异矩阵群（6.8）

下一步

- 四元数通往 Section 7.8 的拉格朗日四平方和定理（four-square theorem）。
- 把矩阵视为自有一套规则的代数物件，是第 7 章的主题。
- 像高斯整数（Gaussian integers）这类数系的算术，当中因数分解（factorisation）可能失效，是 Section 1.7 的主题。

如果你想在上大学之前再走远一点

- **Mario Livio, *The Equation That Couldn't Be Solved***：五次方程、伽罗瓦与群论的诞生。
- **Nathan Carter, *Visual Group Theory***：用图像认识群，是初次接触的理想读物。

- **M. A. Armstrong, *Groups and Symmetry***：建基于几何的大学入门课程。
- **Ian Stewart, *Galois Theory***：留待你准备好了解整个故事的时候。

伽罗瓦在一场决斗之后去世，年仅二十岁，只留下数十页手稿，数学界花了几十年才读懂。他看出，一个方程最深层的秘密并不在它的系数之中，而在它的解的对称性之中。这个洞见——给不同的东西起同一个名字，并研究它们共有的结构——如今贯穿整个数学。你已经在多项式、多边形、四元数和矩阵中见过它发挥作用；到了大学，你几乎处处都会遇见它。

延伸练习答案

以下只是简短的答案和提示，并非完整解答：这里只列出结果和关键步骤，方便你核对自己的做法。请先认真做过每一道练习，然后才看答案。

6.1 以单位根为根的多项式

1. $(x-1)(x+1) = x^2-1$ 而 $(x^2+x+1)(x^2-x+1) = x^4+x^2+1$，两者之积为 x^6-1。同理，$x^{12}-1 = (x^6-1)(x^6+1)$，而 $x^6+1 = (x^2+1)(x^4-x^2+1) = \mathrm{C}_4\mathrm{C}_{12}$。

2. $2p$ 的因数（divisor）为 1、2、p、$2p$，所以 $\mathrm{C}_{2p} = \frac{x^{2p}-1}{(x-1)(x+1)\mathrm{C}_p(x)} = \frac{x^p+1}{x+1}$。另一方面，$\mathrm{C}_p(-x) = \frac{(-x)^p-1}{-x-1} = \frac{x^p+1}{x+1}$，因为 p 是奇数。当 $p=3$ 时：$\mathrm{C}_6 = x^2-x+1 = \mathrm{C}_3(-x)$。

3. 不超过 p^k 而又与 p^k 有公因数的数，就是那 p^{k-1} 个 p 的倍数。本原（primitive）p^k 次单位根是 $x^{p^k}-1$ 的根之中不属于 $x^{p^{k-1}}-1$ 的根的那些，所以 $\mathrm{C}_{p^k} = \frac{x^{p^k}-1}{x^{p^{k-1}}-1} = \mathrm{C}_p\left(x^{p^{k-1}}\right)$。因此 $\mathrm{C}_9 = x^6+x^3+1$。

4. $\mathrm{C}_{15} = \frac{(x^{15}-1)(x-1)}{(x^3-1)(x^5-1)}$。把 $(x-1)(x^2+x+1)(x^4+x^3+x^2+x+1)$ 与题中给出的八次式相乘并展开，验证结果为 $x^{15}-1$。

6.2 高斯的正十七边形

1. $\eta+\eta'$ 是全部六个非实七次单位根之和，即 -1。展开 $\eta\eta'$ 得到九项 α^{a+b}，其指数为 $4,6,0,5,0,1,0,2,3$（模 7）：即三个 1，以及每个非实根各出现一次，所以 $\eta\eta' = 3-1 = 2$。因此 η、η' 就是 $\frac{-1\pm\mathrm{i}\sqrt{7}}{2}$。$\eta$ 的实部给出余弦之和。它的虚部是 $\sin\frac{2\pi}{7} + \sin\frac{4\pi}{7} - \sin\frac{\pi}{7}$，由于 $\sin\frac{4\pi}{7} = \sin\frac{3\pi}{7} > \sin\frac{\pi}{7}$，这个值为正；所以它等于 $+\frac{\sqrt{7}}{2}$。

2. 设 $\zeta = \mathrm{e}^{2\pi\mathrm{i}/5}$，则两个周期（period）$\zeta+\zeta^4 = 2\cos\frac{2\pi}{5}$ 和 $\zeta^2+\zeta^3$ 之和为 -1，之积为 -1，所以它们是 $t^2+t-1=0$ 的根，而其中的正根给出 $\cos\frac{2\pi}{5} = \frac{\sqrt{5}-1}{4}$。作图方法：在单位圆内，两条直角边为 1 和 $\frac{1}{2}$ 的直角三角形，斜边长为 $\frac{\sqrt{5}}{2}$；减去 $\frac{1}{2}$ 再取一半，便得 $\cos\frac{2\pi}{5}$。在轴上这个距离处作垂线，它与圆相交于一个与 $(1,0)$ 相邻的顶点；再用这条弦沿圆周逐步截取。

3. 两者准确至小数点后五位都是 -1.28078。

4. 3 的各次幂依次为 $3,9,10,13,5,15,11,16,14,8,7,4,12,2,6,1$：十六个数全部出现。2 的各次幂在八步之后重复（$2^8 = 256 \equiv 1$），给出 $\{1,2,4,8,9,13,15,16\}$，即模 17 的

非零平方数。

6.3 以 $\sqrt{2}+\sqrt{3}+\sqrt{5}$ 为根的多项式

1. $y^3 = 2\sqrt{2}+6\sqrt{3}+9\sqrt{2}+3\sqrt{3} = 11\sqrt{2}+9\sqrt{3}$, 而 $9y = 9\sqrt{2}+9\sqrt{3}$, 所以 $y^3-9y = 2\sqrt{2}$。于是 $\sqrt{3} = y-\sqrt{2}$ 和 $\sqrt{6} = \sqrt{2}\sqrt{3}$ 也都是 y 的多项式。

2. 设 $x-\sqrt{2} = \sqrt[3]{3}$ 并取立方：$x^3+6x-3 = \sqrt{2}(3x^2+2)$。两边平方得 $x^6-6x^4-6x^3+12x^2-36x+1=0$。

3. 可能的有理根只有 ±1, 而两者都不是根。由于没有 x^3 项, 分解成两个首一 (monic) 二次式时必定形如 $(x^2+ax+b)(x^2-ax+c)$, 其中 $a(c-b)=0$、$b+c-a^2=-10$ 及 $bc=1$。若 $a=0$, 则 $b+c=-10$ 且 $bc=1$, 在整数范围内不可能; 若 $b=c=\pm1$, 则 $a^2=12$ 或 8，同样不可能。所以这个四次式不可约（irreducible），而 $\sqrt{2}+\sqrt{3}$ 的次数（degree）为 4。

4. 由棣美弗 (De Moivre) 定理, $\cos 3\theta = 4\cos^3\theta - 3\cos\theta$; 取 $\theta = 20^\circ$, 得 $\frac{1}{2} = 4c^3-3c$。候选的有理根 $\pm1, \pm\frac{1}{2}, \pm\frac{1}{4}, \pm\frac{1}{8}$ 全都不成立。所以 $\cos 20^\circ$ 的次数为 3，但每个可作图（constructible）长度的次数都是 2 的幂，因此无法由 60° 作出 20°。

6.4 解四次方程

1. $y_1y_2+y_2y_3+y_3y_1$ 是十二项 $\alpha^2\beta\gamma$（一个根的平方乘以另外两个根）之和。而 $\left(\sum\alpha\right)\left(\sum\alpha\beta\gamma\right)$ 则是这十二项再加上 $4\alpha\beta\gamma\delta$。所以所求之和为 $(-p)(-r)-4s$。

2. 比较系数得 $a+b=1$、$a\alpha+b\beta=0$ 及 $a\alpha^2+b\beta^2=-p$，当 $\alpha\neq\beta$ 时有解，而这正是那两个条件所保证的：$p\neq0$ 使关于 t 的方程是真正的二次方程（若 $p=0$，它会退化成 $-qt=0$，只有单一个根 0），而 $q^2\neq4p^3$ 则使它的判别式（discriminant）q^2-4p^3 不为零。对于 $x^3-24x+48$：$p=8$，$q=48$，而 $8t^2-48t+64=0$ 给出 $\alpha=2$，$\beta=4$。于是 $\left(\frac{x-2}{x-4}\right)^3 = -\frac{b}{a} = \frac{\alpha}{\beta} = \frac{1}{2}$，所以 $x-2 = 2^{-1/3}(x-4)$，整理后即得题中给出的根，$x\approx-5.6946$。

3. 预解三次式 (resolvent cubic) 为 $y^3-7y^2+16y-12 = (y-3)(y-2)^2$。由根 $y=3$ 得出 $(x^2-2x+1)(x^2-2x+2)$，所以各根为 1（二重根）和 $1\pm\mathrm{i}$。

4. 在每一对之内可以对换或不对换（4 种选择），两对之间也可以互换或不互换（2 种选择）：恒等置换、$(\alpha\beta)$、$(\gamma\delta)$、$(\alpha\beta)(\gamma\delta)$、$(\alpha\gamma)(\beta\delta)$、$(\alpha\delta)(\beta\gamma)$、$(\alpha\gamma\beta\delta)$ 和 $(\alpha\delta\beta\gamma)$。这与正方形的对称群（symmetry group）结构完全相同。

6.5 从公理出发的微分法

1. 常数仍然给出 0,而由规则 (i) 及归纳法(induction)可得 $\Diamond x^n = x^{n-1}\Diamond x + x\Diamond x^{n-1} = nx^n$。再由线性（linearity）得 $\Diamond\mathrm{h} = x\mathrm{h}'$。

2. $\Diamond \mathrm{f} = \frac{\mathrm{f}}{\mathrm{g}} \Diamond \mathrm{g} + \mathrm{g} \Diamond \frac{\mathrm{f}}{\mathrm{g}}$，所以 $\Diamond \frac{\mathrm{f}}{\mathrm{g}} = \frac{\mathrm{g} \Diamond \mathrm{f} - \mathrm{f} \Diamond \mathrm{g}}{\mathrm{g}^2}$。

3. 若 $\mathrm{f} = (x-a)^2 \mathrm{g}$，则 $\mathrm{f}' = (x-a)\big(2\mathrm{g} + (x-a)\mathrm{g}'\big)$。$x^n - 1$ 的重根（repeated root）必定也是 nx^{n-1} 的根，所以只能是 0；但 0 并不是 $x^n - 1$ 的根。

4. $(2+\varepsilon)^2 + 1 = 5 + 4\varepsilon$，而 $(5+4\varepsilon)^3 = 125 + 300\varepsilon$。所以 $f(2) = 125$ 及 $f'(2) = 300$；用链式法则（chain rule）则得 $3(x^2+1)^2 \cdot 2x = 3 \cdot 25 \cdot 4 = 300$。

6.6 子群的并集

1. 设 $V = \{e, a, b, c\}$：$a^2 = b^2 = c^2 = e$，而 a、b、c 之中任意两个之积都是第三个。每个元素都是自己的逆元（inverse）。在 $\{1, \mathrm{i}, -1, -\mathrm{i}\}$ 中，i 的逆元是 $-\mathrm{i}$，所以这两个群并不相同。

2. 假如 g 属于某个真子群（proper subgroup）H，那么 g 的每个幂也都属于它，于是 H 就是整个群。

3. 1 既不是偶数，也不是 3 的倍数。更一般地，$\mathbb{Z}$ 是循环群（cyclic group，由 1 生成），所以根据上一题，它不是任意多个真子群的并集（union）。

4. 直线 $y = mx$ 与 $y = x^2$ 只相交于 $x = 0$ 和 $x = m$ 两处，而直线 $x = 0$ 与它只相交于原点。所以 k 条直线最多只能包含 k 个形如 (t, t^2) 而 $t \neq 0$ 的点，但这样的点有无穷多个。

6.7 四元数群

1. $C \circ B = (\mathrm{i}, 0) \circ (0, 1) = (0, \mathrm{i}) = D$，而 $B \circ C = (0, 1) \circ (\mathrm{i}, 0) = (0, \mathrm{i}^*) = (0, -\mathrm{i}) = H$。

2. A 的阶（order）为 1，E 的阶为 2；其余六个元素的平方都是 E，所以阶为 4。正方形的对称群有五个阶为 2 的元素（四个反射和一个半周旋转），所以这两个群并不同构（isomorphic）。

3. $\det M = z_1 z_1^* + z_2 z_2^* = |z_1|^2 + |z_2|^2$，这是四个实数平方之和。由于 $M(z)M(w)$ 是 $z \circ w$ 的矩阵，把行列式（determinant）相乘，便可把 $(a_1^2 + \cdots + a_4^2)(b_1^2 + \cdots + b_4^2)$ 写成 $|u_1|^2 + |u_2|^2$，同样是四个平方之和。

4. 各子群为 $\{A\}$、$\{A, E\}$、$\{A, E, C, G\}$、$\{A, E, B, F\}$、$\{A, E, D, H\}$ 和 Q_8。除 A 以外，每个元素都有某个幂等于 E，所以除 $\{A\}$ 以外的任何子群都包含 E。三个阶为 4 的子群合起来包含了全部八个元素。

6.8 由奇异矩阵组成的群

1. $\mathbf{M}_a \mathbf{M}_b = \mathbf{M}_{2ab}$，而 $2(2ab) = (2a)(2b)$。这个映射是一对一（one-to-one）的，并且映成（onto）全体非零实数。

2. $\begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix}\begin{pmatrix} c & d \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} ac & ad \\ 0 & 0 \end{pmatrix}$，其中 $ac \neq 0$，所以这个集合是封闭（closed）的。每个 $\begin{pmatrix} 1 & e \\ 0 & 0 \end{pmatrix}$ 都是左单位元（left identity）。但若要 $\mathbf{XE} = \mathbf{X}$，其中 $\mathbf{E} = \begin{pmatrix} e & f \\ 0 & 0 \end{pmatrix}$，便需要 $e = 1$ 及 $f = b/a$，而后者取决于 $\mathbf{X}$：所以没有右单位元，因而没有单位元，也就不成群。

3. $\mathbf{0}$、$\mathbf{I}$，以及满足 $bc = a(1-a)$ 的矩阵 $\begin{pmatrix} a & b \\ c & 1-a \end{pmatrix}$（迹（trace）为 1，行列式为 0）。这些就是把平面沿一条直线的方向投影（projection）到另一条直线上的变换。

4. 所有元素均为 a 的矩阵 $\mathbf{J}_a$ 满足 $\mathbf{J}_a\mathbf{J}_b = \mathbf{J}_{nab}$。单位元是 $\mathbf{J}_{1/n}$，而 $\mathbf{J}_a$ 的逆元是 $\mathbf{J}_{1/(n^2a)}$。

章

7

线性代数与几何

矩阵不是一张数字表。它是空间的一种变换，而且自有其生命。

本章内容

7.1	对矩阵不成立的法则	STEP II 1993
7.2	只伸缩，不转向	STEP III 1990
7.3	甚么才算行列式	STEP II 2019
7.4	回到起点的矩阵	STEP II 2020
7.5	三维空间中的旋转	STEP III 2022
7.6	双曲旋转与相对论	STEP III 1991
7.7	永恒不变的量	STEP III 2024
7.8	四个平方数	STEP III 1994
7.9	只有五种柏拉图立体	STEP III 1999
A	练习答案	第 495 页

空间的变换

我们思考时不依赖基底，写作时也不依赖基底，但到了紧要关头，我们就关上办公室的门，拼命地用矩阵计算。

— Irving Kaplansky，谈他与 Paul Halmos 处理线性代数的方式

在中学，矩阵（matrix）通常以解联立方程的工具出现，或者是变换点的一套步骤：把矩阵乘以一个列向量（column vector），然后读出新的坐标。两者都有用，但它们掩盖了矩阵之所以成为数学中最重要概念之一的原因。矩阵是一个*线性变换*（linear transformation）：一种把空间拉伸、旋转、反射和错切的方式，而且矩阵可以相加、相乘、取幂和求导，就像数一样。只不过它们又不像数。矩阵的乘法取决于次序，两个矩阵可以相乘得零，而且它们有隐藏的不变量（invariant）——迹（trace）、行列式（determinant）、特征值（eigenvalue）——无论矩阵写成甚么样子，这些量都能刻划它的本质。

线性代数为何重要

线性代数（linear algebra）大概是应用最广的数学分支。每一个电脑图像引擎，由电子游戏到电影动画，都是一部做矩阵乘法的机器：角色的每一次旋转、每一个镜头角度、每一个透视投影，都是一个矩阵。互联网上的搜索引擎和推荐系统建基于特征向量（eigenvector）。工程师利用特征值求出桥梁和建筑物的振动频率；统计学家利用它们找出数据中的主要规律；量子力学完全以线性代数的语言写成。而现代的机器学习，核心就是数量极其庞大的矩阵乘法。

线性代数无处不在，是有原因的。线性问题是我们能够彻底解决的问题，而几乎每一个非线性问题，只要放大到足够近，看起来都是线性的。这正是微积分的想法，也意味着几乎所有事物的局部性态，都是用线性代数这种语言来描述的。

你在大学会遇到甚么

线性代数与分析（analysis）并列为大学一年级的两大支柱。你会遇到*向量空间*（vector space，即可以做加法和纯量乘法的集合，平面和三维空间只是最初的例子）以及它们之间的*线性映射*（linear map），并且学到矩阵其实就是用坐标写出来的线性映射。你会学习行列式、特征值和特征向量，以及*谱定理*（spectral theorem），它指出对称矩阵（symmetric matrix）可以藉一个旋转来对角化（diagonalise）。往后的课程处处用到这些内容：在微分方程（第 8 章）中，在几何中（旋转和反射在那里构成群），在描述连续对称的*李群*（Lie

group）理论中，以及在泛函分析（functional analysis）中，也就是无限多维的线性代数。

贯穿本章的问题

三大问题

1. **矩阵在哪些方面像数，又在哪些方面不同？**哪些算术法则仍然成立，哪些不再成立？
2. **矩阵究竟做了甚么？**我们怎样才能看出一个变换的几何，又有哪些数能够不依赖坐标而刻划它？
3. **甚么保持不变？**哪些量在变换之下，或在运动之下保持不变，它们又告诉我们甚么？

本章由矩阵代数的法则开始（Section 7.1），然后转到变换的几何：特征向量、行列式、有限阶矩阵、三维空间中的旋转，以及狭义相对论中的双曲旋转（hyperbolic rotation）。接着是随时间演化的矩阵的守恒量（conserved quantity）、由行列式得出的四平方恒等式（four-square identity），最后以一个计数论证作结，即欧拉（Euler）公式，它把柏拉图立体（Platonic solids）分类。

本章的概念

§	主题	概念
7.1	不成立的法则	非交换性（non-commutativity）；零因子（zero divisor）
7.2	特征向量	变换不会转动的方向
7.3	行列式	由可乘性刻划；有向面积
7.4	有限阶矩阵	迹、行列式与凯莱–哈密顿（Cayley–Hamilton）定理
7.5	三维旋转	由叉积（cross product）得出罗德里格斯（Rodrigues）公式
7.6	洛伦兹推进（Lorentz boosts）	双曲角可以相加
7.7	守恒量	交换子（commutator）的迹为零
7.8	四平方	行列式的可乘性
7.9	柏拉图立体	算两次（double counting）与欧拉公式

如何阅读本章

对本章的每一个矩阵，都把它所做的事画出来。取单位正方形，或单位圆，或几个向量，然后画出它们的像的草图。花两分钟画一幅草图，往往在你动手计算之前就让你看到答案（一个特征向量、一个反例、一条不变直线），事后也有助你检查计算。

7.1 对矩阵不成立的法则

题目

STEP II 1993，第 6 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/93-s2-q6`

引子

矩阵（matrix）可以相加，也可以相乘，所以我们很容易把它们当作数来看待。这道题正是要考验这种想法：它列出四条对数成立的“法则”，问哪些对矩阵仍然成立。结果大部分都不成立，而它们不成立的方式很有启发性。矩阵乘法不满足交换律（not commutative）；有些非零矩阵的乘积是零；而一个矩阵的二次方程可以有无穷多个解。

这些“失效”并不是缺陷，反而正是矩阵有用的原因：矩阵描述的是空间的变换（transformation），而以不同的次序进行两个变换，确实会得到不同的结果。1925 年，海森堡（Werner Heisenberg）发现量子力学中的物理量的表现像矩阵而不像数，而 $\mathbf{AB} = \mathbf{BA}$ 不成立，正是不确定性原理（uncertainty principle）的数学形式。

动笔之前

想一想 对实数而言，为甚么 $ab = 0$ 可推出 $a = 0$ 或 $b = 0$？这个论证用到了数的哪一项性质？这项性质对矩阵成立吗？

若 $ab = 0$ 而 $a \neq 0$，两边除以 a 便得 $b = 0$。这个论证需要每个非零的数都有逆（inverse）。非零矩阵却不一定有逆（行列式（determinant）为零的矩阵就没有），所以这个论证行不通。同时也要记住几何上的图像：矩阵是平面的一个变换，而乘积 $\mathbf{AB}$ 的意思是“先做 $\mathbf{B}$，再做 $\mathbf{A}$”。

逐步解题

阶段 1　能否有 $\mathbf{AB}=0$ 但 $\mathbf{BA}\neq 0$?

想一想　从几何角度去想。你希望 $\mathbf{B}$ 把所有东西送到一条直线上，然后 $\mathbf{A}$ 把这条直线压成零。但次序调转时，$\mathbf{A}$ 不应把所有东西都压到会被 $\mathbf{B}$ 消去的地方。

设 $\mathbf{A}=\begin{pmatrix}1&0\\0&0\end{pmatrix}$，它是到 x 轴上的投影（projection），把 y 轴消去；又设 $\mathbf{B}=\begin{pmatrix}0&0\\1&0\end{pmatrix}$，它把 (x,y) 送到 $(0,x)$，即送到 y 轴上。于是

$$\mathbf{AB}=\begin{pmatrix}1&0\\0&0\end{pmatrix}\begin{pmatrix}0&0\\1&0\end{pmatrix}=\begin{pmatrix}0&0\\0&0\end{pmatrix},\qquad \mathbf{BA}=\begin{pmatrix}0&0\\1&0\end{pmatrix}\neq\mathbf{0}.$$

这个断言是**错**的（图 7.1.1）。 ■

阶段 2　平方差

小心展开，保持每个乘积的次序：

$$(\mathbf{A}-\mathbf{B})(\mathbf{A}+\mathbf{B})=\mathbf{A}^2+\mathbf{AB}-\mathbf{BA}-\mathbf{B}^2.$$

这个式子等于 $\mathbf{A}^2-\mathbf{B}^2$，当且仅当 $\mathbf{AB}=\mathbf{BA}$。取第 1 阶段的矩阵，$\mathbf{AB}-\mathbf{BA}=-\mathbf{BA}\neq\mathbf{0}$，所以这个断言是**错**的。 ■

阶段 3　$\mathbf{AX}=0$ 的非零解

想一想　若 $\det\mathbf{A}\neq 0$，你可以对 $\mathbf{AX}=\mathbf{0}$ 做甚么？若 $\det\mathbf{A}=0$，你能否找到一个非零的列向量（column vector）$\mathbf{v}$，使得 $\mathbf{Av}=\mathbf{0}$?

若 $\det\mathbf{A}\neq 0$，则 $\mathbf{A}^{-1}$ 存在，而 $\mathbf{X}=\mathbf{A}^{-1}\mathbf{AX}=\mathbf{0}$：没有非零解。

若 $\det\mathbf{A}=0$，写 $\mathbf{A}=\begin{pmatrix}a&b\\c&d\end{pmatrix}$，其中 $ad=bc$。向量 $\begin{pmatrix}b\\-a\end{pmatrix}$ 和 $\begin{pmatrix}d\\-c\end{pmatrix}$ 满足 $\mathbf{A}\begin{pmatrix}b\\-a\end{pmatrix}=\begin{pmatrix}ab-ba\\cb-da\end{pmatrix}=\mathbf{0}$，另一个向量亦同理。由于 $\mathbf{A}\neq\mathbf{0}$，两者之中至少有一个非零；把它记作 $\mathbf{v}$。那么，两列都是 $\mathbf{v}$ 的矩阵 $\mathbf{X}$ 是非零的，并且 $\mathbf{AX}=\mathbf{0}$。所以这个断言是**对**的。 ■

阶段 4　有多少个平方根?

停一停，试一试 找出非零的 $\mathbf{A}$ 和 $\mathbf{B}$，使得 $\mathbf{X}^2+\mathbf{AX}+\mathbf{B}=\mathbf{0}$ 有无穷多个解。（提示：把左边变成完全平方 $(\mathbf{X}-\mathbf{I})^2$，再想想平方为零的矩阵。）

取 $\mathbf{A}=-2\mathbf{I}$ 和 $\mathbf{B}=\mathbf{I}$，两者都非零。方程变成 $(\mathbf{X}-\mathbf{I})^2=\mathbf{0}$。任何满足 $\mathbf{N}^2=\mathbf{0}$ 的矩阵 $\mathbf{N}$ 都给出一个解 $\mathbf{X}=\mathbf{I}+\mathbf{N}$，而这样的 $\mathbf{N}$ 有无穷多个：

$$\mathbf{N}=\begin{pmatrix}0 & t\\ 0 & 0\end{pmatrix},\qquad \mathbf{N}^2=\mathbf{0}\ \text{对任意实数}t.$$

所以对每一个 t，$\mathbf{X}=\begin{pmatrix}1 & t\\ 0 & 1\end{pmatrix}$ 都是解，这个断言是**错**的。 ■

回顾

- **次序很重要。**每次展开都必须保持乘积的次序；$\mathbf{AB}$ 和 $\mathbf{BA}$ 通常并不相同。
- **零因子（zero divisors）。**非零矩阵相乘可以得零，这就破坏了“二次方程最多有两个根”背后的论证。
- **从变换的角度去想。**只要想像矩阵对平面做了甚么，每个反例都很容易找到。

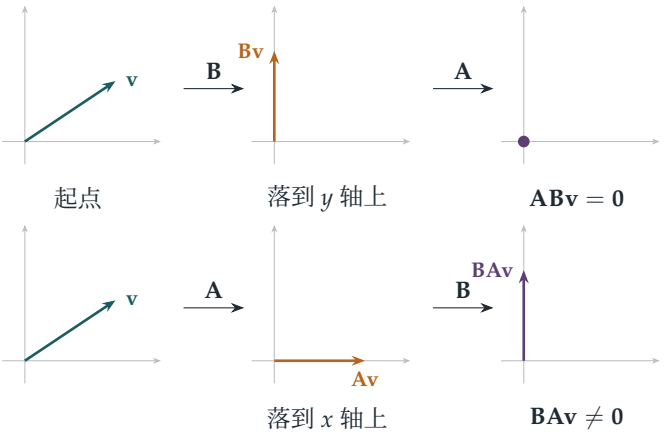


图 7.1.1 为甚么 $\mathbf{AB}=\mathbf{0}$ 但 $\mathbf{BA}\neq\mathbf{0}$。上：$\mathbf{B}$ 把每个向量送到 y 轴上，然后 $\mathbf{A}$（到 x 轴上的投影）把它压成零。下：次序调转时，$\mathbf{A}$ 保留 x 分量，$\mathbf{B}$ 再把它移到 y 轴上，它在那里得以保存下来。

背后的数学

矩阵作为一种代数

凯莱（Arthur Cayley）在 1858 年引入矩阵代数，把矩阵看成可以相加和相乘的单一对象，而不仅仅是系数表。他立刻注意到乘法不满足交换律，并证明了*凯莱—哈密顿定理*（*Cayley–Hamilton theorem*）：每个方阵都满足它自己的特征方程（characteristic equation）。对于迹（trace）为 t、行列式为 d 的 2×2 矩阵，这就是说 $\mathbf{M}^2 - t\mathbf{M} + d\mathbf{I} = \mathbf{0}$，本章会反复用到这个事实（Section 7.4）。2×2 实矩阵构成一个环（*ring*）：可以加、减、乘，但乘法不满足交换律，而且存在*零因子*，即乘积为零的非零元素。

无穷多个平方根

第 4 阶段的结论还可以变得更加惊人。方程 $\mathbf{X}^2 = \mathbf{I}$ 有明显的解 $\pm\mathbf{I}$，但平面上对任何一条通过原点的直线的反射（reflection）$\mathbf{X} = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}$ 也都是解，因为反射两次等于甚么也没做。这就是 $\mathbf{I}$ 的一整个圆那么多的平方根。系数为数的 n 次多项式最多有 n 个根，是因为数没有零因子；矩阵却有，于是这个根的数目上限便不再成立。

物理学中的非交换性

在经典力学中，位置 x 和动量 p 都是数，并且 $xp = px$。在量子力学中，它们变成算子（operator，即无穷矩阵），满足 $xp - px = \mathrm{i}\hbar$，其中 $\hbar$ 是普朗克（Planck）常数除以 2π。单凭这一条非交换关系，就可推出海森堡的不确定性原理：位置和动量不可能同时被准确得知。海森堡在 1925 年发现量子力学的法则时，并不知道矩阵是甚么；是玻恩（Max Born）认出海森堡所发明的那条奇怪的乘法法则就是矩阵乘法。

思维技巧

在新的情境中检验熟悉的法则

遇到一种新的对象时，要检查哪些熟悉的法则仍然成立，并弄清楚每条法则究竟依赖哪一项性质。

法则“$ab = 0$ 可推出 $a = 0$ 或 $b = 0$”依赖除法；“$(a-b)(a+b) = a^2 - b^2$”依赖交换律。知道每条法则建基于甚么，你就知道甚么时候仍然可以使用它。

延伸练习

1 **甚么时候可交换？** 求所有与 $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ 可交换的 2×2 矩阵。

2 **零的平方根。** 求所有满足 $\mathbf{N}^2 = \mathbf{0}$ 的 2×2 实矩阵 $\mathbf{N}$。（提示：利用凯莱–哈密顿定理证明 $\operatorname{tr}\mathbf{N} = \det\mathbf{N} = 0$。）

3 **反射。** 证明每个反射矩阵 $\begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}$ 的平方都是 $\mathbf{I}$，并求出它所对应的反射轴。

4 **凯莱–哈密顿定理。** 以直接计算证明每个 2×2 矩阵都满足 $\mathbf{M}^2 - (\operatorname{tr}\mathbf{M})\mathbf{M} + (\det\mathbf{M})\mathbf{I} = \mathbf{0}$。

延伸阅读

- **Gilbert Strang, *Introduction to Linear Algebra***（Wellesley-Cambridge）。最受欢迎的线性代数入门教科书，着重理解。
- **Sheldon Axler, *Linear Algebra Done Right***（Springer；可在网上免费阅读）。取向较为抽象，以线性映射而非矩阵为中心。
- **Jim Baggott, *The Quantum Story***（Oxford University Press）。量子力学的历史，包括海森堡的矩阵力学。

7.2 只伸缩，不转向

题目

STEP III 1990，第 6 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/90-s3-q6`

引子

矩阵变换（matrix transformation）通常会把每个向量转到新的方向。不过，许多矩阵都有几个特别的方向，它完全不会转动这些方向，只会沿着它们拉伸或压缩。这些方向上的向量就是矩阵的*特征向量*（eigenvector，源自德文 *eigen*，意思是“自身的”或“特有的”），而拉伸的倍数就是它的*特征值*（eigenvalue）。当特征向量的数目足够时（像本题这类对称矩阵就必定足够），这个变换便一目了然：在合适的坐标下，它只不过是沿两条轴的拉伸。

本题要找出某个矩阵的特征向量，再用它们来理解这个矩阵对一个圆有甚么作用。特征向量是整个应用数学中最有用的概念之一。它们描述桥梁的固有振动、旋转物体的主轴、原子的能级、一团数据中最重要的方向，以及 Google 最初的 PageRank 算法怎样为网页排名。

动笔之前

? **想一想** T 使一个向量“方向不变，只乘以一个纯量”，这是甚么意思？试把它写成一条等式。

意思是存在某个数 λ 使得 $T\mathbf{v} = \lambda\mathbf{v}$：$\mathbf{v}$ 是特征值为 λ 的特征向量。留意题目中的两个向量 $(1, 2)$ 和 $(2, -1)$ 互相垂直。这并非巧合：这个矩阵是*对称*（symmetric）的，而对称矩阵总有互相垂直的特征向量。

逐步解题

阶段 1　特征向量

$$T\begin{pmatrix}1\\2\end{pmatrix}=\frac{2}{5}\begin{pmatrix}9-4\\-2+12\end{pmatrix}=\frac{2}{5}\begin{pmatrix}5\\10\end{pmatrix}=2\begin{pmatrix}1\\2\end{pmatrix},\qquad T\begin{pmatrix}2\\-1\end{pmatrix}=\frac{2}{5}\begin{pmatrix}20\\-10\end{pmatrix}=4\begin{pmatrix}2\\-1\end{pmatrix}.$$

所以 $(1,2)$ 被拉伸至 2 倍，而 $(2,-1)$ 被拉伸至 4 倍。 ■

阶段 2　圆的像

> **想一想**　要求出 E 的方程，可先用 X 和 Y 表示 x 和 y（即求矩阵的逆），再代入 $x^2+y^2=1$。

矩阵 $\frac{2}{5}\begin{pmatrix}9&-2\\-2&6\end{pmatrix}$ 的行列式（determinant）是 $\frac{4}{25}(54-4)=8$，它的逆矩阵（inverse）是 $\frac{1}{20}\begin{pmatrix}6&2\\2&9\end{pmatrix}$。所以 $x=\frac{6X+2Y}{20}$ 及 $y=\frac{2X+9Y}{20}$，而 $x^2+y^2=1$ 变成

$$(6X+2Y)^2+(2X+9Y)^2=400\quad\Longrightarrow\quad 40X^2+60XY+85Y^2=400,$$

两边除以 5，便得 $8X^2+12XY+17Y^2=80$。 ■

> **想一想**　E 是甚么形状？它的面积是多少？试利用特征向量作答。

单位圆在每个方向上的半径都是 1。在 $(2,-1)$ 的方向上，它被拉伸至 4 倍；在与之垂直的 $(1,2)$ 方向上，则被拉伸至 2 倍。所以 E 是一个椭圆（ellipse），半轴长为 4 和 2，分别沿着两个特征向量的方向（图 7.2.1），其面积为 $\pi\times4\times2=8\pi$。（换个说法：任何线性映射（linear map）都会把面积乘以 $|\det T|=8$，而圆的面积是 π。）

阶段 3　X 的最大值

> **想一想**　在 C 上，$X=\frac{2}{5}(9x-2y)$。当 $x^2+y^2=1$ 时，$9x-2y$ 最大可以是多少？

$9x-2y$ 是 $(9,-2)$ 与单位向量 (x,y) 的点积（dot product），所以它的最大值是 $|(9,-2)|=\sqrt{85}$，在 (x,y) 指向 $(9,-2)$ 的方向时取得。因此 X 的最大值是 $\frac{2}{5}\sqrt{85}=2\sqrt{\frac{85}{25}}=2\sqrt{\frac{17}{5}}$。 ■

阶段 4　切线

停一停，试一试 C 上的点 $\frac{1}{5}(3,4)$ 的切线（tangent）是 $3x+4y=5$。线性映射会把切线变成切线。试利用这一点求出 E 的切线。

把 $x=\frac{6X+2Y}{20}$、$y=\frac{2X+9Y}{20}$ 代入 $3x+4y=5$：

$$\frac{3(6X+2Y)+4(2X+9Y)}{20}=5 \quad\Longrightarrow\quad 26X+42Y=100,$$

即 $13X+21Y=50$。 ■

（验算：该点本身的像是 $T(\frac{3}{5},\frac{4}{5})=(\frac{38}{25},\frac{36}{25})$，而 $13\cdot\frac{38}{25}+21\cdot\frac{36}{25}=\frac{494+756}{25}=50$。）

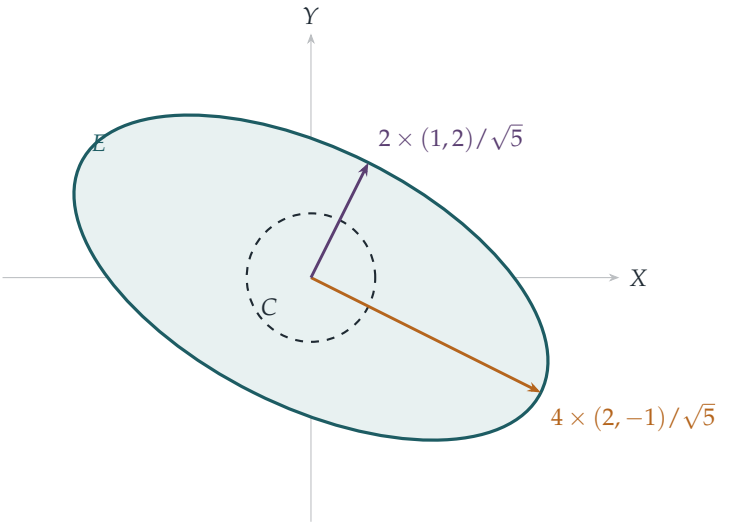


图 7.2.1 单位圆 C（虚线）及它在 T 之下的像 E。沿着特征向量 $(2,-1)$，圆被拉伸至 4 倍；沿着与之垂直的特征向量 $(1,2)$，则被拉伸至 2 倍。所以 E 是一个椭圆，半轴长为 4 和 2，分别沿着这两个方向，面积为 8π。

回顾

- **特征向量是天然的坐标轴。**在特征向量所给出的坐标下，T 只是分别拉伸至 4 倍和 2 倍，它对圆的作用显而易见。
- **行列式量度面积。**E 的面积是 C 的面积的 $|\det T|$ 倍。
- **线性映射保持相切关系。**把圆的切线加以变换，几乎不用计算便得到椭圆的切线。

背后的数学

谱定理

每个实对称矩阵都有一整组互相垂直的特征向量，而且特征值都是实数。这就是*谱定理*（spectral theorem），是大学第一门线性代数（linear algebra）课程的核心结果之一。它告诉我们，每个对称的线性映射，在合适的垂直坐标下，都只是沿各条轴的拉伸。换个说法，每个像 $8X^2 + 12XY + 17Y^2$ 这样的二次型（quadratic form），都可以经旋转化成 $\lambda_1 u^2 + \lambda_2 v^2$ 的形式，所以每条有心圆锥曲线（central conic）都是轴互相垂直的椭圆或双曲线（*主轴定理*，principal axis theorem）。这个想法可追溯至 1820 年代的柯西（Cauchy），他用它来求二次曲面的主轴，以及固体内应力的主轴。

这一切之所以成立，全靠对称性。不对称的矩阵就未必这样合作：旋转 90° 会转动每一个方向，所以它完全没有实特征向量；而错切（shear）$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ 只有一个特征向量方向，所以它不是沿任何一对轴的拉伸。

无处不在的特征向量

以下是特征向量出现的其中几个地方。

- **振动。**由质点和弹簧组成的系统会以*简正模态*（normal mode）振动，即各部分以同一频率同步运动的模式。这些模态是某个矩阵的特征向量，而频率则来自特征值（Section 11.2）。
- **数据。**一团数据点分布得最散的方向，就是其协方差矩阵（covariance matrix）的特征向量。找出这些方向的方法称为*主成分分析*（principal component analysis），是统计学和机器学习的基本工具之一（Section 13.7）。
- **量子力学。**量度一个物理量可能得到的结果，是某个矩阵（或算子）的特征值，而原子的能级也是特征值。
- **网上搜索。**Google 的 PageRank 是这个搜索引擎背后的原始构想：有一个描述网页之间连结的巨型矩阵，PageRank 就按它那个特征值为 1 的特征向量的各个分量来为网页排名。

思维技巧

选择配合变换的坐标

要理解一个线性变换，就找出它不会转动的方向。在这些坐标下，它会变得简单。

这与 Section 5.6 中选费马点（Fermat point）作原点，或 Section 2.8 中以变量代换来简化映射，是同一个想法：选对了坐标，结构便一目了然。

延伸练习

1 **求特征值。**证明 $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ 的特征值是 $\lambda^2 - (a+d)\lambda + (ad-bc) = 0$ 的根，并验证题目中的矩阵给出 $\lambda = 2$ 和 4。

2 **互相垂直的特征向量。**证明若 $\mathbf{M}$ 是对称矩阵，且 $\mathbf{Mu} = \lambda\mathbf{u}$、$\mathbf{Mv} = \mu\mathbf{v}$，其中 $\lambda \neq \mu$，则 $\mathbf{u}\cdot\mathbf{v} = 0$。（提示：用两种方法计算 $\mathbf{v}\cdot\mathbf{Mu}$。）

3 **旋转。**证明除非 θ 是 π 的倍数，否则旋转矩阵 $\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ 没有实特征向量。它的复特征值是甚么？

4 **幂。**利用特征向量，求 $T^n \begin{pmatrix} 1 \\ 0 \end{pmatrix}$。（提示：把 $(1,0)$ 写成 $(1,2)$ 和 $(2,-1)$ 的组合。）

延伸阅读

- **Gilbert Strang, *Introduction to Linear Algebra***（Wellesley-Cambridge）。第 6 章是特征值及其应用的出色入门。
- **Kurt Bryan & Tanya Leise, "The $25,000,000,000 eigenvector: the linear algebra behind Google"**，*SIAM Review* 48 (2006)。浅白易读，可免费取阅。
- **3Blue1Brown, *Essence of Linear Algebra***（影片系列，可在网上免费观看）。以精美的动画呈现的，正是本节的几何概念。

7.3 甚么才算行列式

题目

STEP II 2019，第 8 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2019 年 STEP 试卷（zip 文件）内的 STEP II 试卷：

`www.ocr.org.uk/Images/696590-2019-past-papers.zip`

引子

2×2 矩阵（matrix）的行列式（determinant）是 $ad - bc$。这条公式从何而来？它为甚么如此重要？一个答案是：它是矩阵把面积放大或缩小的倍数。另一个答案，也就是本题要探讨的，是行列式基本上是矩阵上*唯一*能把矩阵乘法变成普通乘法的函数：$\det(\mathbf{MN}) = \det \mathbf{M} \det \mathbf{N}$。

题目给你一个未知函数 f，它具有这一个性质（另加两个用来排除平凡情况的宽松条件），然后要求你一步一步推导出它必然有怎样的表现。与 Section 6.5 的导子一样，这是一道单凭性质来确定一个对象的练习。到最后，f 被迫在两行对调时变号、在奇异矩阵（singular matrix）上取值为零、在错切（shear）上等于 1：这正是行列式的表现。

动笔之前

? **想一想** 题目中每个特殊矩阵各有甚么作用？对一般的 $\mathbf{M} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$，计算 **JM**、**MJ**、**KM** 和 **PM**。

JM 把 **M** 的两行对调；**MJ** 把它的两列对调。**KM** 把第二行乘以 k。**PM** 把第二行加到第一行上。这些就是高斯消元法（Gaussian elimination）中的*初等行运算*（elementary row operations）。由于 f 把乘积变成乘积，只要知道 f 在这些特殊矩阵上的值，就知道 f 对每种行运算有何反应。

逐步解题

阶段 1　单位矩阵

$\mathrm{f}(\mathbf{M}) = \mathrm{f}(\mathbf{MI}) = \mathrm{f}(\mathbf{M})\,\mathrm{f}(\mathbf{I})$。由于 $\mathrm{f}(\mathbf{M}) \neq 0$，两边相除即得 $\mathrm{f}(\mathbf{I}) = 1$。 ■

阶段 2　对调

$\mathbf{J}^2 = \mathbf{I}$，所以 $\mathrm{f}(\mathbf{J})^2 = \mathrm{f}(\mathbf{I}) = 1$，由此得 $\mathrm{f}(\mathbf{J}) = \pm 1$。由于 $\mathrm{f}(\mathbf{J}) \neq 1$，故 $\mathrm{f}(\mathbf{J}) = -1$。

现在 $\begin{pmatrix} c & d \\ a & b \end{pmatrix} = \mathbf{JM}$，所以它的值是 $\mathrm{f}(\mathbf{J})\mathrm{f}(\mathbf{M}) = -\mathrm{f}(\mathbf{M})$。而 $\begin{pmatrix} d & c \\ b & a \end{pmatrix} = \mathbf{JMJ}$（先对调两行，再对调两列），所以它的值是 $\mathrm{f}(\mathbf{J})^2\mathrm{f}(\mathbf{M}) = \mathrm{f}(\mathbf{M})$。 ■

阶段 3　线性相关的两行

> **想一想**　先考虑一个两行相等的矩阵 $\mathbf{B}$。$\mathbf{JB}$ 是甚么？这对 $\mathrm{f}(\mathbf{B})$ 说明了甚么？

若 $\mathbf{B}$ 的两行都是 (a, b)，把它们对调不会带来任何改变：$\mathbf{JB} = \mathbf{B}$。所以 $\mathrm{f}(\mathbf{B}) = \mathrm{f}(\mathbf{J})\mathrm{f}(\mathbf{B}) = -\mathrm{f}(\mathbf{B})$，这就迫使 $\mathrm{f}(\mathbf{B}) = 0$。

现在若 $\mathbf{A}$ 的两行是 (a, b) 和 $k(a, b)$，则 $\mathbf{A} = \mathbf{KB}$，其中 $\mathbf{B}$ 的两行都等于 (a, b)。所以 $\mathrm{f}(\mathbf{A}) = \mathrm{f}(\mathbf{K})\mathrm{f}(\mathbf{B}) = 0$。 ■（这包括 $k = 0$ 的情况，此时第二行为零。）

阶段 4　错切

> **停一停，试一试**　计算 $\mathbf{P}^2$ 和 $\mathbf{K}^{-1}\mathbf{PK}$。选取 k 使两者相等。由此可得 $\mathrm{f}(\mathbf{P})$ 是甚么？

$\mathbf{P}^2 = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$，而当 $k \neq 0$ 时，

$$\mathbf{K}^{-1}\mathbf{PK} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{k} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & k \end{pmatrix} = \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}.$$

取 $k = 2$，则 $\mathbf{K}^{-1}\mathbf{PK} = \mathbf{P}^2$。对两边取 f，并利用 $\mathrm{f}(\mathbf{K}^{-1})\mathrm{f}(\mathbf{K}) = \mathrm{f}(\mathbf{I}) = 1$，得

$$\mathrm{f}(\mathbf{P}) = \mathrm{f}(\mathbf{P})^2,$$

所以 $\mathrm{f}(\mathbf{P}) = 0$ 或 1。但 $\mathbf{P}$ 是可逆的 (invertible)，而 $\mathrm{f}(\mathbf{P})\mathrm{f}(\mathbf{P}^{-1}) = \mathrm{f}(\mathbf{I}) = 1$，所以 $\mathrm{f}(\mathbf{P}) \neq 0$。因此 $\mathrm{f}(\mathbf{P}) = 1$。 ■

回顾

- **特殊元素揭示函数的面貌。**把乘法性质应用于 **I**、**J**、**K** 和 **P**，便得出 f 在每个矩阵上的值，以及它对每种行运算的影响。
- **与自身的变换相等的元素。**两行相等的矩阵经行对调后保持不变，而行对调会改变 f 的符号；所以 f 在该处必为零。
- **共轭（conjugation）保持 f 不变。**$\mathrm{f}(\mathbf{K}^{-1}\mathbf{P}\mathbf{K}) = \mathrm{f}(\mathbf{P})$，这个事实在代数中一再用到。

背后的数学

每个乘性函数都是行列式的函数

题目在 $\mathrm{f}(\mathbf{P}) = 1$ 处便停下，但这个论证是可以完成的。每个可逆矩阵都是初等矩阵（elementary matrices，即行对调、行缩放和错切）的乘积，而题目已说明 f 在每一种上的表现。由此可知 $\mathrm{f}(\mathbf{M}) = \phi(\det \mathbf{M})$，其中 ϕ 是某个单一实数变量的函数，并满足 $\phi(xy) = \phi(x)\phi(y)$。行列式本身就是 $\phi(x) = x$ 的情况，但 $\phi(x) = x^3$ 或 $\phi(x) = |x|^{1/2}\operatorname{sign}(x)$ 同样可行。换言之，*行列式是矩阵上唯一真正崭新的乘性函数*。用群论（group theory）的语言来说，可逆矩阵所成的群之中，凡是交换群（commutative group）所能看见的一切，行列式都已尽数捕捉。

行列式作为面积

从几何上看，2×2 矩阵把单位正方形变成一个平行四边形，而 $\det \mathbf{M}$ 就是它的*有向面积*（signed area）：若定向（orientation）保持不变则为正，若定向反转则为负（图 7.3.1）。这样，乘法性质的意义便一目了然：相继进行两个变换，面积的比例因子便会相乘。行对调会反转定向，所以 $\det \mathbf{J} = -1$；错切把正方形推斜成面积相同的平行四边形，所以 $\det \mathbf{P} = 1$；而两行线性相关的矩阵会把正方形压扁，所以它的行列式为零。题目的每一步都有对应的图像。



图 7.3.1 行列式作为有向面积。两个基向量（basis vectors）的像（琥珀色、梅紫色）张成一个平行四边形。错切保持面积不变；对调基向量会反转定向；线性相关的两行（或两列）把正方形压缩到一条直线上。相继进行两个变换，有向面积便会相乘，这正是性质 $\mathrm{f}(\mathbf{M}\mathbf{N}) = \mathrm{f}(\mathbf{M})\mathrm{f}(\mathbf{N})$。

高维的行列式

对于 $n \times n$ 矩阵，行列式是有向体积的比例因子，并由三个性质所刻划：它对每一行都是线性的，两行对调时它会变号，而且它在单位矩阵上等于 1。单凭这三点，你就能推导出熟悉的展开公式，而高斯消元法则提供了一个高效的计算方法。行列式出现在多重积分的换元公式（雅可比行列式，Jacobian）之中，出现在逆矩阵的公式之中，并且如我们在 Section 5.4 所见，也出现在多项式的判别式（discriminant）之中。

思维技巧

用特殊输入探测未知函数

当一个函数只能通过某个性质来认识时，把这个性质应用于你能想到的最简单的特殊输入，看看它会迫使甚么结果成立。

在这里，单位矩阵、一个对调、一个缩放和一个错切，已足以确定几乎一切。选取那些等于自身的平方、逆或共轭的输入，尤其有效。

延伸练习

1 **其他例子。** 证明 $\mathrm{f}(\mathbf{M}) = (\det \mathbf{M})^3$ 满足题目的所有条件，并证明 $\mathrm{f}(\mathbf{M}) = |\det \mathbf{M}|$ 满足乘法性质，但不满足 $\mathrm{f}(\mathbf{J}) \neq 1$。

2 **列运算。** 证明对于任何第二列是第一列的倍数的矩阵，f 的值同样为零。

3 **缩放。** 设 $\mathbf{K}_k = \begin{pmatrix} 1 & 0 \\ 0 & k \end{pmatrix}$。证明 $\mathrm{f}\begin{pmatrix} k & 0 \\ 0 & 1 \end{pmatrix} = \mathrm{f}(\mathbf{K}_k)$，并证明 $\mathrm{f}(\mathbf{K}_k)\mathrm{f}(\mathbf{K}_l) = \mathrm{f}(\mathbf{K}_{kl})$。

4 **归结为行列式。** 较难 证明每个可逆的 2×2 矩阵都可以写成形如 $\mathbf{J}$、$\mathbf{K}_k$ 和 $\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$ 的矩阵的乘积。由此推导出 $\mathrm{f}(\mathbf{M})$ 只取决于 $\det \mathbf{M}$。

延伸阅读

- **Sheldon Axler, *Linear Algebra Done Right***（Springer；网上免费阅览）。从行列式的性质出发来定义行列式，与本题的精神一脉相承。
- **Gilbert Strang, *Introduction to Linear Algebra***（Wellesley-Cambridge）。第 5 章从三条规则出发建立行列式的理论。

- **3Blue1Brown, “The determinant”**（影片，网上免费观看）。以精美的动画呈现行列式的面积诠释。

7.4 回到起点的矩阵

题目

STEP II 2020，第 6 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2020 年 STEP 试卷（zip 文件）内的 STEP II 试卷：

`www.ocr.org.uk/Images/696589-2020-past-papers.zip`

引子

有些变换（transformation）只要重复足够多次，就会令一切回到原位。旋转 90° 要四步，反射则只要两步。哪些 2×2 矩阵（matrix）具有这种性质，即对某个 n 有 $\mathbf{M}^n=\mathbf{I}$？又怎样判断？

这道题令人意外之处，在于只需两个数就足以判定：迹（trace）和行列式（determinant）。你不必知道那四个元素，只需知道 $a+d$ 和 $ad-bc$。而且答案对现实世界有一个惊人的推论：它解释了为甚么晶体可以有 2 重、3 重、4 重和 6 重对称，却永远没有 5 重对称。这条晶体学“定律”屹立了两个世纪，直至 1982 年的一项发现才被打破，而该发现更赢得了诺贝尔奖。

动笔之前

想一想 每个 2×2 矩阵都满足 $\mathbf{M}^2-(\operatorname{tr}\mathbf{M})\mathbf{M}+(\det\mathbf{M})\mathbf{I}=\mathbf{0}$（这是凯莱—哈密顿定理（Cayley–Hamilton theorem）；把它乘开来验证一下）。这对处理有关 $\mathbf{M}^2$ 的问题有甚么帮助？

它把 $\mathbf{M}^2$ 用 $\mathbf{M}$ 和 $\mathbf{I}$ 表示出来：$\mathbf{M}^2=t\mathbf{M}-d\mathbf{I}$，其中 $t=\operatorname{tr}\mathbf{M}$，$d=\det\mathbf{M}$。于是，关于 $\mathbf{M}^2$ 的条件就变成一个把 $\mathbf{M}$、t 和 d 联系起来的条件。这就是整道题的关键。

逐步解题

阶段 1　平方的迹

$$\mathbf{M}^2 = \begin{pmatrix} a^2+bc & ab+bd \\ ac+cd & bc+d^2 \end{pmatrix},$$ 所以 $\operatorname{tr}(\mathbf{M}^2) = a^2 + 2bc + d^2 = (a+d)^2 - 2(ad-bc)$。 ■

阶段 2　$\pm\mathbf{I}$ 的平方根

> **想一想**　假设 $\mathbf{M}^2 = \mathbf{I}$。把它代入 $\mathbf{M}^2 = t\mathbf{M} - d\mathbf{I}$。若 $t \neq 0$，这说明 $\mathbf{M}$ 是怎样的矩阵?

若 $\mathbf{M}^2 = \mathbf{I}$，则 $\mathbf{I} = t\mathbf{M} - d\mathbf{I}$，所以 $t\mathbf{M} = (1+d)\mathbf{I}$。若 $t \neq 0$，则 $\mathbf{M}$ 是 $\mathbf{I}$ 的倍数，设为 $c\mathbf{I}$，而 $c^2 = 1$，所以 $\mathbf{M} = \pm\mathbf{I}$，但这已被排除。因此 $t = 0$，于是 $(1+d)\mathbf{I} = \mathbf{0}$，所以 $d = -1$。

反过来，若 $t = 0$ 且 $d = -1$，则 $\mathbf{M}^2 = t\mathbf{M} - d\mathbf{I} = \mathbf{I}$，而且 $\mathbf{M} \neq \pm\mathbf{I}$，因为 $\operatorname{tr}(\pm\mathbf{I}) = \pm 2 \neq 0$。 ■

至于 $\mathbf{M}^2 = -\mathbf{I}$：代入后得 $t\mathbf{M} = (d-1)\mathbf{I}$。若 $t \neq 0$，则 $\mathbf{M} = c\mathbf{I}$，其中 $c^2 = -1$，这对实数 c 是不可能的。所以 $t = 0$ 且 $d = 1$。反过来，由 $t = 0$ 和 $d = 1$ 可得 $\mathbf{M}^2 = -\mathbf{I}$。 ■

阶段 3　四次方根

> **停一停，试一试**　设 $\mathbf{N} = \mathbf{M}^2$，则 $\mathbf{M}^4 = \mathbf{I}$ 即是 $\mathbf{N}^2 = \mathbf{I}$。根据第 (ii) 部分，要么 $\mathbf{N} = \pm\mathbf{I}$，要么 $\operatorname{tr}\mathbf{N} = 0$ 且 $\det\mathbf{N} = -1$。为甚么第二种情况不可能发生?

$\det\mathbf{N} = \det(\mathbf{M}^2) = (\det\mathbf{M})^2 \geq 0$，所以 $\det\mathbf{N} \neq -1$。因此 $\mathbf{N} = \mathbf{M}^2 = \pm\mathbf{I}$。逆命题显然成立。 ■

至于 $\mathbf{M}^4 = -\mathbf{I}$：把第 (ii) 部分应用于 $\mathbf{N} = \mathbf{M}^2$，可知它成立当且仅当 $\operatorname{tr}(\mathbf{M}^2) = 0$ 且 $\det(\mathbf{M}^2) = 1$，即 $t^2 - 2d = 0$ 且 $d^2 = 1$。由于 $t^2 = 2d \geq 0$，必须有 $d = 1$，于是 $t^2 = 2$。所以

$$\mathbf{M}^4 = -\mathbf{I} \iff \det\mathbf{M} = 1 \text{ 且 } \operatorname{tr}\mathbf{M} = \pm\sqrt{2}. \qquad \blacksquare$$

阶段 4　一个不是旋转的 I 的八次方根

任何满足 $\mathbf{M}^4 = -\mathbf{I}$ 的矩阵都有 $\mathbf{M}^8 = \mathbf{I}$。旋转 45° 的矩阵的迹是 $\sqrt{2}$，行列式是 1；我们想找另一个迹和行列式都相同、但不是旋转的矩阵。取

$$\mathbf{M} = \begin{pmatrix} \sqrt{2} & -1 \\ 1 & 0 \end{pmatrix}, \qquad \operatorname{tr}\mathbf{M} = \sqrt{2}, \quad \det\mathbf{M} = 1.$$

根据第 3 步，$\mathbf{M}^4 = -\mathbf{I}$，所以 $\mathbf{M}^8 = \mathbf{I}$。但 $\mathbf{M}$ 既不是旋转也不是反射，因为旋转和反射都保持长度不变，而 $\mathbf{M}\begin{pmatrix}1\\0\end{pmatrix} = \begin{pmatrix}\sqrt{2}\\1\end{pmatrix}$ 的长度是 $\sqrt{3} \neq 1$。 ■

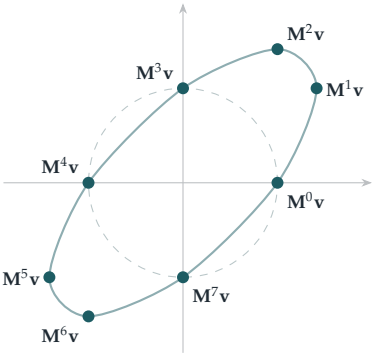


图 7.4.1 $\mathbf{v} = (1,0)$ 在 $\mathbf{M} = \begin{pmatrix}\sqrt{2} & -1\\1 & 0\end{pmatrix}$ 作用下的轨道（orbit）。八步之后它回到起点。这些点并不在同一个圆（虚线）上，所以 $\mathbf{M}$ 不是旋转，但它们确实位于一个椭圆上：$\mathbf{M}$ 是在斜坐标下看到的 45° 旋转。

回顾

- **凯莱－哈密顿定理把幂化为线性组合（linear combination）。** 2×2 矩阵的每一个幂都是 $\mathbf{M}$ 和 $\mathbf{I}$ 的组合，其系数只取决于迹和行列式。
- **平方的行列式是非负的。** 这个简单的事实在第 (iii) 部分排除了一整种情况。
- **迹和行列式是不变量（invariant）。** 两个迹和行列式都相同的矩阵（且都不是 $\mathbf{I}$ 的倍数），不论外表是否相似，取幂时的表现都一样。

背后的数学

相似矩阵

为甚么说第 4 步的 $\mathbf{M}$ 是“伪装了的旋转”？若存在可逆（invertible）矩阵 $\mathbf{S}$ 使得 $\mathbf{M} = \mathbf{SRS}^{-1}$，则称两个矩阵 $\mathbf{M}$ 和 $\mathbf{R}$ 相似（similar）：它们在不同的坐标下表示同一个变换。相似矩阵有相同的迹和行列式，而且（对于不是 $\mathbf{I}$ 的倍数的 2×2 矩阵）逆命题也成立。所以任何迹为 $\sqrt{2}$、行列式为 1 的实矩阵都与 45° 的旋转相似，而在 $\mathbf{S}$ 所给出的斜坐标下，它就只是在旋转（图 7.4.1）。一般而言，满足 $\mathbf{M}^n = \mathbf{I}$ 的 2×2 实矩阵必定与某个旋转或反射相似。

为甚么晶体没有五重对称

晶体是原子有规律、重复的排列。晶体的旋转对称把原子所成的格（lattice）映射到它自身，所以在以这个格为基础的坐标下，它由一个元素全是整数的矩阵给出。因此它的迹是整数。但是旋转角为 θ 的旋转，在任何坐标下的迹都是 $2\cos\theta$，而 $2\cos\theta$ 只有在等于 $-2,-1,0,1$ 或 2 时才是整数，也就是对应阶（order）为 2、3、4 和 6（以及 1）的旋转。五重对称需要 $2\cos 72^\circ = \frac{\sqrt{5}-1}{2}$，而这不是整数。这就是*晶体学限制定理*（crystallographic restriction theorem）。1982 年，Dan Shechtman 观察到一种看似具有五重对称的合金，起初还遭人嘲笑；后来才知道那是一种*准晶体*（quasicrystal），有序但没有周期性，就像彭罗斯铺砖（Penrose tilings）一样，因此不在该定理的涵盖范围之内。他于 2011 年获颁诺贝尔化学奖。

思维技巧

> **寻找能刻划本质行为的不变量**
>
> *当题目涉及一个有很多参数的对象时，问问自己：是否只需这些参数的几个组合，即那些转换观点后仍然不变的组合，就已经足够？*
>
> 在这里，四个元素化约为两个不变量，即迹和行列式，因为这个问题（“$\mathbf{M}^n = \mathbf{I}$ 是否成立?”）并不取决于坐标。找到合适的不变量，往往是解决分类问题的关键。

延伸练习

1. **三阶的旋转。**试做 STEP II 2021，第 7 题（载于 OCR 提供下载的 2021 年试卷，`www.ocr.org.uk/Images/696588-2021-past-papers.zip`）。它从一个具有不寻常性质的旋转矩阵出发，最后得出一个结论，适用于所有立方为单位矩阵（identity matrix）的 2×2 实矩阵。

2. **六阶。**求迹和行列式须满足的条件，使得 $\mathbf{M}^6 = \mathbf{I}$，但 $\mathbf{M}^2 \neq \mathbf{I}$ 且 $\mathbf{M}^3 \neq \mathbf{I}$。

3. **晶体学限制。**证明：若 $\mathbf{M}$ 的元素全是整数，$\det\mathbf{M} = 1$，且对某个 n 有 $\mathbf{M}^n = \mathbf{I}$，则 $\operatorname{tr}\mathbf{M} \in \{-2,-1,0,1,2\}$。找出一个阶为 6 的整数矩阵。

4. **以凯莱－哈密顿定理求幂。**利用 $\mathbf{M}^2 = t\mathbf{M} - d\mathbf{I}$，把 $\mathbf{M}^3$ 和 $\mathbf{M}^4$ 表示成 $p\mathbf{M} + q\mathbf{I}$ 的形式，并借此重新推导 $\mathbf{M}^4 = -\mathbf{I}$ 的条件。

延伸阅读

- **M. A. Armstrong, *Groups and Symmetry***（Springer）。内容包括晶体学限制和墙纸图案的分类。
- **Marjorie Senechal, *Quasicrystals and Geometry***（Cambridge University Press）。准晶体与非周期有序结构的数学。
- **Gilbert Strang, *Introduction to Linear Algebra***（Wellesley-Cambridge）。相似矩阵与凯莱—哈密顿定理。

7.5 三维空间中的旋转

题目

STEP III 2022，第 7 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2022 年 STEP 试卷（zip 文件）内的 STEP III 试卷：

`www.ocr.org.uk/Images/696587-2022-past-papers.zip`

引子

在二维空间中，旋转（rotation）一个点很容易：乘以 $e^{i\theta}$ 便可。到了三维空间，事情就困难得多。没有一个复数可以让我们直接相乘，绕不同轴的旋转不可交换（commute），而要写出“绕通过 $(1,2,2)$ 的轴旋转 $37°$”的矩阵，看来更是令人却步。

本题利用叉积（cross product），为任意的轴和任意的角构造出这个矩阵。所得的结果就是*罗德里格旋转公式*（Rodrigues′ rotation formula），由罗德里格（Olinde Rodrigues）于 1840 年发表。机械人技术、电脑图像和太空船导航每天都用到它，而它亦揭示了一个优美的事实：与 $\mathbf{n}$ 作叉积，就是“绕 $\mathbf{n}$ 旋转 $90°$”，而每一个绕 $\mathbf{n}$ 的旋转都由它构成，正如 $e^{i\theta} = \cos\theta + i\sin\theta$ 是由乘以 i 构成一样。

动笔之前

想一想 当 $\mathbf{r}$ 垂直于 $\mathbf{n}$ 时，$\mathbf{n}\times\mathbf{r}$ 在几何上是甚么样子？当 $\mathbf{r}$ 平行于 $\mathbf{n}$ 时又如何？

若 $\mathbf{r}\parallel\mathbf{n}$，则 $\mathbf{n}\times\mathbf{r}=\mathbf{0}$。若 $\mathbf{r}\perp\mathbf{n}$，则 $\mathbf{n}\times\mathbf{r}$ 的长度与 $\mathbf{r}$ 相同（因为 $|\mathbf{n}|=1$，而夹角是 $90°$），与两者皆垂直，而且根据右手定则（right-hand rule），它就是 $\mathbf{r}$ 绕 $\mathbf{n}$ 逆时针旋转 $90°$ 的结果（沿 $\mathbf{n}$ 向下看）。因此 f 消去 $\mathbf{r}$ 沿 $\mathbf{n}$ 方向的分量（component），并把垂直的分量旋转 $90°$。这幅图像足以解释整条题目。

逐步解题

阶段 1　作两次叉积

$\mathrm{f}(\mathbf{r}) = \mathbf{n}\times\mathbf{r} = (bz-cy,\ cx-az,\ ay-bx)$。再与 $\mathbf{n}=(a,b,c)$ 作一次叉积，$\mathrm{f}(\mathrm{f}(\mathbf{r}))$ 的 x 分量是

$$b(ay-bx)-c(cx-az) = -x(b^2+c^2)+aby+acz. \qquad \blacksquare$$

由于 $a^2+b^2+c^2=1$，上式等于 $-x(1-a^2)+aby+acz = a(ax+by+cz)-x$，这正是 $(\mathbf{n}.\mathbf{r})\mathbf{n}-\mathbf{r}$ 的 x 分量。其余分量可由对称性得出，所以 $\mathrm{f}(\mathrm{f}(\mathbf{r})) = (\mathbf{n}.\mathbf{r})\mathbf{n}-\mathbf{r}$。 $\blacksquare$

> **想一想**　设 $\mathbf{r}=\mathbf{r}_{\parallel}+\mathbf{r}_{\perp}$，其中 $\mathbf{r}_{\parallel}=(\mathbf{n}.\mathbf{r})\mathbf{n}$ 是沿 $\mathbf{n}$ 方向的分量。用这两部分表示，$\mathrm{f}(\mathrm{f}(\mathbf{r}))$ 是甚么？

$\mathrm{f}(\mathrm{f}(\mathbf{r})) = \mathbf{r}_{\parallel}-\mathbf{r} = -\mathbf{r}_{\perp}$。施行 f 两次，会消去平行的部分，并把垂直的部分绕 $\mathbf{n}$ 旋转 180°，亦即把它反向（图 7.5.1）。这就是 $\mathrm{i}^2=-1$ 的类比。

阶段 2　旋转

> **想一想**　计算 $\mathrm{g}(\mathbf{n})$。然后对于 $\mathbf{r}\perp\mathbf{n}$ 的情况计算 $\mathrm{g}(\mathbf{r})$，并利用 $\mathrm{f}(\mathrm{f}(\mathbf{r}))=-\mathbf{r}$ 对这样的 $\mathbf{r}$ 成立这一点。

$\mathrm{f}(\mathbf{n}) = \mathbf{n}\times\mathbf{n} = \mathbf{0}$，所以 $\mathrm{g}(\mathbf{n})=\mathbf{n}$：轴保持不动。若 $\mathbf{r}\perp\mathbf{n}$，则 $\mathrm{f}(\mathrm{f}(\mathbf{r}))=-\mathbf{r}$，于是

$$\mathrm{g}(\mathbf{r}) = \mathbf{r}+\sin\theta\,(\mathbf{n}\times\mathbf{r})-(1-\cos\theta)\mathbf{r} = \cos\theta\,\mathbf{r}+\sin\theta\,(\mathbf{n}\times\mathbf{r}).$$

由于 $\mathbf{n}\times\mathbf{r}$ 是 $\mathbf{r}$ 绕 $\mathbf{n}$ 转过 90° 的结果，长度不变，所以上式就是 $\mathbf{r}$ 绕 $\mathbf{n}$ 旋转 θ 的结果，与平面上的 $(\cos\theta+\mathrm{i}\sin\theta)z$ 如出一辙。由于 g 是线性（linear）的，而每个向量都可写成 $\mathbf{n}$ 的倍数与一个垂直于 $\mathbf{n}$ 的向量之和，所以 g 就是**绕通过原点、方向为 $\mathbf{n}$ 的轴旋转角 θ 的旋转**（从 $\mathbf{n}$ 的尖端望去为逆时针方向）。 $\blacksquare$

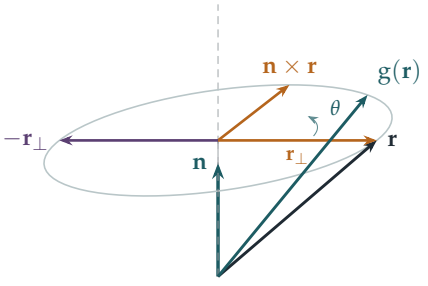


图 7.5.1　罗德里格公式。向量 $\mathbf{r}$ 分解为沿轴 $\mathbf{n}$ 的部分和与轴垂直的部分 $\mathbf{r}_{\perp}$，后者扫出一个圆。与 $\mathbf{n}$ 作叉积，会把 $\mathbf{r}_{\perp}$ 转过 90°（琥珀色）；作两次叉积则得 $-\mathbf{r}_{\perp}$（紫红色）。旋转后的向量 $\mathrm{g}(\mathbf{r})$ 是 $\mathbf{r}_{\parallel}+\cos\theta\,\mathbf{r}_{\perp}+\sin\theta\,(\mathbf{n}\times\mathbf{r})$。

阶段 3 反射

> **停一停，试一试** 利用第 (i) 部分化简 $\mathrm{h}(\mathbf{s}) = -\mathbf{s} - 2\mathrm{f}(\mathrm{f}(\mathbf{s}))$。它对沿 $\mathbf{n}$ 方向的向量有甚么作用？对垂直于 $\mathbf{n}$ 的向量又有甚么作用？

$\mathrm{h}(\mathbf{s}) = -\mathbf{s} - 2\big((\mathbf{n}.\mathbf{s})\mathbf{n} - \mathbf{s}\big) = \mathbf{s} - 2(\mathbf{n}.\mathbf{s})\mathbf{n}$。垂直于 $\mathbf{n}$ 的向量保持不变，而 $\mathrm{h}(\mathbf{n}) = -\mathbf{n}$。所以 h 是**对通过原点且垂直于 n 的平面的反射（reflection）**。 ■

回顾

- **分解为平行和垂直的部分。**一旦把 $\mathbf{r}$ 分解为沿 $\mathbf{n}$ 方向和与之垂直的分量，每一步都变得清晰。
- **线性减省工夫。**要辨认一个线性映射（linear map），只需知道它对几个精心挑选的向量有甚么作用便足够。
- **复数的类比。**在垂直于 $\mathbf{n}$ 的向量上，f 的表现就像乘以 i：$\mathrm{f}^2 = -1$，而 $\mathrm{g} = \cos\theta + \sin\theta\,\mathrm{f}$。

背后的数学

欧拉旋转定理

三维空间中每一个固定原点的旋转，都是绕某条通过原点的轴的旋转。欧拉（Euler）于 1775 年证明了这一点。用矩阵的语言来说，每个 3×3 旋转矩阵都有一个特征值（eigenvalue）为 1 的特征向量（eigenvector），也就是旋转轴。罗德里格公式给出其逆命题：对每一条轴 $\mathbf{n}$ 和每一个角 θ，旋转矩阵是

$$\mathbf{R} = \mathbf{I} + \sin\theta\,\mathbf{K} + (1-\cos\theta)\,\mathbf{K}^2, \qquad \mathbf{K} = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix},$$

其中 $\mathbf{K}$ 是 $\mathbf{r} \mapsto \mathbf{n}\times\mathbf{r}$ 的矩阵。由于 $\mathbf{K}^3 = -\mathbf{K}$，指数级数（exponential series）$e^{\theta\mathbf{K}} = \mathbf{I} + \theta\mathbf{K} + \frac{\theta^2}{2!}\mathbf{K}^2 + \cdots$ 恰好化简成这条公式，它就是 $e^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta$ 的三维版本。旋转与“无穷小旋转”$\mathbf{K}$ 的指数之间的这种联系，正是李群（Lie groups）理论的起点。

旋转不可交换

把一本书平放，先绕垂直轴旋转 90°，再绕水平轴旋转 90°；然后从头再来，以相反的次序做这两个旋转。书最终的位置并不相同。三维空间中旋转的复合（composition）不符

合交换律，与二维的情况不同。这正是 Section 6.7 的四元数（quaternion）有用的原因之一：绕 $\mathbf{n}$ 旋转 θ 对应于四元数 $\cos\frac{\theta}{2} + \sin\frac{\theta}{2}(a\mathrm{i} + b\mathrm{j} + c\mathrm{k})$，而复合旋转就是把四元数相乘，不可交换性已包含在其中。罗德里格本人在 1840 年的论文中，已找到旋转的复合公式，比哈密顿（Hamilton）发明四元数还早三年。

思维技巧

分解为不变的部分

要理解一个变换，可把空间分成变换处理得简单的几块：在这里，就是轴（保持不动）和与轴垂直的平面（被旋转）。

这与 Section 7.2 中特征向量的想法相同：找出变换所保持的子空间（subspace），然后在每一块上分别研究它。

延伸练习

1 **平面上的旋转。**试做 STEP III 2017, 第 2 题（`step.maths.org/questions/17-s3-q2`）。该题利用复数，证明平面上的两个旋转复合起来仍是一个旋转，只有一种例外。找出这个例外，以及此时复合的结果是甚么。

2 **一个具体的旋转。**求绕方向为 $(1,1,1)$ 的轴旋转 120° 的矩阵。证明它把各坐标轴循环置换。

3 **$\mathbf{K}$ 的幂。**证明 $\mathbf{K}^2\mathbf{r} = (\mathbf{n}.\mathbf{r})\mathbf{n} - \mathbf{r}$ 及 $\mathbf{K}^3 = -\mathbf{K}$，并由指数级数推导出罗德里格公式。

4 **两次反射构成一个旋转。**证明先后对两个通过原点、互成 α 角的平面作反射，结果是绕两平面的交线旋转 2α。

延伸阅读

- **Simon Altmann, *Rotations, Quaternions, and Double Groups***（Dover）。关于三维旋转的权威论著，当中包括罗德里格的贡献。
- **Simon Altmann, "Hamilton, Rodrigues, and the quaternion scandal"**, *Mathematics Magazine* 62 (1989)。一篇简短而生动的历史叙述。
- **John Stillwell, *Naive Lie Theory***（Springer）。为大学本科生而写的旋转、四元数与李群入门。

7.6 双曲旋转与相对论

题目

STEP III 1991，第 6 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/91-s3-q6`

引子

旋转矩阵（rotation matrix）$\begin{pmatrix} \cos u & -\sin u \\ \sin u & \cos u \end{pmatrix}$ 把点沿着圆 $x^2 + y^2 = 1$ 移动，而先旋转 u 再旋转 w，就等于旋转 $u + w$。把 cos 和 sin 换成 cosh 和 sinh，再调整一个正负号，便得到本题的矩阵。它改为把点沿着双曲线（hyperbola）$x^2 - y^2 = 1$ 移动，而且它的“角”同样可以相加。这就是*双曲旋转*（hyperbolic rotation）。

这并非只是一件趣闻。若以 x 代表时间、y 代表距离（采用光速为 1 的单位），这个矩阵就是*劳仑兹推进*（Lorentz boost）：它是爱因斯坦（Einstein）狭义相对论（special relativity）中发现的变换，联系着两个以恒定相对速度运动的观察者各自的量度结果。“双曲角”u 称为*快度*（rapidity），而快度可以相加这一事实，正好解释了我们在 Section 3.8 遇过的那条奇怪的速度相加定律。

动笔之前

想一想 $\cosh u$ 和 $\sinh u$ 满足哪一条恒等式，与 $\cos^2 u + \sin^2 u = 1$ 相对应？$\cosh(u + v)$ 和 $\sinh(u + v)$ 的加法公式又是甚么？

$\cosh^2 u - \sinh^2 u = 1$，以及

$$\cosh(u + v) = \cosh u \cosh v + \sinh u \sinh v,$$
$$\sinh(u + v) = \sinh u \cosh v + \cosh u \sinh v.$$

你可以由定义 $\cosh u = \frac{1}{2}(e^u + e^{-u})$ 和 $\sinh u = \frac{1}{2}(e^u - e^{-u})$ 验证这些公式。点

$(\cosh v, \sinh v)$ 位于双曲线 $x^2 - y^2 = 1$ 上，正如 $(\cos v, \sin v)$ 位于圆上一样。

逐步解题

阶段 1　双曲线保持不变

$$\begin{aligned} x'^2 - y'^2 &= (x\cosh u + y\sinh u)^2 - (x\sinh u + y\cosh u)^2 \\ &= (x^2 - y^2)(\cosh^2 u - \sinh^2 u) = x^2 - y^2, \end{aligned}$$

因为交叉项 $2xy\cosh u \sinh u$ 互相抵消。所以 T 保持 $x^2 - y^2$ 不变，特别是把双曲线 $x^2 - y^2 = 1$ 映射到它自身。 ■

> **想一想**　哪些通过原点的直线会被映射到自身？（试找特征向量（eigenvector）。）

$T\begin{pmatrix}1\\1\end{pmatrix} = (\cosh u + \sinh u)\begin{pmatrix}1\\1\end{pmatrix} = e^u\begin{pmatrix}1\\1\end{pmatrix}$，而 $T\begin{pmatrix}1\\-1\end{pmatrix} = e^{-u}\begin{pmatrix}1\\-1\end{pmatrix}$。所以直线 $y = x$ 和 $y = -x$ 各自被映射到自身：一条被拉长 e^u 倍，另一条被缩短至 e^{-u} 倍。这两条直线正是双曲线的渐近线（asymptote）。 ■

阶段 2　双曲角相加

斜率（gradient）为 $\tanh v$ 的直线，其方向向量为 $(\cosh v, \sinh v)$。它的像的方向为

$$T\begin{pmatrix}\cosh v\\ \sinh v\end{pmatrix} = \begin{pmatrix}\cosh u\cosh v + \sinh u\sinh v\\ \sinh u\cosh v + \cosh u\sinh v\end{pmatrix} = \begin{pmatrix}\cosh(u+v)\\ \sinh(u+v)\end{pmatrix},$$

其斜率为 $\tanh(u+v)$。线性映射（linear map）把平行线变成平行线，所以无论直线是否通过原点，这个结论都成立。因此 $v' = v + u$。 ■

阶段 3　夹角何时保持不变？

> **想一想**　斜率分别为 m_1、m_2 的两条直线，其夹角 ψ 满足 $\tan\psi = \left|\frac{m_1 - m_2}{1 + m_1 m_2}\right|$。当 $m_i = \tanh v_i$ 时，化简这个式子。（提示：把所有项通分，以 $\cosh v_1 \cosh v_2$ 为分母。）

利用 $\tanh v_1 - \tanh v_2 = \dfrac{\sinh(v_1 - v_2)}{\cosh v_1 \cosh v_2}$ 和 $1 + \tanh v_1 \tanh v_2 = \dfrac{\cosh(v_1 + v_2)}{\cosh v_1 \cosh v_2}$，得

$$\tan\psi = \left|\frac{\sinh(v_1 - v_2)}{\cosh(v_1 + v_2)}\right|.$$

变换之后，$v_1 - v_2$ 不变，而 $v_1 + v_2$ 则变成 $v_1 + v_2 + 2u$。

> **停一停，试一试** 甚么时候 $\cosh(v_1 + v_2) = \cosh(v_1 + v_2 + 2u)$?

假设两条直线不平行（$v_1 \neq v_2$），则两个夹角相等，当且仅当 $\cosh(v_1 + v_2) = \cosh(v_1 + v_2 + 2u)$。由于 cosh 是偶函数，并且在 $[0, \infty)$ 上严格递增，这表示 $v_1 + v_2 = \pm(v_1 + v_2 + 2u)$。取正号得 $u = 0$，而这已被排除，所以

$$v_1 + v_2 = -u. \qquad \blacksquare$$

阶段 4 图像

当 $v_1 + v_2 = -u$ 时，变换后的角为 $v_1' = v_1 + u = -v_2$ 和 $v_2' = v_2 + u = -v_1$。所以 ℓ_1 的像的斜率为 $-\tanh v_2$：它是 ℓ_2 对 x 轴的反射，而 ℓ_2 的像则是 ℓ_1 的反射。对于通过原点的直线，这个变换把两条直线对调并加以反射，这就解释了为何它们之间的夹角不变（图 7.6.1）。

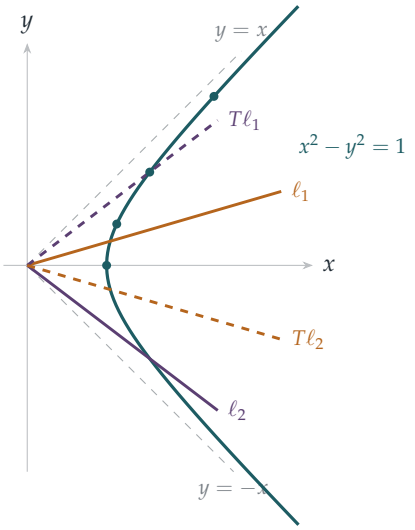


图 7.6.1 把 $u = 0.7$ 的双曲旋转施于通过原点的直线 ℓ_1、ℓ_2，其中 $v_1 = 0.3$，$v_2 = -1.0$，因此 $v_1 + v_2 = -u$。ℓ_1 的像（紫红色虚线）是 ℓ_2 的镜像，而 ℓ_2 的像（琥珀色虚线）是 ℓ_1 的镜像，所以两线之间的夹角保持不变。双曲线上的蓝绿色点按双曲角等距分布（$v = 0, 0.5, 1, 1.5$）；T 把每一点沿曲线再移前 0.7，到达新的位置。

回顾

- **不变的二次型（quadratic form）。** T 保持 $x^2 - y^2$ 不变，正如旋转保持 $x^2 + y^2$ 不变。

- **特征向量就是渐近线。**保持不变的直线是 $y = \pm x$，分别按 $e^{\pm u}$ 的倍数拉长和缩短。
- **双曲角可以相加。**以 $\tanh v$ 把方向参数化之后，T 的作用便化为简单的加法 $v \mapsto v + u$。

背后的数学

狭义相对论

在爱因斯坦的狭义相对论（1905 年）中，以恒定相对速度运动的不同观察者，对时间和距离的量度并不一致，但他们对 $t^2 - x^2$ 这个量（取光速 $c = 1$）却意见一致，这个量称为时空间隔（spacetime interval）。保持它不变的变换就是劳仑兹变换（Lorentz transformation），而在一个空间维度的情况下，它们恰好就是本题的矩阵，其中 x 为时间、y 为距离。一个以速度 $v = \tanh u$ 运动的观察者，所用的坐标与我们的坐标就是由这个矩阵联系起来，其中 $\cosh u = \frac{1}{\sqrt{1-v^2}}$，即时间膨胀（time dilation）中著名的因子 γ。闵可夫斯基（Hermann Minkowski）在 1908 年看出，这使时空成为一种四维几何，其“距离”之中带有一个负号：“从今以后，空间本身和时间本身，都注定要消退成为影子。”

题目的物理意义

- **不变直线** $y = \pm x$ 是光线的路径。它们对每个观察者都相同：光速对所有人都一样，这正是相对论的奠基原理。
- **双曲线** $x^2 - y^2 = 1$ 是与原点相距一个单位“原时”（proper time）的所有事件所成的集合，而原时是由以各种可能的恒定速度运动的观察者所量度的。
- **快度相加。**通过原点而斜率为 $\tanh v$ 的直线，是一个以速度 $\tanh v$ 运动的物体的路径。快度为 u 的推进把它的速度变为 $\tanh(v+u) = \frac{\tanh v + \tanh u}{1 + \tanh v \tanh u}$，这就是 Section 3.8 中爱因斯坦的速度相加公式。速度不能直接相加，快度却可以。

而在 Section 5.8 中，我们见过同一条加法公式，它是庞加莱圆盘（Poincaré disc）中的距离法则：相对论中所有可能的速度所成的集合，以快度来量度，就是一个双曲空间（hyperbolic space）。

思维技巧

找出合适的参数化

选用一个参数来描述你的对象，使变换在这个参数之下变得简单。

若以斜率描述方向，T 的作用是一条繁复的公式；若以 v 描述方向，其中斜率为 $\tanh v$，它就只是加上 u。同样的想法，曾令旋转在使用角度时变得简单，也曾在 Section 2.8 中令映射 $x \mapsto x^2 - 2$ 变得简单。

延伸练习

1 **推进的复合。**证明 $T(u)T(w) = T(u+w)$，其中 $T(u)$ 是题目中的矩阵。$T(u)^{-1}$ 是甚么？

2 **时间膨胀。**一艘太空船以光速的 0.8 倍航行。求它的快度 u 和 $\cosh u$。若地球上过了 10 年，船上过了多少时间？（利用 $t_{\text{船}} = t_{\text{地球}} / \cosh u$。）

3 **双曲面积。**证明由双曲线 $x^2 - y^2 = 1$、x 轴，以及由原点至 $(\cosh v, \sinh v)$ 的线段所围成的面积为 $\frac{v}{2}$，正如圆的相应面积为 $\frac{\theta}{2}$。为甚么 T 必定保持这类面积不变？

4 **挤压。**把坐标转换为 $p = x + y$，$q = x - y$。证明 T 变成 $p \mapsto e^u p$，$q \mapsto e^{-u} q$，而双曲线则变成 $pq = 1$。试在这组坐标下，以几何方式描述这个变换。

延伸阅读

- **Edwin Taylor & John Wheeler, *Spacetime Physics***（W. H. Freeman）。以快度和时空图讲解狭义相对论，适合具备高级程度（A-level）数学基础的学生。
- **Tevian Dray, *The Geometry of Special Relativity***（CRC Press）。把相对论看成双曲三角学，与本节的精神完全一致。
- **Albert Einstein, *Relativity: The Special and the General Theory***（多种版本）。爱因斯坦亲自撰写的通俗读物，至今仍然十分易读。

7.7 永恒不变的量

题目

STEP III 2024，第 5 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2024 年 STEP 试卷（zip 文件）内的 STEP III 试卷：

`www.ocr.org.uk/Images/725753-2024-past-papers.zip`

引子

物理学中一些最深刻的事实是*守恒定律*（conservation laws）：无论运动多么复杂，能量、动量和角动量始终保持不变。当一个系统有足够多的守恒量时，它的运动可以精确地解出；守恒量太少时，运动则可能是混沌的。

这道题展示了一种制造守恒量的绝妙方法。若一个矩阵按照 $\frac{\mathrm{d}\mathbf{M}}{\mathrm{d}t} = \mathbf{MN} - \mathbf{NM}$ 随时间变化，那么无论它的元素变化得多么剧烈，它的迹（trace）和行列式（determinant），以至它的特征值（eigenvalues），都永远不变。这样的一对 $(\mathbf{M}, \mathbf{N})$ 称为*拉克斯对*（Lax pair），以彼得．拉克斯（Peter Lax）命名。他在 1968 年提出这个想法，用来解释*孤立子*（solitons）的惊人表现：孤立子是一种孤立波，彼此穿过之后丝毫不变。

动笔之前

想一想 表达式 $\mathbf{MN} - \mathbf{NM}$ 称为 $\mathbf{M}$ 与 $\mathbf{N}$ 的*交换子*（commutator）。它在甚么时候等于零？根据第 (i) 部分，它的迹是多少？

它等于零，当且仅当 $\mathbf{M}$ 与 $\mathbf{N}$ 可交换。而根据第 (i) 部分，它的迹是 $\mathrm{tr}(\mathbf{MN}) - \mathrm{tr}(\mathbf{NM}) = 0$，恒常如此。这一个事实，就是题目中每一条守恒定律的起点。

逐步解题

阶段 1　迹

设 $\mathbf{M}=\begin{pmatrix}a & b\\ c & d\end{pmatrix}$ 及 $\mathbf{N}=\begin{pmatrix}p & q\\ r & s\end{pmatrix}$，则 $\mathbf{MN}$ 的对角线元素是 $ap+br$ 和 $cq+ds$，所以 $\operatorname{tr}(\mathbf{MN})=ap+br+cq+ds$。把 $\mathbf{M}$ 与 $\mathbf{N}$ 的角色对调，得 $pa+qc+rb+sd$，结果相同。显然 $\operatorname{tr}(\mathbf{M}+\mathbf{N})=\operatorname{tr}\mathbf{M}+\operatorname{tr}\mathbf{N}$。 ■

阶段 2　行列式的导数

想一想　用积法则求 $\det\mathbf{M}=ad-bc$ 的导数。然后利用 $\mathbf{M}^{-1}=\frac{1}{\det\mathbf{M}}\begin{pmatrix}d & -b\\ -c & a\end{pmatrix}$，计算 $\mathbf{M}^{-1}\frac{\mathrm{d}\mathbf{M}}{\mathrm{d}t}$ 及其迹。

以字母上的一点表示导数，则 $\frac{\mathrm{d}}{\mathrm{d}t}\det\mathbf{M}=\dot a d+a\dot d-\dot b c-b\dot c$。而

$$\mathbf{M}^{-1}\dot{\mathbf{M}}=\frac{1}{\det\mathbf{M}}\begin{pmatrix}d & -b\\ -c & a\end{pmatrix}\begin{pmatrix}\dot a & \dot b\\ \dot c & \dot d\end{pmatrix}$$

的对角线元素是 $\frac{d\dot a-b\dot c}{\det\mathbf{M}}$ 和 $\frac{-c\dot b+a\dot d}{\det\mathbf{M}}$。两者之和是 $\frac{1}{\det\mathbf{M}}\frac{\mathrm{d}}{\mathrm{d}t}\det\mathbf{M}$。 ■

阶段 3　三个守恒量

想一想　对 $\dot{\mathbf{M}}=\mathbf{MN}-\mathbf{NM}$ 取迹。至于行列式，则用第 (ii) 部分：$\mathbf{M}^{-1}(\mathbf{MN}-\mathbf{NM})$ 是甚么？

迹：$\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{tr}\mathbf{M}=\operatorname{tr}\dot{\mathbf{M}}=\operatorname{tr}(\mathbf{MN})-\operatorname{tr}(\mathbf{NM})=0$。

行列式：根据第 (ii) 部分，

$$\frac{1}{\det\mathbf{M}}\frac{\mathrm{d}}{\mathrm{d}t}\det\mathbf{M}=\operatorname{tr}(\mathbf{M}^{-1}(\mathbf{MN}-\mathbf{NM}))=\operatorname{tr}\mathbf{N}-\operatorname{tr}(\mathbf{M}^{-1}\mathbf{NM})=\operatorname{tr}\mathbf{N}-\operatorname{tr}\mathbf{N}=0,$$

当中用到第 (i) 部分的以下形式：$\operatorname{tr}(\mathbf{M}^{-1}(\mathbf{NM}))=\operatorname{tr}((\mathbf{NM})\mathbf{M}^{-1})=\operatorname{tr}\mathbf{N}$。

平方的迹：$\operatorname{tr}(\mathbf{M}^2)=(\operatorname{tr}\mathbf{M})^2-2\det\mathbf{M}$（见 Section 7.4），由于两项都是常数，所以它也是常数。 ■

阶段 4　一个显式解

停一停，试一试　设 $\mathbf{N}=t\mathbf{I}+t\mathbf{E}$，其中 $\mathbf{E}=\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix}$，计算 $\mathbf{MN}-\mathbf{NM}$，并写出 a、b、c、d 所满足的微分方程。按以下次序求解：先解 c，然后 a 和 d，最后 b。

$t\mathbf{I}$ 这部分与任何矩阵都可交换，所以 $\mathbf{MN}-\mathbf{NM}=t(\mathbf{ME}-\mathbf{EM})$。现在 $\mathbf{ME}=\begin{pmatrix}0 & a\\ 0 & c\end{pmatrix}$

而 $\mathbf{EM} = \begin{pmatrix} c & d \\ 0 & 0 \end{pmatrix}$，所以

$$\dot{a} = -tc, \qquad \dot{b} = t(a-d), \qquad \dot{c} = 0, \qquad \dot{d} = tc.$$

因此 $c = C$ 是常数；于是 $a = A - \frac{1}{2}Ct^2$ 及 $d = D + \frac{1}{2}Ct^2$；继而 $\dot{b} = t(A - D - Ct^2)$，所以 $b = B + \frac{1}{2}(A-D)t^2 - \frac{1}{4}Ct^4$。故此

$$\mathbf{M} = \begin{pmatrix} A - \frac{1}{2}Ct^2 & B + \frac{1}{2}(A-D)t^2 - \frac{1}{4}Ct^4 \\ C & D + \frac{1}{2}Ct^2 \end{pmatrix}. \qquad \blacksquare$$

验算：迹是 $A + D$，而行列式是 $(A - \frac{1}{2}Ct^2)(D + \frac{1}{2}Ct^2) - C\big(B + \frac{1}{2}(A-D)t^2 - \frac{1}{4}Ct^4\big) = AD - BC$，一如所料是常数（图 7.7.1）。

阶段 5 另一条方程

想一想 若 $\dot{\mathbf{M}} = \mathbf{MN}$，则 $\mathbf{N} = \mathbf{M}^{-1}\dot{\mathbf{M}}$，而根据第 (ii) 部分，$\operatorname{tr}\mathbf{N}$ 就是 $\det\mathbf{M}$ 的对数导数（logarithmic derivative）。$\mathbf{M}$ 的迹能否保持不变，而它的行列式却在改变?

当然可以。取 $\mathbf{M} = \begin{pmatrix} 1 & t \\ t & 1 \end{pmatrix}$，其中 $|t| < 1$（此时 $\det\mathbf{M} = 1 - t^2 \neq 0$）。它的迹恒等于 2。定义 $\mathbf{N} = \mathbf{M}^{-1}\dot{\mathbf{M}}$，使得 $\dot{\mathbf{M}} = \mathbf{MN}$：

$$\mathbf{N} = \frac{1}{1-t^2}\begin{pmatrix} 1 & -t \\ -t & 1 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \frac{1}{1-t^2}\begin{pmatrix} -t & 1 \\ 1 & -t \end{pmatrix},$$

这是一个非奇异（non-singular）矩阵，而且当 $t \neq 0$ 时，$\operatorname{tr}\mathbf{N} = \frac{-2t}{1-t^2} \neq 0$。所以答案是**否定的**，$\operatorname{tr}\mathbf{N}$ 不一定等于零。 $\blacksquare$

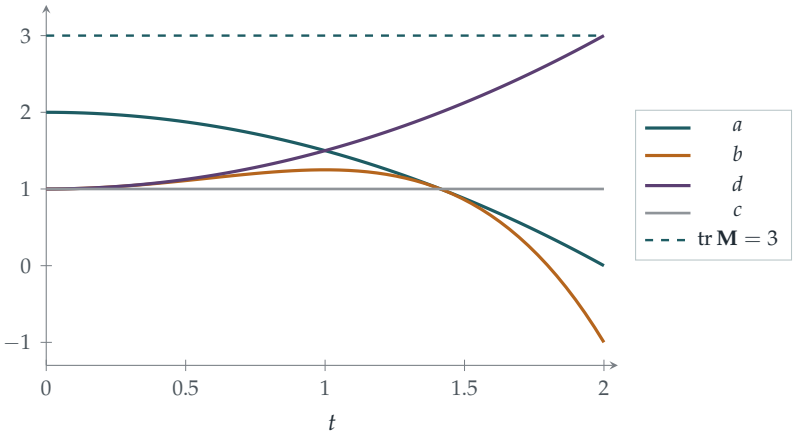


图 7.7.1　第 4 阶段中 $\mathbf{M}(t)$ 的各个元素，其中 $A = 2$、$B = 1$、$C = 1$、$D = 1$。四个元素之中除了 c 以外全都在变，但迹始终是 3，行列式始终是 1。所以特征值 $\frac{3\pm\sqrt{5}}{2}$ 永远不变：这个流是等谱（isospectral）的。

回顾

- **交换子的迹等于零。**由 $\operatorname{tr}(\mathbf{MN}) = \operatorname{tr}(\mathbf{NM})$ 出发，每一条守恒定律都随之而来。
- **对数导数。**第 (ii) 部分把行列式的导数化为一个迹，而迹是线性的，容易处理。
- **在显式解中检查守恒。**我们确认了 $\mathbf{M}(t)$ 的迹和行列式都是常数，借此验算它的公式。

背后的数学

孤立子与拉克斯对

1834 年，苏格兰工程师 John Scott Russell 策马追踪一个沿爱丁堡附近的联合运河（Union Canal）前进的孤立波，追了超过一英里，波形始终没有改变。这类波由科特韦赫－德弗里斯（*Korteweg–de Vries*，简称 *KdV*）方程描述。到了 1960 年代，Norman Zabusky 与 Martin Kruskal 的电脑实验显示，两个这样的“孤立子”可以相撞，然后原封不动地离开，仿佛它们是粒子一样。要解释这个现象，需要无穷多个守恒量。彼得. 拉克斯在 1968 年证明，KdV 方程可以写成 $\dot{L} = LP - PL$，其中 L 和 P 是一对微分算子（differential operators）；于是，与这道题完全一样，L 的特征值是守恒的，而它们正好提供了那无穷多条守恒定律。拥有拉克斯对的系统称为可积（integrable）系统，它们属于少数可以精确求解的非线性系统。

等谱流

由于 $\mathbf{M}$ 的迹和行列式决定了它的特征多项式（characteristic polynomial）$\lambda^2 - (\operatorname{tr}\mathbf{M})\lambda + \det\mathbf{M}$，$\mathbf{M}$ 的特征值不随时间改变。流 $\dot{\mathbf{M}} = [\mathbf{M}, \mathbf{N}]$ 使 $\mathbf{M}$ 在特征值固定的矩阵所组成的集合之内移动。事实上，解总是具有 $\mathbf{M}(t) = \mathbf{S}(t)^{-1}\mathbf{M}(0)\mathbf{S}(t)$ 的形式，其中 $\mathbf{S}(t)$ 是某个可逆矩阵：这个矩阵始终与它的初始值相似（similar）。同一个想法贯穿整个物理学：在量子力学中，可观测量（observable）$\mathbf{A}$ 按 $\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}t} = \frac{\mathrm{i}}{\hbar}[\mathbf{H}, \mathbf{A}]$ 演化，其中 $\mathbf{H}$ 是能量，所以凡是与能量可交换的量都是守恒的。

雅可比公式

第 (ii) 部分对任何大小的可逆矩阵都成立，称为雅可比公式（Jacobi's formula）。最稳妥的写法是不用对数，写成 $\frac{\mathrm{d}}{\mathrm{d}t}\det\mathbf{M} = \det\mathbf{M}\,\operatorname{tr}\left(\mathbf{M}^{-1}\dot{\mathbf{M}}\right)$，因为 $\det\mathbf{M}$ 可能是负数；对数形式 $\frac{\mathrm{d}}{\mathrm{d}t}\ln\det\mathbf{M} = \operatorname{tr}\left(\mathbf{M}^{-1}\dot{\mathbf{M}}\right)$ 需要 $\det\mathbf{M} > 0$（一般情况下则用 $\ln|\det\mathbf{M}|$）。由此可得一个推论：若 $\dot{\mathbf{M}} = \mathbf{A}\mathbf{M}$，其中 $\mathbf{A}$ 是常数矩阵，则 $\det\mathbf{M}(t) = e^{t\operatorname{tr}\mathbf{A}}\det\mathbf{M}(0)$。在微分方程的研究中，这个结果告诉你相空间（phase space）中的体积如何增长或缩小；对于经典力学的方程，这个迹等于零，体积保持不变，这个结果称为刘维尔定理（Liouville's theorem）。

思维技巧

> **找出不变的东西**
>
> 当某样东西在演化时，试找出由它构成而又保持不变的量。这些量限制了运动，并往往揭示出它的结构。
>
> 在这里，迹和行列式之所以守恒，是因为变化率是一个交换子。在力学中，守恒量来自对称性（诺特定理，Noether's theorem），而找出守恒量往往是解题的第一步。

延伸练习

1 **验算解。**通过求导，验证第 4 阶段的矩阵 $\mathbf{M}(t)$ 满足 $\dot{\mathbf{M}} = \mathbf{M}\mathbf{N} - \mathbf{N}\mathbf{M}$。

2 **相似性。**证明：若 $\mathbf{M}(t) = \mathbf{S}^{-1}\mathbf{M}_0\mathbf{S}$，其中 $\mathbf{M}_0$ 是常数矩阵，而 $\mathbf{S}(t)$ 满足 $\dot{\mathbf{S}} = \mathbf{S}\mathbf{N}$，则 $\dot{\mathbf{M}} = \mathbf{M}\mathbf{N} - \mathbf{N}\mathbf{M}$。（提示：$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{S}^{-1} = -\mathbf{S}^{-1}\dot{\mathbf{S}}\mathbf{S}^{-1}$。）

3 **体积增长。**设 $\mathbf{M}(t)=\begin{pmatrix} e^{2t} & 0 \\ 0 & e^{-t} \end{pmatrix}$。求满足 $\dot{\mathbf{M}}=\mathbf{M}\mathbf{N}$ 的 $\mathbf{N}$，并验证 $\frac{\mathrm{d}}{\mathrm{d}t}\ln\det\mathbf{M}=\operatorname{tr}\mathbf{N}$。

4 **更高次幂。**直接证明：当 $\dot{\mathbf{M}}=\mathbf{M}\mathbf{N}-\mathbf{N}\mathbf{M}$ 时，对每个正整数 k 都有 $\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{tr}(\mathbf{M}^k)=0$。

延伸阅读

- **P. G. Drazin & R. S. Johnson, *Solitons: An Introduction***（Cambridge University Press）。大学本科程度的标准入门书，内容包括拉克斯对。
- **Thierry Dauxois & Michel Peyrard, *Physics of Solitons***（Cambridge University Press）。书中讲述了 Scott Russell 那个孤立波的故事。
- **Dwight Neuenschwander, *Emmy Noether's Wonderful Theorem***（Johns Hopkins University Press）。讲解对称性如何带来守恒定律。

7.8 四个平方数

题目

STEP III 1994，第 8 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/94-s3-q8`

引子

7 不是两个平方数之和，也不是三个平方数之和，却是四个平方数之和：$7 = 4 + 1 + 1 + 1$。1770 年，拉格朗日（Joseph-Louis Lagrange）证明了每一个正整数都是四个平方数之和（容许使用 0^2）。这条定理早在 1621 年已由 Bachet 提出猜想，费马（Fermat）亦声称自己证得，但一如以往没有留下证明；欧拉（Euler）为此钻研了四十年，始终差一步未能完成。

关键的工具由欧拉于 1748 年发现，正正就在这道题目之中：两个四平方和的积，本身也是一个四平方和。这表示只需对质数（prime number）证明拉格朗日定理，其余情况靠相乘便可得出。题目利用复矩阵（complex matrix）的行列式（determinant），寥寥数行便推导出这条恒等式（identity），而这些矩阵其实就是 Section 6.7 中的四元数（quaternion），只是换了个面貌。

动笔之前

想一想 回想两平方的版本：$(a^2 + b^2)(p^2 + q^2)$ 是两个平方数之和，因为对复数 $z = a + b\mathrm{i}$、$w = p + q\mathrm{i}$ 有 $|z|^2|w|^2 = |zw|^2$。矩阵可以怎样把这个结果推广到四个平方？

想法是找出一种由四个实数构成的“数”，它的“大小的平方”是 $a^2 + b^2 + c^2 + d^2$，而且具有可乘性（multiplicative）。矩阵的行列式正有可乘性：$\det(\mathbf{AB}) = \det \mathbf{A} \det \mathbf{B}$。因此我们要找一个由 a, b, c, d 构成的矩阵，其行列式为 $a^2 + b^2 + c^2 + d^2$，并且这类矩阵相乘后仍然是同一形式。

逐步解题

阶段 1　行列式是四个平方数之和

想一想　$\begin{pmatrix} z_1 & z_2 \\ -z_2^* & z_1^* \end{pmatrix}$ 的行列式是甚么？若 $z_1 = a + b\mathrm{i}$ 及 $z_2 = c + d\mathrm{i}$，以 a, b, c, d 表示又是甚么？

$\det = z_1 z_1^* + z_2 z_2^* = |z_1|^2 + |z_2|^2 = a^2 + b^2 + c^2 + d^2$。同样地，取 $z_3 = p + q\mathrm{i}$ 及 $z_4 = r + s\mathrm{i}$，第二个矩阵的行列式为 $p^2 + q^2 + r^2 + s^2$。

阶段 2　乘积具有相同形式

相乘得

$$\begin{pmatrix} z_1 & z_2 \\ -z_2^* & z_1^* \end{pmatrix}\begin{pmatrix} z_3 & z_4 \\ -z_4^* & z_3^* \end{pmatrix} = \begin{pmatrix} z_1 z_3 - z_2 z_4^* & z_1 z_4 + z_2 z_3^* \\ -z_2^* z_3 - z_1^* z_4^* & -z_2^* z_4 + z_1^* z_3^* \end{pmatrix}.$$

设 $w_1 = z_1 z_3 - z_2 z_4^*$ 及 $w_2 = z_1 z_4 + z_2 z_3^*$，则下面一行是 $(-w_2^*, w_1^*)$，所以乘积是 $\begin{pmatrix} w_1 & w_2 \\ -w_2^* & w_1^* \end{pmatrix}$，形式相同。它的行列式是 $|w_1|^2 + |w_2|^2$，而由于行列式可以相乘，

$$(a^2 + b^2 + c^2 + d^2)(p^2 + q^2 + r^2 + s^2) = |w_1|^2 + |w_2|^2.$$

阶段 3　读出四个平方

停一停，试一试　把 $w_1 = (a + b\mathrm{i})(p + q\mathrm{i}) - (c + d\mathrm{i})(r - s\mathrm{i})$ 及 $w_2 = (a + b\mathrm{i})(r + s\mathrm{i}) + (c + d\mathrm{i})(p - q\mathrm{i})$ 展开，写成实部（real part）和虚部（imaginary part）。

$$w_1 = (ap - bq - cr - ds) + \mathrm{i}(aq + bp + cs - dr),$$
$$w_2 = (ar - bs + cp + dq) + \mathrm{i}(as + br - cq + dp).$$

所以

$$L_1 = ap - bq - cr - ds, \qquad L_2 = aq + bp + cs - dr,$$
$$L_3 = ar - bs + cp + dq, \qquad L_4 = as + br - cq + dp,$$

每一个对 a, b, c, d 及对 p, q, r, s 都是线性（linear）的，并且 $(a^2 + b^2 + c^2 + d^2)(p^2 + q^2 + r^2 + s^2) = L_1^2 + L_2^2 + L_3^2 + L_4^2$。 ■

例如，$7 = 2^2 + 1^2 + 1^2 + 1^2$ 而 $3 = 1^2 + 1^2 + 1^2 + 0^2$。取 $(a, b, c, d) = (2, 1, 1, 1)$ 及 $(p, q, r, s) = (1, 1, 1, 0)$，由公式得 $(L_1, L_2, L_3, L_4) = (0, 2, 4, 1)$，而果然 $0^2 + 2^2 + 4^2 + 1^2 = 21 = 7 \times 3$。

回顾

- **可乘性是关键。**这条恒等式来自 $\det(\mathbf{AB}) = \det\mathbf{A}\det\mathbf{B}$，正如两平方恒等式来自 $|zw| = |z||w|$ 一样。
- **对乘法封闭（closure）。**$\begin{pmatrix} z_1 & z_2 \\ -z_2^* & z_1^* \end{pmatrix}$ 这个特殊形式在相乘之下保持不变，结果才会是另一个四平方和。
- **细心记账十分重要。**最后一步虽属例行运算，却容易出错；正负号和共轭（conjugate）都必须小心处理。

背后的数学

拉格朗日四平方和定理

拉格朗日的证明大致如下。根据欧拉恒等式，只需证明每个质数 p 都是四个平方数之和。先用一个计数论证，证明存在某个倍数 mp（其中 $m < p$）是四个平方数之和。然后用一个巧妙的“递降法”（descent）证明：若 mp 是四个平方数之和而 $1 < m < p$，则对某个更小的 m'，$m'p$ 也是四个平方数之和；如此重复下去，便会到达 $m = 1$。（可与 Section 1.5 中费马的递降法比较。）1834 年，雅可比（Jacobi）更进一步，求出把 n 写成四个平方数之和的方法的确切数目：n 的所有非 4 的倍数的因数之和的 8 倍。例如，1 有 8 种写法（± 1 可放在四个位置中的任何一个）。

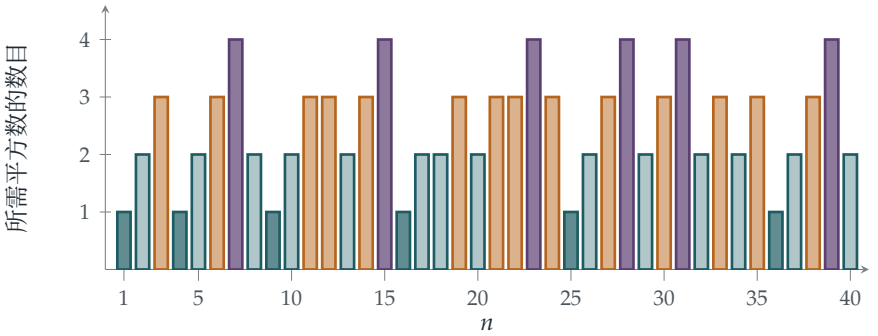


图 7.8.1 由 1 至 40 的每个 n 最少需要多少个平方数才能组成。根据拉格朗日定理，每个数最多只需四个。需要四个的数（紫红色）是 $7, 15, 23, 28, 31, 39, \ldots$：根据勒让德（Legendre）和高斯（Gauss）的一条定理，它们正好是形如 $4^k(8m+7)$ 的数。

二、三、四

不同维数之间的规律相当微妙。两平方和来自复数，而一个数是两个平方数之和，当且仅当它的质因数中形如 $4k+3$ 的都以偶数次幂出现（Section 1.6）。四平方和来自四元数，而每个数都是四平方和。但三平方和却完全没有乘法恒等式（并不存在一个“大小”具可乘性的三维数系；这正是哈密顿（Hamilton）研究三元组时失败的原因，见 Section 6.7），相应的定理也较难：勒让德和高斯证明了，除非 $n = 4^k(8m+7)$，否则 n 是三个平方数之和（图 7.8.1）。四以上，八维的八元数（octonion）给出一条八平方恒等式，由 Degen 和 Graves 发现；而胡尔维兹（Hurwitz）于 1898 年证明，这类恒等式只在 1、2、4 和 8 个平方的情况下存在。

华林问题

拉格朗日定理是华林（Edward Waring）于 1770 年提出的一个问题的第一个情况：对每个 k，是否存在一个数 $g(k)$，使得每个正整数都是最多 $g(k)$ 个 k 次幂之和？希尔伯特（David Hilbert）于 1909 年证明了确实存在。现在我们知道 $g(2) = 4$、$g(3) = 9$（例如 $23 = 2\cdot 2^3 + 7\cdot 1^3$ 需要九个立方数）以及 $g(4) = 19$。

思维技巧

寻找具可乘性的结构

要证明某个性质在乘法之下得以保持，可寻找一个代数结构（复数、四元数、矩阵），使该性质就是其中某个对象的“大小”，而大小具有可乘性。

一旦找到这样的结构，乘积恒等式便唾手可得，而一个关于所有数的问题，也就化为一个关于质数的问题。

延伸练习

1 **一个实例。**利用 $L_1, \dots, L_4$ 的公式，取 $(a,b,c,d) = (p,q,r,s) = (2,1,1,1)$，把 $49 = 7\times 7$ 写成四个平方数之和，并验算你的答案。它与显而易见的 $49 = 7^2+0^2+0^2+0^2$ 是否相同？

2 **四平方中的两平方。**证明若 $c = d = r = s = 0$，这条恒等式便简化为 Section 1.6 的两平方恒等式。

3 **需要四个平方的数。**考虑平方数模 8 的余数，证明形如 $8m+7$ 的数都不是三个平方数之和。

4 **雅可比公式。**数一数把 2 表示成四个平方数之和的方法有多少种（次序和正负号不同均视为不同），并验证答案是 $8\times(1+2)=24$。

延伸阅读

- **Harold Davenport, *The Higher Arithmetic***（Cambridge University Press）。第 V 章证明了拉格朗日四平方和定理。
- **John Conway & Derek Smith, *On Quaternions and Octonions***（A. K. Peters）。讲述四元数的算术，包括利用四元数整数证明四平方和定理。
- **Aigner & Ziegler, *Proofs from THE BOOK***，“Representing numbers as sums of two squares”（把数表示成两个平方数之和）一章。与本节相配的两平方版本。

7.9 只有五种柏拉图立体

题目

STEP III 1999，第 4 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/99-s3-q4`

引子

正多边形（regular polygon）有无穷多种：三角形、正方形、五边形，如此类推，任何边数都有。你或会以为正多面体（regular polyhedron）也有无穷多种。其实恰好只有五种：正四面体（tetrahedron）、正方体（cube）、正八面体（octahedron）、正十二面体（dodecahedron）和正二十面体（icosahedron）。柏拉图（Plato）把其中四种与四大元素（火、土、气、水）对应，第五种则对应宇宙；欧几里得（Euclid）的《几何原本》（*Elements*）是有史以来影响最深远的数学著作，全书最后就是证明再没有其他正多面体；开普勒（Kepler）更用它们建构了一个太阳系模型。

本题只用寥寥数行就证明了这个分类，所依据的只是欧拉（Euler）在 1750 年发现的一条非凡公式：对任何凸多面体（convex polyhedron），$V - E + F = 2$。一条关于计数的公式，化成了一个关于形状的定理，而 2 这个数原来正是球面的“指纹”。

动笔之前

想一想 用正方体验证欧拉公式：它有多少个顶点（vertex）、多少条棱（edge）、多少个面（face）？以正方形为底的棱锥体（pyramid）又如何？

正方体有 $V = 8$，$E = 12$，$F = 6$，而 $8 - 12 + 6 = 2$。以正方形为底的棱锥体有 5 个顶点、8 条棱和 5 个面，而 $5 - 8 + 5 = 2$。任何没有洞的多面体，不论多么不规则，这条公式都成立。

想一想　在正多面体中，每个面有 m 条棱，而每条棱是两个面的公共边。由此可得 E 与 F 之间的甚么方程？同样地，E 与 V 之间又有甚么方程？

用两种方法计数棱与面的关联：F 个面各有 m 条棱，而每条棱被数了两次（它所属的两个面各数一次），所以 $mF = 2E$。同样地，每个顶点连着 n 条棱，而每条棱有两个端点，所以 $nV = 2E$。这种“用两种方法计数同一样东西”的技巧称为双重计数（double counting）。

逐步解题

阶段 1　F 的公式

由 $mF = 2E$ 和 $nV = 2E$，可得 $E = \frac{mF}{2}$ 和 $V = \frac{mF}{n}$。代入欧拉公式，

$$\frac{mF}{n} - \frac{mF}{2} + F = 2 \quad\Longrightarrow\quad F\,\frac{2m - mn + 2n}{2n} = 2 \quad\Longrightarrow\quad F = \frac{4n}{2m + 2n - mn}.$$

而 $4 - (n-2)(m-2) = 4 - (mn - 2m - 2n + 4) = 2m + 2n - mn = h$。所以 $F = \frac{4n}{h}$。∎

阶段 2　m 和 n 可以取哪些值？

想一想　m（每个面的边数）和 n（每个顶点处的面数）最小可以是多少？又既然 $F > 0$，h 必须是正还是负？

多边形最少要有 3 条边，所以 $m \geq 3$；立体的每个顶点处最少要有 3 个面相交，所以 $n \geq 3$。由于 $F > 0$，我们需要 $h > 0$，即 $(n-2)(m-2) < 4$。两个因子都至少是 1，所以 $(m-2, n-2)$ 只可能是 $(1,1)$、$(1,2)$、$(2,1)$、$(1,3)$ 和 $(3,1)$。

停一停，试一试　就这五种可能逐一计算 h，然后求 F，再求 $E = \frac{mF}{2}$ 和 $V = \frac{mF}{n}$。指出每种情况是哪一个立体。

m	n	h	F	E	V	立体
3	3	3	4	6	4	正四面体
3	4	2	8	12	6	正八面体
4	3	2	6	12	8	正方体
3	5	1	20	30	12	正二十面体
5	3	1	12	30	20	正十二面体

可能的情况只有这些，所以正多面体最多只有五种；而这五种全部存在（欧几里得说明了如何逐一构作）。∎

回顾

- **双重计数。**用两种方法计数棱与面、棱与顶点的关联，得出 $mF = 2E$ 和 $nV = 2E$。
- **由正值得出不等式。**$F > 0$ 这个简单的要求，把无穷多个候选者筛剩五个。
- **存在性要另行处理。**这个论证只说明最多有五种；每一种是否真的存在，还须实际构作出来才能确定。

背后的数学

欧拉公式与拓扑学

欧拉在 1750 年提出 $V - E + F = 2$，但他的证明有漏洞；这条公式其后经历了一段引人入胜的历史，当中有一次又一次的尝试证明、反例（counterexample）和修补。一个“画框”形的多面体，中间穿了一个正方形的洞，有 $V - E + F = 0$；由两个正方体在一个顶点处相连而成的多面体，则有 $V - E + F = 3$。拉卡托斯（Imre Lakatos）把这段历史写成一本名著 *Proofs and Refutations*（《证明与反驳》），探讨数学概念如何在反例的磨砺下变得精确。最终的答案是：$V - E + F$ 只取决于曲面在拓扑学（topology）那种宽松意义下的形状：凡是可以变形成球面的曲面，它等于 2；环面（torus，即冬甩形）是 0；有 g 个洞的曲面则是 $2 - 2g$。这个数称为欧拉示性数（Euler characteristic），是拓扑学最早的概念之一；拓扑学研究的，就是经拉伸和弯曲后仍然不变的性质。

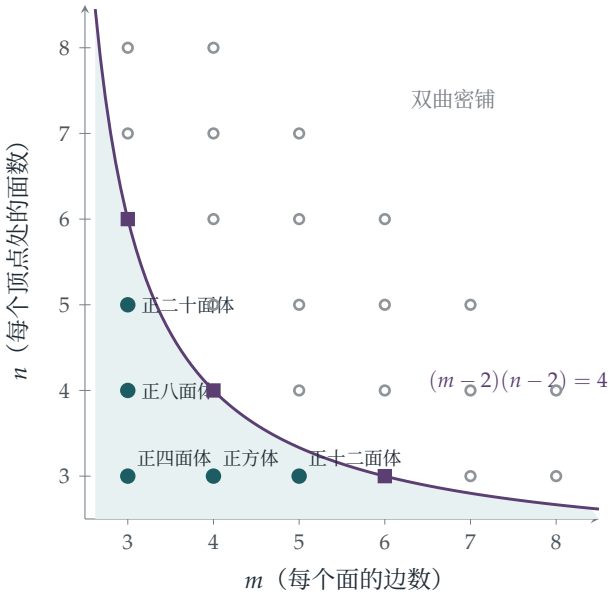


图 7.9.1　满足 $m, n \geq 3$ 的所有可能数对 (m, n)。在曲线 $(m-2)(n-2)=4$ 下方（阴影部分）的是五种柏拉图立体（蓝绿色）。在曲线上的是平面的三种密铺（紫红色方点）：分别用三角形（$m=3$，$n=6$）、正方形（4,4）和六边形（6,3）。曲线上方的每一点（空心圆圈）都对应双曲平面的一种正则密铺。

球面以外

当 $h=0$ 或 $h<0$ 时又会怎样？当 $(m-2)(n-2)=4$ 时，有三组解：(3,6)、(4,4) 和 (6,3)，即六个三角形、四个正方形或三个六边形在一点相接，各角之和恰好是 360°。这就是平面的三种正则密铺（regular tiling），也就是浴室地砖和蜂巢的图案。而当 $(m-2)(n-2)>4$ 时，一个顶点处各角之和大于 360°，这在平面上不可能出现，但在 Section 5.8 的双曲平面（hyperbolic plane）上却完全可行。双曲平面有无穷多种正则密铺，图 7.9.1 中曲线上方的每一点各对应一种；艾雪（Escher）的木刻版画 *Circle Limit*（《圆极限》）系列，画的就是其中几种。所以，柏拉图立体的分类其实是一个三分法（trichotomy）的第一种情况，这个三分法（球面、平面、双曲平面）贯穿整个几何学。

对称群

柏拉图立体的对称构成群（group）：正四面体有 12 个旋转，正方体和正八面体有 24 个（两者互为对偶（dual）：把面与顶点互换，$m \leftrightarrow n$），正十二面体和正二十面体则有 60 个。这个阶（order）为 60 的正二十面体群是一个“单”群（simple group）；正如我们在 Section 6.4 提过，正是它令五次方程（quintic）不能用根式（radicals）求解。克莱因（Felix Klein）写了一整本书 *Lectures on the Icosahedron*（《正二十面体讲义》，1884），把

正二十面体与五次方程的解法联系起来。

思维技巧

双重计数

用两种不同的方法计数同一组东西，然后令两个结果相等。

这里所计数的，是所有（棱，面）配对所成的集合，其中棱在该面上。双重计数是组合数学（combinatorics）中最强大的工具之一，在第 9 章会一再出现。

延伸练习

1 **对偶性。**证明把 m 和 n 互换，会令 V 和 F 互换，而 E 保持不变。哪几对柏拉图立体互为对偶？哪一个是自对偶的？

2 **足球。**传统的足球由 12 个五边形和 20 个六边形组成，每个顶点处有三个面相接。利用双重计数和欧拉公式，验证 $V - E + F = 2$。再证明更一般的结果：任何这样由五边形和六边形组成、每个顶点处有三个面相接的多面体，都恰好有 12 个五边形。

3 **环面。**造一个“画框”形的多面体（一个正方形方块，中间有一个正方形的洞），数出 V、E 和 F。验证 $V - E + F = 0$。

4 **顶点处的角。**正 m 边形的内角（interior angle）是 $180°(1 - \frac{2}{m})$。证明 n 个这样的角能围着凸立体的一个顶点拼合（总和小于 360°），当且仅当 $(m-2)(n-2) < 4$，从而得出这个分类的第二个证明。

延伸阅读

- **David Richeson, *Euler's Gem***（Princeton University Press）。讲述 $V - E + F = 2$ 的历史，以及它如何引出拓扑学。非常适合高中第六学级（sixth form）的学生阅读。
- **Imre Lakatos, *Proofs and Refutations***（Cambridge University Press）。一场关于欧拉公式和证明本质的对话。
- **H. S. M. Coxeter, *Regular Polytopes***（Dover）。论述正多面体及其高维同类的经典著作。

我们学到了甚么

本章开首我们说过，矩阵（matrix）是空间的一种变换（transformation），自有其生命；我们还提出了三个问题：矩阵与数有何相似、有何不同，矩阵究竟做了甚么，以及甚么保持不变。以下是本章九道题目所揭示的答案。

1. 矩阵与数有何相似，又有何不同？

矩阵可以相加和相乘，但 Section 7.1 指出了哪些熟悉的法则不再成立。乘法不符合交换律（commutative law），所以 $(\mathbf{A}-\mathbf{B})(\mathbf{A}+\mathbf{B})$ 未必等于 $\mathbf{A}^2-\mathbf{B}^2$；矩阵中存在零因子（zero divisor），所以 $\mathbf{AB}=\mathbf{0}$ 并不保证 $\mathbf{BA}=\mathbf{0}$；而一条二次矩阵方程可以有无限多个根。每一项法则的失效都有几何解释，可以从这些矩阵对平面的作用来理解。至关重要的是，有一项性质保留了下来：行列式（determinant）可乘，即 $\det(\mathbf{AB})=\det\mathbf{A}\det\mathbf{B}$。Section 7.3 说明这项性质基本上刻划了行列式，而 Section 7.8 则用它证明了欧拉（Euler）的四平方恒等式（four-square identity），这正是拉格朗日（Lagrange）定理的关键一步；该定理指出每个正整数都是四个平方数之和。

2. 矩阵究竟做了甚么？

有几节通过找出合适的观点来回答这个问题。

- **特征向量（eigenvector）**（Section 7.2）是矩阵不会转动的方向。在以特征向量为轴的坐标下，对称矩阵（symmetric matrix）只不过是一个伸缩，而圆的像是一个椭圆，其轴和面积都可以直接读出。
- **行列式**（Section 7.3）是带正负号的面积比例因子：错切（shear）保持面积不变，对换会把定向（orientation）反转，奇异矩阵（singular matrix）则把平面压扁。
- **迹（trace）与行列式**（Section 7.4）通过凯莱–哈密顿定理（Cayley–Hamilton theorem），决定一个矩阵会否在若干步之后回到单位矩阵（identity matrix）。满足 $\mathbf{M}^8=\mathbf{I}$ 的矩阵未必是旋转；它是在歪斜坐标下看到的旋转。又由于整数矩阵的迹是整数，晶体不可能有五重对称。
- **三维空间中的旋转**（Section 7.5）可由向量积（cross product）通过罗德里格斯公式（Rodrigues′ formula）构造出来，这是 $e^{i\theta}=\cos\theta+i\sin\theta$ 在三维的对应版本。

- **双曲旋转（hyperbolic rotation）**（Section 7.6）使点沿双曲线移动，并把它们的双曲角相加。它们就是狭义相对论中的洛伦兹变换（Lorentz boost），而它们的不变直线（invariant line）就是光的路径。

3. 甚么保持不变？

不变量（invariant）无处不在。旋转保持 $x^2 + y^2$ 不变，洛伦兹变换保持 $x^2 - y^2$ 不变，而相似矩阵（similar matrices）有相同的迹和行列式。在 Section 7.7 中，一个按 $\dot{\mathbf{M}} = \mathbf{MN} - \mathbf{NM}$ 演化的矩阵，不论其元素如何变化，它的迹和行列式都保持不变，因而它的特征值（eigenvalue）也保持不变：这就是孤立子（soliton）各条守恒定律背后的机制。而在 Section 7.9 中，不变量是一个数，$V - E + F = 2$，它对每个凸多面体（convex polyhedron）都成立，并把正多面体限定为仅有五种。这个数称为欧拉示性数（Euler characteristic），是拓扑学（topology）中最早出现的不变量之一。

4. 线性代数工具箱

方法	你在哪里用过
几何反例	失效的法则（7.1）
以特征向量为自然坐标轴	椭圆（7.2）；洛伦兹变换（7.6）
用特殊矩阵作试探	行列式（7.3）
凯莱–哈密顿定理	有限阶（7.4）
分解为平行和垂直部分	罗德里格斯公式（7.5）
选择合适的参数	快度（rapidity）（7.6）
$\mathrm{tr}(\mathbf{MN}) = \mathrm{tr}(\mathbf{NM})$	守恒量（7.7）
行列式的可乘性	四平方和（7.8）
重复计数（double counting）	柏拉图立体（Platonic solids）（7.9）

下一步

- 在 Section 11.2 中，特征值和特征向量支配着弹簧与质点系统的振动。
- 在 Section 13.7 中，协方差矩阵（covariance matrix）的特征向量会在统计学中出现。
- 微分方程（differential equation）和动力系统（dynamical system）是第 8 章的主题。

如果你想在入读大学之前再走远一点

- **3Blue1Brown, *Essence of Linear Algebra***（线性代数的本质）：一套免费的影片系列，展示矩阵究竟做了甚么。请先看这个。

- **Gilbert Strang, *Introduction to Linear Algebra***（线性代数导论）：经典的入门课程，十分着重理解和应用。
- **Sheldon Axler, *Linear Algebra Done Right***（线性代数应该这样学）：较为抽象的处理手法，以线性映射（linear map）为中心；可在网上免费阅读。
- **David Richeson, *Euler's Gem***（欧拉的瑰宝）：欧拉公式与拓扑学的诞生。

Kaplansky 曾打趣说，他和 Halmos 思考时不用坐标，但关起门来却“拼命地”用矩阵计算。这句话正好道出本章想说明的事：矩阵既具体又可以计算，但每一项计算背后都有几何（一个伸缩、一个旋转、一个投影、一个洛伦兹变换），还有一个不变量，它不取决于你怎样把它写下来。学会在这两种观点之间来回转换，是数学家所能拥有的最宝贵技能之一。

延伸练习答案

以下只是简短的答案和提示，并非完整解答：这里给出结果和关键步骤，让你核对自己的做法。请先认真做过每道练习，然后才看答案。

7.1 对矩阵不成立的法则

1. 设 $\mathbf{X}=\begin{pmatrix}a & b\\ c & d\end{pmatrix}$，比较 $\mathbf{XP}$ 与 $\mathbf{PX}$，必有 $c=0$ 及 $a=d$：所求的矩阵（matrix）为 $\begin{pmatrix}a & b\\ 0 & a\end{pmatrix}$。

2. 由凯莱—哈密顿定理（Cayley–Hamilton theorem）得 $\mathbf{N}^2=t\mathbf{N}-d\mathbf{I}$，所以 $t\mathbf{N}=d\mathbf{I}$。若 $t\neq 0$，则 $\mathbf{N}$ 是 $\mathbf{I}$ 的倍数，而 $\mathbf{N}^2=\mathbf{0}$ 迫使 $\mathbf{N}=\mathbf{0}$，与 $t\neq 0$ 矛盾。所以 $t=0$，从而 $d=0$。答案是：$\begin{pmatrix}a & b\\ c & -a\end{pmatrix}$，其中 $a^2+bc=0$。

3. 平方后的对角线元素为 $\cos^2\theta+\sin^2\theta=1$，非对角线元素为 0。保持不变的向量（vector）是 $(\cos\frac{\theta}{2},\sin\frac{\theta}{2})$，所以它是对通过原点、倾角为 $\frac{\theta}{2}$ 的直线的反射（reflection）。

4. 设 $\mathbf{M}=\begin{pmatrix}a & b\\ c & d\end{pmatrix}$，则 $\mathbf{M}^2=\begin{pmatrix}a^2+bc & b(a+d)\\ c(a+d) & d^2+bc\end{pmatrix}$；减去 $(a+d)\mathbf{M}$ 再加上 $(ad-bc)\mathbf{I}$，便得 $\mathbf{0}$。

7.2 只伸缩，不转向

1. $\mathbf{Mu}=\lambda\mathbf{u}$ 且 $\mathbf{u}\neq\mathbf{0}$，必须有 $\det(\mathbf{M}-\lambda\mathbf{I})=0$，这就是该二次方程。这里迹（trace）为 6，行列式（determinant）为 $\frac{4}{25}(54-4)=8$，所以 $\lambda^2-6\lambda+8=0$：$\lambda=2$（对应 $(1,2)$）和 4（对应 $(2,-1)$）。

2. $\mathbf{v}\cdot\mathbf{Mu}=\lambda\,\mathbf{v}\cdot\mathbf{u}$，又因为 $\mathbf{M}$ 是对称矩阵（symmetric matrix），$\mathbf{v}\cdot\mathbf{Mu}=\mathbf{Mv}\cdot\mathbf{u}=\mu\,\mathbf{v}\cdot\mathbf{u}$。所以 $(\lambda-\mu)\,\mathbf{u}\cdot\mathbf{v}=0$。

3. 特征方程（characteristic equation）$\lambda^2-2\lambda\cos\theta+1=0$ 的判别式（discriminant）为 $-4\sin^2\theta$，除非 θ 是 π 的倍数，否则必为负数。特征值（eigenvalue）是 $\mathrm{e}^{\pm\mathrm{i}\theta}$。

4. $(1,0)=\frac{1}{5}(1,2)+\frac{2}{5}(2,-1)$，所以 $T^n(1,0)=\frac{2^n}{5}(1,2)+\frac{2\cdot 4^n}{5}(2,-1)=\frac{1}{5}\big(2^n+4^{n+1},\,2^{n+1}-2\cdot 4^n\big)$。当 $n=1$ 时，结果是 $(\frac{18}{5},-\frac{4}{5})$，正如所料。

7.3 甚么才算行列式

1. $(\det \mathbf{MN})^3 = (\det \mathbf{M})^3 (\det \mathbf{N})^3$，而 $(\det \mathbf{J})^3 = -1 \neq 1$。函数 $|\det \mathbf{M}|$ 具有可乘性（multiplicative），但 $|\det \mathbf{J}| = 1$。
2. $\begin{pmatrix} a & \lambda a \\ c & \lambda c \end{pmatrix} = \begin{pmatrix} a & 0 \\ c & 0 \end{pmatrix} \begin{pmatrix} 1 & \lambda \\ 0 & 0 \end{pmatrix}$，而第二个因子的第二行是第一行的 0 倍，所以它的 f 值为 0。
3. $\begin{pmatrix} k & 0 \\ 0 & 1 \end{pmatrix} = \mathbf{J}\mathbf{K}_k\mathbf{J}$，所以它的值是 $\mathrm{f}(\mathbf{J})^2 \mathrm{f}(\mathbf{K}_k) = \mathrm{f}(\mathbf{K}_k)$。另外 $\mathbf{K}_k\mathbf{K}_l = \mathbf{K}_{kl}$。
4. 行简化（row reduction）可把任何可逆矩阵（invertible matrix）写成初等矩阵（elementary matrix）的乘积：行对换（$\mathbf{J}$）、行倍乘（$\mathbf{K}_k$ 或 $\mathbf{J}\mathbf{K}_k\mathbf{J}$），以及错切（shear）（$\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$ 或它经 $\mathbf{J}$ 共轭（conjugate）后所得的矩阵）。错切的 $\mathrm{f} = 1$（做法如第 (iv) 部分，利用 $\mathbf{K}^{-1}\mathbf{P}\mathbf{K}$）。此外，$\mathbf{J}$ 是若干错切与 $\mathbf{K}_{-1}$ 的乘积，所以 $\mathrm{f}(\mathbf{K}_{-1}) = \mathrm{f}(\mathbf{J}) = -1$。因此 $\mathrm{f}(\mathbf{M}) = \mathrm{g}(\det \mathbf{M})$，其中 $\mathrm{g}(k) = \mathrm{f}(\mathbf{K}_k)$；若 $\mathbf{M}$ 是奇异矩阵（singular matrix），则由第 (iii) 部分得 $\mathrm{f}(\mathbf{M}) = 0$。

7.4 回到起点的矩阵

1. 由 $|1 + \mathrm{e}^{\mathrm{i}\phi}|^2 = 2 + 2\cos\phi = 1$ 得 $\phi = \pm\frac{2\pi}{3}$，所以 $\mathbf{R}^3 = \mathbf{I}$。至于 $\mathbf{S}$：它的特征值是 1 的立方根。假如两个特征值都是 1，则由于 $x^3 - 1$ 的根各不相同，$\mathbf{S}$ 可对角化（diagonalisable），因而等于 $\mathbf{I}$。实矩阵的非实数特征值必以共轭对的形式出现，所以特征值是 $\mathrm{e}^{\pm 2\pi\mathrm{i}/3}$：$\det \mathbf{S} = 1$ 及 $\operatorname{tr} \mathbf{S} = -1$。
2. 特征值必定是六次本原单位根（primitive sixth roots of unity）$\mathrm{e}^{\pm\mathrm{i}\pi/3}$（若选取其他六次单位根并与其共轭配对，都会得出 $\mathbf{M}^2 = \mathbf{I}$ 或 $\mathbf{M}^3 = \mathbf{I}$）。所以 $\operatorname{tr} \mathbf{M} = 1$ 及 $\det \mathbf{M} = 1$。
3. 特征值是乘积为 1 的单位根，所以它们是 $\mathrm{e}^{\pm\mathrm{i}\theta}$（或同为 ± 1），而迹 $2\cos\theta$ 是 $[-2, 2]$ 内的整数。阶（order）为 6 的例子：$\begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}$。
4. $\mathbf{M}^3 = (t^2 - d)\mathbf{M} - td\mathbf{I}$ 及 $\mathbf{M}^4 = (t^3 - 2td)\mathbf{M} - d(t^2 - d)\mathbf{I}$。要使 $\mathbf{M}^4 = -\mathbf{I}$（$\mathbf{M}$ 不可能是 $\mathbf{I}$ 的倍数），须有 $t(t^2 - 2d) = 0$ 及 $d(t^2 - d) = 1$。$t = 0$ 给出 $-d^2 = 1$，不可能成立；所以 $t^2 = 2d$，$d^2 = 1$，于是 $d = 1$，$t^2 = 2$。

7.5 三维空间中的旋转

1. $z \mapsto a + \mathrm{e}^{\mathrm{i}\theta}(z - a)$。复合变换是 $z \mapsto \mathrm{e}^{\mathrm{i}(\theta+\phi)}z + k$，其中 $k = \mathrm{e}^{\mathrm{i}\phi}a(1 - \mathrm{e}^{\mathrm{i}\theta}) + b(1 - \mathrm{e}^{\mathrm{i}\phi})$；若 $\mathrm{e}^{\mathrm{i}(\theta+\phi)} \neq 1$，它有不动点（fixed point）$c = \frac{k}{1 - \mathrm{e}^{\mathrm{i}(\theta+\phi)}}$。若 $\theta + \phi = 2\pi$，它就是位

移为 $(b-a)(1-\mathrm{e}^{-\mathrm{i}\theta})$ 的平移（translation）。

2. $\begin{pmatrix}0&0&1\\1&0&0\\0&1&0\end{pmatrix}$：它保持 $(1,1,1)$ 不变，迹为 $0=1+2\cos 120^\circ$，并把 $\mathbf{i}\to\mathbf{j}\to\mathbf{k}\to\mathbf{i}$。

3. $\mathbf{K}^2\mathbf{r}=\mathbf{n}\times(\mathbf{n}\times\mathbf{r})=(\mathbf{n}.\mathbf{r})\mathbf{n}-\mathbf{r}$，于是 $\mathbf{K}^3\mathbf{r}=\mathbf{n}\times\big((\mathbf{n}.\mathbf{r})\mathbf{n}-\mathbf{r}\big)=-\mathbf{K}\mathbf{r}$。在级数（series）中，奇数次幂给出 $(\theta-\frac{\theta^3}{3!}+\cdots)\mathbf{K}=\sin\theta\,\mathbf{K}$，偶数次幂（由 $\mathbf{K}^2$ 起）给出 $(\frac{\theta^2}{2!}-\frac{\theta^4}{4!}+\cdots)\mathbf{K}^2=(1-\cos\theta)\mathbf{K}^2$。

4. 复合变换保持交线上的每一点不变，而且行列式为 $(-1)^2=1$。在每个垂直于该交线的平面上，它是对两条夹角为 α 的直线先后作反射的乘积，即转角为 2α 的旋转（rotation）。

7.6 双曲旋转与相对论

1. $T(u)T(w)$ 的元素是 $\cosh u\cosh w+\sinh u\sinh w=\cosh(u+w)$ 和 $\sinh u\cosh w+\cosh u\sinh w=\sinh(u+w)$。所以 $T(u)^{-1}=T(-u)$。

2. 由 $\tanh u=0.8$ 得 $u=\frac12\ln\frac{1.8}{0.2}=\ln 3\approx 1.099$ 及 $\cosh u=\frac12(3+\frac13)=\frac53$。在飞船上，过去的时间是 $10\times\frac35=6$ 年。

3. 面积是 $\frac12\oint(x\,\mathrm{d}y-y\,\mathrm{d}x)$。两条直边没有贡献；沿着双曲线（hyperbola），取 $x=\cosh s$，$y=\sinh s$，被积函数（integrand）是 $(\cosh^2 s-\sinh^2 s)\,\mathrm{d}s=\mathrm{d}s$，得出 $\frac{v}{2}$。T 的行列式为 $\cosh^2 u-\sinh^2 u=1$，所以它保持面积不变；它把双曲线映射到自身，并把参数为 v 的点移到参数为 $v+u$ 的点。

4. $x'+y'=(\cosh u+\sinh u)(x+y)=\mathrm{e}^u p$ 及 $x'-y'=\mathrm{e}^{-u}q$；另外 $x^2-y^2=pq$。在这组坐标下，T 是一个挤压映射（squeeze）：沿渐近线（asymptote）$y=x$ 拉伸 e^u 倍，沿 $y=-x$ 压缩至 e^{-u} 倍，使各点沿着每条双曲线 $pq=$ 常数滑动。

7.7 永恒不变的量

1. $\dot a=-Ct=-tc$，$\dot d=Ct=tc$，$\dot c=0$，以及 $\dot b=(A-D)t-Ct^3=t(a-d)$，与 $\mathbf{MN}-\mathbf{NM}=t\begin{pmatrix}-c&a-d\\0&c\end{pmatrix}$ 吻合。

2. $\dot{\mathbf{M}}=-\mathbf{S}^{-1}\dot{\mathbf{S}}\mathbf{S}^{-1}\mathbf{M}_0\mathbf{S}+\mathbf{S}^{-1}\mathbf{M}_0\dot{\mathbf{S}}=-\mathbf{N}\mathbf{M}+\mathbf{M}\mathbf{N}$，当中用到 $\dot{\mathbf{S}}=\mathbf{S}\mathbf{N}$。

3. $\mathbf{N}=\mathbf{M}^{-1}\dot{\mathbf{M}}=\begin{pmatrix}2&0\\0&-1\end{pmatrix}$，而 $\ln\det\mathbf{M}=2t-t=t$，其导数（derivative）为 $1=\operatorname{tr}\mathbf{N}$。

4. 利用迹的循环性质（cyclic property），$\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{tr}(\mathbf{M}^k)=\sum_j\operatorname{tr}(\mathbf{M}^j\dot{\mathbf{M}}\mathbf{M}^{k-1-j})=$

$k\,\mathrm{tr}(\mathbf{M}^{k-1}\dot{\mathbf{M}})$；再用一次循环性质，$\mathrm{tr}(\mathbf{M}^{k-1}(\mathbf{MN}-\mathbf{NM})) = \mathrm{tr}(\mathbf{M}^k\mathbf{N}) - \mathrm{tr}(\mathbf{M}^{k-1}\mathbf{NM}) = 0$。

7.8 四个平方数

1. $(L_1, L_2, L_3, L_4) = (1,4,4,4)$，而 $1+16+16+16 = 49$。这个表示法与 $7^2+0^2+0^2+0^2$ 不同。

2. 取 $c = d = r = s = 0$：$L_1 = ap - bq$，$L_2 = aq + bp$，$L_3 = L_4 = 0$，即 $(a^2+b^2)(p^2+q^2) = (ap-bq)^2 + (aq+bp)^2$。

3. 平方数模（modulo）8 只可能是 0、1 或 4。三个这样的数之和模 8 可得 $0,1,2,3,4,5,6$，但永远不会是 7。

4. $2 = (\pm1)^2 + (\pm1)^2 + 0^2 + 0^2$：选出两个非零的位置（6 种方法）及它们的正负号（4 种方法），共 24 种。由于 2 不能被 4 整除，雅可比（Jacobi）公式给出 $8\sigma(2) = 8\times 3 = 24$。

7.9 只有五种柏拉图立体

1. 设 $h = 2m + 2n - mn$，则 $V = \frac{4m}{h}$，$E = \frac{2mn}{h}$ 及 $F = \frac{4n}{h}$；把 m 和 n 对调，V 和 F 也随之对调。正方体与正八面体（octahedron）互为对偶（dual），正十二面体（dodecahedron）与正二十面体（icosahedron）亦然；正四面体（tetrahedron）则是自对偶（self-dual）的。

2. $F = 32$，$E = \frac{1}{2}(12\cdot 5 + 20\cdot 6) = 90$ 及 $V = \frac{1}{3}(180) = 60$，而 $60 - 90 + 32 = 2$。一般而言，若有 P 个五边形和 H 个六边形，则 $E = \frac{1}{2}(5P+6H)$ 及 $V = \frac{1}{3}(5P+6H)$，所以 $V - E + F = -\frac{1}{6}(5P+6H) + P + H = \frac{P}{6}$。这个值等于 2，所以 $P = 12$。

3. 把顶面和底面（两者都有洞）各分成四个梯形（trapezium）。于是 $V = 16$，$F = 4+4+4+4 = 16$ 及 $E = 12+12+8 = 32$，所以 $V - E + F = 0$。

4. $180^\circ n(1 - \frac{2}{m}) < 360^\circ$ 即 $n(m-2) < 2m$，也就是 $mn - 2m - 2n < 0$，或 $(m-2)(n-2) < 4$。当 $m, n \geq 3$ 时，解为 $(3,3)$、$(3,4)$、$(4,3)$、$(3,5)$ 和 $(5,3)$。

章

8

微分方程与动力系统

绝大多数微分方程都解不出来。学问就在于：解不出来，也要读懂它。

本章内容

8.1	当解不唯一	STEP III 1994
8.2	不求解而估界	STEP III 2013
8.3	收获种群	STEP II 1995
8.4	盘旋趋向圆	STEP III 2015
8.5	格力犬追逐格力犬	STEP III 1988
8.6	量子振子	STEP III 1988
8.7	可以解出的混沌	STEP I 2014
8.8	周期三蕴含混沌	STEP III 2025
A	练习答案	第 561 页

变化的法则

大自然对积分的种种困难一笑置之。

— 相传为拉普拉斯（Pierre-Simon Laplace）所言

在中学，微分方程（differential equation）是拿来解的。你把变量分离，或者找一个积分因子（integrating factor），又或者猜一个形如 $e^{\lambda x}$ 的解，最后得到一条公式。课本里的方程都是特意挑选的，所以这些方法才行得通。可惜的是，来自现实世界的微分方程（三颗行星的运动、天气、捕食者与猎物的种群数量、大幅度摆动的单摆），几乎全都没有可以用我们熟悉的函数写出来的解。

本章要谈的，就是这时候该怎么办。我们不再问“公式是甚么?”，而是问一些不靠公式也能回答的问题。解到底存在吗？是否只有一个？它会保持有界（bounded）吗？它会趋于一个稳定状态，还是趋于一个周而复始的循环，抑或永远游移不定？长远来说它会怎样？这种由“求解”转为“理解”的观点，始于 1880 年代的庞加莱（Henri Poincaré），并由此开创了今天称为*动力系统*（dynamical systems）的学科。

微分方程为何重要

微分方程是科学的语言。牛顿（Newton）第二定律是一条微分方程；麦克斯韦（Maxwell）的电磁学方程组、量子力学的薛定谔（Schrödinger）方程，以及用来预测天气和设计飞机的流体流动方程，也都是微分方程。流行病学家用它们为疾病的传播建立模型，生态学家用来描述鱼类资源的数量，经济学家用来描述经济增长，工程师则用来描述各种控制系统的响应，由恒温器到自动驾驶仪无一例外。凡是一个量的变化率取决于它当前的状态，就有一条微分方程，而且通常是一条没有人能精确解出的方程。

与这些连续模型并列的还有离散模型，其中时间是一步一步推进的：今年的种群数量决定明年的种群数量。这些*迭代映射*（iterated maps）写起来比较简单，但最后两节会说明，它们的行为可以同样狂野。1970 年代的重大发现之一，就是非常简单的法则也能产生*混沌*（chaos）：这种行为完全是确定的，实际上却无法预测。

你在大学会遇到甚么

大学一年级的微分方程课程会讲授标准的求解方法，但也会引入本章的这些问题。你会遇到存在性与唯一性定理（existence and uniqueness theorems），它们告诉你解在甚么时候存在，又在甚么时候由初始值决定。你会学习绘画*相图*（phase portraits），即把所

有解一次过画出来的图，并学习通过线性化（linearising）把平衡点（equilibria）分类为稳定或不稳定。往后的课程会走得远得多：*非线性动力学*（nonlinear dynamics）研究极限环（limit cycles）、分歧（bifurcations）和混沌；*数学生物学*（mathematical biology）把这些概念应用于种群和流行病；而偏微分方程（partial differential equations）、傅里叶级数（Fourier series）和正交函数（orthogonal functions）的理论（在 Section 8.6 可以略窥一二）则是量子力学和大部分应用数学的基础。

贯穿本章的问题

三大问题

1. **现在能决定将来吗？**一条微分方程配上一个初始值，在甚么时候恰好只有一个解？
2. **不求解，我们能知道甚么？**在没有公式的情况下，怎样为解定出上下界、找出它最终停留在哪里，或者描述它的形状？
3. **长远来说会怎样？**解会趋于稳定状态、趋近一个循环，还是呈现混沌的行为？我们又怎样分辨？

本章首先讨论唯一性，以及唯一性不成立时会发生甚么事（Section 8.1），然后发展一些不必求解也能理解方程的工具：不会增加的类能量的量、平衡点及其稳定性，以及配合问题对称性的坐标。有两节把这些想法应用于物理问题：在正方形内互相追逐的几只狗，以及量子谐振子（quantum harmonic oscillator）。本章最后讨论离散映射：在那里，一条公式揭示了混沌，而单单一个长度为三的循环，就迫使任何长度的循环都必然存在。

本章的概念

§	课题	概念
8.1	唯一性	包络线（envelopes）与奇解（singular solutions）；黏合
8.2	能量方法	一个永不增加的量
8.3	种群模型	平衡点、稳定性与临界点（tipping point）
8.4	极限环	极坐标与伯努利代换（Bernoulli substitution）
8.5	追逐曲线（pursuit curves）	对称性把四个物体化约为一个
8.6	正交函数	朗斯基行列式（Wronskian）与正交性
8.7	混沌	变量代换：逻辑斯谛映射（logistic map）其实是乔装了的倍角映射
8.8	周期轨道（periodic orbits）	区间覆盖；沙可夫斯基定理（Sharkovskii's theorem）

如何阅读本章

在动手解任何东西之前，先画草图。对于微分方程 $y' = F(y)$，标出 F 在哪里为正、为负和为零，并在 y 轴上画上箭头；对于映射 $x \mapsto f(x)$，把它的图像连同直线 $y = x$ 一起画出来，再用蛛网图法（cobweb method）追踪几个数列。这些图只需一分钟就能画好，而且往往在代数运算开始之前，就已经告诉你答案应该是甚么模样。

8.1 当解不唯一

题目

STEP III 1994，第 4 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/94-s3-q4`

引子

微分方程（differential equation）是一条告诉你某个量如何变化的规则：牛顿（Newton）定律告诉行星，在知道自己目前位置的情况下该如何运动。古典物理学的图像是，这样就已足够：只要知道世界此刻的状态，定律便决定了它的整个未来。拉普拉斯（Pierre-Simon Laplace）曾设想有这样一个智者，只要知道每一个位置和速度，便能预测一切。这幅图像建基于一个数学假设：一条微分方程加上一个初始条件（starting condition），恰好只有一个解。

这道题说明这个假设可以不成立。看似平平无奇的方程 $(y')^2 = 4y$，有些解会沿着 x 轴走，然后在自己选定的一刻离轴而去。经过某些点的解有无限多个。知道现在，并不能决定未来。找出唯一性（uniqueness）究竟在甚么时候成立，是大学微分方程课程中最先遇到的定理之一。

动笔之前

想一想 把两边开平方：$y' = \pm 2\sqrt{y}$。试试 $y = (x-c)^2$。它满足这条方程吗？有没有更简单的解？

若 $y = (x-c)^2$，则 $y' = 2(x-c)$，而 $(y')^2 = 4(x-c)^2 = 4y$。所以对每一个 x，每一条抛物线（parabola）$y = (x-c)^2$ 都是解。而 $y = 0$ 也是解：$(0)^2 = 4 \times 0$。请记住最后这一个解；它是整道题的关键。

逐步解题

阶段 1　经过一点的解

想一想　哪些抛物线 $y=(x-c)^2$ 经过 (a,b^2)？

我们需要 $(a-c)^2=b^2$，所以 $a-c=\pm b$，即 $c=a-b$ 或 $c=a+b$。由于 $b\neq 0$，这两个值并不相同，于是得出两个解

$$y=(x-a+b)^2 \qquad 及 \qquad y=(x-a-b)^2. \qquad \blacksquare$$

（我们还应该检查有没有遗漏：在 $y>0$ 的区域上，$y'=\pm 2\sqrt{y}$ 而正负号固定，并且 $\frac{\mathrm{d}}{\mathrm{d}x}\sqrt{y}=\frac{y'}{2\sqrt{y}}=\pm 1$，所以 $\sqrt{y}=\pm(x-c)$，得出的正是这些抛物线。因此在给定的点附近，只要 y 保持为正，就只有这两个解。不过，解一旦到达 x 轴，便可以与零解黏合起来，我们将在第 3 阶段看到这一点，而这样会多出无限多个对所有 x 都有定义的解。题目所说的“两个解”，就是这两条抛物线。）

阶段 2　两个点与一个共同的解

经过原点的解必定是 $y=x^2$（即 $c=0$ 的那条抛物线）。它在 $x=1$ 和 $x=-1$ 处达到高度 1。所以取 $(a_1,1)=(1,1)$ 和 $(a_2,1)=(-1,1)$。

经过 $(1,1)$（即 $a=1$，$b=\pm 1$）的解是 $y=x^2$ 和 $y=(x-2)^2$。经过 $(-1,1)$ 的解是 $y=(x+2)^2$ 和 $y=x^2$。两种情况下，经过原点的解都是同一条曲线 $y=x^2$。 $\blacksquare$

阶段 3　整个曲线族与一条公切线

停一停，试一试　就不同的 c 值，画出几条抛物线 $y=(x-c)^2$ 的草图。哪一条直线与它们每一条都相切？这条直线本身是不是解？

每一条抛物线 $y=(x-c)^2$ 都在其顶点 $(c,0)$ 与 x 轴相切，所以 x 轴是整个曲线族的公切线（common tangent），亦即它们的*包络线*（envelope）（图 8.1.1）。而 x 轴，即 $y=0$，本身也是这条微分方程的解。它不是那些抛物线的其中一条；它是另外一个独立的解，称为*奇解*（singular solution）。

阶段 4　为何重要

想一想　从原点出发。哪些解经过原点？你能否构造一个解，先沿着 x 轴走一段，然后离开？

经过原点的有 $y=x^2$ 和 $y=0$。但事实远不止于此。由于每条抛物线都与 x 轴相切（两

者在该处的斜率都是 0），你可以把各段平滑地黏合起来：对任意 $c_1 \le 0 \le c_2$，

$$y = \begin{cases} (x-c_1)^2 & x < c_1, \\ 0 & c_1 \le x \le c_2, \\ (x-c_2)^2 & x > c_2 \end{cases}$$

是一个可微函数（differentiable function），处处满足 $(y')^2 = 4y$，并且经过原点。经过 x 轴上每一点的解都有无限多个：一个解可以随意在 $y = 0$ 上停留多久，然后才离开。这条公切线正正就是唯一性失效的地方。 ■

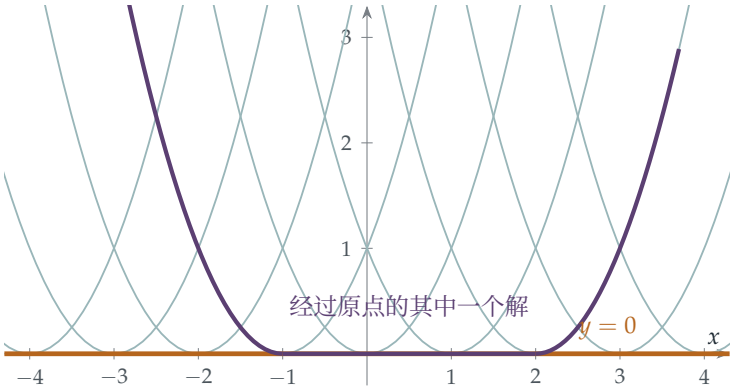


图 8.1.1 $(y')^2 = 4y$ 的解 $y = (x-c)^2$（蓝绿色）全都与 x 轴相切；这条轴是它们的包络线，本身也是解（琥珀色）。把各段黏合起来可得出更多的解，例如紫红色的那一个：它沿着一条抛物线落到轴上，沿轴而行，然后在自己选定的时刻再次离开。经过 x 轴上任何一点的解都有无限多个。

回顾

- **先找简单的解。**试一试 $y = (x-c)^2$ 和 $y = 0$，便看清了整个局面。
- **包络线给出奇解。**一族解的公切线，本身可以是一个不属于该族的解。
- **黏合。**两个解相切之处，可以把它们接合成一个新的可微解。非唯一性就是这样蔓延开去的。

背后的数学

解在甚么时候是唯一的？

皮卡—林德洛夫定理（Picard–Lindelöf theorem，由皮卡（Émile Picard）和林德洛夫（Ernst Lindelöf）约于 1890–1894 年证明）指出，方程 $y' = F(x, y)$ 连同 $y(x_0) = y_0$ 在 x_0 附近恰好有一个解，条件是 F 对 y 满足利普希茨（Lipschitz）条件：存在常数 L 使得 $|F(x, y_1) - F(x, y_2)| \le L|y_1 - y_2|$。粗略地说，当 y 变化时，F 不可以变化得无限快。我们的方程是 $y' = \pm 2\sqrt{y}$，而 $\sqrt{y}$ 在 $y = 0$ 处的斜率是无限大：它在该处不满足利普希茨条件，而那里正是唯一性失效的地方。（皮亚诺（Giuseppe Peano）在 1886 年证明了单凭 F 的连续性（continuity）已能保证解*存在*，但不能保证解唯一。）皮卡–林德洛夫定理的证明用到一个迭代（iteration），它凭借 Section 2.2 的压缩映射原理（contraction mapping principle）而收敛。

漏水的水桶

以下是一个物理例子。一个底部有孔的水桶按照托里拆利定律（Torricelli's law）流干：水的高度 h 满足 $h' = -k\sqrt{h}$。让时间向前走，一切都没有问题：水桶流干，然后一直是空的。但让时间倒流：假如你发现一个空桶，这条方程无法告诉你它是甚么时候流干的。在所有时间都有 $h = 0$ 的解，以及那些直到某一刻仍然盛满、然后才流干的解，全都与“现在是空的”相符。这种非唯一性，正是“关于过去的信息已经失去”这个事实在数学上的投影。

Norton 圆顶

2003 年，哲学家 John Norton 描述了一个无摩擦的圆顶，其形状使得一个恰好放在顶端的球所遵从的牛顿定律，形式为 $r'' = \sqrt{r}$。其中一个解是球永远停在顶端。但还有其他的解：球在顶端停留任意长的时间，然后在没有任何原因的情况下，自发地朝任意方向滚下。这个例子至今仍有争论：古典力学到底是否决定论（deterministic）的？从数学上看，这正是本题所见的那种非唯一性。

思维技巧

不要假设唯一性

找到一个解，并不表示它是唯一的解。要问一问是否还有其他解，尤其是当方程涉及不光滑的函数时。

在大学里，存在性与唯一性定理（existence and uniqueness theorems）会确切地告诉你需要甚么条件。在这里，一个在零处斜率无限大的平方根，已足以破坏唯一性。

延伸练习

1 **两种增长率。**试做 STEP II 2018，第 8 题（`step.maths.org/questions/18-s2-q8`）。该题用代换法（substitution）解两条增长方程。然后想一想，为甚么 $y = 0$ 也是具有相同初始值的解，而这对唯一性又说明了甚么。

2 **漏水的水桶。**解 $h' = -k\sqrt{h}$，其中 $h(0) = h_0 > 0$。水桶在甚么时候变空？写出两个不同的解，使得对某个 T 有 $h(T) = 0$，并且对所有 $t \geq T$ 有 $h(t) = 0$。

3 **恢复唯一性。**证明 $y' = 2y$ 连同 $y(0) = 0$ 只有 $y = 0$ 这一个解。（提示：考虑 $\frac{\mathrm{d}}{\mathrm{d}x}(ye^{-2x})$。）$2y$ 具有甚么性质，是 $2\sqrt{y}$ 所没有的？

4 **立方。**求 $y' = 3y^{2/3}$ 经过原点的所有解。

延伸阅读

- **Morris Hirsch, Stephen Smale & Robert Devaney, *Differential Equations, Dynamical Systems, and an Introduction to Chaos***（Academic Press）。讲述存在性与唯一性，以及本章其余的大部分内容。
- **George Simmons, *Differential Equations with Applications and Historical Notes***（CRC Press）。一本经典著作，富于历史掌故；书中讨论了奇解与包络线。
- **John Norton, “The dome: an unexpectedly simple failure of determinism”**, *Philosophy of Science* 75 (2008)。篇幅短小，容易阅读。

8.2 不求解而估界

题目

STEP III 2013，第 7 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/13-s3-q7`

引子

大多数微分方程（differential equation）都是解不出来的。$y'' + y^3 = 0$ 的解无法用我们熟悉的函数写成公式，第 (iii) 部分那头怪物就更不用说了。然而，我们仍然可以对这些解说出很多东西：它们保持有界（bounded）、它们会振荡、它们会渐趋平稳。诀窍在于找出一个由解及其导数（derivative）构成、而且永不增加的量。这样一来，解就被困住了。

在物理学中，这个量通常就是能量（energy）。在碗中滚动的球不可能爬到比起始高度更高的地方，因为能量守恒，而在有摩擦时能量更会减少。本题把同样的推理用于三条方程，其中最后一条是特意挑选的，要你自己发明那个“能量”。这个方法由李雅普诺夫（Aleksandr Lyapunov）于 1892 年提出，是研究微分方程最重要的工具之一。

动笔之前

想一想 在第 (i) 部分，把 y 看成质点的位置，把 x 看成时间。那么 $y'' = -y^3$ 就是牛顿定律，其中力为 $-y^3$。动能（kinetic energy）和势能（potential energy）分别是甚么？

动能是 $\frac{1}{2}(y')^2$（取质量为 1），而力 $-y^3$ 来自势函数 $V(y) = \frac{1}{4}y^4$，因为力等于 $-V'(y)$。总能量是 $\frac{1}{2}(y')^2 + \frac{1}{4}y^4$，即 E 的一半。能量守恒（conservation of energy）所说的，正正就是 E 为常数。

逐步解题

阶段 1　守恒

由链式法则（chain rule），

$$\frac{\mathrm{dE}}{\mathrm{d}x} = 2y'y'' + 2y^3y' = 2y'(y'' + y^3) = 0.$$

所以 E 是常数，等于它在 $x = 0$ 时的值：$\mathrm{E} = 0 + \frac{1}{2} = \frac{1}{2}$。由于 $(y')^2 \geq 0$，我们得到 $\frac{1}{2}y^4 \leq \frac{1}{2}$，所以 $y^4 \leq 1$，即 $|y| \leq 1$。 ■

这个解在 -1 与 1 之间振荡，在 (y, y') 平面上描出闭曲线 $(y')^2 + \frac{1}{2}y^4 = \frac{1}{2}$（图 8.2.1）。

阶段 2　耗散

> **想一想**　对 $\mathrm{E} = (v')^2 + 2\cosh v$ 求导。利用微分方程把 v'' 换掉。结果的正负号是甚么？

$$\frac{\mathrm{dE}}{\mathrm{d}x} = 2v'v'' + 2\sinh v\, v' = 2v'(v'' + \sinh v) = 2v'(-xv') = -2x(v')^2 \leq 0 \quad \text{当} x \geq 0.$$

xv' 这一项的作用有如摩擦，会带走能量。所以当 $x \geq 0$ 时，$\mathrm{E}(x) \leq \mathrm{E}(0)$。在 $x = 0$ 处，$\cosh(\ln 3) = \frac{3+\frac{1}{3}}{2} = \frac{5}{3}$，所以 $\mathrm{E}(0) = \frac{10}{3}$。因此

$$2\cosh v \leq (v')^2 + 2\cosh v = \mathrm{E}(x) \leq \tfrac{10}{3}, \qquad \text{所以} \qquad \cosh v(x) \leq \tfrac{5}{3}.$$ ■

阶段 3　发明能量

第 (iii) 部分没有给任何提示。我们需要一个不会增加的函数 $\mathrm{E} = (w')^2 + F(w)$。

> **想一想**　若 $\mathrm{E} = (w')^2 + F(w)$，则 $\mathrm{E}' = 2w'w'' + F'(w)w'$。$F'(w)$ 应该是甚么，才可以利用方程把 w'' 换掉，最后只剩下 $(w')^2$ 的倍数？

我们希望 $\mathrm{E}' = 2w'\big(w'' + w\cosh w + 2\sinh w\big)$，所以需要 $F'(w) = 2(w\cosh w + 2\sinh w)$。由于 $\frac{\mathrm{d}}{\mathrm{d}w}(w\sinh w) = \sinh w + w\cosh w$，我们可以取

$$F(w) = 2w\sinh w + 2\cosh w, \qquad \mathrm{E} = (w')^2 + 2w\sinh w + 2\cosh w.$$

于是，利用方程可得

$$\frac{\mathrm{dE}}{\mathrm{d}x} = -2(5\cosh x - 4\sinh x - 3)(w')^2.$$

停一停，试一试 证明对所有 x 都有 $5\cosh x - 4\sinh x - 3 \geq 0$。（把它用 e^x 表示，然后找出一个完全平方。）然后完成整个论证。

$5\cosh x - 4\sinh x - 3 = \frac{5}{2}(e^x + e^{-x}) - 2(e^x - e^{-x}) - 3 = \frac{1}{2}e^x - 3 + \frac{9}{2}e^{-x} = \frac{1}{2}\big(e^{x/2} - 3e^{-x/2}\big)^2 \geq 0$。所以 E 永不增加。在 $x = 0$ 处，$\mathrm{E}(0) = \frac{1}{2} + 0 + 2 = \frac{5}{2}$。由于对任何 w 都有 $w\sinh w \geq 0$（两个因子同号），

$$2\cosh w \leq (w')^2 + 2w\sinh w + 2\cosh w = \mathrm{E}(x) \leq \tfrac{5}{2}, \qquad \text{所以} \qquad \cosh w(x) \leq \tfrac{5}{4}. \quad \blacksquare$$

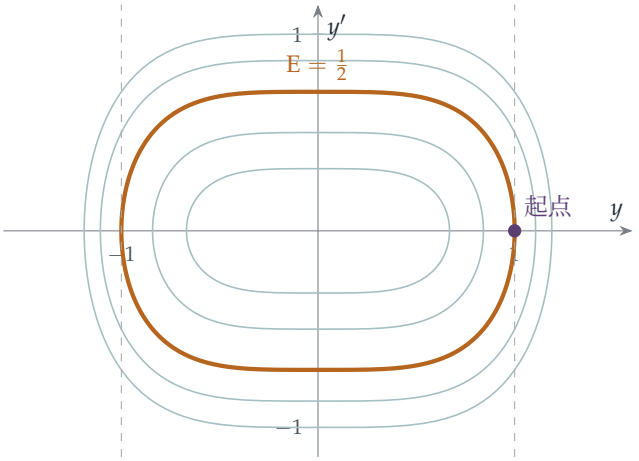


图 8.2.1 能量 $\mathrm{E} = (y')^2 + \frac{1}{2}y^4$ 在 (y, y') 平面上的等高线。由于 E 是常数，第 (i) 部分的解从 $y = 1$、$y' = 0$ 出发，永远留在曲线 $\mathrm{E} = \frac{1}{2}$（琥珀色）上，因而永不离开带状区域 $|y| \leq 1$（虚线）。我们根本不必找出 y 的公式，就已经知道这一点。

回顾

- **乘以 y'。**在每一部分，E′ 都是 $2y'$ 乘以方程的左方，再加上余下的项。正是这个余下的项决定了 E 是守恒还是递减。
- **能量限制位置。**由于动能部分 $(y')^2$ 非负，E 的上界就是势能部分的上界，从而也给出 y 的界。
- **留意隐藏的平方。**第 (iii) 部分的阻尼系数看来可能会变号，但它其实是一个经过伪装的完全平方。

背后的数学

李雅普诺夫函数

李雅普诺夫在 1892 年的博士论文中研究运动的*稳定性*（stability）：如果一个系统受到轻微扰动而偏离平衡，它会回到平衡，还是愈走愈远？他的“第二方法”是找出一个函数，它在平衡点处取最小值，而且沿着解永不增加；这种函数现在称为*李雅普诺夫函数*（Lyapunov function）。这样，从平衡点附近出发的解就会一直留在平衡点附近。这个方法的美妙之处，在于它不需要解的公式，只需要微分方程本身，而这正是本题所展示的。现今，从飞机控制系统、电网以至生态系统，各种各样事物的稳定性都是用李雅普诺夫函数来证明的。

物理与工程中的能量

第 (i) 部分是一个非线性振子（nonlinear oscillator）：它像弹簧，但回复力与位移的立方成正比。与简谐振子不同，它的周期取决于振幅（振荡愈大，速度愈快），但能量守恒仍然把它困住。第 (ii) 部分加入了随时间增强的阻尼（damping），能量因而逐渐流失。在工程学中，人们用能量方法来证明结构不会把自己摇散，以及控制系统不会振荡至失控。在偏微分方程（partial differential equation）的理论中，这类“能量估计”是证明解存在而且性质良好的主要工具；有关流体流动方程的问题，包括千禧年大奖难题（Millennium Prize Problems）之一的纳维－斯托克斯（Navier–Stokes）问题，归根究底就是能否找到这样的界。

思维技巧

找出一个不能增加的量

如果你解不出一条方程，就找一个由解构成、而且导数有确定正负号的量。它会给出解的界。

这种寻找一方面由物理引导（“能量是甚么?”），另一方面由代数引导（“我要加上甚么，导数才会化简?”）。第 (iii) 部分示范了在没有物理直觉可以依靠时，如何走代数这条路。

延伸练习

1 **简谐运动。**对于 $y'' + y = 0$，其中 $y(0) = A$、$y'(0) = 0$，求守恒的能量，并由此推出 $|y| \le A$。用精确解核对。

2 **单摆。**单摆的角度 θ 满足 $\theta'' + \sin\theta = 0$（采用适当的单位）。证明 $(\theta')^2 - 2\cos\theta$ 是常数。若单摆在角度 θ_0 处由静止开始，其中 $|\theta_0| \le \pi$，证明 $|\theta|$ 永不超过 $|\theta_0|$。若 $|\theta_0| > \pi$，哪里会出问题？

3 **阻尼。**对于 $y'' + 2ky' + y = 0$，其中 $k > 0$，证明 $(y')^2 + y^2$ 递减。它是否必定趋向零？

4 **逃逸。**对于 $y'' - y^3 = 0$，其中 $y(0) = 1$、$y'(0) = 0$，求一个守恒量，并证明它不能给出 y 的界。哪里出了问题？

延伸阅读

- **Steven Strogatz, *Nonlinear Dynamics and Chaos***（CRC Press）。第 6 章和第 7 章讲述相平面、保守系统和李雅普诺夫函数。
- **D. W. Jordan & P. Smith, *Nonlinear Ordinary Differential Equations***（Oxford University Press）。一本标准的大学本科教科书，载有大量能量方法的例子。
- **Keith Devlin, *The Millennium Problems***（Basic Books）。书中有一篇关于纳维－斯托克斯问题的浅白介绍。

8.3 收获种群

题目

STEP II 1995，第 8 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/95-s2-q8`

引子

食物供应无限的种群会以指数增长（exponential growth）的方式扩张。现实中的种群却总会用尽空间：当种群数量接近环境所能维持的上限，即其*环境承载力*（carrying capacity）时，增长便会放缓。描述这个现象的最简单模型是*逻辑斯谛方程*（logistic equation），它出现在本题的前半部分，由培养皿中的细菌到新科技的普及，都可以用它来描述。

后半部分所问的，是一个在现实中关系重大的问题：假如我们以稳定的速率收获（harvest）这个种群，会发生甚么事？在某个阈值（threshold）以下，种群会调整到一个新的、较低的平衡点（equilibrium）。但只要收获率越过这个阈值，哪怕只是超出少许，平衡点便会消失，种群数量随即崩溃至零。这就是渔业崩溃背后的数学，也是设立捕鱼配额的原因。

动笔之前

想一想 在动手求解之前，先把增长率 $(2K - x)x$ 当作 x 的函数，画出它的略图。它在哪里为零？在哪里为正？这告诉你 x 会怎样变化？

增长率在 $x = 0$ 和 $x = 2K$ 处为零，在两者之间为正，超过 $2K$ 则为负。所以，由 0 与 $2K$ 之间出发的种群会向 $2K$ 增长，而由 $2K$ 以上出发的种群则会向它缩减。环境承载力是 $2K$，而增长在 $x = K$ 时最快，那时的增长率是 K^2。不必解方程，你也能预测长远的变化。

逐步解题

阶段 1　解逻辑斯谛方程

想一想　分离变量（separate the variables）：$\displaystyle\int \frac{\mathrm{d}x}{(2K-x)x} = \int \mathrm{d}t$。把左边拆成部分分式（partial fractions）。

$\dfrac{1}{(2K-x)x} = \dfrac{1}{2K}\left(\dfrac{1}{x} + \dfrac{1}{2K-x}\right)$，所以

$$\frac{1}{2K}\ln\frac{x}{2K-x} = t + C.$$

当 $t=0$ 时，$x=K$，所以 $\ln 1 = 0$，而 $C=0$。于是 $\frac{x}{2K-x} = e^{2Kt}$，由此得

$$x = \frac{2Ke^{2Kt}}{1+e^{2Kt}} = \frac{2K}{1+e^{-2Kt}} = K + K\frac{1-e^{-2Kt}}{1+e^{-2Kt}}. \qquad \blacksquare$$

这个分式就是 $\tanh(Kt)$，所以 $x = K + K\tanh Kt$：这是一条由 K 上升并趋向 $2K$ 的 S 形曲线。当 $t\to\infty$ 时，$e^{-2Kt}\to 0$，而 $x\to 2K$。

阶段 2　收获

以速率 L 移走细菌，便得

$$\frac{\mathrm{d}x}{\mathrm{d}t} = (2K-x)x - L.$$

想一想　对 $(2K-x)x - L$ 进行配方（complete the square）。它的零点在哪里？

$(2K-x)x - L = K^2 - L - (x-K)^2 = \alpha^2 - (x-K)^2 = \big(\alpha-(x-K)\big)\big(\alpha+(x-K)\big)$。平衡点是 $x = K \pm \alpha$。

停一停，试一试　现在作代换（substitution）$y = x - K + \alpha$。y 满足甚么方程？把它与原来的逻辑斯谛方程比较一下。

设 $y = x - K + \alpha$，则 $x - K = y - \alpha$，所以 $\alpha + (x-K) = y$，而 $\alpha - (x-K) = 2\alpha - y$。因此

$$\frac{\mathrm{d}y}{\mathrm{d}t} = (2\alpha - y)y,$$

这又是逻辑斯谛方程，只是 K 换成了 α。初始时 $y(0) = K - K + \alpha = \alpha$，正好对应第一部分的 $x(0) = K$。所以根据第 1 阶段的结果（以 α 代替 K），$y(t) = \alpha + \alpha\tanh\alpha t \to 2\alpha$，因此

$$x(t) = K + \alpha\tanh\alpha t \to K + \alpha. \qquad \blacksquare$$

回顾

- **先作定性分析。**还未动手求解，增长率的正负号已经预测了长远的变化。
- **配方揭示平衡点。**有收获的方程的平衡点在 $K \pm \alpha$。
- **代换把新问题化为旧问题。**把变量平移之后，有收获的方程便变成了早已解出的逻辑斯谛方程。

背后的数学

临界点

有收获的方程有两个平衡点，$K+\alpha$ 和 $K-\alpha$。上面的一个是*稳定*（stable）的：种群数量稍高于或稍低于它，都会回到它那里。下面的一个是*不稳定*（unstable）的：种群数量只要稍低于 $K-\alpha$，便会一直下跌，直至灭绝。随着收获率 L 增加，$\alpha = \sqrt{K^2 - L}$ 会缩小，两个平衡点互相靠近；到 $L = K^2$ 时两者合而为一，而当 $L > K^2$ 时则完全没有平衡点，任何种群都会崩溃（图 8.3.1）。这是一个*鞍结分岔*（saddle-node bifurcation），是最简单的一种*临界点*（tipping point）。它的可怕之处，在于事前没有任何明显的预兆：当 L 接近 K^2 时，平衡种群数量只是缓缓下降，然后平衡点突然不复存在。事后才减少收获未必有用，因为种群数量可能已经跌到不稳定的阈值以下。

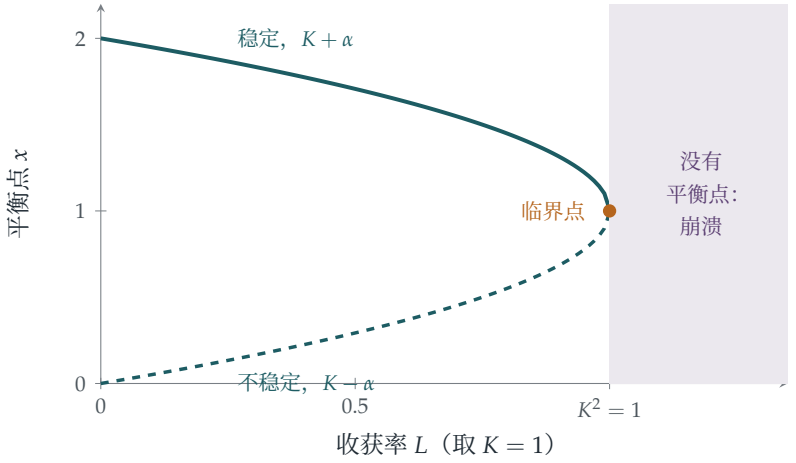


图 8.3.1 平衡种群数量与收获率 L 的关系，其中 $K = 1$。随着 L 增加，稳定平衡点（实线）与不稳定阈值（虚线）互相靠近，并在 $L = K^2$ 处相遇。越过这一点之后，平衡点不再存在，无论种群有多大都会崩溃。

最大可持续产量

种群所能承受的最大收获量是多少？答案是 $L = K^2$，即增长率 $(2K - x)x$ 的峰值，在种群数量维持于 K（环境承载力的一半）时达到。二十世纪中叶，渔业科学采纳了这个“最大可持续产量”（maximum sustainable yield）作为管理目标。可是，恰好按最大值收获，就等于把系统恰好放在临界点上，只要稍为高估了 K，或者遇上一个坏年头，便会引致崩溃。纽芬兰对开的大浅滩（Grand Banks）鳕鱼渔场，五百年来一直是全世界最丰饶的渔场之一，却在 1990 年代初崩溃；1992 年，加拿大政府颁布禁捕令，令数以万计的人失去工作。鱼类资源至今仍未完全恢复。

其他地方的逻辑斯谛增长

逻辑斯谛方程由 Pierre-François Verhulst 于 1838 年提出，用来为人类人口建立模型。同一条 S 形曲线，也描述了疫症初期的传播、新科技的普及、自催化化学反应以及肿瘤的生长。它的离散时间版本，即逻辑斯谛映射（logistic map），表现却截然不同：它可以是混沌（chaotic）的，我们会在 Section 8.7 看到。

思维技巧

平衡点及其稳定性

对于方程 $x' = f(x)$，平衡点就是 f 的零点，而 f 在各零点之间的正负号，告诉你解会向哪个方向移动。

若 f 在某个平衡点由正变负，该平衡点是稳定的；若由负变正，则是不稳定的。画出 f 的图像，往往不必求解任何东西，便能回答最重要的问题。

延伸练习

1. **由低处出发。**在有收获且 $K^2 > L$ 的情况下，证明若 $x(0) < K - \alpha$，则种群数量会在有限时间内降至零。（提示：当 $y(0) < 0$ 时，y 会怎样变化?）

2. **过度收获。**若 $L > K^2$，设 $\beta = \sqrt{L - K^2}$。在 $x(0) = 2K$ 的条件下解 $x' = -(x - K)^2 - \beta^2$，并求种群数量何时降至零。

3. **按比例收获。**假设改为移走种群中某个固定的比例：$x' = (2K - x)x - hx$。求平衡点，并证明对所有 $h < 2K$，种群都能存活。为甚么这个方案较为安全?

4. **半程时间。**在没有收获的情况下，求种群数量达到 $\frac{3}{2}K$ 的时间。

延伸阅读

- **Steven Strogatz, *Nonlinear Dynamics and Chaos***（CRC Press）。第 3 章讨论分岔，并以这个收获模型作为习题。
- **Nicholas Britton, *Essential Mathematical Biology***（Springer）。为大学本科生而写的种群模型。
- **Mark Kurlansky, *Cod: A Biography of the Fish That Changed the World***（Penguin）。鳕鱼渔业及其崩溃的历史。

8.4 盘旋趋向圆

题目

STEP III 2015，第 8 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/15-s3-q8`

引子

自然界中有许多系统，无论起始状态如何，最终都会进入同一种规律的节奏：心脏的跳动、蟋蟀的鸣叫、稳定拉奏的小提琴弦、捕食者与猎物的种群数量周期。稍加扰动，它们又会回到原来的节奏。用微分方程（differential equation）的语言来说，它们的解会趋近一条闭合曲线，称为**极限环**（limit cycle）。

这道题目正是要构造一个极限环。方程以 x 和 y 表示时看似无从入手，但换成极坐标（polar coordinates）后，就化简成可以求解的形式，而且各个解原来都会从内侧和外侧盘旋着贴近单位圆。当中的关键一步，是选取与问题的对称性（symmetry）相配的坐标，这也是本书一直沿用的手法。

动笔之前

想一想 两条方程其实都暗藏 $x\,\mathrm{d}x + y\,\mathrm{d}y$ 和 $x\,\mathrm{d}y - y\,\mathrm{d}x$ 这两个组合。它们在极坐标中是甚么？

由于 $r^2 = x^2 + y^2$，求导后得 $x\,\mathrm{d}x + y\,\mathrm{d}y = r\,\mathrm{d}r$。另外 $x\,\mathrm{d}y - y\,\mathrm{d}x = r^2\,\mathrm{d}\theta$（这就是 Section 4.9 中的面积公式）。有了这两条恒等式（identity），变量代换（change of variables）几乎毫不费力：只要把方程改写成只含这两个组合的形式即可。

逐步解题

阶段 1　以极坐标处理第 (i) 部分

把方程写成 $(y+x)\,\mathrm{d}y-(y-x)\,\mathrm{d}x=0$，再重新整理：

$$(x\,\mathrm{d}x+y\,\mathrm{d}y)+(x\,\mathrm{d}y-y\,\mathrm{d}x)=0.$$

> **想一想**　展开算式，验证以上的整理无误。然后运用上面两条恒等式。

$(y+x)\,\mathrm{d}y-(y-x)\,\mathrm{d}x=y\,\mathrm{d}y+x\,\mathrm{d}y-y\,\mathrm{d}x+x\,\mathrm{d}x$，正是上面的算式。化成极坐标形式就是 $r\,\mathrm{d}r+r^2\,\mathrm{d}\theta=0$，两边除以 $r^2\,\mathrm{d}\theta$（其中 $r>0$），得

$$\frac{\mathrm{d}r}{\mathrm{d}\theta}+r=0. \qquad \blacksquare$$

方程的解是 $r=Ce^{-\theta}$：这些是对数螺线（logarithmic spiral），随着 θ 增加而向原点卷入，随着 θ 减少而向无穷远处卷出。

阶段 2　第 (ii) 部分

> **停一停，试一试**　新方程是在旧方程之上，加入含有 $x^2+y^2=r^2$ 的额外项。用同样的方法改写，并求出 $\frac{\mathrm{d}r}{\mathrm{d}\theta}$。

方程是 $\bigl(y+x-xr^2\bigr)\,\mathrm{d}y-\bigl(y-x-yr^2\bigr)\,\mathrm{d}x=0$，即

$$\bigl[(y+x)\,\mathrm{d}y-(y-x)\,\mathrm{d}x\bigr]-r^2\bigl(x\,\mathrm{d}y-y\,\mathrm{d}x\bigr)=0.$$

第一个括号和之前一样，等于 $r\,\mathrm{d}r+r^2\,\mathrm{d}\theta$，第二个括号则是 $r^2\,\mathrm{d}\theta$。因此 $r\,\mathrm{d}r+r^2(1-r^2)\,\mathrm{d}\theta=0$，由此得

$$\frac{\mathrm{d}r}{\mathrm{d}\theta}=r^3-r.$$

> **想一想**　这条方程对 r 而言并非线性（linear），但换成 $u=r^{-2}$ 后就变成线性。u 满足甚么方程？

设 $u=r^{-2}$，则 $\frac{\mathrm{d}u}{\mathrm{d}\theta}=-2r^{-3}\frac{\mathrm{d}r}{\mathrm{d}\theta}=-2r^{-3}(r^3-r)=-2+2u$。所以 $\frac{\mathrm{d}u}{\mathrm{d}\theta}-2u=-2$，其解为 $u=1+Ae^{2\theta}$。因此

$$r^2=\frac{1}{1+Ae^{2\theta}}. \qquad \blacksquare$$

阶段 3　三类解

- $A = 0$：$r = 1$。单位圆是一个解。
- $A > 0$：恒有 $r < 1$。当 $\theta \to -\infty$ 时，$r \to 1$；当 $\theta \to +\infty$ 时，$r \to 0$。曲线在圆与原点之间盘旋。
- $A < 0$：$r > 1$，而且公式只在 $1 + Ae^{2\theta} > 0$ 时才有意义，即 $\theta < \frac{1}{2}\ln\left(-\frac{1}{A}\right)$。当 $\theta \to -\infty$ 时，$r \to 1$；当 θ 趋近该上界时，$r \to \infty$。曲线在圆与无穷远之间盘旋。

所以当 $\theta \to -\infty$ 时，每条解曲线都会从内侧或外侧趋近单位圆（图 8.4.1）。

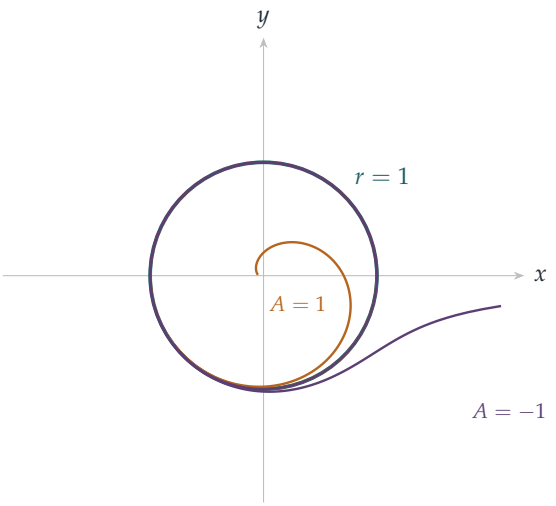


图 8.4.1 第 (ii) 部分的解。单位圆（蓝绿色）是一个解。$A = 1$ 的解（琥珀色）位于圆内，从圆出发（当 $\theta \to -\infty$ 时）盘旋着走向原点；$A = -1$ 的解（紫红色）位于圆外，从圆出发盘旋着走向无穷远。反方向来看（θ 递减），每个解都会缠绕到圆上：这就是极限环。

回顾

- **善用对称性。**这些方程对所有方向一视同仁，而极坐标令这一点显露出来。
- **认出标准组合。**$x\,dx + y\,dy = r\,dr$ 和 $x\,dy - y\,dx = r^2\,d\theta$ 完成了大部分工作。
- **代换可使方程线性化。**非线性方程 $r' = r^3 - r$ 换成 $u = r^{-2}$ 后变成线性（这是一条*伯努利方程*（Bernoulli equation））。

背后的数学

极限环

庞加莱（Henri Poincaré）在 1880 年代研究平面上微分方程的定性行为时，引入了极限环。不妨把题目中的方程理解为描述一个点在平面上运动的路径。单位圆是一条孤立的闭轨道（closed orbit），附近的每条轨道都会缠绕着向它靠近（或远离它，视乎时间的方向而定）。1926 年，荷兰工程师 Balthasar van der Pol 在一个装有真空管的电路中发现了极限环，并提出类似的方程可以描述心脏的跳动。凡是具有自持而稳定节奏的系统，都可以用极限环作为模型：生理时钟、神经脉冲、捕食者与猎物种群数量的周期。

平面上没有混沌

庞加莱—本迪克松定理（Poincaré–Bendixson theorem）指出，对于平面上的光滑系统，如果一条轨道一直留在某个有界区域内，又不趋近任何平衡点（equilibrium），它就必定趋近一条闭轨道。所以在二维的情况下，长期行为只限于平衡点和环：混沌（chaos）没有容身之处，因为轨道不能相交。混沌至少需要三维，正如洛伦兹（Edward Lorenz）在 1963 年以他那著名的三变量对流模型（“蝴蝶”）所发现的那样。至于离散时间的情况，我们会在 Section 8.7 看到，即使只有一维，混沌也可能出现。

一个未解决的问题

1900 年，希尔伯特（David Hilbert）在他的第十六问题中问道：当 P 和 Q 是 n 次多项式时，系统 $x' = P(x, y)$，$y' = Q(x, y)$ 最多可以有多少个极限环？即使是 $n = 2$ 的情况，答案至今仍然未知：已知有些系统具有四个极限环，但是否可能有更多，则无人知晓。在希尔伯特的问题之中，这是极少数至今仍然完全悬而未决的一个。

思维技巧

转换坐标以配合对称性

如果问题具有旋转对称性，就用极坐标；如果具有其他对称性，就找出与之相配的坐标。

合适的坐标把一条难以处理的方程变成可分离变量（separable）的方程，并令解的几何形状（圆和螺线）一目了然。

延伸练习

1 **螺线。**对于第 (i) 部分的解 $r = e^{-\theta}$，证明曲线与每条由原点出发的射线都以相同的角度 45° 相交。（这就是对数螺线又称为等角螺线（equiangular spiral）的原因。）

2 **极坐标恒等式。**证明 $x\,dx + y\,dy = r\,dr$ 和 $x\,dy - y\,dx = r^2\,d\theta$。

3 **随时间变化的系统。**考虑系统 $\dot{r} = r(1 - r^2)$，$\dot{\theta} = 1$。证明当 $t \to \infty$ 时，每个满足 $r(0) > 0$ 的解都趋近单位圆。把这个系统以 x 和 y 表示。

4 **两个环。**找出一个形如 $\dot{r} = f(r)$，$\dot{\theta} = 1$ 的系统，使它有两个极限环，分别位于 $r = 1$ 和 $r = 2$，其中一个是吸引的（attracting），另一个是排斥的（repelling）。

延伸阅读

- **Steven Strogatz, *Nonlinear Dynamics and Chaos***（CRC Press）。第 7 章讲述极限环和庞加莱–本迪克松定理。
- **Steven Strogatz, *Sync***（Hyperion）。一本讲述自然界中的节奏与同步现象的科普读物。
- **James Gleick, *Chaos: Making a New Science***（Penguin）。讲述洛伦兹与发现混沌的故事。

8.5 格力犬追逐格力犬

题目

STEP III 1988，第 10 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/88-s3-q10`

引子

四只狗分别站在一块正方形田地的四个角上，每只狗都追赶牠左边的那一只。牠们会在哪里相遇？要花多少时间？所走的路径又是怎样的？这些路径是优美的螺线，绕着中心转无穷多圈，然而每只狗只跑了有限的距离，恰好等于正方形的边长。

这是一个追逐问题（pursuit problem），这类问题最早由 Pierre Bouguer 于 1732 年研究，当时考虑的是一艘船追赶海盗船的情形。令问题得以解决的诀窍是对称性（symmetry）：发生在一只狗身上的事，同样发生在其余各只身上，所以四只狗始终构成一个正方形。一旦看出这一点，微分方程（differential equation）几乎就自然而然地写出来了。

动笔之前

? **想一想** 根据对称性，四只狗在每一刻构成甚么形状？它的中心在哪里？

把整个情形绕正方形的中心旋转四分之一圈（即 $A \to B \to C \to D \to A$），情形保持不变，所以四只狗始终构成一个正方形，而且中心始终是同一点 O。这个正方形一边旋转一边缩小，但始终是正方形。特别地，A 和 B 与 O 的距离始终同为 r，而 OB 始终比 OA 超前四分之一圈。

逐步解题

阶段 1　运动的方向

想一想　A 朝着 B 跑。在三角形 OAB 中，AB 与 AO（指向中心的方向）所成的角是多少?

三角形 OAB 中 $OA = OB$，且在 O 处是直角，所以其余两个角都是 45°。因此 A 的速度（velocity）大小为 v，方向与向内的径向（radial）方向成 45°，并且偏向 B 的一方，即 θ 增加的方向（假设我们标记角度时，令 B 位于 A 的逆时针方向）。

阶段 2　微分方程

想一想　把 A 的速度分解为径向和横向（transverse）两部分。在极坐标（polar coordinates）中，径向速度是 $\dot{r}$，横向速度是 $r\dot{\theta}$。

径向分量（向内）是 $v\cos 45°$，横向分量是 $v\sin 45°$，所以

$$\dot{r} = -\frac{v}{\sqrt{2}}, \qquad r\dot{\theta} = \frac{v}{\sqrt{2}}.$$

两式相除，得 $\dfrac{\mathrm{d}r}{\mathrm{d}\theta} = \dfrac{\dot{r}}{\dot{\theta}} = -r$，所以 $r = \lambda e^{-\theta}$，其中若最初 $\theta = 0$，则 λ 是最初与中心的距离。 ■

阶段 3　n 只狗

停一停，试一试　对于正 n 边形，三角形 OAB 中在 O 处的角是多少？AB 与 AO 所成的角是多少？重复第 2 阶段的做法。

在 O 处的角是 $\frac{2\pi}{n}$，所以两个相等的底角是 $\frac{1}{2}(\pi - \frac{2\pi}{n}) = \frac{\pi}{2} - \frac{\pi}{n}$。速度与向内的半径所成的角就是这个角，所以

$$\dot{r} = -v\cos\left(\frac{\pi}{2} - \frac{\pi}{n}\right) = -v\sin\frac{\pi}{n}, \qquad r\dot{\theta} = v\sin\left(\frac{\pi}{2} - \frac{\pi}{n}\right) = v\cos\frac{\pi}{n}.$$

因此 $\dfrac{\mathrm{d}r}{\mathrm{d}\theta} = -r\tan\dfrac{\pi}{n}$，于是

$$r = \lambda e^{-\theta\tan(\pi/n)}. \qquad ■$$

当 $n = 4$ 时，$\tan\frac{\pi}{4} = 1$，这就回到正方形的情形。

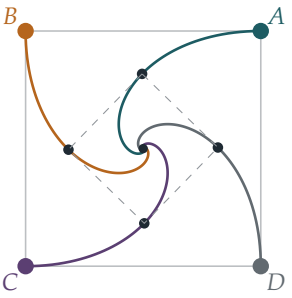


图 8.5.1 四只狗从正方形的四个角出发互相追逐。每只狗都沿着一条对数螺线 $r = \lambda e^{-\theta}$（相对于牠自己的起始方向）前进，绕着中心转无穷多圈。在每一刻，四只狗都构成一个正方形（虚线所示为其中一个中间时刻），这个正方形一边旋转一边缩小，直至牠们在中心相遇。

回顾

- **对称性把多个物体化为一个。**四只互相影响的狗，变成了一个遵从简单法则的质点（particle），因为整个布局始终是正方形。
- **绕中心的运动宜用极坐标。**速度的径向和横向分量 $\dot{r}$ 和 $r\dot{\theta}$ 直接给出了方程。
- **相除以消去时间。**比值 $\dot{r}/\dot{\theta}$ 给出了路径的形状，而毋须知道何时到达每一点。

背后的数学

无穷多圈，有限距离

径向速率 $\dot{r} = -v\sin\frac{\pi}{n}$ 是常数，所以各只狗在时间 $T = \frac{r_0}{v\sin(\pi/n)}$ 之后到达中心，每只狗跑过的距离是 $vT = \frac{r_0}{\sin(\pi/n)}$。对于边长为 s 的正 n 边形，$r_0 = \frac{s}{2\sin(\pi/n)}$，所以每只狗跑过的距离是

$$\frac{s}{2\sin^2(\pi/n)}.$$

对于四只狗，这恰好是 s：每只狗跑过的距离正好是一条边的长度。（这里有一个巧妙的看法：B 的速度始终与 AB 垂直，所以它不会改变距离 AB；只有 A 以速率 v 沿着 AB 笔直前进的运动才会缩短两者的距离，而这个距离最初是 s。）对于三只狗，距离是 $\frac{2}{3}s$；对于六只狗，距离是 $2s$。然而在每一种情形下，螺线都绕着中心转无穷多圈，因为当 $r \to 0$ 时 $\theta \to \infty$。一条长度有限却转了无穷多圈的曲线并不是悖论：那些圈缩小得足够快。

对数螺线

曲线 $r = \lambda e^{-k\theta}$ 称为对数螺线（logarithmic spiral）或等角螺线（equiangular spiral）：它与每一条从中心出发的射线相交的角度都相同。它出现在大自然之中，例如鹦鹉螺的壳和某些星系的旋臂，因为保持形状不变的生长就会产生这种曲线。雅各布·伯努利（Jacob Bernoulli）对它极为着迷，甚至要求把它刻在自己的墓碑上，并附上格言 *Eadem mutata resurgo*（“纵然改变，我仍将依旧复起”），指的是它在任何尺度下看起来都一样。结果石匠刻的却是一条阿基米德螺线（Archimedean spiral）。

实际中的追逐

“纯追逐”（pure pursuit），即始终笔直朝着目标前进，是导弹或捕食者最简单的导引法则。但它并不是最好的：朝着目标将会到达的位置前进（“比例导引”，proportional navigation）更有效率，而现代导弹以及某些捕食者（例如蜻蜓和隼）实际上正是这样做的。追逐与逃避的数学在二十世纪自成一门学科，名为微分博弈（differential games）。

思维技巧

利用对称性简化多体问题

如果一个系统具有对称性，而其法则又不破坏这种对称性，那么对称性就会一直保持下去，你只需追踪其中一部分。

这个想法能解决力学和动力学中的许多问题；它是对称与守恒（conservation）之间深刻联系的一个简单例子。

延伸练习

1. **巧妙的论证。**仔细解释为何在四只狗的问题中，距离 AB 恰好以速率 v 减少，并由此推出四只狗在时间 s/v 之后相遇。

2. **三只狗。**三只狗从边长为 s 的等边三角形的三个角出发。求每条路径的方程、牠们相遇所需的时间，以及每只狗跑过的距离。

3. **转了多少圈？**设四只狗出发时距离中心 100 m，牠们转了完整一圈之后距离中心多近？转了多少个完整的圈之后，牠们与中心的距离在 1 mm 以内？

4. **两只狗。**当 $n = 2$（两只狗面对面）时会怎样？公式是否仍然有意义？

延伸阅读

- **Paul Nahin, *Chases and Escapes: The Mathematics of Pursuit and Evasion***（Princeton University Press）。关于追逐曲线的一切，从 Bouguer 到导弹。
- **Eli Maor, *e: The Story of a Number***（Princeton University Press）。内容包括对数螺线和伯努利的墓碑。
- **D'Arcy Thompson, *On Growth and Form***（Cambridge University Press）。讲述大自然中的数学的经典著作，包括螺线式的生长。

8.6 量子振子

题目

STEP III 1988，第 7 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/88-s3-q7`

引子

系在弹簧上的质点，振动的能量可以随你喜欢：把它拉开少许也好，拉开很远也好。但在 1900 年，普朗克（Max Planck）发现，要解释高温物体所发出的光的颜色，唯一的办法是假定振动中的原子只能具有某些特定的能量，每级相差一个固定的大小。在原子的尺度上，能量是一份一份的，称为量子（quanta）。二十五年后，量子力学（quantum mechanics）解释了个中原因。

这道题目正包含了这个解释。题中的微分方程（differential equation）经过单位转换后，就是一个量子粒子处于弹簧式势能中的薛定谔方程（Schrödinger equation），亦即量子谐振子（quantum harmonic oscillator）。只有当能量取某些特殊的值时，它才有性质良好的解，即在无穷远处衰减至零的解；这些能量值以 $n = 0, 1, 2, \ldots$ 标记，彼此等距排列。而且不同的 n 所对应的解是正交（orthogonal）的：它们乘积的积分为零，就像傅里叶级数（Fourier series）中的 $\sin mx$ 与 $\sin nx$ 一样。量子谐振子是研究分子、固体和光的物理学的起点。

动笔之前

想一想 为甚么像 $e^{-\lambda x^2}$ 这样的函数会是一个合理的猜测？当 $x \to \pm\infty$ 时，$e^{-\lambda x^2}$ 及其导数会怎样？

当 $|x|$ 很大时，方程主要由 $y'' \approx \omega^2 x^2 y$ 支配，而高斯函数（Gaussian）$e^{-\lambda x^2}$ 在 x 很大时满足 $y'' \approx 4\lambda^2 x^2 y$，只要 $4\lambda^2 = \omega^2$，两者便吻合。若 $\lambda > 0$，高斯函数及其所有导数衰减至 0 的速度比 x 的任何次幂都要快。因此边界条件（boundary conditions）会自动

成立；真正要下工夫的，是令函数符合方程。

逐步解题

阶段 1　基态

想一想　把 $y_0 = e^{-\lambda x^2}$ 求导两次，代入 $n = 0$ 时的方程，然后比较 x^2 的系数和常数项。

$y_0' = -2\lambda x e^{-\lambda x^2}$，而 $y_0'' = (4\lambda^2 x^2 - 2\lambda)e^{-\lambda x^2}$。$n = 0$ 时的方程要求

$$4\lambda^2 x^2 - 2\lambda - \omega^2 x^2 = -\omega$$

对所有 x 成立。比较系数（match coefficients）：$4\lambda^2 = \omega^2$ 及 $2\lambda = \omega$。取 $\lambda = \tfrac{\omega}{2}$ 时两式同时成立（取 $\lambda > 0$，使得 $y_0 \to 0$）。而且当 $x \to \pm\infty$ 时，y_0、$y_0' \to 0$。 ■

阶段 2　第一激发态

$y_1 = x e^{-\lambda x^2}$ 满足 $y_1' = (1 - 2\lambda x^2)e^{-\lambda x^2}$ 及 $y_1'' = (4\lambda^2 x^3 - 6\lambda x)e^{-\lambda x^2}$。$n = 1$ 时的方程要求

$$4\lambda^2 x^3 - 6\lambda x - \omega^2 x^3 = -3\omega x,$$

而同样取 $\lambda = \tfrac{\omega}{2}$ 时此式成立。y_1 与 y_1' 同样在 $\pm\infty$ 处衰减至零。 ■

阶段 3　一个恒等式

想一想　用积法则（product rule）对 $y_m y_n' - y_n y_m'$ 求导。四项之中有两项互相抵销。然后利用微分方程代换 y_n'' 和 y_m''。

$\dfrac{\mathrm{d}}{\mathrm{d}x}(y_m y_n' - y_n y_m') = y_m' y_n' + y_m y_n'' - y_n' y_m' - y_n y_m'' = y_m y_n'' - y_n y_m''$。由方程可得 $y_n'' = (\omega^2 x^2 - (2n+1)\omega) y_n$，对 m 亦有类似的式子。所以

$$y_m y_n'' - y_n y_m'' = y_m y_n[\omega^2 x^2 - (2n+1)\omega - \omega^2 x^2 + (2m+1)\omega] = 2(m-n)\omega y_m y_n.$$ ■

阶段 4　正交性

停一停，试一试　把恒等式由 $-\infty$ 积分至 ∞。左方是甚么？

左方积分后得 $\big[y_m y_n' - y_n y_m'\big]_{-\infty}^{\infty} = 0 - 0 = 0$，因为每个 y 和 y' 在两端都趋向 0。所以

$2(m-n)\omega\int_{-\infty}^{\infty} y_m y_n\,\mathrm{d}x = 0$，而由于 $m \neq n$ 及 $\omega > 0$，

$$\int_{-\infty}^{\infty} y_m(x)y_n(x)\,\mathrm{d}x = 0. \qquad \blacksquare$$

对于 y_0 和 y_1，这一点很容易直接看出：$y_0 y_1 = xe^{-\omega x^2}$ 是奇函数（odd function）。但这个恒等式毋须知道函数本身是甚么，便一举证明了每一对解都正交。

回顾

- **比较系数。**代入猜测的形式，并要求方程对所有 x 成立，便确定了 λ。
- **朗斯基行列式（Wronskian）的技巧。**$y_m y_n' - y_n y_m'$ 这个组合求导后，其中的二阶导数可以用方程代换，余下的式子十分简单。
- **边界条件消去边界项。**解在无穷远处衰减，令左方积分后为零，从而得出正交性。

背后的数学

薛定谔方程

1926 年，薛定谔（Erwin Schrödinger）提出，量子粒子可用*波函数*（wave function）$\psi(x)$ 来描述；若粒子的能量为 E，并处于势能 $V(x)$ 之中，则波函数满足

$$-\frac{\hbar^2}{2m}\psi'' + V(x)\psi = E\psi.$$

对于弹簧，$V(x) = \frac{1}{2}m\Omega^2 x^2$，把 x 重新标度后，这正是题目中的方程，其中 $E = \left(n+\frac{1}{2}\right)\hbar\Omega$。物理上要求 ψ 在无穷远处衰减至零（这样粒子才确实身处某处），而只有这些能量才能满足这个要求：这就是*量子化*（quantisation），是由数学自然推导出来的结果。留意最低的能量并不是零，而是 $\frac{1}{2}\hbar\Omega$：量子谐振子永远不可能完全静止。这种*零点能*（zero-point energy）真实存在，而且可以量度；正因如此，液态氦在常压下即使到了绝对零度也不会凝固。

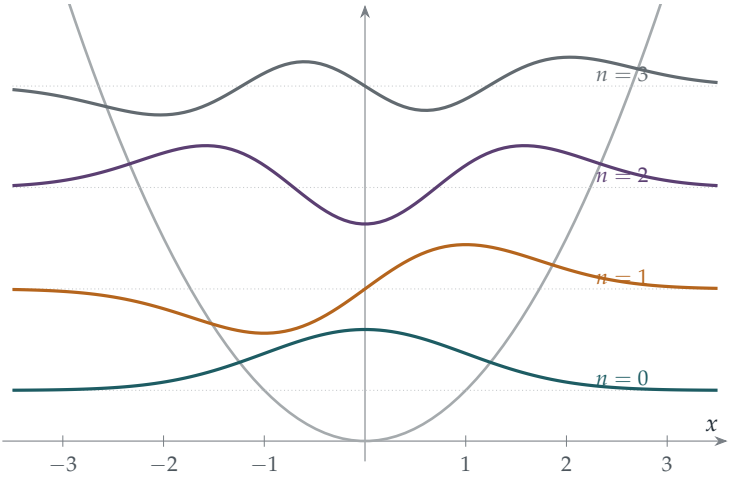


图 8.6.1　量子谐振子（取 $\omega = 1$）。抛物线是势能。每个波函数 y_n 都画在其能级 $n + \frac{1}{2}$ 的高度上；各能级等距排列，而最低的能级高于零。函数 y_n 穿过零 n 次，而且其中任意两个都正交。

埃尔米特函数与施图姆－刘维尔理论

通解为 $y_n = H_n(\sqrt{\omega}x)e^{-\omega x^2/2}$，其中 H_n 是一个 n 次多项式，称为*埃尔米特多项式*（Hermite polynomial）（图 8.6.1）。正交性并非这条方程独有。只要一条方程可以写成 $(py')' + qy = -\lambda wy$ 并配上适当的边界条件，属于不同 λ 值的解便互相正交，所用的论证与第 4 阶段完全相同。这就是*施图姆－刘维尔理论*（Sturm–Liouville theory，1830 年代），它涵盖了傅里叶级数中的正弦函数（Section 3.7）、鼓面的振动、静电学中的勒让德多项式（Legendre polynomials），以及氢原子的波函数。Section 4.7 的特征值问题（eigenvalue problem）是另一个例子：在那里求得的临界值就是特征值 λ，而那些正交函数就是系统固有的“模态”（modes）。

思维技巧

由恒等式得出正交性

要证明相关方程的两个解正交，可找出一个恰当导数（exact derivative），使它的积分因边界条件而为零，而它本身又等于两个解的乘积的某个倍数。

这比直接计算积分有力得多：它一次过适用于每一对解，即使你无法把解写出来也行。

延伸练习

1 **下一个态。**证明 $y_2 = (4\lambda x^2 - 1)e^{-\lambda x^2}$（其中 $\lambda = \frac{\omega}{2}$）满足 $n = 2$ 时的方程。利用 $\int_{-\infty}^{\infty} e^{-\omega x^2}\,\mathrm{d}x = \sqrt{\pi/\omega}$ 及 $\int_{-\infty}^{\infty} x^2 e^{-\omega x^2}\,\mathrm{d}x = \frac{1}{2\omega}\sqrt{\pi/\omega}$，直接验证 $\int_{-\infty}^{\infty} y_0 y_2\,\mathrm{d}x = 0$。

2 **没有其他能量。**把 n 换成一般的数 ν，并尝试 $y = e^{-\lambda x^2}$。证明只有当 $\nu = 0$ 时它才是解。（事实上，只有当 ν 是非负整数时，才存在衰减的解。）

3 **正弦波。**由 $y'' = -n^2 y$ 及 $y(0) = y(\pi) = 0$ 出发，用同样的方法证明当 $m \neq n$ 时 $\int_0^{\pi} \sin mx \sin nx\,\mathrm{d}x = 0$。

4 **阶梯。**证明若 y_n 是方程的解，则 $y_{n+1} = \omega x y_n - y_n'$ 是把 n 换成 $n+1$ 后的方程的解。利用这个结果，由 y_0 生成 y_1 和 y_2。（这就是量子力学中的“升算符”（raising operator）。）

延伸阅读

- **David Griffiths, *Introduction to Quantum Mechanics***（Cambridge University Press）。大学本科物理的标准教科书；第 2 章讨论谐振子。
- **费曼（Richard Feynman）, *QED: The Strange Theory of Light and Matter***（Princeton University Press）。由量子力学的一位大师撰写的科普读物。
- **Mary Boas, *Mathematical Methods in the Physical Sciences***（Wiley）。为数学和物理学生讲解埃尔米特多项式和施图姆－刘维尔理论。

8.7 可以解出的混沌

题目

STEP I 2014，第 6 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/14-s1-q6`

引子

规则 $u_{n+1} = ku_n(1-u_n)$ 称为*逻辑斯谛映射*（logistic map），是描述资源有限的种群的最简单模型：u_n 是第 n 年的种群数量占环境所能承载的最大数量的比例，k 是繁殖率，而因子 $1-u_n$ 则代表过度拥挤的影响。1976 年，生物学家罗伯特·梅（Robert May）在《自然》（*Nature*）上发表了一篇短文，题为“Simple mathematical models with very complicated dynamics”（动力学行为非常复杂的简单数学模型）。他指出，当 k 逐渐增大至 4 时，这条只有一行的规则会由趋于稳定，变为来回振荡，再变为看似完全随机。它成为了混沌（chaos）理论的奠基例子之一。

混沌系统通常无法求解：u_n 没有公式，要求出 u_{100}，唯一的方法就是把它之前的各项逐一算出来。本题找到的正是那个例外。在 $k=4$ 这个最混沌的取值上，公式确实存在，而且它恰好解释了这个数列为何无法预测。

动笔之前

想一想 你以前在哪里见过 $4s(1-s)$ 这个式子，其中 $s=\sin^2\theta$？哪一条恒等式把 $\sin^2\theta$、$\cos^2\theta$ 和 $\sin 2\theta$ 联系起来？

若 $s=\sin^2\theta$，则 $1-s=\cos^2\theta$，所以 $4s(1-s)=4\sin^2\theta\cos^2\theta=(2\sin\theta\cos\theta)^2=\sin^2 2\theta$。映射 $s\mapsto 4s(1-s)$ 其实就是乔装了的倍角公式（double-angle formula）。

逐步解题

阶段 1　猜出规律

设 $u_0 = \sin^2\theta$，由上面的恒等式可得

$$u_1 = 4\sin^2\theta\cos^2\theta = \sin^2 2\theta, \qquad u_2 = 4\sin^2 2\theta\cos^2 2\theta = \sin^2 4\theta.$$

> **想一想**　每一步之后，角度有甚么变化？你对 u_n 有甚么猜想？

角度每次都加倍，所以我们猜想（conjecture）$u_n = \sin^2(2^n\theta)$。

阶段 2　证明猜想

> **想一想**　一个数列由“从一项到下一项”的规则定义，而我们要证明的命题涉及数列的每一项。哪一种证明方法最合适？

归纳法（induction）。命题在 $n = 0$ 时成立，因为 $u_0 = \sin^2\theta = \sin^2(2^0\theta)$。假设对某个 n 有 $u_n = \sin^2(2^n\theta)$，则

$$u_{n+1} = 4\sin^2(2^n\theta)(1 - \sin^2(2^n\theta)) = \left(2\sin(2^n\theta)\cos(2^n\theta)\right)^2 = \sin^2(2^{n+1}\theta),$$

这正是 $n+1$ 时的命题。所以对所有 $n \geq 0$，都有 $u_n = \sin^2(2^n\theta)$。 ■

留意 $[0,1]$ 内的每一个初始值 u 都可以写成 $\sin^2\theta$（取 $\theta = \arcsin\sqrt{u}$），所以对于 $[0,1]$ 内的任何初始值，这个公式都解出了 $k = 4$ 的逻辑斯谛映射。

阶段 3　变量代换

> **想一想**　把 $v_n = \alpha u_n + \beta$ 代入 v_{n+1} 的规则。你需要结果是 $u_{n+1} = 4u_n - 4u_n^2$。这会给出多少个条件？你又有多少个未知数？

代入后得

$$\alpha u_{n+1} + \beta = -p(\alpha u_n + \beta)^2 + q(\alpha u_n + \beta) + r,$$

所以

$$u_{n+1} = -p\alpha\, u_n^2 + (q - 2p\beta)\, u_n + \frac{-p\beta^2 + q\beta + r - \beta}{\alpha}.$$

要使它等于 $4u_n - 4u_n^2$，需要满足三个条件：

$$p\alpha = 4, \qquad q - 2p\beta = 4, \qquad -p\beta^2 + (q-1)\beta + r = 0.$$

头两个条件确定了 $\alpha = 4/p$ 和 $\beta = (q-4)/(2p)$。这时已经没有未知数剩下，所以第三条方程是对 p、q 和 r 的一个真正的条件。代入 β 并乘以 $4p$，得

$$-(q-4)^2 + 2(q-1)(q-4) + 4pr = 0,$$

展开后得 $-(q^2-8q+16) + (2q^2-10q+8) + 4pr = 0$，即 $q^2 - 2q - 8 + 4pr = 0$，亦即

$$4pr = 8 + 2q - q^2. \qquad \blacksquare$$

> **停一停，试一试** 三条方程，却只有两个未知数：这就是出现条件的原因。你能否在未计算之前，就看出为甚么一般的二次映射不应该总是与 $(*)$ 等价？（想想线性的变量代换能改变甚么，又不能改变甚么。）

阶段 4 例子

对于 $v_{n+1} = -v_n^2 + 2v_n + 2$，我们有 $p = 1$、$q = 2$、$r = 2$，而 $4pr = 8 = 8 + 4 - 4$，所以条件成立。于是 $\alpha = 4$ 及 $\beta = -1$，所以 $v_n = 4u_n - 1$，其中 u_n 满足 $(*)$。由于 $v_0 = 1$，故 $u_0 = \frac{1}{2} = \sin^2\frac{\pi}{4}$，所以

$$u_n = \sin^2\!\Big(\frac{2^n\pi}{4}\Big), \qquad v_n = 4\sin^2\!\Big(\frac{2^n\pi}{4}\Big) - 1 = 1 - 2\cos\!\Big(\frac{2^n\pi}{2}\Big).$$

由此得 $v_0 = 1$、$v_1 = 3$，而对所有 $n \geq 2$ 都有 $v_n = -1$。（直接验算：$-1+2+2 = 3$，然后 $-9+6+2 = -1$，再然后 $-1-2+2 = -1$。） $\blacksquare$

这个初始值是特意选定的，好让角度落在 π 的倍数上，数列从此不再变化。这样一来，有趣的部分就被掩盖了，而那正是接下来要谈的。

回顾

- **识破乔装。**一旦写成 $s = \sin^2\theta$，映射 $s \mapsto 4s(1-s)$ 就变成了角度加倍。
- **先猜想，再用归纳法证明。**头两项提示了规律，归纳法则证明了它。
- **数一数方程和未知数。**线性代换有两个自由参数，二次式有三个系数，所以必定剩下一个条件。

背后的数学

数列为何无法预测

公式 $u_n = \sin^2(2^n\theta)$ 看来已是故事的结局：任何一项都可以直接算出。但请细看它说了甚么。由于 $\sin^2$ 的周期（period）是 π，所以只有 $2^n\theta$ 模（modulo）π 的值才有影响。设

$\theta = \pi x$，其中 $0 \le x < 1$，并把 x 写成二进制（binary）：

$$x = 0.b_1 b_2 b_3 \ldots \quad (\text{底为}2).$$

乘以 2 会把二进制小数点向右移一位，而把结果取模 1，就是丢掉越过小数点的那个数位。所以 u_n 只取决于 x 由 b_{n+1} 起的二进制数位。逻辑斯谛映射每走一步，就多读取初始值的一个数位，同时丢掉最前面的一个。

这就是*对初始条件的敏感依赖*（sensitive dependence on initial conditions），也是混沌的决定性特征。两个初始值若在首 20 个二进制位上相同（约六个小数位），它们产生的数列一般会在大约 20 步之内吻合，其后便毫无关系（图 8.7.1）。要预测 u_{100}，你需要知道 u_0 准确至 100 个二进制位。初始角度的任何误差，无论多小，每一步都会加倍。（种群数量 u 本身的误差并非每次都以完全相同的倍数增长，因为这个映射对 $[0,1]$ 某些部分的拉伸比其他部分大，但平均而言，它的增长率是一样的。）公式帮不上忙：它只是令这个机制显现出来。

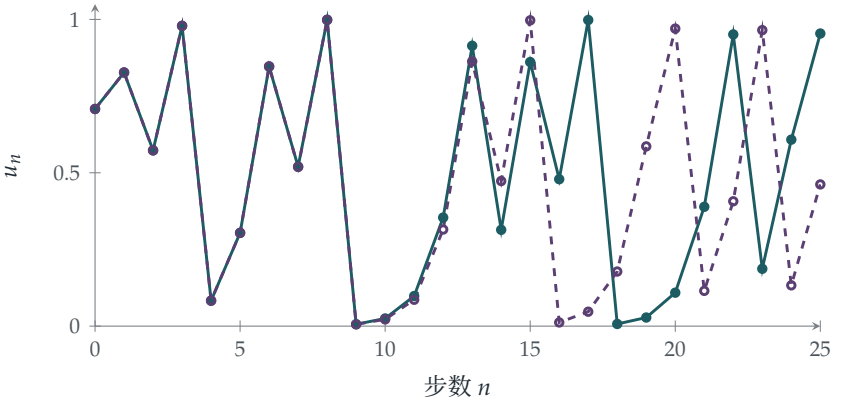


图 8.7.1　$k = 4$ 的逻辑斯谛映射的两个数列，分别由 $\sin^2(1)$（蓝绿色）和 $\sin^2(1.00001)$（紫红色虚线）开始。在大约头十步内两者无法分辨，到第 12 步已有明显差别，其后便毫无关系：到第 12 步，角度之差 10^{-5} 已经加倍了十二次，变成约 $2^{12} \times 10^{-5} \approx 0.04$。

周期点无处不在

这个公式还能找出映射的每一个循环（cycle）。数列在 N 步之后回到初始值，当且仅当 $\sin^2(2^N\theta) = \sin^2\theta$，即 $2^N\theta \equiv \pm\theta \pmod{\pi}$。所以 $\theta = m\pi/(2^N - 1)$ 或 $\theta = m\pi/(2^N + 1)$ 给出一个周期整除 N 的点。例如 $\theta = \pi/5$ 满足 $4\theta \equiv -\theta$，所以 $\sin^2\frac{\pi}{5} \to \sin^2\frac{2\pi}{5} \to \sin^2\frac{4\pi}{5} = \sin^2\frac{\pi}{5}$ 是一个 2-循环。每一个周期都有对应的周期点（periodic point），而且它们在 $[0,1]$ 中稠密（dense）：在每一个区间内，无论多小，都有一个初始值使数列重复出现。然而每一个周期点都是不稳定的，因为任何误差都会加倍。稠密的周期点，却把附近的一切都推开，这是混沌的第二个标志。

通往混沌之路

对于较小的 k，并没有公式可用，而数列的行为会随 k 剧烈变化（图 8.7.2）。当 $1 < k < 3$ 时，每一个由 $(0,1)$ 内出发的数列都会趋于不动点（fixed point）$1 - 1/k$。在 $k = 3$ 时，不动点变得不稳定，由一个 2-循环取而代之（这是 STEP II 2002 Q5 的主题；见延伸练习）。在 $k = 1 + \sqrt{6} \approx 3.449$ 时，2-循环让位予 4-循环，然后是 8-循环，周期加倍来得愈来愈快，直至 $k \approx 3.5699$ 时，循环的长度变成无限，混沌由此开始。在混沌区域之内还有一些有序的窗口：最明显的一个在 $k = 3.83$ 附近，其中有一个 3-循环，那是 Section 8.8 的主题。

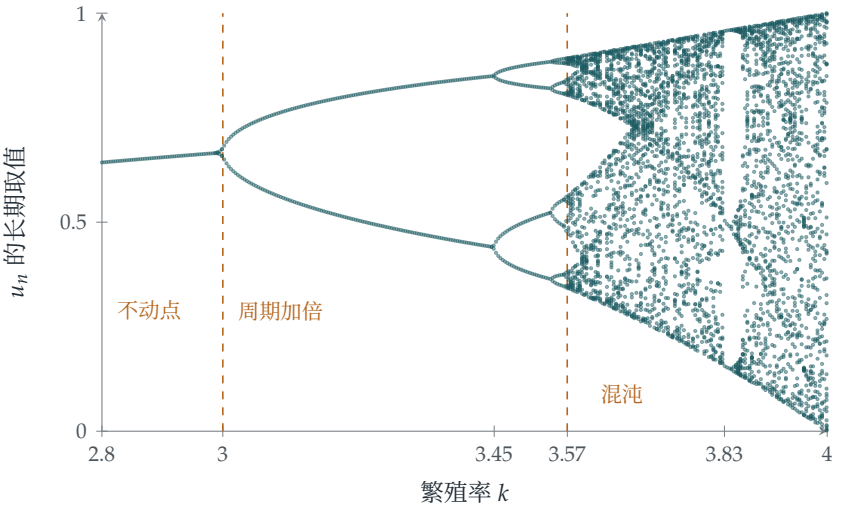


图 8.7.2 逻辑斯谛映射的分岔图（bifurcation diagram）。对每一个 k，由 $u_0 = 0.31416$ 开始把映射迭代 300 次，然后画出其后的 40 个值。单一个不动点在 $k = 3$ 时分裂成一个 2-循环，然后是 4、8、...，直至在 $k = 3.57$ 附近进入混沌。$k = 3.83$ 附近那道浅色的垂直空隙，就是含有稳定 3-循环的窗口。在 $k = 4$ 时，这些值填满了整个 $[0,1]$。

费根鲍姆常数

1975 年，费根鲍姆（Mitchell Feigenbaum）在洛斯阿拉莫斯（Los Alamos）工作时，用一部可编程的袖珍计算机，量度了周期加倍时的各个 k 值：$k_1 = 3$、$k_2 \approx 3.4495$、$k_3 \approx 3.5441$、$k_4 \approx 3.5644$、...。他留意到这些间距以一个固定的倍数缩小，

$$\frac{k_n - k_{n-1}}{k_{n+1} - k_n} \to 4.6692\ldots,$$

然后更令人惊讶地发现，同一个数 4.6692... 也出现在 $x \mapsto k\sin(\pi x)$ 之中，以及其他所有只有单一个峰、而峰顶像抛物线般圆滑（即二次极大值，quadratic maximum）的光

滑映射之中。这个常数对这一整类映射是*普适*（universal）的；至于峰顶较平坦的映射，例如形状像 $-x^4$ 的那种，则各有自己的普适常数。其后，人们在滴水的水龙头、受热的流体和电子电路的实验中都量度到这个常数。要解释个中原因，需要借用相变（phase transition）物理学的想法，而严格的证明要到 1982 年才由 Oscar Lanford 借助电脑完成。

同一招，用三次

第 (ii) 部分的代换是*共轭*（conjugacy）的一个例子：若一个变量代换能把一个映射变成另一个，这两个映射便称为共轭，而它们的动力学行为完全相同。本题证明了每一个满足该条件的二次映射，都与 $k = 4$ 的逻辑斯谛映射共轭，而后者又通过 $u = \sin^2\theta$，成为乔装了的角度加倍。（严格来说，最后这一环是*半共轭*（semiconjugacy）：θ 和 $\pi - \theta$ 给出相同的 u，所以这个变量代换并非一对一（one-to-one）。逻辑斯谛映射真正共轭的对象是*帐篷映射*（tent map），即当 $t \le \frac{1}{2}$ 时 $t \mapsto 2t$、当 $t \ge \frac{1}{2}$ 时 $t \mapsto 2 - 2t$，所用的代换是 $u = \sin^2\frac{\pi t}{2}$，它在 $[0,1]$ 上是一对一的。）在 Section 2.8 中，你遇过 $x \mapsto x^2 - 2$，它由 $x = 2\cos\theta$ 解出，角度每一步加倍；它同样满足该条件（不妨试试）。所以这些全都是同一个映射的不同乔装。在 Section 8.8 中，这个家族的第三位成员将会登场：一个由直线构成的映射，即*帐篷映射*。

思维技巧

变量代换，看清结构

一个好的代换可以把复杂的规则变成简单的规则。设法找出一个变量代换，把问题变成你已经明白的问题。

在这里，$u = \sin^2\theta$ 把一个二次映射变成了角度加倍，而一个线性代换则把一整族二次式归结为一个。在大学里，共轭是一件核心工具：在动力系统（dynamical systems）中，在线性代数中（把矩阵对角化，就是选取一组令它变得简单的坐标），以及在群论中，都是如此。

延伸练习

1 **逻辑斯谛映射的循环。**试做 STEP II 2002，第 5 题（`step.maths.org/questions/02-s2-q5`）。然后证明只有当 $k > 3$ 时才存在 2-循环。

2 **平方减二。**证明 $v_{n+1} = v_n^2 - 2$ 满足第 (ii) 部分的条件（把它写成 $-pv_n^2 + qv_n + r$，其中 $p = -1$）。求 α 和 β，并验证你的答案与 $v_n = 2\cos(2^n\phi)$ 一致。

3 **一个 3-循环。**利用公式 $u_n = \sin^2(2^n\theta)$，求 $u_{n+1} = 4u_n(1-u_n)$ 的一个 3-循环。（提示：试试 $\theta = \pi/7$，或 $\theta = \pi/9$。）周期恰好为 3 的点有多少个?

4 **丢失数位。**电脑储存数字时, 大约准确至 53 个二进制位。如果你在电脑上迭代 $u_{n+1} = 4u_n(1-u_n)$，大约多少步之后，计算出来的值便与真正的数列毫无关系? 到那时，计算出来的数列是否就没有用处?

延伸阅读

- **Robert May，“Simple mathematical models with very complicated dynamics”**，*Nature* 261 (1976)。经典之作，至今仍然十分易读。
- **James Gleick，*Chaos: Making a New Science***（Penguin）。讲述梅、费根鲍姆以及混沌被发现的故事。
- **Steven Strogatz，*Nonlinear Dynamics and Chaos***（CRC Press），第 10 章。一维映射、逻辑斯谛映射和费根鲍姆常数，程度相当于大学一年级。
- **Ian Stewart，*Does God Play Dice?***（Penguin）。一本介绍混沌的科普读物，其中有一章对逻辑斯谛映射讲解得很好。

8.8 周期三蕴含混沌

题目

STEP III 2025，第 2 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2025 年 STEP 试卷（zip 文件）内的 STEP III 试卷：

`www.ocr.org.uk/Images/750486-2025-past-papers.zip`

引子

1975 年，李天岩（Tien-Yien Li）与约克（James Yorke）在《美国数学月刊》（*American Mathematical Monthly*）发表了一篇论文，题目是数学史上最著名的标题之一：“Period three implies chaos”（周期三蕴涵混沌）。他们证明了：如果一个由区间映到自身的连续函数（continuous function）有一个周期（period）为 3 的点（即一个恰好经过三步后返回原处的起始值），那么它就有*任何*周期的点：1、2、4、5、17、1000，无一例外。这篇论文亦把“混沌”（chaos）一词带进了数学。

这道 2025 年的题目，正是让你亲手走一遍这个定理。函数 $\mathrm{f}(x) = 7 - 2|x|$ 由两条直线组成，所以一切都可以精确计算。你会找出它的 2-循环（cycle）和 3-循环，并在过程中看到为甚么其中一种循环存在，就迫使其余的循环也必然存在。

动笔之前

想一想 $y = 7 - 2|x|$ 的图像是甚么样子的？它在哪里与直线 $y = x$ 相交？它把区间 $[-7, 7]$ 送到哪里？

图像是一个倒转的 V 形，顶点在 $(0, 7)$，斜率（gradient）为 ± 2。它与 $y = x$ 相交于 $x = 7 - 2x$（其中 $x \geq 0$）之处，得 $x = \frac{7}{3}$；以及 $x = 7 + 2x$（其中 $x < 0$）之处，得 $x = -7$。这些就是*不动点*（fixed point）：从这里出发，数列永远不会移动。此外 $\mathrm{f}(-7) = \mathrm{f}(7) = -7$ 而 $\mathrm{f}(0) = 7$，所以 f 把 $[-7, 7]$ 映满自身：先把它拉长至两倍，再对折起来。

逐步解题

阶段 1　$\mathrm{f}\circ\mathrm{f}$ 的图像

想一想　当 x 由 -7 走到 7 时，$\mathrm{f}(x)$ 由 -7 升至 7，再降回 -7。那么 $\mathrm{f}(\mathrm{f}(x))$ 会怎样变化？

当 x 由 -7 走到 0 时，$\mathrm{f}(x)$ 走遍整个 $[-7,7]$，所以 $\mathrm{f}(\mathrm{f}(x))$ 描出一个完整的 V 形：由 -7 升至 7，然后再降下来。当 x 由 0 走到 7 时，情况相同。因此在 $[-7,7]$ 上，$\mathrm{f}\circ\mathrm{f}$ 的图像是一个倒转的 W 形：四段斜率为 ± 4 的直线段，每段都在高度 -7 与 7 之间来回，顶点在 $x=\pm\frac{7}{2}$（该处 $\mathrm{f}(x)=0$），并在 $x=0$ 处下降至 -7（图 8.8.1）。在 $[-7,7]$ 之外，两个图像都急速下降。

阶段 2　解 $\mathrm{f}(\mathrm{f}(x))=x$

想一想　四段直线段每一段都与对角线 $y=x$ 恰好相交一次。其中两个交点你早已知道，为甚么？怎样才能不必展开多重绝对值而找出另外两个？

不动点 $\frac{7}{3}$ 和 -7 当然满足 $\mathrm{f}(\mathrm{f}(x))=x$。任何其他的解 p 都属于一个 2-循环 $p\to q\to p$，其中 $q=\mathrm{f}(p)\neq p$。试试各种正负号的组合。如果 p 和 q 都是正数，则 $q=7-2p$ 及 $p=7-2q$，这迫使 $p=q=\frac{7}{3}$；如果两者都是负数，同样的论证给出 -7。所以必定一正一负，设 $p>0>q$：

$$q=7-2p,\qquad p=7+2q.$$

于是 $p=7+14-4p$，所以 $p=\frac{21}{5}$ 及 $q=-\frac{7}{5}$，而这确实满足 $q<0$。

在 $[-7,7]$ 之外并没有解：当 $x>7$ 时，有 $\mathrm{f}(\mathrm{f}(x))=21-4x<x$；当 $x<-7$ 时，有 $\mathrm{f}(\mathrm{f}(x))=21+4x<x$。所以四个解是

$$x=-7,\quad -\frac{7}{5},\quad \frac{7}{3},\quad \frac{21}{5}.\qquad\blacksquare$$

数列的周期为 2，当且仅当 $a=\frac{21}{5}$ 或 $a=-\frac{7}{5}$。（在不动点处，数列是常数数列，周期为 1。）

阶段 3　迟了才进入循环

取 $a=\frac{28}{5}$：$u_1=7-\frac{56}{5}=-\frac{21}{5}$，$u_2=7-\frac{42}{5}=-\frac{7}{5}$，$u_3=7-\frac{14}{5}=\frac{21}{5}$，$u_4=-\frac{7}{5}$，如此类推。由 u_2 开始，数列就是那个 2-循环，但 $u_0=\frac{28}{5}$ 和 $u_1=-\frac{21}{5}$ 并不在循环之内。$\blacksquare$

想一想　你自己可以怎样找到 $\frac{28}{5}$？试从后倒推：f 把哪些数送到 $-\frac{7}{5}$？

$f(x) = -\frac{7}{5}$ 即是 $|x| = \frac{21}{5}$，所以 $x = \pm\frac{21}{5}$。其中一个在循环之内；另一个，即 $-\frac{21}{5}$，就是一个入口。然后 $f(x) = -\frac{21}{5}$ 即是 $|x| = \frac{28}{5}$。由于 f 是二对一的，$(-7,7)$ 内每一点都有两个原像（preimage），所以每个循环都有一整棵由起始值组成的树，最终汇入这个循环。

阶段 4 三循环

> **停一停，试一试** 计算由 $a = 1$ 和 $a = -\frac{7}{9}$ 出发的数列。

由 $a = 1$ 出发：$1 \to 5 \to -3 \to 1$。由 $a = -\frac{7}{9}$ 出发：$-\frac{7}{9} \to \frac{49}{9} \to -\frac{35}{9} \to -\frac{7}{9}$。两者都是 3-循环。

> **想一想** $f(f(f(x))) = x$ 可以有多少个解？像第 1 阶段那样，想想它的图像的形状。

重复第 1 阶段的论证，每多施行一次 f，直线段的数目就增加一倍。所以在 $[-7,7]$ 上，$f \circ f \circ f$ 的图像是由八段组成的锯齿形，斜率为 ± 8，每段都走遍由 -7 到 7 的整个高度（图 8.8.2）。每一段都与对角线恰好相交一次，所以在 $[-7,7]$ 内恰好有八个解，而（一如之前）区间之外则没有解。我们已经找到八个：两个不动点，以及两个 3-循环的六个点。所以解是

$$x = -7,\ \frac{7}{3},\ 1,\ 5,\ -3,\ -\frac{7}{9},\ \frac{49}{9},\ -\frac{35}{9}. \qquad \blacksquare$$

阶段 5 迟了才进入 3-循环

由循环 $1 \to 5 \to -3$ 从后倒推。我们想要 u_3 在循环之内，但 u_2 不在。1 的原像满足 $|x| = 3$，所以 $x = \pm 3$；-3 在循环之内，故取 $u_2 = 3$。3 的原像满足 $|x| = 2$；取 $u_1 = 2$。2 的原像满足 $|x| = \frac{5}{2}$；取 $u_0 = \frac{5}{2}$。于是

$$\tfrac{5}{2} \to 2 \to 3 \to 1 \to 5 \to -3 \to 1 \to \cdots,$$

而 $\frac{5}{2}$、2、3 全都不在循环之内。所以 $a = \frac{5}{2}$ 符合要求。 $\blacksquare$

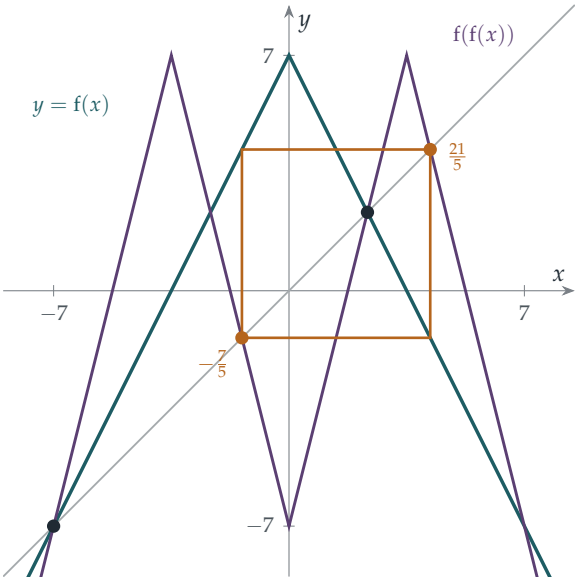


图 8.8.1　$y = \mathrm{f}(x)$（蓝绿色）、$y = \mathrm{f}(\mathrm{f}(x))$（紫红色）和 $y = x$（灰色）的图像。紫红色的图像与对角线相交四次：在不动点 -7 和 $\frac{7}{3}$（黑色），以及在 2-循环 $\frac{21}{5} \leftrightarrow -\frac{7}{5}$（琥珀色）。琥珀色的正方形是这个 2-循环的蛛网图（cobweb）：向上走到 f 的图像，横向走到对角线，然后重复。

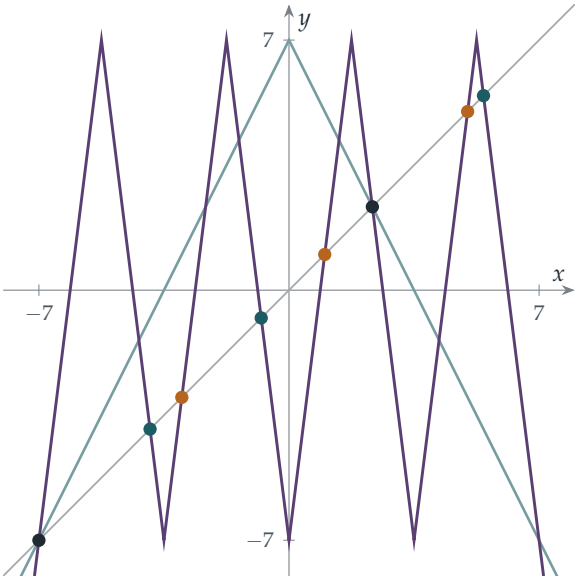


图 8.8.2　$y = \mathrm{f}(\mathrm{f}(\mathrm{f}(x)))$ 的图像（紫红色）在 $[-7, 7]$ 上是由八段组成的锯齿形，斜率为 ± 8，每段与对角线相交一次。八个交点分别是两个不动点（黑色）、3-循环 $1 \to 5 \to -3$（琥珀色）和 3-循环 $-\frac{7}{9} \to \frac{49}{9} \to -\frac{35}{9}$（蓝绿色）。

回顾

- **凭形状计数。**我们没有展开多重绝对值，而是计数图像中直线段的数目。这告诉我们要找多少个解，也就知道甚么时候已经全部找齐。
- **利用循环的结构。**$f(f(x)) = x$ 的解若非不动点，就是成对出现；$f(f(f(x))) = x$ 的解若非不动点，就是三个一组出现。
- **从后倒推。**第 (i)(d) 和 (ii)(c) 部分中那些“迟来者”，都是靠原像构造出来的。

背后的数学

任何长度的循环

第 4 阶段的计数方法适用于任何步数。把 f^n 记作施行 f 共 n 次。在 $[-7,7]$ 上，f^n 的图像有 2^n 段直线段，每段都走遍整个高度，所以 $f^n(x) = x$ 恰好有 2^n 个解。其中有些解的周期较短，是某个整除 n 的 d，但对每个 d 而言，这样的解最多只有 2^d 个，而且

$$2^n > 2^{n/2} + 2^{n/3} + \cdots + 2^1 \quad \text{（对所有满足 } d < n \text{ 的 } n \text{ 的因数求和）}$$

因为所有小于 n 的因数都不超过 $n/2$，而 $2^1 + 2^2 + \cdots + 2^{\lfloor n/2 \rfloor} < 2^{\lfloor n/2 \rfloor + 1} \le 2^n$。所以对每个 n，都存在周期恰好为 n 的点。这个函数有任何长度的循环。

为甚么周期三迫使一切出现

李天岩与约克的定理指出，这并非我们这个函数独有的性质：任何由区间映到自身的连续函数 g，只要有一个 3-循环，就有任何长度的循环。证明所用的不外乎介值定理（intermediate value theorem），而我们的 3-循环正好展示了证明的思路。

> **想一想** 设 $I_1 = [-3,1]$ 及 $I_2 = [1,5]$，即循环 $1 \to 5 \to -3$ 各点之间的两个区间。f 把它们分别送到哪里？

在 I_2 上，f 由 $f(1) = 5$ 递减至 $f(5) = -3$，所以 $f(I_2) = [-3,5] = I_1 \cup I_2$。在 I_1 上，f 由 $f(-3) = 1$ 升至 $f(0) = 7$，再降至 $f(1) = 5$，所以 $f(I_1) = [1,7]$，它包含 I_2。留意我们只用到循环的三个点：任何具有 3-循环 $a \to b \to c$、$a < b < c$ 的连续函数，都以同样的方式覆盖它的区间。我们可以用一幅图来概括哪个区间覆盖哪个区间（图 8.8.3）：

$$I_1 \to I_2, \qquad I_2 \to I_1, \qquad I_2 \to I_2.$$

定理于是由以下两个事实推出，两者都是介值定理的推论。

- 如果 $g(I) \supseteq I$，则 g 在 I 内有一个不动点。（对 $g(x) - x$ 应用介值定理。）

▶ 如果 $g(I) \supseteq J$，则存在一个较小的闭区间 $I' \subseteq I$，使得 $g(I') = J$ 恰好成立。

要找一个周期为 n 的点，可沿着图中的回路 $I_2 \to I_2 \to \cdots \to I_2 \to I_1 \to I_2$ 走：$n-1$ 步留在 I_2 之内，另有一步走出去到 I_1 再回来。反复运用第二个事实，我们可以把 I_2 收窄成一个由起始点组成的区间，这些起始点的轨道（orbit）都依循这条路线，而 g^n 把这个区间映满整个 I_2。第一个事实给出一点 x，满足 $g^n(x) = x$，并且只在第 $n-1$ 步到访 I_1。它的周期不可能小于 n，否则对 I_1 的到访便会在更早的时候重复出现。（这里有一个微妙之处，因为 I_1 和 I_2 共有一点 1；验证这不会造成问题，留作习题。）

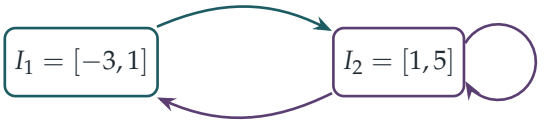


图 8.8.3　3-循环的覆盖图（covering diagram）。由一个区间指向另一个区间的箭嘴，表示前者的像包含后者。图中每一条回路都给出函数的一个循环：回路 $I_2 \to I_2$ 给出一个不动点，$I_1 \to I_2 \to I_1$ 给出一个 2-循环，而先绕 I_2 处的回路 $n-1$ 次再到访 I_1，则给出一个长度为 n 的循环。

沙可夫斯基定理

李天岩与约克并不是最先发现的人。早在 1964 年，乌克兰数学家沙可夫斯基（Oleksandr Sharkovskii）已经证明了一个完备得多的结果，以俄文发表于《乌克兰数学杂志》（*Ukrainian Mathematical Journal*），其后十年间西方对此一无所知。把正整数排列成

$$3,\ 5,\ 7,\ 9,\ \ldots,\ 2\cdot 3,\ 2\cdot 5,\ \ldots,\ 2^2\cdot 3,\ 2^2\cdot 5,\ \ldots,\ \ldots,\ 2^3,\ 2^2,\ 2,\ 1:$$

先是由 3 开始的奇数，然后是奇数的两倍，再然后是奇数的四倍，如此类推，最后以 2 的幂按递减次序作结。沙可夫斯基定理指出：如果区间上的一个连续函数有一个长度为 m 的循环，那么凡是在这个列表中排在 m 之后的长度，它都有相应的循环。由于 3 排在最前，周期 3 便蕴涵所有周期。而且这个定理是最佳的：对每个 m，都存在一个函数，它有长度为 m 的循环，却没有列表中排在更前的长度的循环。列表末端 2 的幂，正正就是 Section 8.7 中逻辑斯蒂映射（logistic map）的倍周期级联（period-doubling cascade），而 3-循环在那里出现于 $k = 1 + \sqrt{8}$。

帐篷映射与逻辑斯蒂映射

作代换 $x = 14y - 7$，它把 $[0,1]$ 映满 $[-7,7]$。简单计算可见，f 变成了帐篷映射（tent map）

$$T(y) = \begin{cases} 2y & 0 \le y \le \frac{1}{2}, \\ 2 - 2y & \frac{1}{2} \le y \le 1, \end{cases}$$

而再作代换 $u = \sin^2(\pi y/2)$，就把帐篷映射变成逻辑斯蒂映射 $u \mapsto 4u(1-u)$。所以这道题目中的函数，其实就是 Section 8.7 中那个混沌的逻辑斯蒂映射，只是被拉直了。它的 $\mathrm{f}^3(x) = x$ 的八个解，对应于 $\sin^2(8\theta) = \sin^2\theta$ 在 $[0, \frac{\pi}{2}]$ 内的八个解。

思维技巧

以介值定理证明存在性

要证明解存在，不一定要把它找出来。只需证明一个连续函数变号，或者一个区间被映射后盖过自身。

在这道题目中，每一个解都可以算出来。但李天岩与约克的定理适用于我们根本无从计算的函数，而且它只用介值定理和一幅显示哪个区间覆盖哪个区间的图，就证明了无穷多个循环的存在。在大学里，这种论证方式，即不作构造而证明存在，会出现在分析（analysis）（介值定理本身）、拓扑学（topology）（不动点定理）以及动力系统（dynamical systems）的各个角落。

延伸练习

1 **一个 4-循环。**利用覆盖图预测 f 的一个 4-循环应该在哪里（三步在 I_2，一步在 I_1）。然后把它精确地求出来。（提示：这个循环有三个正的点和一个负的点；写下四条线性方程。）

2 **微妙之处。**在“周期 3 蕴涵周期 n”的证明中，解释为甚么当 $n \geq 4$ 时，所找到的点 x 不可能是点 1（或原来那个 3-循环的任何一点）。

3 **有周期 5 而没有周期 3。**草绘一个在 $[1,5]$ 上的连续函数的图像，它满足 $g(1) = 3$、$g(3) = 4$、$g(4) = 2$、$g(2) = 5$、$g(5) = 1$，并以直线连接这些点。就区间 $[1,2]$、$[2,3]$、$[3,4]$、$[4,5]$ 画出它的覆盖图。你能找到一条长度为 3 的回路吗？（沙可夫斯基定理说它没有 3-循环。）

4 **帐篷映射。**验证 $x = 14y - 7$ 把 f 变成帐篷映射。然后证明帐篷映射把以二进制写成的 y 加倍，并且每当首位数字是 1 时就把所有数字反转；再利用这一点找出它的 3-循环。

延伸阅读

► **Tien-Yien Li & James Yorke, “Period three implies chaos”**, *American Mathematical*

Monthly 82 (1975)。篇幅短小，懂得介值定理便可读懂。

- **Robert Devaney, *A First Course in Chaotic Dynamical Systems*** (CRC Press)。为大学本科生而写；涵盖帐篷映射、逻辑斯蒂映射和沙可夫斯基定理，并附完整证明。
- **Keith Burns & Boris Hasselblatt, "The Sharkovsky theorem: a natural direct proof"**, *American Mathematical Monthly* 118 (2011)。一个运用覆盖图的现代证明。

我们学到了甚么

本章开首我们说过，大多数微分方程（differential equation）都无法求解，而真正的功夫在于即使解不出来，也能理解它们。我们提出了三个问题：现在是否决定未来？不求解的话，我们能知道些甚么？长远而言会发生甚么事？以下是本章八道题目所揭示的答案。

1. 现在是否决定未来？

通常是，但并非必然。在 Section 8.1 中，方程 $(y')^2 = 4y$ 经过 x 轴上每一点都有无穷多个解：一个解可以随意在 $y = 0$ 停留任意长的时间，然后才离开。罪魁祸首是平方根：它在零处的斜率为无穷大，破坏了皮卡–林德勒夫定理（Picard–Lindelöf theorem）中的利普希茨条件（Lipschitz condition）。同一机制也解释了为何一个漏水的水桶无法告诉你它是何时漏空的，而它亦是诺顿圆顶（Norton's dome）背后的原理——那是古典力学中一个关于决定论（determinism）的难题。

即使未来已被决定，它也未必可以预测。Section 8.7 的逻辑斯谛映射（logistic map）有一条精确的公式 $u_n = \sin^2(2^n\theta)$，然而每走一步，起始角 θ 的任何误差都会加倍，而起始值的误差平均而言也增长得同样快。原则上的决定论，并不等于实际上的可预测性。

2. 不求解的话，我们能知道些甚么？

能知道的非常多。

- **能量。**在 Section 8.2 中，把方程乘以 y' 便得出一个不可能增加的量，从而为一条无人能解的方程的解定出了界限。这类量称为*李雅普诺夫函数*（Lyapunov function），是证明稳定性（stability）的主要工具。
- **平衡点与稳定性。**在 Section 8.3 中，增长率的正负号预示了鱼群的命运，并显示捕捞量一旦超过某个临界值，鱼群便会突然崩溃。
- **合适的坐标。**在 Section 8.4 中，极坐标（polar coordinates）把一个耦合方程组化为只含 r 的单一方程，显示每个解都盘旋着趋向一个圆。在 Section 8.5 中，对称性把四只互相追逐的狗简化为一只，再以相除消去时间，便得出牠们路径的形状。
- **由恒等式看出结构。**在 Section 8.6 中，朗斯基行列式（Wronskian）证明了某方程的解互相正交（orthogonal），而无须对其中任何一个解进行积分。

- **计数与覆盖。**在 Section 8.8 中，计数图像的段数便找出了所有循环（cycle），而介值定理（intermediate value theorem）则证明了任意长度的循环都存在。

3. 长远而言会发生甚么事？

本章遇到了各种各样的长期行为。解或会趋于稳定的平衡点（Section 8.3）、趋近一个不断重复的循环（Section 8.4）、在有限时间内盘旋收于一点（Section 8.5），又或者表现出混沌（chaos）行为（Section 8.7）。就逻辑斯谛映射而言，单单一个参数便可令系统经历以上各种情况：先是一个不动点（fixed point），然后是一连串由费根鲍姆（Feigenbaum）普适常数支配的倍周期（period doubling），最后是混沌，其间夹杂着一些有序的窗口，例如 3-循环。而 Section 8.8 则证明了，一旦出现 3-循环，任意长度的循环都必定存在——这是沙可夫斯基（Sharkovskii）整数排序的第一步。

4. 微分方程与动力学的工具箱

方法	你在哪里用过
寻找奇解（singular solution）	非唯一性（8.1）
一个不可能增加的量	定界（8.2）
增长率的正负号	捕捞（8.3）
配合对称性的坐标	极限环（limit cycle）（8.4）；追逐（8.5）
以代换法线性化	伯努利方程（Bernoulli equation）（8.4）
朗斯基行列式	正交性（8.6）
共轭（conjugacy）	逻辑斯谛映射（8.7）
归纳法（induction）	$u_n = \sin^2(2^n\theta)$（8.7）
计数图像的段数	f^n 的循环（8.8）
介值定理	周期三（8.8）

下一步

- 能量与守恒定律会在力学中再次出现，由 Section 10.1 开始。
- 含多个未知函数的线性微分方程可用特征向量（eigenvector）求解，它们支配着 Section 11.2 中的振动系统；至于无法以初等函数求解的非线性单摆，则见于 Section 11.3。
- 在 Section 12.15 中，反复迭代一个映射以找出其不动点，可判定一个姓氏会否消亡；而随机数列的长期行为则是 Section 12.16 的主题。

如果你想在上大学之前再进一步

- **Steven Strogatz, *Nonlinear Dynamics and Chaos***：动力系统（dynamical systems）最好的入门书，应用和图解十分丰富。Strogatz 根据此书讲授的课堂录影可在网上免费观看。
- **James Gleick, *Chaos: Making a New Science***：讲述混沌是如何被发现的故事。
- **George Simmons, *Differential Equations with Applications and Historical Notes***：经典的入门课本，历史材料丰富。
- **Robert Devaney, *A First Course in Chaotic Dynamical Systems***：迭代映射，附有证明，程度相当于大学一年级。

拉普拉斯（Laplace）曾经设想，只要准确知道现在，便可以算出整个未来。本章显示了这个梦想在何处止步。唯一性可以失效；而即使唯一性成立，混沌也可以把最微小的不确定性不断放大，直至无法预测。但本章同时显示，没有公式也能理解很多东西：界限、平衡点、循环及其稳定性，以至混沌本身的结构。由“解方程”转为“理解方程的解”，是大学数学课程要求你作出的最重要的思维转变之一。

延伸练习答案

以下只是简短的答案和提示，并非完整解答：只列出结果和关键步骤，方便你核对自己的做法。请先认真做过每道练习，然后才翻看。

8.1 当解不唯一

1. 设 $y = v^2$：$2vv' = \alpha v - \beta v^2$，所以 $v' = \frac{1}{2}(\alpha - \beta v)$，而 $y_1 = \frac{\alpha^2}{\beta^2}\left(1 - \mathrm{e}^{-\beta t/2}\right)^2$。设 $y = v^3$：$v' = \frac{1}{3}(\alpha - \beta v)$，而 $y_1 = \frac{\alpha^3}{\beta^3}\left(1 - \mathrm{e}^{-\beta t/3}\right)^3$。常数函数 $y = 0$ 令等式两边都等于零；推导时除以 v，就把这个解丢掉了。

2. $2\sqrt{h} = 2\sqrt{h_0} - kt$，所以 $h = \left(\sqrt{h_0} - \frac{1}{2}kt\right)^2$，水桶在 $T = \frac{2\sqrt{h_0}}{k}$ 时流干。$h \equiv 0$，以及当 $t \le T$ 时 $h = \frac{k^2}{4}(T - t)^2$、当 $t \ge T$ 时 $h = 0$ 的函数，两者都满足方程，而且由 T 起都等于零：看着一个空桶，你无法知道它是何时流干的。

3. $\frac{\mathrm{d}}{\mathrm{d}x}(y\mathrm{e}^{-2x}) = (y' - 2y)\mathrm{e}^{-2x} = 0$，所以 $y = C\mathrm{e}^{2x}$，而 $y(0) = 0$ 迫使 $C = 0$。函数 $2y$ 满足利普希茨（Lipschitz）条件（它的变化至多为 $2|y_1 - y_2|$），而 $2\sqrt{y}$ 在 0 处的斜率为无穷大。

4. 对任意 $b \le 0 \le a$（容许 $b = -\infty$ 或 $a = \infty$）：当 $x < b$ 时 $y = (x - b)^3$，当 $b \le x \le a$ 时 $y = 0$，当 $x > a$ 时 $y = (x - a)^3$。$a = b = 0$ 的情况就是 $y = x^3$。

8.2 不求解而估界

1. $\frac{\mathrm{d}}{\mathrm{d}x}\left((y')^2 + y^2\right) = 2y'(y'' + y) = 0$，所以 $(y')^2 + y^2 = A^2$，因而 $|y| \le A$。精确解是 $A\cos x$。

2. $\frac{\mathrm{d}}{\mathrm{d}t}\left((\theta')^2 - 2\cos\theta\right) = 2\theta'(\theta'' + \sin\theta) = 0$。由静止开始，$(\theta')^2 = 2(\cos\theta - \cos\theta_0) \ge 0$，所以 $\cos\theta \ge \cos\theta_0$；当 $|\theta_0| < \pi$ 时，连续性（continuity）令 θ 留在 $[-|\theta_0|, |\theta_0|]$ 之内，而当 $|\theta_0| = \pi$ 时，摆倒立平衡，并一直停在那里。若 $|\theta_0| > \pi$，例如 $\theta_0 = \frac{3\pi}{2}$，则 $\theta''(0) = -\sin\frac{3\pi}{2} = 1 > 0$，所以 θ 起初递增：摆其实位于角度 $\theta_0 - 2\pi = -\frac{\pi}{2}$，并在由该处量起的 $\pm\frac{\pi}{2}$ 之间摆动。

3. $\frac{\mathrm{d}}{\mathrm{d}t}\left((y')^2 + y^2\right) = 2y'(y'' + y) = -4k(y')^2 \le 0$。它确实趋向零，但单凭单调性（monotonicity）并不足以证明这一点：你还需要其他工具，例如显式解，其指数 $-k \pm \sqrt{k^2 - 1}$ 的实部为负。

4. $(y')^2 - \frac{1}{2}y^4$ 是常数，等于 $-\frac{1}{2}$。由此得 $(y')^2 = \frac{1}{2}(y^4 - 1)$，但这给不出任何界：“位

能” $-\frac{1}{2}y^4$ 没有可以困住 y 的位阱，而事实上 y 会在有限时间内趋向无穷（blow up）。

8.3 收获种群

1. 方程是 $x' = \alpha^2 - (x-K)^2$，可化为 $y' = y(2\alpha - y)$。设 $w = \frac{1}{y}$，则 $w' = 1 - 2\alpha w$，所以 $w = \frac{1}{2\alpha} + Ce^{-2\alpha t}$。若 $y(0) < 0$，则 $w(0) < 0 < \frac{1}{2\alpha}$，所以 w 在有限时间内经过 0，而 $y \to -\infty$；x 在此之前已到达 0。

2. 设 $z = x - K$，则 $z' = -(z^2 + \beta^2)$，所以 $\tan^{-1}\frac{z}{\beta} = \tan^{-1}\frac{K}{\beta} - \beta t$。种群数量在 $t = \frac{2}{\beta}\tan^{-1}\frac{K}{\beta}$ 时降至零（$z = -K$）。

3. $x' = x(2K - h - x)$ 的平衡点（equilibria）是 0 和 $2K - h$。当 $h < 2K$ 时，这是一条环境承载量（carrying capacity）为正的逻辑斯谛方程（logistic equation），每个正数解都趋向 $2K - h$。捕获量随种群数量下降而减少，所以永远不会把种群数量逼至零。

4. $x = K + K\tanh Kt$，而 $\tanh Kt = \frac{1}{2}$ 给出 $t = \frac{1}{K}\tanh^{-1}\frac{1}{2} = \frac{\ln 3}{2K}$。

8.4 盘旋趋向圆

1. 曲线与半径之间的夹角 ψ 满足 $\tan\psi = \frac{r}{\mathrm{d}r/\mathrm{d}\theta} = -1$，在每一点都相同：曲线与每条射线都以 45° 相交。

2. $x\,\mathrm{d}x + y\,\mathrm{d}y = \frac{1}{2}\mathrm{d}(x^2 + y^2) = r\,\mathrm{d}r$。另外 $x\,\mathrm{d}y - y\,\mathrm{d}x = r\cos\theta(\sin\theta\,\mathrm{d}r + r\cos\theta\,\mathrm{d}\theta) - r\sin\theta(\cos\theta\,\mathrm{d}r - r\sin\theta\,\mathrm{d}\theta) = r^2\,\mathrm{d}\theta$。

3. 设 $s = r^2$，则 $\dot{s} = 2s(1-s)$，是一条逻辑斯谛方程，所以对任意 $s(0) > 0$ 都有 $s \to 1$。写成笛卡儿（Cartesian）形式：$\dot{x} = x(1 - x^2 - y^2) - y$，$\dot{y} = y(1 - x^2 - y^2) + x$。

4. $\dot{r} = r(r-1)(r-2)$，$\dot{\theta} = 1$。在 0 与 1 之间，$\dot{r} > 0$；在 1 与 2 之间，$\dot{r} < 0$；超过 2 之后，$\dot{r} > 0$。所以 $r = 1$ 是吸引的，而 $r = 2$ 是排斥的。

8.5 格力犬追逐格力犬

1. 由对称性可知，几只狗始终构成一个正方形。A 沿 AB 以速率 v 奔跑，令距离以速率 v 缩短；B 沿 BC 奔跑，而这个方向垂直于 AB，所以在一阶近似下不会改变距离。因此 AB 恰好以速率 v 缩短，由 s 减至 0，需时 $\frac{s}{v}$。

2. $r = \lambda e^{-\sqrt{3}\theta}$。距离以速率 $v - v\cos 120° = \frac{3}{2}v$ 缩短，所以牠们在时间 $\frac{2s}{3v}$ 后相遇，每只各跑了 $\frac{2s}{3}$（这与 $\frac{s}{2\sin^2(\pi/n)}$ 在 $n = 3$ 时的值吻合）。

3. 每转一圈，r 就乘以 $e^{-2\pi} \approx \frac{1}{535}$：转一圈后约为 19 cm。转两圈后约为 0.35 mm，所以转足两圈便已足够。

4. 两只狗笔直地向对方奔跑，在时间 $\frac{s}{2v}$ 后于中点相遇。公式在此失效：$\tan\frac{\pi}{2}$ 为无穷大，正好对应完全没有转弯这个事实。

8.6 量子振子

1. 这只需直接求导（虽然颇为繁琐）；另一个做法是由第 4 题推出。取 $2\lambda=\omega$，则 $\int y_0y_2=\int(2\omega x^2-1)\mathrm{e}^{-\omega x^2}\,\mathrm{d}x=2\omega\cdot\frac{1}{2\omega}\sqrt{\frac{\pi}{\omega}}-\sqrt{\frac{\pi}{\omega}}=0$。

2. $y''=(4\lambda^2x^2-2\lambda)y$，所以方程要求 $4\lambda^2=\omega^2$ 及 $2\lambda=(2\nu+1)\omega$。取 $\lambda=\frac{\omega}{2}>0$（使解衰减），得 $\nu=0$。

3. $\frac{\mathrm{d}}{\mathrm{d}x}(y_my_n'-y_ny_m')=y_my_n''-y_ny_m''=(m^2-n^2)y_my_n$。由 0 积分至 π：因为 y_m 和 y_n 在两端都等于零，左边得 0，而 $m^2\neq n^2$。

4. 记 $c=(2n+1)\omega$，于是 $y_n''=(\omega^2x^2-c)y_n$。把 $z=\omega xy_n-y_n'$ 求导两次，并运用上式（及其导数），得 $z''-\omega^2x^2z=-(c+2\omega)z$，而 $c+2\omega=(2n+3)\omega$。由 $y_0=\mathrm{e}^{-\omega x^2/2}$ 出发：先得 $2\omega x\mathrm{e}^{-\omega x^2/2}$，再得 $2\omega(2\omega x^2-1)\mathrm{e}^{-\omega x^2/2}$，它们分别是 y_1 和 y_2 的倍数。

8.7 可以解出的混沌

1. (i) 用归纳法（induction）：若 $0<x<1$，则 $0<x(1-x)\leq\frac{1}{4}$，所以 $0<kx(1-x)<1$。(ii) $a=ka(1-a)$ 给出 $k=\frac{1}{1-a}$。(iii) 由于 $x_1=ka(1-a)=ab$，条件 $x_2=a$ 即 $kab(1-ab)=a$；代入 $k=\frac{b}{1-a}$ 后，化为 $ab^3-b^2+1-a=0$。根 $b=1$ 对应不动点（fixed point）（$x_1=a$），所以把它提出来：$ab^2+(a-1)b+(a-1)=0$；解出 b，便得 $k=\frac{b}{1-a}$。等价的说法是：2-循环（cycle）的两点是 $k^2x^2-k(k+1)x+(k+1)=0$ 的两个根，其判别式（discriminant）$k^2(k+1)(k-3)$ 只在 $k>3$ 时为正。

2. $p=-1$，$q=0$，$r=-2$：$4pr=8=8+2q-q^2$。比较系数得 $\alpha=\frac{4}{p}=-4$ 及 $\beta=\frac{q-4}{2p}=2$，所以 $v_n=2-4\sin^2(2^n\theta)=2\cos(2^{n+1}\theta)$，即 $2\cos(2^n\phi)$，其中 $\phi=2\theta$。

3. $u_3=u_0$ 要求 $\sin^2 8\theta=\sin^2\theta$，所以 $8\theta=\pm\theta+k\pi$。取 $\theta=\frac{\pi}{7}$：得循环 $\sin^2\frac{\pi}{7}\to\sin^2\frac{2\pi}{7}\to\sin^2\frac{4\pi}{7}\to\sin^2\frac{8\pi}{7}=\sin^2\frac{\pi}{7}$。取 $\theta=\frac{\pi}{9}$：$\sin^2\frac{\pi}{9}$，$\sin^2\frac{2\pi}{9}$，$\sin^2\frac{4\pi}{9}$。映射（map）$u\mapsto u_3$ 有 $2^3=8$ 个不动点；其中两个（0 和 $\frac{3}{4}$）是原映射本身的不动点，余下 6 个点的周期恰好是 3：即两个 3-循环。

4. 每一步都把 θ 加倍，因而损失一个二进制数位的准确度，所以大约 50 步之后，计算所得的数值与真实数值已毫无关系。但这些数值并非没有用处：计算所得的数列始终贴近一条由稍为不同的起点出发的真实轨道（orbit）（这就是影随（shadowing）性质），所以它的统计特性仍然正确。

8.8 周期三蕴含混沌

1. 设 a、b、$c>0$ 而 $d<0$：$b=7-2a$，$c=7-2b$，$d=7-2c$，$a=7+2d$。解得 $17a=49$，所以循环是 $\frac{49}{17}\to\frac{21}{17}\to\frac{77}{17}\to-\frac{35}{17}\to\frac{49}{17}$，其中三点在 $I_2=[1,5]$ 内，一点在 $I_1=[-3,1]$ 内。

2. 所找到的点的行程（itinerary）是 $I_2, I_2, \ldots, I_2, I_1$（其中有 $n-1\geq 3$ 步在 I_2 内）。循环 $1\to 5\to -3$ 不可能依循这个行程：-3 不在 I_2 内，所以 1（它的第二个像是 -3）、5（它的第一个像是 -3）以及 -3 本身，都不可能是该点。

3. 覆盖图(covering diagram)是 $[1,2]\to[3,4],[4,5]$；$[2,3]\to[4,5]$；$[3,4]\to[2,3],[3,4]$；$[4,5]\to[1,2]$。长度为 3 的回路只有 $[3,4]\to[3,4]\to[3,4]$，它给出的是不动点，而不是 3-循环。

4. 设 $x=14y-7$，则 $\frac{f(x)+7}{14}=1-|2y-1|$，即帐篷映射（tent map）。加倍会把二进制数位向左移一位；若移出的数位是 1，结果便是 $2-2y$，即把余下的数位逐一反转。3-循环有 $\frac{2}{7}\to\frac{4}{7}\to\frac{6}{7}$ 和 $\frac{2}{9}\to\frac{4}{9}\to\frac{8}{9}$（即 $x=-3,1,5$ 和 $x=-\frac{35}{9},-\frac{7}{9},\frac{49}{9}$）；以二进制表示，$\frac{2}{7}=0.\overline{010}$ 而 $\frac{2}{9}=0.\overline{001110}$。

章

9

计数与机会

计数是最古老的数学，但要数得巧妙，至今仍是最困难的事情之一。

本章内容

9.1	直线分割平面	STEP II 1990
9.2	小型版的拉姆齐定理	STEP I 1989
9.3	为篱笆桩着色	STEP II 1995
9.4	把数列看成函数	STEP I 2016
9.5	凑零钱	STEP I 1997
9.6	奇怪的骰子	STEP I 1993
9.7	相等的幂和	STEP I 1992
9.8	没有人拿到自己的恤衫	STEP II 2016
9.9	生日问题	STEP II 2007
9.10	总能找赎	STEP I 2011
9.11	集齐一套	STEP II 1994
9.12	随机子集	STEP III 2019
9.13	做了手脚的骰子	STEP I 1999
A	练习答案	第 653 页

计数的艺术

组合数学是拓扑学的贫民窟。

— J. H. C. Whitehead，说于 1950 年代。今天已经没有人抱持这种看法。

人人都会数数。但若要问：用硬币凑出 £1 有多少种方法？安排十个人入座，使得没有人坐在自己的位子上，有多少种方法？用四种颜色为一幅地图着色，又有多少种方法？单纯的计数很快就行不通了。情况多得无法逐一列出，而且各种情况还会互相重叠。组合数学（combinatorics）是关于计数与排列的数学，它是一门艺术：把这类问题整理得井井有条，使我们不必逐一列举也能解决。

正如 Whitehead 那句话所反映的，在很长一段时间里，组合数学只被视为一堆巧妙技巧的集合，而不是一门严肃的学科。这种情况在二十世纪下半叶改变了，推动力一方面来自计算机科学（computer science），因为计数一个算法（algorithm）所需的步骤数目，本身就是一个组合问题；另一方面则来自人们发现它与代数、几何和概率（probability）之间有深刻的联系。今天，它是数学中最活跃的领域之一。2022 年，许埈珥（June Huh）获颁数学界的最高荣誉菲尔兹奖（Fields Medal），部分原因是他运用代数几何（algebraic geometry）的思想，证明了一个 1968 年提出的猜想，这个猜想所涉及的，正是 Section 9.3 中的色多项式（chromatic polynomial）。

计数为何重要

概率最简单的形式就是计数：当所有结果的可能性均等时，一个事件的概率等于有利结果的数目除以结果的总数。因此，每一条关于纸牌、骰子和彩票的问题，都是计数问题；本章后半部分取材自概率试卷，正是这个缘故。但计数的重要性远远不止于游戏。计算机科学家要计数一个算法所需的运算次数，以及它可能出错的方式。化学家要计数分子可能具有的结构。生物学家要计数一条 DNA 链可以折叠的方式。密码学家要计数攻击者必须尝试的密钥数目，以及（如 Section 9.9 所示）要有多少个，才会出现两个相同的情况。

你在大学会遇到甚么

大多数大学都会在一年级或二年级开设组合数学或离散数学（discrete mathematics）的课程，其主题正是本章的主题。你会学到基本的计数原理（乘法原理（product rule）、双射（bijection）、容斥原理（inclusion–exclusion）、鸽巢原理（pigeonhole principle）），以及怎

样建立和求解递推关系（recurrence relation）。你会遇到生成函数（generating function），它把数列变成函数，把计数变成代数。你会研究图（graph），从着色（colouring）到拉姆齐理论（Ramsey theory），还有偏序集（partially ordered set），例如子集所构成的格（lattice）。概率课程也会运用同样的工具，再加上期望值的线性性质（linearity of expectation），来分析随机结构。再往后，组合数学会与代数（对称函数、表示论）、数论（分拆、加性组合）以及计算机科学（算法分析、复杂度理论）连系起来。

贯穿本章的问题

三大问题

1. **怎样不逐一列举而完成计数？**有哪些结构（递推关系、双射、生成函数）能让你准确地计数庞大的集合？
2. **甚么时候某件事必然发生？**甚么时候单凭数量之大，就足以迫使某种规律、巧合或配对出现？
3. **计数与概率怎样互相帮助？**甚么时候最好用计数来求概率，甚么时候又最好用概率来求数目？

本章分为两半。前七节取材自纯数学试卷，先讲递推关系（直线分割平面、为篱笆桩着色）和鸽巢原理（拉姆齐定理），然后建立生成函数，并用它来计数硬币、设计奇特的骰子，以及把数字分成幂和相等的队伍。后六节是概率试卷中的经典计数问题：错排（derangement）与 $1/e$ 的出现、生日问题、选票问题及其反射原理（reflection principle）、优惠券收集问题（coupon collector）、随机子集，以及为何不可能把骰子灌铅，使得每个总和出现的可能性均等。

本章的思想

§	课题	思想
9.1	平面的区域	多加一个时，有甚么改变
9.2	拉姆齐理论	鸽巢原理
9.3	图着色	耦合递推关系；色多项式
9.4	生成函数	数列变成函数
9.5	分拆（partition）	通过系数来计数
9.6	Sicherman 骰子	多项式的因式分解
9.7	Thue–Morse	重根带来平衡的和
9.8	错排	容斥原理；$1/e$
9.9	巧合	补事件；平方根定律
9.10	选票问题	反射原理
9.11	优惠券收集问题	期望值的线性性质；$n \ln n$
9.12	随机子集	每次只看一个元素
9.13	灌铅骰子	以算术—几何平均不等式（AM–GM）证明不可能

怎样阅读本章

永远从小的情况入手。画三条直线，列出三张牌的六种次序，数一数凑出 5p 的方法。小的情况能给你数据去猜想，最后也可以用来检验你的公式。但在找到理由之前，不要轻信任何规律：Section 9.1 中有一个著名的数列 $1, 2, 4, 8, 16, \ldots$，它的下一项是 31。

9.1 直线分割平面

题目

STEP II 1990，第 4 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/90-s2-q4`

引子

用 n 刀直线把一个薄饼切开，各块不必大小相同，最多可以切成多少块？一刀得 2 块，两刀得 4 块，三刀得 7 块（不是 8 块：不妨试试），四刀得 11 块。数列 $1, 2, 4, 7, 11, 16, \ldots$ 称为*懒惰宴会承办人数列*（lazy caterer's sequence），而本题要求的正是它的公式。

这道题是一种贯穿整个组合数学（combinatorics）的技巧的第一课：不要试图一次过把所有东西数完，而应该问：多加一个对象时，有甚么改变？而答案原来是欧拉公式（Euler's formula）$V - E + F = 2$ 的一个特殊情况，你在 Section 7.9 讨论柏拉图立体时已经见过这条公式。

动笔之前

停一停，试一试 仔细画出 $n = 1, 2, 3, 4$ 的情况，并保持各直线处于一般位置（general position）（没有两条平行，没有三条共点）。数一数每种情况中线段和区域的数目。你能猜出当中的规律吗？

你应该会得到：

n	0	1	2	3	4
线段	0	1	4	9	16
区域	1	2	4	7	11

线段的数目看来是 n^2。区域的数目则依次增加 $1, 2, 3, 4$：相邻两项之差正是自然数。

逐步解题

阶段 1　数线段

> **想一想**　集中看其中一条直线。其余的直线中有多少条与它相交？它们把这条直线切成多少段？

每条直线都与其余 $n-1$ 条直线逐一相交（没有一条与它平行），而且交于 $n-1$ 个不同的点（没有三条直线交于同一点）。这 $n-1$ 个点把该直线切成 n 段：两端各有一段无限长的，中间有 $n-2$ 段有限长的。共有 n 条直线，所以共有

$$n \times n = n^2 \text{ 条线段.} \qquad \blacksquare$$

这是直接计数，无需使用归纳法（induction）：我们是逐条直线去数线段的。

阶段 2　加一条直线时会怎样？

> **想一想**　假设已画好 $n-1$ 条直线，现在加上第 n 条。它被其余直线切成 n 段。每一段对它所穿过的区域有甚么影响？

画出新直线时，它的 n 段各自穿过一个原有的区域，并把该区域一分为二（图 9.1.1）。不同的线段位于不同的区域内，因为新直线要由一个区域走到另一个区域，就必须越过一条原有的直线。因此，加上第 n 条直线后，区域的数目恰好增加 n：

$$R(n) = R(n-1) + n.$$

阶段 3　解递归关系

由 $R(0) = 1$（没有直线，只有一个区域）开始，

$$R(n) = 1 + (1 + 2 + \cdots + n) = 1 + \frac{n(n+1)}{2} = \frac{n^2 + n + 2}{2}.$$

若要写成归纳法证明：公式对 $n = 1$ 成立，因为 $\frac{1}{2}(1+1+2) = 2$；而若 $R(n-1) = \frac{1}{2}\big((n-1)^2 + (n-1) + 2\big) = \frac{1}{2}(n^2 - n + 2)$，则 $R(n) = \frac{1}{2}(n^2 - n + 2) + n = \frac{1}{2}(n^2 + n + 2)$。
$\blacksquare$

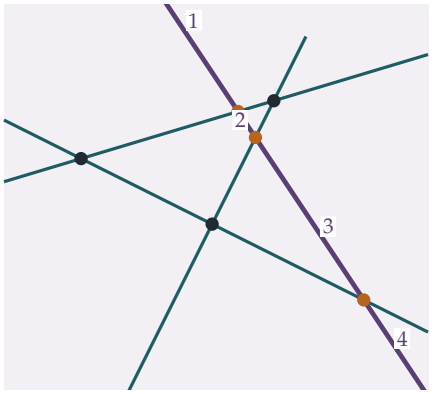


图 9.1.1 三条直线（蓝绿色）把平面分成 7 个区域。第四条直线（紫红色）与其余三条相交于琥珀色的点，这些点把它切成四段，编号为 1 至 4。每一段把一个区域一分为二，所以第四条直线增加了 4 个区域，合共 11 个。

回顾

- **每次只数一部分。**线段是逐条直线去数的；区域则是按每条新直线所增加的数目去数的。
- **把几何上的一步化为递归关系。**“每条新线段把一个区域一分为二”这项观察，化成了 $R(n) = R(n-1) + n$。
- **一般位置十分重要。**若有平行的直线，或有三条直线共点，区域的数目就会较少。试找出每项假设在哪里用过。

背后的数学

乔装的欧拉公式

以下是第二个证明，它不仅推导出答案，更解释了答案从何而来。把这些直线组成的图形看成一幅地图：交点是顶点（vertices），线段是边（edges），区域是面（faces）。图中有 $V = \binom{n}{2}$ 个顶点（每一对直线对应一个）和 $E = n^2$ 条边。无限长的线段只有一个端点，所以我们额外加上一个“在无穷远处”的顶点，让它们全部在那里相遇；这样，整幅图就成为画在球面上的地图（想像把平面包在一个球上，无穷远点位于球顶）。由 Section 7.9 的欧拉公式 $V - E + F = 2$ 可得

$$\left(\binom{n}{2} + 1\right) - n^2 + F = 2, \qquad \text{所以} \qquad F = n^2 - \frac{n(n-1)}{2} + 1 = \frac{n^2 + n + 2}{2}.$$

同一个数，却来自完全不同的方向。这些区域还可以整齐地分成两类：其中 $2n$ 个是无界的（每两条相邻的无限射线之间有一个），其余 $\frac{1}{2}(n^2 + n + 2) - 2n = \binom{n-1}{2}$ 个是有界的。

二项式系数与高维空间

这条公式可以写成

$$R(n) = \binom{n}{0} + \binom{n}{1} + \binom{n}{2},$$

而这并非巧合。把图形转动，使得没有一条直线是水平的。一个下方有界的区域必有一个最低点，而这个点必定是交点；每个交点又恰好是一个区域（即它正上方的那个）的最低点：这样就得到 $\binom{n}{2}$ 个区域，每一对直线对应一个。其余每个区域都向下无限延伸，因而会与一条画在所有交点下方很远处的水平线相遇；那 n 条直线把这条水平线切成 $n+1$ 个区间，每个这样的区域内各有一个。合共 $\binom{n}{2} + n + 1$ 个。在三维空间中，n 个处于一般位置的平面把空间切成

$$\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \binom{n}{3} = \frac{n^3 + 5n + 6}{6}$$

块（即蛋糕数（cake numbers）$1, 2, 4, 8, 15, 26, \ldots$），而这个规律在任何维数都成立。证明用的是同一个归纳法：新加的平面被其余平面切成若干区域，其排列方式就如平面上的直线所分成的区域，而每个这样的区域都把一块空间一分为二。

施泰纳与超平面排列

直线和平面所分割出的区域数目的公式，是施泰纳（Jakob Steiner）在 1826 年发现的。今天，它是*超平面排列*（hyperplane arrangements）理论的起点；这套理论计数任意一组直线、平面及其高维类比所形成的区域数目，包括并非处于一般位置的情况。Thomas Zaslavsky 在 1975 年证明了一条非凡的定理：利用一个记录了哪些超平面相交的多项式，就可以数出任何排列的区域数目。同样的想法也出现在机器学习之中：由“ReLU”单元构成的神经网络，会用超平面把输入空间切成多个区域，而区域的数目正好衡量它所能表示的函数有多复杂。

思维技巧

问一问：多加一个时，有甚么改变

要计数由 n 个对象组成的构形，可把它与 $n-1$ 个对象的构形比较。计数其中的改变，往往比计数整体容易得多。

这是把归纳法用作发现的工具，而不仅是证明的工具。它是组合数学中最常见的手法之一，能把几何问题化为可以求解的递归关系（recurrence relation）。

延伸练习

1 **圆。**证明：若 n 个圆中每两个都交于两点，且没有三个圆共点，则它们把平面分成 $n^2 - n + 2$ 个区域。

2 **平行线。**若 n 条直线中有 k 条互相平行（其余方面均处于一般位置），共有多少个区域？

3 **蛋糕数。**用归纳法证明：n 个处于一般位置的平面把空间分成 $\frac{1}{6}(n^3 + 5n + 6)$ 个区域。

4 **会失效的规律。**在圆周上取 n 个点，并把每一对点用弦连起来，其中没有三条弦在圆内交于同一点。当 $n = 1, 2, 3, 4, 5$ 时，分别得到 $1, 2, 4, 8, 16$ 个区域。数一数 $n = 6$ 时的区域数目。（提示：使用欧拉公式；每四个点给出一个交点。）证明区域的数目是 $1 + \binom{n}{2} + \binom{n}{4}$。

延伸阅读

- **Ronald Graham, Donald Knuth & Oren Patashnik, *Concrete Mathematics***（Addison-Wesley），第 1 章。开首讨论的正是这个问题，还有“薄饼”和“之字形”的变化题。
- **Miklós Bóna, *A Walk Through Combinatorics***（World Scientific）。一本平易近人的入门教材，设有归纳法、图和欧拉公式等章节。
- **Richard Guy, “The strong law of small numbers”**, *American Mathematical Monthly* 95 (1988)。收录了许多看似明显、其实不成立的规律，包括练习中的圆问题。

9.2 小型版的拉姆齐定理

题目

STEP I 1989，第 4 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/89-s1-q4`

引子

在任何一个六人派对上，要么有三个人彼此都认识，要么有三个人彼此都不认识。无论这六个人是谁，这一点都必定成立。换成五个人，就没有这个保证了。这个小小的事实是*拉姆齐定理*（*Ramsey's theorem*）的第一个情形，它道出了一条贯穿现代组合数学（combinatorics）的原则：*完全的无序是不可能的*。任何足够大的结构，无论怎样着色或排列，都必定包含一个高度有序的部分。

证明只需四行。但顺理成章的下一个问题（要多少人才能保证有四个互相认识的人或四个互不认识的人？五个呢?）却通向数学中一些最困难的未解难题。

动笔之前

想一想 那六个点之所以重要，只在于哪些点对以哪种颜色相连：几何只会令人分心。一共有多少条线段？每个点又连出多少条？

一共有 $\binom{6}{2} = 15$ 条线段，每个点连出 5 条。（一般位置（general position）这个条件只是确保任意三点都构成一个真正的三角形。这道题其实是关于六个顶点上的*完全图*（*complete graph*）K_6，它的边以两种颜色着色。）

逐步解题

阶段 1 五条线段，两种颜色

> **想一想**　从 A 连出五条线段，每条不是金色就是银色。关于它们的颜色，你可以肯定甚么?

其中至少有三条颜色相同。假如金色的至多两条，银色的也至多两条，那么线段至多只有四条。这就是鸽巢原理（*pigeonhole principle*）：把五件物件放进两个盒子，必有一个盒子里至少有三件。不妨设（有需要时把两种颜色的名称对调）AB、AC 和 AD 全是金色。

阶段 2　看看另一端

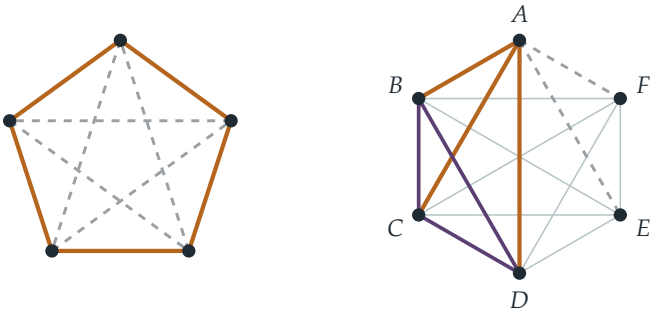
> **想一想**　现在看看三角形 BCD。如果它有任何一条边是金色，会怎样？如果一条都不是呢?

如果三角形 BCD 有任何一条边是金色，比如 BC，那么 ABC 的三条边 AB、AC 和 BC 全是金色。如果 BCD 没有一条边是金色，那么 BC、CD 和 DB 全是银色，BCD 就是一个银色三角形。无论哪种情况，都有一个单色三角形（图 9.2.1，右）。 ■

阶段 3　五个点并不足够

> **想一想**　若只有五个点，每个点只连出四条线段。怎样着色，才能使每个点连出的线段恰好每种颜色各两条?

把五个点放成一个正五边形（如果想要它们处于一般位置，可以把它们稍微提离平面）。把五边形的五条边涂成金色，五条对角线涂成银色（图 9.2.1，左）。金色的边构成一个五圈，当中没有三角形；银色的边构成一个五角星，它同样是一个五圈，当中也没有三角形。所以并没有单色三角形。 ■



五个点：没有单色三角形　　六个点：看看 BCD

图 9.2.1　左：五个点，五边形为金色（实线），五角星为银色（虚线），没有单色三角形。右：六个点时，从 A 连出的五条线段中有三条颜色相同，比如连到 B、C、D 的三条都是金色。这时三角形 BCD（紫红色）决定了结果：只要有一条边是金色，就与 A 构成一个金色三角形；没有金色的边，BCD 就是一个银色三角形。

回顾

- **撇除无关的东西。**空间中的点变成了一个着色的图；重要的只有边的颜色。
- **鸽巢原理。**五条边、两种颜色，迫使其中三条同色。
- **对合适的对象分情况讨论。**一旦找到三条金色的边，它们另一端的三角形就决定了一切。
- **紧确性（sharpness）。**五边形的例子说明“六”这个数不能再降低。一条好的定理，总会附带一个例子，说明它已是最佳的结果。

背后的数学

拉姆齐数

拉姆齐数（*Ramsey number*）$R(s,t)$ 是满足以下条件的最小的 N：把 K_N 的边以金、银两色着色，无论怎样着色，都必定包含一个金色的 K_s（即 s 个点，它们之间的所有边都是金色）或一个银色的 K_t。这道题已证明了 $R(3,3)=6$：六个点必定可行，而五边形说明五个点并不可行。

弗兰克·拉姆齐（Frank Ramsey）在 1928 年证明了这些数必定存在：对任意的 s 和 t，总有某个 N 符合要求。证明用的是归纳法（induction），把上面的论证加以推广，并得出

$R(s,t) \le R(s-1,t) + R(s,t-1)$。但要找出确切的数值却异常困难：

$$R(3,3) = 6, \qquad R(4,4) = 18, \qquad 43 \le R(5,5) \le 46.$$

$R(4,4)$ 的值是在 1955 年求得的。尽管已动用电脑作大规模搜索，$R(5,5)$ 至今仍然未知；上界 46 到 2024 年才获证明。电脑无法直接检查 K_{43} 的每一种着色：一共有 2^{903} 种，远远多于宇宙中原子的数目。

艾狄胥与外星人

保罗·艾狄胥（Paul Erdős）对拉姆齐理论（Ramsey theory）发展的贡献无人能及，他很喜欢讲这个故事。假设外星人入侵地球，并威胁说除非我们告诉他们 $R(5,5)$ 的值，否则就毁灭地球。我们应当调动所有电脑和数学家，这样或许能把它找出来。但假如他们问的是 $R(6,6)$，我们就应该设法消灭外星人。艾狄胥还在 1947 年证明了 $R(k,k) > 2^{k/2}$，所用的方法令人惊叹：他证明了，在少于 $2^{k/2}$ 个点上随机着色的图，有正的概率不包含任何单色的 K_k，所以好的着色必定存在，尽管没有人知道怎样把它构造出来。这种概率方法（*probabilistic method*）如今是组合数学中最强大的工具之一。

乱中有序

拉姆齐定理有很多近亲，它们表达的都是同一个想法：足够大的结构无法避开秩序。范德瓦尔登定理（*Van der Waerden's theorem*）说，如果把所有整数以有限种颜色着色，就存在任意长的单色等差数列（arithmetic progression）。艾狄胥—塞凯赖什定理（*Erdős–Szekeres theorem*）说，任何由 n^2+1 个相异的数组成的数列，都包含一个长度为 $n+1$ 的递增或递减子数列（subsequence）。而拉姆齐本人感兴趣的是一个逻辑问题：他是在一篇关于判定逻辑命题真假的论文中，把这条定理作为引理（lemma）来证明的。

思维技巧

鸽巢原理

如果把多于 kn 件物件放进 n 个盒子，则必有一个盒子里有多于 k 件物件。

这条原理显而易见，但要用得好却是一门艺术：功夫在于选定甚么是物件、甚么是盒子。在这道题中，物件是某一点上的五条边，盒子是两种颜色。在大学里，鸽巢原理的论证会出现在数论（以分数逼近无理数）、分析（波尔查诺–魏尔斯特拉斯定理（Bolzano–Weierstrass theorem））之中，更遍及整个组合数学。

延伸练习

1 **至少两个。**证明在 K_6 的任何双色着色中，其实至少有两个单色三角形。（提示：数一数“双色角”的数目，即在同一个顶点上的一对不同颜色的边。若一个点有 g 条金色的边和 $5-g$ 条银色的边，它就是 $g(5-g) \leq 6$ 个这样的角的顶点。每个非单色的三角形恰好包含两个这样的角。）

2 **三种颜色。**如果把 K_{17} 的边以三种颜色着色，证明其中有一个单色三角形。（提示：有一个点至少有六条同色的边。）

3 **三或四。**证明 $R(3,4) \leq 10$：在 K_{10} 的任何金、银着色中，都有一个金色三角形或一个银色的 K_4。（提示：一个点要么至少有 4 条金色的边，要么至少有 6 条银色的边。）事实上 $R(3,4) = 9$。

4 **单调子数列。**就 $n = 2$ 的情形证明艾狄胥–塞凯赖什定理：任何由 5 个相异的数组成的数列，都包含一个长度为 3 的递增或递减子数列。找出一个由 4 个数组成、两者皆无的数列。

延伸阅读

- **Miklós Bóna, *A Walk Through Combinatorics***（World Scientific），组合数学入门书，可读其中关于鸽巢原理和拉姆齐理论的章节。
- **Ronald Graham & Joel Spencer, “Ramsey theory”**，*Scientific American*（1990 年 7 月）。由两位顶尖研究者撰写的科普介绍。
- **Noga Alon & Joel Spencer, *The Probabilistic Method***（Wiley）。讲述艾狄胥这套方法的标准著作；第一章不难读懂。
- **Paul Hoffman, *The Man Who Loved Only Numbers***（Fourth Estate）。艾狄胥的传记。

9.3 为篱笆桩着色

题目

STEP II 1995，第 2 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/95-s2-q2`

引子

为排成一列的柱子髹色很容易：第一根有 3 种选择，其后每根各有 2 种（只要不与前一根同色即可），共有 $3 \times 2^{n-1}$ 种髹法。把这一列弯成一个圆圈，问题却变得出奇地棘手，因为最后一根柱子既要与相邻的前一根不同色，又要与第一根不同色，而这究竟排除了一种颜色还是没有排除任何颜色，取决于前面所有柱子的髹法。

题目用一个经典技巧绕过这个难处：按最后一根柱子的颜色把计数分成几种情况，再找出各种情况之间如何互相转化。答案 $2^n + 2(-1)^n$ 是*色多项式*（chromatic polynomial）的第一个例子，而色多项式正是四色定理（four colour theorem）背后的数学的一部分。

动笔之前

停一停，试一试 数一数为围成一圈的 3 根柱子髹色有多少种方法，4 根又有多少种。$3 \times 2^{n-1}$ 有没有多算？多算了多少？

围成一圈的三根柱子两两相邻，所以需要三种不同的颜色：共 $3! = 6$ 种方法。围成一圈的四根柱子：18 种方法（你可以在最后验证这一点）。与排成一列时的 $3 \times 2^2 = 12$ 和 $3 \times 2^3 = 24$ 比较：圆圈少了的，正是最后一根柱子与第一根同色的那些着色。

逐步解题

阶段 1 两类着色

想一想 考虑 $n+1$ 根柱子的一种着色（colouring），其中第一根和最后一根都是红色。把最后一根柱子拿走。余下的首 n 根柱子的着色有甚么特点?

第 n 根柱子与第 $n+1$ 根相邻，而后者是红色，所以第 n 根不是红色。拿走第 $n+1$ 根后，剩下的是 n 根柱子的一种着色，其中第一根是红色而最后一根不是红色：这正是 s_n 种着色之一。反过来，s_n 种着色中的任何一种都恰好有一种方法延伸，就是把第 $n+1$ 根柱子髹成红色。所以两个集合完全对应，$r_{n+1} = s_n$。

想一想 $r_n + s_n$ 是甚么?

每一种第一根柱子为红色的着色，其最后一根柱子不是红色就是非红色，所以 $r_n + s_n$ 数的是 n 根柱子中第一根为红色的全部着色。第一根之后，每根柱子有 2 种选择，所以 $r_n + s_n = 2^{n-1}$。因此

$$r_{n+1} + r_n = s_n + r_n = 2^{n-1}. \qquad \blacksquare$$

阶段 2 用数学归纳法（induction）证明公式

当 $n = 1$ 时只有一根柱子，髹成红色，所以 $r_1 = 1$，而公式给出 $\frac{1}{3}(1+2) = 1$。假设公式对 r_n 成立。那么

$$r_{n+1} = 2^{n-1} - r_n = 2^{n-1} - \frac{2^{n-1} + 2(-1)^{n-1}}{3} = \frac{2 \cdot 2^{n-1} - 2(-1)^{n-1}}{3} = \frac{2^n + 2(-1)^n}{3},$$

这正是 r_{n+1} 的公式。 $\blacksquare$

停一停，试一试 这条公式从何而来?递归关系（recurrence relation）$r_{n+1} + r_n = 2^{n-1}$ 就像一条微分方程（differential equation）：解是一个“特解”（particular solution）$c \cdot 2^n$ 加上一个满足 $r_{n+1} + r_n = 0$ 的“余函数”（complementary function）$A(-1)^n$。求 c 和 A。

阶段 3 圆圈

想一想 在 n 根柱子围成的圆圈中，固定第 1 根柱子的颜色，例如红色。现在在第 n 根与第 1 根柱子之间把圆圈剪开，拉直成一列。哪些直列着色对应于合乎要求的圆圈着色?

拉直之后，圆圈变成由红色开始的一列 n 根柱子，其中最后一根还必须与第 1 根不同色：最后一根不是红色。这些正好就是那 s_n 种着色。第 1 根柱子可以是 3 种颜色中的任何一种，所以圆圈着色的数目是

$$3s_n = 3(2^{n-1} - r_n) = 3 \cdot 2^{n-1} - 2^{n-1} - 2(-1)^{n-1} = 2^n + 2(-1)^n. \qquad \blacksquare$$

验算：$n=3$ 给出 $8-2=6$，而 $n=4$ 给出 $16+2=18$。

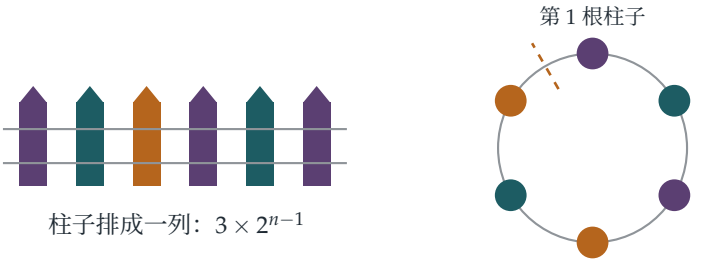


图 9.3.1 左：排成一列时，第一根之后的每根柱子只须与前一根不同色。右：围成一圈时，最后一根柱子还必须与第 1 根柱子不同色。沿虚线把圆圈剪开，圆圈着色就变成首尾两根柱子不同色的直列着色：即题目中所数的 s_n 种着色之一。

回顾

- **分成互相转化的情况。**记下最后一根柱子是否红色，便得到两个数列 r_n 和 s_n，它们由简单的递归关系联系起来。
- **双射（bijection）。**$r_{n+1}=s_n$ 是靠把着色一一配对来证明的，而不是靠计算。
- **剪开圆圈。**一个循环的问题变成了一个对两端附加条件的直线问题。

背后的数学

色多项式

柱子和它们之间的相邻关系构成一个*图*（graph）：排成一列时是一条路径（path），围成一圈时是一个圈（cycle）。*正常着色*（proper colouring）给相邻的顶点（vertex）不同的颜色。用 k 种颜色的正常着色数目是 k 的一个多项式，称为该图的*色多项式*，由伯克霍夫（George Birkhoff）于 1912 年引入。用 k 种颜色重复题目中的论证，可得

$$\text{路径：}k(k-1)^{n-1},\qquad \text{圈：}(k-1)^n+(-1)^n(k-1).$$

取 $k=3$，圈的公式就是 $2^n+2(-1)^n$。伯克霍夫希望通过证明每幅地图所对应的图的色多项式在 $k=4$ 时为正数，从而证明四色定理（每幅地图都可以用四种颜色着色，使相邻的国家颜色不同）。这个方法始终未能成功；定理最终在 1976 年由阿佩尔（Kenneth Appel）和哈肯（Wolfgang Haken）借助大量电脑运算证明。色多项式至今仍是一个活跃的研究领域，并与统计物理（statistical physics）有出人意表的联系：在那里它们以描述磁性的*波茨模型*（Potts model）的形式出现。

用矩阵数出同一个答案

还有另一种方法可以看出答案，用的是第 7 章的线性代数（linear algebra）。设 $\mathbf{M}$ 为对角线上是 0、其余位置是 1 的 3×3 矩阵：当颜色 j 可以跟在颜色 i 之后时，第 (i,j) 项为 1。那么 $\mathbf{M}^{n-1}$ 的第 (i,j) 项数的是由颜色 i 开始、以颜色 j 结束的直列着色，所以 r_n 是 $\mathbf{M}^{n-1}$ 的一个对角线项。圆圈着色是经过 n 步后回到起点的一个序列，所以圆圈着色的数目就是 $\mathbf{M}^n$ 的迹（trace）。由于 $\mathbf{M} = \mathbf{J} - \mathbf{I}$，其中 $\mathbf{J}$ 的每一项都是 1，$\mathbf{M}$ 的特征值（eigenvalue）是 $3-1=2$（特征向量（eigenvector）为 $(1,1,1)$）和 -1（两重），而 $\mathbf{M}^n$ 的迹是各特征值的 n 次方之和：

$$\operatorname{tr}\mathbf{M}^n = 2^n + (-1)^n + (-1)^n = 2^n + 2(-1)^n.$$

这种*转移矩阵*（transfer matrix）方法在组合数学（combinatorics）和统计物理中随处可见，用来数算逐步构造出来的排列。

思维技巧

建立互相耦合的递归关系

当一个数列单靠自己无法满足简单的递归关系时，不妨多记录一点资料（这里是最后一根柱子的颜色），再找出联系各种情况的递归关系。

这就是*状态*（state）的概念：只记下继续下去所必须知道的资料，不多也不少。它是电脑科学中动态规划（dynamic programming）的基础，也是概率中马可夫链（Markov chains）（Section 12.16）和转移矩阵方法的基础。

延伸练习

1 **k 种颜色。**用 k 种颜色重做这道题。证明 $r_{n+1} = s_n$ 仍然成立，并且 $r_{n+1} + r_n = (k-1)^{n-1}$，求 r_n，并由此推导出圈的色多项式 $(k-1)^n + (-1)^n(k-1)$。

2 **没有两根红柱相邻。**把排成一列的 n 根柱子髹成红色或白色，而且没有两根相邻的柱子同为红色，共有多少种方法？（你应该会得到斐波那契数（Fibonacci numbers）。）

3 **轮子。**加一根位于中心的柱子，与圆圈上的每根柱子都相邻（一个“轮子”）。用 4 种颜色为它着色有多少种方法？（*提示*：中心用去一种颜色，圆圈必须使用其余三种。）

4 **项链。**如果圆圈的两种着色在其中一种可由另一种旋转而得时视为相同，那么当 $n = 4$ 时有多少种不同的着色？$n = 5$ 时呢？（这是*伯恩赛德引理*（Burnside's lemma）的开端。）

延伸阅读

- **Robin Wilson, *Introduction to Graph Theory***（Pearson）。标准的图论入门教材；涵盖着色和四色问题。
- **Robin Wilson, *Four Colours Suffice***（Penguin）。四色定理的故事，为一般读者而写。
- **Herbert Wilf, *generatingfunctionology***（可在网上免费阅读）。讲述递归关系及其解法，包括转移矩阵方法。

9.4 把数列看成函数

题目

STEP I 2016，第 8 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/16-s1-q8`

引子

用计算机算一算 $1/89$：$0.011235955\ldots$。开首的数字是 $0, 1, 1, 2, 3, 5$，正是斐波那契数（Fibonacci numbers）。这是巧合吗？不是，而这道题目正好解释了原因。

这个想法是组合数学（combinatorics）中最强大的想法之一。取一个无穷数列，把各项用作幂级数（power series）的系数，从而把整个数列包装成单一个函数。这样，关于数列的事实就变成关于函数的事实，你便可以用代数和微积分来处理。Herbert Wilf 有一句令人难忘的话：“生成函数（generating function）是一条晾衣绳，我们把一列数字挂在上面展示。”

动笔之前

想一想 哪个数列的生成函数是 $\dfrac{1}{1-x}$？$\dfrac{1}{1-2x}$ 又如何？如果把一个数列的生成函数乘以 x，数列会有甚么变化？

等比级数（geometric series）$\frac{1}{1-x} = 1 + x + x^2 + \cdots$ 是 $1, 1, 1, \ldots$ 的生成函数，而 $\frac{1}{1-2x} = 1 + 2x + 4x^2 + \cdots$ 则是 2 的各次幂的生成函数。乘以 x 会把每一项向后移一位：$x \cdot \frac{1}{1-x} = x + x^2 + \cdots$ 对应的数列是 $0, 1, 1, 1, \ldots$。

逐步解题

阶段 1　自然数与三角形数

想一想　用二项级数（binomial series）展开 $(1-x)^{-2}$。x^n 的系数是甚么？

由二项级数可得

$$(1-x)^{-2} = 1 + 2x + 3x^2 + 4x^3 + \cdots = \sum_{n\ge 0} (n+1)x^n,$$

所以 $x(1-x)^{-2} = x + 2x^2 + 3x^3 + \cdots = \sum nx^n$：这就是 $u_n = n$ 的生成函数。

同样地，$(1-x)^{-3}$ 中 x^n 的系数是 $\frac{(-3)(-4)\cdots(-3-n+1)}{n!}(-1)^n = \frac{(n+1)(n+2)}{2}$，所以

$$x(1-x)^{-3} = \sum_{n\ge 1} \frac{n(n+1)}{2}\, x^n = x + 3x^2 + 6x^3 + 10x^4 + \cdots,$$

这就是三角形数（triangle numbers）$u_n = \frac{1}{2}n(n+1)$ 的生成函数。

想一想　怎样用 n 和 $\frac{1}{2}n(n+1)$ 来表示 n^2？

由于 $n^2 = 2\cdot\frac{1}{2}n(n+1) - n$，而生成函数可以像数列一样相加，

$$\sum n^2 x^n = \frac{2x}{(1-x)^3} - \frac{x}{(1-x)^2} = \frac{2x - x(1-x)}{(1-x)^3} = \frac{x(1+x)}{(1-x)^3}. \qquad \blacksquare$$

阶段 2　递推关系变成方程

想一想　把 $u_n = ku_{n-1}$ 乘以 x^n，然后对所有 $n \ge 1$ 求和。两边用 f 表示，分别是甚么？

左边是 $\sum_{n\ge 1} u_n x^n = \mathrm{f}(x) - u_0 = \mathrm{f}(x) - a$。右边是 $kx\sum_{n\ge 1} u_{n-1}x^{n-1} = kx\,\mathrm{f}(x)$。所以 $\mathrm{f}(x) = a + kx\,\mathrm{f}(x)$，于是

$$\mathrm{f}(x) = \frac{a}{1-kx}. \qquad \blacksquare$$

当然，$u_n = ak^n$；把 $\frac{a}{1-kx}$ 展开成等比级数，便可确认这一点。

阶段 3　斐波那契数

对 $u_n = u_{n-1} + u_{n-2}$（其中 $n \ge 2$）作同样的处理：

$$\sum_{n\ge 2} u_n x^n = x\sum_{n\ge 2} u_{n-1}x^{n-1} + x^2 \sum_{n\ge 2} u_{n-2}x^{n-2},$$

即

$$\mathrm{f}(x) - u_0 - u_1 x = x\big(\mathrm{f}(x) - u_0\big) + x^2\,\mathrm{f}(x).$$

代入 $u_0 = 0$ 和 $u_1 = 1$，上式即 $\mathrm{f}(x) - x = x\,\mathrm{f}(x) + x^2\,\mathrm{f}(x)$，所以

$$\mathrm{f}(x) = \frac{x}{1 - x - x^2}. \qquad \blacksquare$$

阶段 4 解释 1/89

> **停一停，试一试** 暂且不理收敛（convergence）的问题（这里没有问题：级数在 $|x| < 0.618$ 时收敛），把 $x = \frac{1}{10}$ 代入 f。你得到甚么？

$$\frac{1}{10}F_1 + \frac{1}{100}F_2 + \frac{1}{1000}F_3 + \cdots = \mathrm{f}\Big(\frac{1}{10}\Big) = \frac{1/10}{1 - 1/10 - 1/100} = \frac{10}{89}.$$

除以 10，得 $\frac{1}{89} = \sum F_n/10^{n+1}$。斐波那契数依次放在相继的小数位上，当它们大于 9 时便会互相重叠（图 9.4.1）。1/89 开首的数字显示出斐波那契数；其后的数字则因进位而变得模糊。

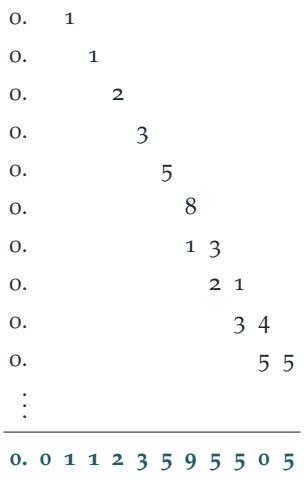


图 9.4.1 把 $F_n/10^{n+1}$ 对 $n = 1, 2, 3, \ldots$ 相加。每个斐波那契数的末位都比前一个向右多移一个小数位。一旦它们有两位数字，便会互相重叠，而进位会改变第七位起的数字：$\frac{1}{89} = 0.01123595505\ldots$。

回顾

- **把数列变成函数。**整个无穷数列变成了一个有理函数（rational function）。

- **递推关系变成方程。**乘以 x^n 再求和，便把规则 $u_n = u_{n-1} + u_{n-2}$ 变成一条关于 f 的方程，可以用代数方法解出。
- **从旧的造出新的。**把生成函数相加、平移和求导，便能得出新的生成函数，无须从头做起。

背后的数学

从生成函数到公式

生成函数不只是把数列编码：它还可以把数列解出来。分母可以因式分解为 $1 - x - x^2 = (1 - \varphi x)(1 - \psi x)$，其中 $\varphi = \frac{1+\sqrt{5}}{2}$ 和 $\psi = \frac{1-\sqrt{5}}{2}$ 是 $t^2 = t + 1$ 的根。由部分分式（partial fractions）可得

$$\frac{x}{1 - x - x^2} = \frac{1}{\sqrt{5}}\left(\frac{1}{1 - \varphi x} - \frac{1}{1 - \psi x}\right),$$

再把每个分式展开成等比级数，

$$F_n = \frac{\varphi^n - \psi^n}{\sqrt{5}}.$$

这就是*比内公式*（*Binet's formula*）。由于 $|\psi| < 1$，第二项非常小，所以 F_n 是最接近 $\varphi^n/\sqrt{5}$ 的整数，而相邻两个斐波那契数之比趋向黄金比（golden ratio）（这是 Section 2.2 的主题）。同一方法可以解任何常系数线性递推关系（linear recurrence with constant coefficients）；练习中有一个来自 STEP I 2004 的例子。

生成函数为何如此有用

真正的威力在相乘时才显现出来。乘积 $f(x)g(x)$ 中 x^n 的系数是 $\sum_k u_k v_{n-k}$，即对所有把 n 拆成两部分的方法求和。要数算由两个部分组成的东西，这正是我们所需要的。接下来的两节会用它来数算找换零钱的方法（Section 9.5），以及找出点数古怪、但总和的分布与普通骰子相同的骰子（Section 9.6）。在概率中，同一想法给出*概率生成函数*（*probability generating function*），它把独立随机变量之和变成乘积。

棣莫弗、欧拉与拉普拉斯

棣莫弗（Abraham de Moivre）在 1720 年代用生成函数来解线性递推关系，所用的方法基本上就是上面的方法。欧拉（Leonhard Euler）在 1748 年用它们来研究整数的分拆（partitions），而拉普拉斯（Pierre-Simon Laplace）则以它们作为其著作 *Théorie analytique des probabilités*（1812）的基础。到了二十世纪，生成函数成为组合数学的核心，而解析方法（把生成函数看成复变量的函数来研究）如今能为复杂得无法精确求解的计数问题给出精确的估计。

思维技巧

转换表示方式

一个关于数列的问题，在把数列变成函数、矩阵或图形之后，可能会变得容易。

生成函数把递推关系翻译成代数，把计数翻译成乘法。在大学里你会遇到许多这类翻译：傅里叶级数（Fourier series）把函数变成数列，拉普拉斯变换（Laplace transforms）把微分方程变成代数，而矩阵则把几何变成算术。

延伸练习

1 **每次两位数字。**试做 STEP II 2022，第 3 题（载于 OCR 提供下载的 2022 年试卷，`www.ocr.org.uk/Images/696587-2022-past-papers.zip`）。它解释了为何 $\frac{1}{89}$ 的小数展开式以斐波那契数开首，并要求找出一个分数，其小数每次两位数字地显示这个规律。

2 **解递推关系。**试做 STEP I 2004，第 7 题（`step.maths.org/questions/04-s1-q7`）。它用生成函数和部分分式解一个线性递推关系，然后把这个方法倒过来做。

3 **逐一数数。**证明 $\frac{1}{9801} = 0.00\,01\,02\,03\,04\ldots$，其中每两位数字为一组，逐一向上数。这个规律最先在哪里出错？为甚么？

4 **平方和。**解释为何 $\frac{1}{1-x}f(x)$ 是部分和 $u_0 + u_1 + \cdots + u_n$ 的生成函数。利用这一点和第 (i) 部分的答案，求 $1^2 + 2^2 + \cdots + n^2$ 的公式。

延伸阅读

- **Herbert Wilf, *generatingfunctionology***（A K Peters；网上免费阅读）。经典的入门书。读完本节后，最适合接着读第 1 章。
- **Ronald Graham, Donald Knuth & Oren Patashnik, *Concrete Mathematics***（Addison-Wesley），第 7 章。
- **Philippe Flajolet & Robert Sedgewick, *Analytic Combinatorics***（Cambridge University Press；网上免费阅读）。远远超出大学一年级的程度，但开首数页展示了这门学问的去向。

9.5 凑零钱

题目

STEP I 1997，第 1 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/97-s1-q1`

引子

这看来是全书最容易的一题：把所有可能逐一列出便是。可是列举很快就会失控。用英国人口袋里的硬币，凑出 20p 有 41 种方法，凑出 50p 有 451 种，凑出 £1 更有 4563 种。这道题真正要求的是一套*系统*，也就是一种整理计数的方法，使得既无遗漏，亦无重复。

这套系统背后是一段优美的数学。每种硬币对应一个等比级数（geometric series），凑出 n 便士的方法数就是这些级数之积的一个系数（coefficient）；把同一个想法一并用于所有正整数，便得到欧拉（Euler）的*分拆*（partition）理论，那是数论中最深刻的课题之一。

动笔之前

想一想 只用 2p 和 1p 硬币，凑出 n 便士有多少种方法？为甚么答案只取决于你用了多少枚 2p 硬币？

一旦选定 2p 硬币的数目，其余的就必须全是 1p 硬币，所以 2p 硬币的每个可能数目 $0, 1, \ldots, \lfloor n/2 \rfloor$ 都恰好对应一种方法。这样共有 $\lfloor n/2 \rfloor + 1$ 种方法，记作 $a(n)$。例如 $a(10) = 6$，$a(5) = 3$。

逐步解题

阶段 1　十便士

想一想　按面值最大的硬币来整理计数。你可以用多少枚 10p 硬币？定下来之后，又可以用多少枚 5p 硬币？

- 一枚 10p 硬币：1 种方法。
- 不用 10p，用两枚 5p：1 种方法。
- 不用 10p，用一枚 5p：余下的 5p 由 2p 和 1p 凑成，有 $a(5) = 3$ 种方法。
- 不用 10p 也不用 5p：10p 全由 2p 和 1p 凑成，有 $a(10) = 6$ 种方法。

总数是 $1 + 1 + 3 + 6 = 11$。 ■

阶段 2　逐层建立

想一想　设 $b(n)$ 为用 5p、2p 和 1p 硬币凑出 n 便士的方法数。怎样用函数 a 表示 $b(n)$？

选定 5p 硬币的数目 k；余下的 $n - 5k$ 便士由 2p 和 1p 硬币凑成。所以

$$b(n) = a(n) + a(n-5) + a(n-10) + \cdots .$$

由此得

n	0	5	10	15	20
$a(n)$	1	3	6	8	11
$b(n) = a(n) + b(n-5)$	1	4	10	18	29

同样地，设 $c(n)$ 为连 10p 硬币也可使用时的方法数；则 $c(n) = b(n) + b(n-10) + \cdots$，所以 $c(10) = 10 + 1 = 11$（与第 1 阶段的结果一致），而 $c(20) = b(20) + b(10) + b(0) = 29 + 10 + 1 = 40$。

阶段 3　二十便士

最后轮到 20p 硬币：我们要么用它（一种方法），要么不用（$c(20)$ 种方法）。所以共有

$$40 + 1 = 41 \text{ 种方法.}$$ ■

停一停，试一试　留意当中的结构：每加入一种新硬币就多一层，而每一层都是把上一层按新硬币的面值为间隔取值再相加。你能看出电脑会怎样为 £1 甚至 £100 进行这项计算吗？

回顾

- **按最大的部分来整理。**先固定面值最大的硬币的数目，计数便有条不紊。
- **建立表格。**每加入一种新硬币，就多出一行，由上一行计算出来。这就是*动态规划*（dynamic programming），是电脑计算这类数目的标准方法。
- **与已知的情况核对。**表格重现了 10p 的 11 种方法。

背后的数学

把硬币看成等比级数

以下这个想法把问题化为代数。所用 5p 硬币的数目可以是 $0, 1, 2, \ldots$，分别贡献 $0, 5, 10, \ldots$ 便士。把这一点记录成级数

$$1 + x^5 + x^{10} + x^{15} + \cdots = \frac{1}{1 - x^5}.$$

对每种硬币都这样做，然后相乘：

$$\frac{1}{(1-x)(1-x^2)(1-x^5)(1-x^{10})(1-x^{20})} = \sum_n w(n)x^n.$$

把各括号乘开时，每一项 x^n 都来自在每个括号中各选一项，也就是为每种硬币各选一个数目，而总值为 n。所以 x^n 的系数 $w(n)$ 正好就是凑出 n 便士的方法数。第 2 阶段的表格，其实就是把这个乘积逐个括号乘开。这正是 Section 9.4 中母函数（generating function）的想法，而乘法正是它大显威力之处：母函数之积所计算的，就是各项选择的组合数目。

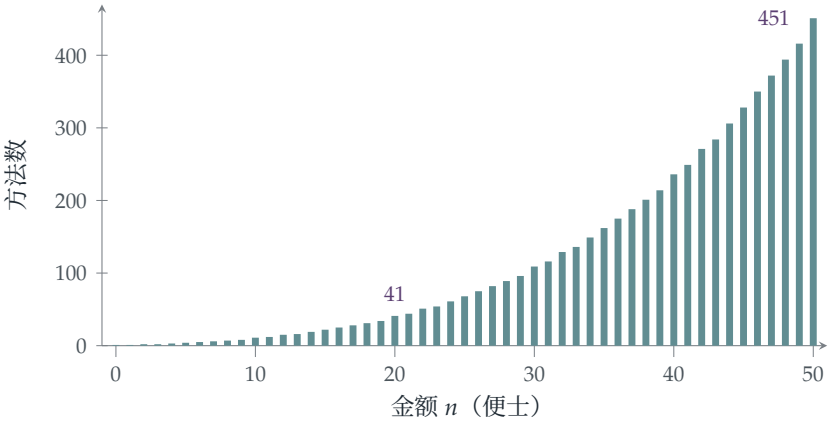


图 9.5.1 用 1p、2p、5p、10p、20p 和 50p 硬币凑出 n 便士的方法数。在 5 和 10 的倍数处出现的跳升，显示新硬币首次可用的位置。有六种硬币时，方法数的增长大致与 n^5 相若：一般而言，增长有如一个多项式，其次数比硬币种数少一。

分拆

假如每个正整数都可以作为硬币的面值，又会怎样？这时，不计次序而把 n 写成正整数之和的方法数，就是*分拆数*（partition number）$p(n)$。例如 $p(4)=5$，因为 $4=3+1=2+2=2+1+1=1+1+1+1$。欧拉在 1748 年看出

$$\sum_{n\ge 0} p(n)x^n = \frac{1}{(1-x)(1-x^2)(1-x^3)(1-x^4)\cdots},$$

并用它发现了一些非凡的恒等式（identity）。其中之一是：把 n 写成*互不相同*的部分之和的方法数，等于把它写成*奇数*部分之和的方法数。（用母函数来证明，只需一行代数：请在练习中一试。）

哈代、拉马努金与 $p(200)$

分拆数增长得很快：$p(10)=42$，$p(100)=190\,569\,292$。1918 年，哈代（G. H. Hardy）与拉马努金（Srinivasa Ramanujan）为 $p(n)$ 找到一条令人惊叹的公式，那是一个无穷级数（infinite series），其首项为

$$p(n) \approx \frac{1}{4n\sqrt{3}}\, e^{\pi\sqrt{2n/3}}.$$

为了检验这条公式，Percy MacMahon 利用欧拉的恒等式，徒手算出 $p(200)=3\,972\,999\,029\,388$。哈代与拉马努金的级数只取首几项，所得结果与正确答案相差不足一。拉马努金还发现 $p(5n+4)$ 必定可被 5 整除，$p(7n+5)$ 必定可被 7 整除，而 $p(11n+6)$ 必定可被 11 整除。至于为何再没有其他这类规律，则要到二十一世纪才得到解释。

思维技巧

藉系数来计数

当一个对象由若干独立的选择构成时，每项选择就是一个因式，而总大小为 n 的对象的数目，就是乘积中 x^n 的系数。

这把计数化为代数。在大学里，它是计数组合学（enumerative combinatorics）的基础，而概率母函数（probability generating function）也正是这样处理独立随机变量（independent random variables）之和的。

延伸练习

1 **两种六。**试做 STEP II 2012，第 1 题（step.maths.org/questions/12-s2-q1）。该题利用等比级数之积来计数。做完之后，请描述一个以这些系数为答案的硬币计数问题。

2 **互异与奇数。**证明 $(1+x)(1+x^2)(1+x^3)\cdots$ 是分拆成互不相同部分的母函数。利用 $1+x^k=\frac{1-x^{2k}}{1-x^k}$ 证明它等于 $\frac{1}{(1-x)(1-x^3)(1-x^5)\cdots}$，并由此推出欧拉定理。就 $n=6$ 验证这个结果。

3 **一条公式。**证明只用 1p、2p 和 5p 硬币凑出 n 便士的方法数，是最接近 $\frac{(n+4)^2}{20}$ 的整数。（先用较小的 n 验证。）

4 **次序有别。**把 1p 和 2p 硬币逐枚投入投币口来凑出 n 便士，若次序有别，共有多少种方法？（这个数列（sequence）你以前见过。）

延伸阅读

- **George Andrews & Kimmo Eriksson, *Integer Partitions***（Cambridge University Press）。一本为大学本科生而写的入门书，篇幅短小，浅白易读。
- **波利亚（George Pólya），“On picture-writing”**，*American Mathematical Monthly* 63 (1956)。波利亚以兑换一美元的问题（共有 292 种方法）来讲解母函数的经典文章。
- **Robert Kanigel, *The Man Who Knew Infinity***（Abacus）。拉马努金的生平，以及他与哈代的合作。

9.6 奇怪的骰子

题目

STEP I 1993，第 1 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/93-s1-q1`

引子

两颗六面骰子，各面标上正整数，除了平常的标法之外，还有没有别的标法，可以令由 2 到 12 的每个点数和出现的次数，都与普通骰子完全相同？1978 年，Martin Gardner 在《科学人》（*Scientific American*）杂志上宣布，这样的标法恰好只有一种：一颗骰子标上 $1,2,2,3,3,4$，另一颗标上 $1,3,4,5,6,8$。这对骰子由 George Sicherman 上校发现，现在市面上有售，称为 *Sicherman* 骰子。用它们来玩一局大富翁（Monopoly），与用普通骰子玩根本无从分辨。

本题是一道热身题：从点数和的列表，推断两颗骰子各面标了甚么数字。要有系统地解决这个问题，以至证明 Sicherman 骰子是唯一的一对奇异骰子，方法都是把每颗骰子化为一个多项式（polynomial）。

动笔之前

想一想 假设第一颗骰子有 p_i 个面显示 i，第二颗有 q_j 个面显示 j，那么掷出点数和 t 共有多少种方法？

方法的数目是 $\sum_i p_i q_{t-i}$：先选定第一颗骰子的点数 i，第二颗就必须显示 $t-i$。就题目中的第一对骰子而言，点数和 6 可以来自 $1+5$（有 1×1 种方法）、$2+4$（2×1 种）、$3+3$（2×2 种），此外再无其他可能，因为第一颗骰子没有 4 或 5，而第二颗没有 0。所以共有 $1+2+4=7$ 种方法。 ■

逐步解题

阶段 1　两端的情况

把两颗骰子称为 A 和 B，并以 a_i、b_i 分别表示两颗骰子上显示 i 的面数。

> **想一想**　点数和不可能是 2，也不可能是 12。这对 1 和 6 的分布有甚么启示？掷出 3 却确实有两种方法。哪一颗骰子上有 1？

掷不出点数和 2，即 $a_1 b_1 = 0$：最多只有一颗骰子上有 1。点数和 3 只可能是 $1+2$ 或 $2+1$，而且有两种方法，所以其中一颗骰子确实有 1；把它称为 A，于是 $b_1 = 0$。那么

$$\text{点数和}3: \quad a_1 b_2 = 2, \qquad \text{所以} (a_1, b_2) = (2,1) \text{ 或} (1,2).$$

同理，掷不出点数和 12，即 $a_6 b_6 = 0$。

阶段 2　逐步向上推

> **想一想**　写出点数和为 4 的方程，记住 $b_1 = 0$。在两种情况下，这个方程分别说明了甚么？

点数和 4 是 $1+3$ 或 $2+2$（由于 $b_1 = 0$，$3+1$ 不可能出现），所以 $a_1 b_3 + a_2 b_2 = 1$。

- **情况 (a)**，$a_1 = 2$，$b_2 = 1$：$2b_3 + a_2 = 1$，所以 $b_3 = 0$ 及 $a_2 = 1$。
- **情况 (b)**，$a_1 = 1$，$b_2 = 2$：$b_3 + 2a_2 = 1$，所以 $a_2 = 0$ 及 $b_3 = 1$。

点数和 5：$a_1 b_4 + a_2 b_3 + a_3 b_2 = 1$。

- 情况 (a)：$2b_4 + 0 + a_3 = 1$，所以 $b_4 = 0$，$a_3 = 1$。至此 A 有 $1,1,2,3$，B 有 2；B 其余五个面都是 5 或 6。
- 情况 (b)：$b_4 + 0 + 2a_3 = 1$，所以 $a_3 = 0$，$b_4 = 1$。至此 A 有 1，B 有 $2,2,3,4$；A 其余五个面取自 $4,5,6$，B 其余两个面取自 $5,6$。

阶段 3　情况 (a) 不成立

> **想一想**　在情况 (a) 中，写出点数和为 7 的方程，并利用 $b_5 + b_6 = 5$。

点数和 7：$a_1 b_6 + a_2 b_5 + a_3 b_4 + a_4 b_3 + a_5 b_2 = 2b_6 + b_5 + 0 + 0 + a_5 = 3$。代入 $b_6 = 5 - b_5$，得 $10 - b_5 + a_5 = 3$，所以 $b_5 = 7 + a_5 \geq 7$。但 B 只有六个面。情况 (a) 不可能成立。

阶段 4　情况 (b) 成立

想一想 在情况 (b) 中，利用点数和 6 和 11，以及 $a_6b_6 = 0$ 这个事实。

点数和 6：$a_1b_5 + a_4b_2 = b_5 + 2a_4 = 4$（其余各个乘积都含有 $a_2 = a_3 = 0$ 或 $b_1 = 0$）。由于 $b_5 \le 2$，而且 $b_5 = 4 - 2a_4$ 是偶数，所以要么 $b_5 = 0, a_4 = 2$，要么 $b_5 = 2, a_4 = 1$。

- 若 $b_5 = 0$，则 B 是 $2,2,3,4,6,6$，于是 $a_6 = 0$，而 A 是 $1,4,4,5,5,5$。但这样点数和 7 就有 $a_1b_6 + a_4b_3 + a_5b_2 = 2 + 2 + 6 = 10$ 种方法，而不是 3 种。
- 若 $b_5 = 2$，则 B 是 $2,2,3,4,5,5$，其中 $b_6 = 0$，而 A 是 $1,4$ 再加上四个取自 5 和 6 的面。点数和 11：$a_6b_5 = 2a_6 = 6$，所以 $a_6 = 3$ 及 $a_5 = 1$。

唯一的可能是

$$A:\ 1,4,5,6,6,6, \qquad B:\ 2,2,3,4,5,5.$$

最后，我们必须验证这对骰子确实给出整个列表，因为我们只用了其中一部分点数和。把乘积展开（或利用下文的多项式），即可确认每一项都吻合。 ■

回顾

- **从两端入手。**最小和最大的点数和只牵涉一两个乘积，所以提供的信息最有力。
- **分情况讨论，逐一排除。**每个点数和都把可能性收窄一些，直到只剩下一种情况。
- **检查充分性（sufficiency）。**推论说明了骰子*必须*是甚么样子；最后的验证则说明这对骰子确实可行。

背后的数学

把骰子看成多项式

把每颗骰子写成一个多项式：每个显示 i 的面对应一项 x^i。普通骰子是 $x + x^2 + x^3 + x^4 + x^5 + x^6$，而题目中的两颗骰子是

$$A(x) = x + x^4 + x^5 + 3x^6, \qquad B(x) = 2x^2 + x^3 + x^4 + 2x^5.$$

相乘时，每一对面贡献 $x^i x^j = x^{i+j}$，所以 $A(x)B(x)$ 中 x^t 的系数（coefficient），就是掷出点数和 t 的方法数目。这又是 Section 9.4 中母函数（generating function）的想法：点数和的列表其实就是乘积 $A(x)B(x)$，而我们逐个情况的推论，正是把它因式分解（factorise）的一种方法。

Sicherman 骰子

两颗普通骰子的乘积是 $(x+x^2+\cdots+x^6)^2$。任何一对点数和分布相同的骰子，其多项式 $P(x)$、$Q(x)$ 必须满足 $P(x)Q(x)=(x+\cdots+x^6)^2$。现在作因式分解：

$$x+x^2+\cdots+x^6=x(1+x)(1+x+x^2)(1-x+x^2),$$

而这些因式不能再进一步分解（成为整数系数的多项式）。所以 P 和 Q 都必须由因式 x、$1+x$、$1+x+x^2$、$1-x+x^2$ 组成，而每个因式合共出现两次。以下三个条件把选择收窄：

- 每颗骰子有 6 个面，所以 $P(1)=Q(1)=6$；各因式在 $x=1$ 时的值分别是 1、2、3、1，所以每颗骰子各得一个 $1+x$ 和一个 $1+x+x^2$；
- 没有任何面显示 0，所以 $P(0)=Q(0)=0$：每颗骰子各得一个因式 x；
- 余下唯一的自由，是两个 $1-x+x^2$ 如何分配。每颗各得一个，就是普通骰子。两个都归同一颗骰子，则得

$$x(1+x)(1+x+x^2)=x+2x^2+2x^3+x^4,$$
$$x(1+x)(1+x+x^2)(1-x+x^2)^2=x+x^3+x^4+x^5+x^6+x^8.$$

这就是 Sicherman 骰子，1, 2, 2, 3, 3, 4 和 1, 3, 4, 5, 6, 8（图 9.6.1），而以上论证亦证明了它们是唯一的另一种可能。关键在于：多项式与整数一样，可以唯一地分解为不可约（irreducible）因式的乘积——数论与组合数学（combinatorics）在此相遇。

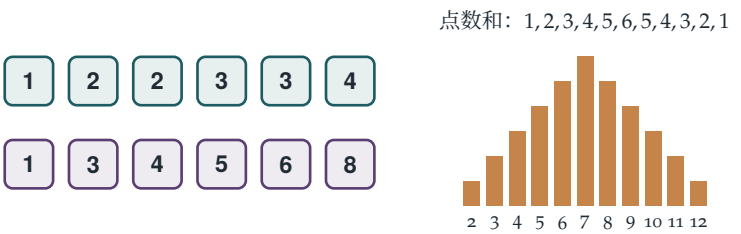


图 9.6.1　Sicherman 骰子。两颗一起掷，由 2 到 12 的每个点数和出现的频数，都与两颗普通骰子完全相同（右）：掷出 2 有一种方法，掷出 3 有两种方法，如此类推，直至掷出 7 有六种方法。

概率母函数

在概率论中，同一个想法以*概率母函数*（probability generating function）的形式出现：取值 $0,1,2,\ldots$ 的随机变量（random variable）X 有 $\mathrm{G}(t)=\mathrm{E}(t^X)=\sum \mathrm{P}(X=k)t^k$，而对于独立（independent）的 X 和 Y，$\mathrm{E}(t^{X+Y})=\mathrm{E}(t^X)\mathrm{E}(t^Y)$。独立随机变量之和，对应于母函数之积。STEP III 2003 Q14（见练习）就是用这个方法找出 Sicherman 骰子，

而 Section 9.13 则用一个相关的论证，说明无论怎样在两颗骰子上做手脚，都不能令所有点数和出现的机会均等。

思维技巧

把计数问题编码成多项式

当物件以数值相加的方式组合时，把每个物件表示成 x 的多项式。组合变成相乘，而有关组合的问题就变成有关因式分解的问题。

在这里，一个关于骰子的问题化成了多项式的唯一分解（unique factorisation）。在大学里，这类编码随处可见：在组合数学中，在概率论中，也在代数中——在代数里，多项式本身就是研究对象，它们构成一个环（ring），其算术与整数的算术相似。

延伸练习

1 **四面体骰子。**试做 STEP III 2003, 第 14 题（`step.maths.org/questions/03-s3-q14`），该题就四面骰子提出 Sicherman 的问题。（提示：$x + x^2 + x^3 + x^4 = x(1 + x)(1 + x^2)$。）

2 **验证列表。**展开 $(x + x^4 + x^5 + 3x^6)(2x^2 + x^3 + x^4 + 2x^5)$，并与题目中的列表对照。

3 **三颗骰子。**是否存在三颗六面骰子，各面标上正整数，不全是标准骰子，而其点数和的分布与三颗标准骰子相同？（提示：$(x + \cdots + x^6)^3$ 的因式现在包括三个 $1 - x + x^2$。）

4 **一枚硬币与一个八面体。**一颗标上 $1, \ldots, 8$ 的八面骰子，和一颗标上 $1, 2$ 的两面“骰子”（即一枚硬币）：能否用另一种方式重新标号，而点数和的分布保持不变？

延伸阅读

- **Martin Gardner, *Penrose Tiles to Trapdoor Ciphers*** (Mathematical Association of America)。收录了关于 Sicherman 骰子的那篇专栏文章。
- **Joseph Gallian & David Rusin, “Cyclotomic polynomials and nonstandard dice”**, *Discrete Mathematics* 27 (1979)。一般理论，适用于任意面数的骰子。
- **Herbert Wilf, *generatingfunctionology*** （可于网上免费阅览），第 1 章。

9.7 相等的幂和

题目

STEP I 1992，第 5 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/92-s1-q5`

引子

把 1 至 16 这些数分成两队，每队八个：

$$\{1,4,6,7,10,11,13,16\} \quad 和 \quad \{2,3,5,8,9,12,14,15\}.$$

两队的和都是 68。两队的平方和都是 748。两队的立方和都是 9248。很难相信这不是纯粹的巧合，但它其实是一个无穷无尽的规律的第三步：只要用正确的方法把 1 至 32 分开，连四次方之和也会相等。

这个分法由 Eugène Prouhet 于 1851 年发现，而它背后的规则，即 *Thue–Morse 数列*（Thue–Morse sequence），出现的地方多得惊人：数论、计算理论、国际象棋，以至最公平的轮流方式。本题通过一个多项式（polynomial）及其导数（derivative）把它发掘出来。

动笔之前

想一想 若多项式 $p(x) = \sum c_k x^k$ 满足 $p(1) = 0$，这对各系数（coefficient）有甚么含意？若 $p'(1) = 0$ 又如何？

$p(1) = \sum c_k$，所以 $p(1) = 0$ 表示各系数之和为零。而 $p'(x) = \sum k c_k x^{k-1}$，所以 $p'(1) = \sum k c_k$。若所有系数都是 ± 1，这两个条件就是说：$+1$ 与 -1 所在的位置数目相同，而且位置之和也相等。

逐步解题

阶段 1　一个三重根

想一想　把 $1-x^2$ 和 $1-x^4$ 因式分解。当中共出现多少个因式 $1-x$?

由于 $1-x^2=(1-x)(1+x)$ 及 $1-x^4=(1-x)(1+x)(1+x^2)$,

$$p_0(x)=(1-x)^3(1+x)^2(1+x^2),$$

所以 $(1-x)^3$ 是一个因式（factor）。 ■

阶段 2　在 1 处的导数

想一想　若 $(1-x)^3$ 整除 p_0，那么 $(1-x)$ 的多少次方整除 p_0'? 接着 $p_1=xp_0'$ 和 p_1' 又如何?

写成 $p_0(x)=(1-x)^3q(x)$。根据积法则（product rule），$p_0'(x)=-3(1-x)^2q(x)+(1-x)^3q'(x)$，它有因式 $(1-x)^2$。所以 $p_0'(1)=0$。于是 $p_1(x)=xp_0'(x)$ 同样有因式 $(1-x)^2$，而用相同的论证可知 p_1' 有因式 $(1-x)$，所以 $p_1'(1)=0$。 ■

阶段 3　读出各系数

展开可得 $(1-x)(1-x^2)=1-x-x^2+x^3$，然后

$$p_0(x)=1-x-x^2+x^3-x^4+x^5+x^6-x^7.$$

想一想　以系数 c_k 表示，$p_0'(1)$ 和 $p_1'(1)$ 分别是甚么?

由于 $p_1(x)=xp_0'(x)=\sum kc_kx^k$，我们有 $p_1'(1)=\sum k^2c_k$。系数在 $k=0,3,5,6$ 时为 $+1$，在 $k=1,2,4,7$ 时为 -1。所以 $p_0'(1)=0$ 和 $p_1'(1)=0$ 即是说

$$0+3+5+6=1+2+4+7\quad(=14),$$
$$0^2+3^2+5^2+6^2=1^2+2^2+4^2+7^2\quad(=70),$$

略去 0 之后，这正是所需的恒等式。 ■

阶段 4　再多一个因式

> **停一停，试一试** 要令立方和也平衡，我们需要第四个因式 $(1-x)$。应该用哪一个多项式取代 p_0?

取 $\mathrm{P}(x)=(1-x)(1-x^2)(1-x^4)(1-x^8)$，它有因式 $(1-x)^4$。它的系数来自 $\mathrm{p}_0(x)-x^8\mathrm{p}_0(x)$：先是 0 至 7 的正负号规律，然后在 8 至 15 重复同一规律，但正负号全部相反：

k	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
c_k	+	−	−	+	−	+	+	−	−	+	+	−	+	−	−	+

把第 2 阶段的论证再用一次，可知对于 $r=0,1,2,3$ 都有 $\sum c_k k^r=0$。取 $+$ 号的位置 $\{0,3,5,6,9,10,12,15\}$ 和 $-$ 号的位置 $\{1,2,4,7,8,11,13,14\}$。当 $r=1,2,3$ 时，来自 0 的一项为零，所以可以略去。因此以下的次序

$$a_1,\dots,a_8=1,2,4,7,8,11,13,14,\qquad a_9,\dots,a_{15}=3,5,6,9,10,12,15$$

符合要求：两半的和都是 60，平方和都是 620，立方和都是 7200。 ■

回顾

- **导数能侦测重根（repeated root）。**因式 $(1-x)^m$ 会迫使首 $m-1$ 阶导数在 $x=1$ 处全部为零。
- **先乘以 x 再求导**把 $\sum kc_k$ 变成 $\sum k^2c_k$，如此类推便得到更高次方。
- **倍增。**每多一个因式 $(1-x^{2^m})$，长度便增加一倍，同时多一个次方达到平衡。

背后的数学

Thue–Morse 数列

$(1-x)(1-x^2)(1-x^4)(1-x^8)\cdots$ 各系数的正负号构成 *Thue–Morse* 数列。以 0 代表 $+$、1 代表 $-$，写出来就是：

$$0\,1\,1\,0\,1\,0\,0\,1\,1\,0\,0\,1\,0\,1\,1\,0\ \dots$$

它有三种等价的描述。

- **倍增：**由 0 开始，不断把手上已有的部分的补（complement，即把 0 与 1 互换）接在后面：$0\to 01\to 0110\to 01101001\to\cdots$。
- **二进制数字：**把 k 写成二进制（binary）时，若当中 1 的个数是偶数，第 k 项便是 0；若是奇数，便是 1。（这就是 x^k 的正负号为 $(-1)^{1\text{的个数，取自}k}$ 的原因：要从乘积中得出 x^k，就要选取那些在二进制中组成 k 的因子 x^{2^m}，而每一个都带来一个负号。）

► **多项式：** $\prod_m (1 - x^{2^m})$ 的各系数，一如题目所示。

按 Thue–Morse 数列把 $0, 1, \ldots, 2^N - 1$ 分开，所得的两个集合对于每个 $r < N$ 都有相等的 r 次方之和（图 9.7.1）。这就是 *Prouhet–Tarry–Escott 问题*（Prouhet–Tarry–Escott problem）：找出两个不同的整数集合，使它们的 r 次方之和在 $r = 1, \ldots, k$ 时全部相等。Prouhet 的解用了 2^{k+1} 个数；究竟最少需要多少个数，是一个著名的未解决问题。

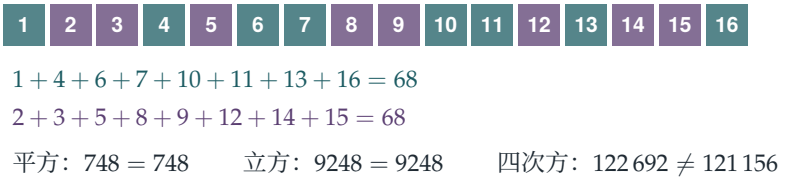


图 9.7.1　按 Thue–Morse 数列分开的 1 至 16。两个集合的和相等、平方和相等、立方和也相等，但四次方之和并不相等；要做到这一点，便需要 1 至 32。由左至右读出颜色，得到的规律是蓝绿–紫–紫–蓝绿、紫–蓝绿–蓝绿–紫、...。

最公平的轮流方式

两个小孩轮流从一堆糖果中挑选，最好的先被选走。若他们以 $ABABAB\ldots$ 的次序交替，那么每一对糖果中较好的一颗总是归 A，这是一种系统性的优势。$ABBA$ 的次序较为公平，而 $ABBABAAB$ 更加公平：这正是 Thue–Morse 数列，而次方和平衡恰恰就是个中原因。若糖果的价值由其位置的一个光滑函数给出，把该函数展开成多项式，便可看到 Thue–Morse 次序会把它的首几项抵销。同样的想法曾被提议用于足球的互射十二码（足球规则制定者自 2017 年起试行的“$ABBA$”次序），而网球决胜局的发球次序亦采用这个方式。

Thue、Morse 与 Euwe

这个数列曾多次被人发现。Prouhet 在 1851 年用过它。Axel Thue 在 1912 年研究没有重复规律的数列时重新发现了它：它永远不会连续三次出现同一段（即它是“无立方”（cube-free）的）。Marston Morse 在 1921 年研究曲面上的测地线（geodesics）时再次发现它。而国际象棋世界冠军尤伟（Max Euwe）在 1929 年用它证明了，当时生效的重复局面和棋规则并不能防止对局无限延续下去。时至今日，它是*自动数列*（automatic sequences）理论中的核心例子；自动数列是指由简单机器读取 n 的二进制数字而计算出来的数列。

思维技巧

把组合性质编码成多项式

次方和是否平衡，关键在于 $+$ 号和 $-$ 号落在哪些位置。以这些正负号作为系数的多项式，可把这个性质转化为一个关于根的命题。

本题从头到尾都没有要求你去搜索：$x=1$ 处的一个重根已经包办了所有工作。利用代数来构造组合（combinatorial）对象，而不是靠反复试验去找，是大学组合数学中一再出现的想法。

延伸练习

1 **平移。**试做 STEP II 2025, 第 8 题（载于 OCR 提供下载的 2025 年试卷，`www.ocr.org.uk/Images/750486-2025-past-papers.zip`）。该题说明怎样把一个令和与平方和平衡的分法，通过平移其中一列再合并，变成一个更大而且连立方和也平衡的分法。

2 **平方的平方。**（同样出自 STEP II, 2025。）集合 $A=\{1,3,4,5,9,11\}$ 和 $B=\{2,6,7,8,10\}$ 的和相等，平方和也相等。对于每个正整数 n，把 $\{n^2,(n+1)^2,\dots,(n+11)^2\}$ 分成两个和相等的集合。（提示：把 0 加入第二个集合。）

3 **平移不变性 (translation invariance)。**证明：若两个数集的 r 次方之和在 $r=0,1,\dots,k$ 时全部相等，那么把每个元素都加上同一个数 c 所得的两个集合亦然。为甚么这表示题目中的集合 $\{1,\dots,15\}$ 同样可以换成任意 16 个连续整数？

4 **公平分糖果。**糖果的价值为 $1,\frac{1}{2},\frac{1}{3},\dots,\frac{1}{8}$。比较 A 在 $ABABABAB$ 和 $ABBABAAB$ 两种次序下所得的总值。

延伸阅读

- **Jean-Paul Allouche & Jeffrey Shallit, “The ubiquitous Prouhet–Thue–Morse sequence”**, 载于 *Sequences and Their Applications*（Springer, 1999）。综述这个数列在各处的出现；可在网上免费阅读。
- **Matt Parker, *Things to Make and Do in the Fourth Dimension***（Penguin）。内容包括等次方和及其他数字趣闻。
- **Joshua Cooper & Aaron Dutle, “Greedy Galois games”**, *American Mathematical Monthly* 120 (2013)。公平轮流与 Thue–Morse 数列。

9.8 没有人拿到自己的恤衫

题目

STEP II 2016，第 12 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/16-s2-q12`

引子

一支十一人的足球队匆忙更衣，各人随手抓起一件球衣就穿。没有任何人穿上自己球衣的机会有多大？再设想那是一支五十人的大军，甚至是一百万人的人群。人多了，出错的人也多了，机会是否随之上升？还是因为有更多人可能碰巧拿对，机会反而下降？出人意料的是，这个机会几乎完全不变：由五个人开始，它与 $1/e \approx 0.368$ 的差距不超过半个百分点。

这是组合概率（combinatorial probability）中最古老的问题，由 Pierre de Montmort 于 1708 年提出，称为 *problème des rencontres*（相遇问题）。本题以容斥原理（inclusion–exclusion principle）解决这个问题；容斥原理是计数互相重叠的事物的工具，而所得的答案，是 e 在数学中最令人意想不到的现身之一。

动笔之前

停一停，试一试 先试 $n = 3$：列出卡 $1, 2, 3$ 的全部六种次序，数一数其中有多少种是没有任何一张卡在自己位置上的。如果你有耐性，再试 $n = 4$（共有 24 种次序）。

当 $n = 3$ 时，只有 231 和 312 没有任何卡在自己的位置上，所以概率是 $\frac{2}{6} = \frac{1}{3}$。当 $n = 4$ 时，24 种次序中有 9 种符合条件，概率为 $\frac{3}{8} = 0.375$。没有任何一张卡留在原位的排列，称为错排（derangement）。

逐步解题

阶段 1　三个事件的容斥原理

想一想　怎样把 $A \cup B \cup C$ 写成两个事件的并集（union），使两个事件的公式可以派上用场？

把 $A \cup B$ 看成单一事件，对它和 C 应用公式：

$$\mathrm{P}(A \cup B \cup C) = \mathrm{P}(A \cup B) + \mathrm{P}(C) - \mathrm{P}((A \cup B) \cap C).$$

现在 $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$，再应用一次公式，得 $\mathrm{P}((A \cap C) \cup (B \cap C)) = \mathrm{P}(A \cap C) + \mathrm{P}(B \cap C) - \mathrm{P}(A \cap B \cap C)$。把 $\mathrm{P}(A \cup B)$ 也展开，再合并各项，便得出结果。 ■

四个事件时，规律依旧：把四个单一事件的概率相加，减去六个两两交集（intersection）的概率，加上四个三重交集的概率，再减去 $\mathrm{P}(A \cap B \cap C \cap D)$。

阶段 2　各个概率

想一想　如果卡 i 在位置 i，其余的卡有多少种排法？如果卡 i 和卡 j 都固定了呢？

共有 $n!$ 种机会均等的次序。如果卡 i 固定在位置 i，其余 $n-1$ 张卡可以任意排列：共 $(n-1)!$ 种排法。所以

$$\mathrm{P}(E_i) = \frac{(n-1)!}{n!} = \frac{1}{n}, \qquad \mathrm{P}(E_i \cap E_j) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)},$$
$$\mathrm{P}(E_i \cap E_j \cap E_k) = \frac{(n-3)!}{n!} = \frac{1}{n(n-1)(n-2)}.$$

一般而言，指定的 k 张卡全部在自己位置上的概率是 $\frac{(n-k)!}{n!}$。

阶段 3　加起来

想一想　在 $\mathrm{P}(E_1 \cup \cdots \cup E_n)$ 的容斥公式中，涉及 k 个事件交集的项有多少个？它们的总和是多少？

从这些事件中选出 k 个，有 $\binom{n}{k}$ 种方法，而每个交集的概率都是 $\frac{(n-k)!}{n!}$。所以各个 k 重项的总和是

$$\binom{n}{k}\frac{(n-k)!}{n!} = \frac{n!}{k!\,(n-k)!} \cdot \frac{(n-k)!}{n!} = \frac{1}{k!}.$$

除了 $k!$ 之外，全部约去。配上正负相间的符号，

$$\mathrm{P}(\text{最少有一张卡在自己的位置上}) = 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n+1}\frac{1}{n!}. \qquad \blacksquare$$

所以出现错排的概率是

$$d_n = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n\frac{1}{n!} = \sum_{k=0}^{n}\frac{(-1)^k}{k!}.$$

阶段 4 恰好一张

想一想 要恰好有一张卡在原位，是哪一张卡？其余的卡又必须怎样？

先选定哪一张卡在自己的位置上（n 种选法）。卡 i 在原位的概率是 $\frac{1}{n}$；在这个条件下，其余 $n-1$ 张卡必须在它们之间形成一个错排，其概率为 d_{n-1}。所以

$$\mathrm{P}(\text{恰好一张}) = n \cdot \frac{1}{n} \cdot d_{n-1} = \sum_{k=0}^{n-1}\frac{(-1)^k}{k!}. \qquad \blacksquare$$

回顾

- **先多算，再修正。**把各个 $\mathrm{P}(E_i)$ 相加，会把有两张卡固定的排列算了两次；减去两两交集的项可以修正，但对三张卡固定的情况又减得太多；如此类推。
- **对称性令各项相等。**每个 k 重交集的概率都相同，所以公式的每一层都是一个数乘以一个二项式系数（binomial coefficient）。
- **化为较小的情况。**“恰好一张”变成了 $n-1$ 张卡的错排。

背后的数学

e 从何而来

级数 $\sum(-1)^k/k!$ 就是指数级数（exponential series）$e^x = \sum x^k/k!$ 在 $x=-1$ 时的值。所以当 $n \to \infty$ 时，

$$d_n \to e^{-1} \approx 0.3679,$$

而恰好有一张卡对上的概率也趋向同一个极限。收敛（convergence）的速度快得惊人（图 9.8.1）：在 $1/n!$ 一项停下来所产生的误差小于 $1/(n+1)!$，所以在 $n=7$ 时，答案已经与 $1/e$ 吻合至小数点后四位。事实上，n 件物件的错排数目 $n!\,d_n$ 恰好是最接近 $n!/e$ 的整数。

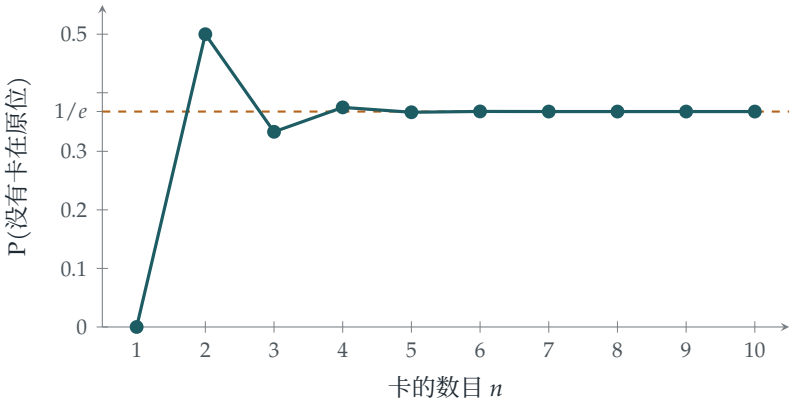


图 9.8.1　把 n 张卡随机洗匀后，没有任何一张卡留在自己位置上的概率。它在 $1/e$ 的上下交替出现，而由 $n = 6$ 开始，在这个比例尺下已经与 $1/e$ 无法分辨。

对上的数目服从泊松分布

结论还不止于此。设 N 为在自己位置上的卡的数目。那么对每个 k，都有 $\mathrm{P}(N = k) \to \frac{e^{-1}}{k!}$：在极限之下，对上的数目服从平均值为 1 的泊松分布（Poisson distribution）。平均值为甚么是 1，不用任何计算也看得出来：N 是 $E_1, \ldots, E_n$ 的指示变量（indicator）之和，每个的期望值（expectation）都是 $\frac{1}{n}$，而即使这些事件并非独立，期望值仍然可以相加。STEP III 2004 Q12（见练习）证明方差（variance）同样是 1。每当有许多机会很微、又近乎独立的事件时，泊松分布就会出现；错排是一个经典例子，也是第 12 章 Section 12.11 的主题。

Montmort 的 *treize* 游戏

Pierre Rémond de Montmort 是一位法国贵族，他在著作 *Essay d'analyse sur les jeux de hazard*（1708）中研究各种博彩游戏。在 *treize*（十三）这个游戏中，玩家派出十三张牌，同时口中数着“一、二、三……”，只要有一次派出的牌与所数的数字相同，便算胜出。Montmort 求出了胜出的概率，并在与尼古拉·伯努利（Nicolaus Bernoulli）的通信中，于 1713 年或之前确立了一般公式。欧拉（Euler）在 1753 年重新发现了这条公式。容斥原理其后由棣莫弗（Abraham de Moivre）加以形式化，而它的一般形式则由西尔维斯特（James Joseph Sylvester）在十九世纪建立。

思维技巧

容斥原理

要计数互相重叠的集合的并集，先把各集合的大小相加，减去两两重叠的部分，加上三重重叠的部分，如此类推。

当对称性令同一大小的重叠部分全部相等时，这个方法最见威力，本题正是如此。在大学里，它会在组合数学（combinatorics）、概率论和数论（number theory）中出现；在数论中，它用来计数不能被一系列质数中任何一个整除的整数，并由此引出欧拉函数（Euler's totient function）和默比乌斯函数（Möbius function）。

延伸练习

1 **球衣。**试做 STEP III 2004，第 12 题（`step.maths.org/questions/04-s3-q12`）。这道题用指示变量计数穿上自己球衣的球员人数，求出平均值和方差，并把结果与泊松分布比较。

2 **递推关系。**设 D_n 为 n 件物件的错排数目。考虑卡 1 去了哪个位置，以及原本在该位置的卡是否去了位置 1，证明 $D_n = (n-1)(D_{n-1} + D_{n-2})$。用 $D_3 = 2$、$D_4 = 9$ 验算。

3 **恰好 k 张。**证明恰好有 k 张卡在自己位置上的概率是 $\frac{1}{k!}d_{n-k}$，并且它趋向 $\frac{e^{-1}}{k!}$。

4 **互质的整数。**用容斥原理计数由 1 至 30 的整数中，不能被 2、3 和 5 任何一个整除的有多少个。你的答案与 $30\left(1-\frac{1}{2}\right)\left(1-\frac{1}{3}\right)\left(1-\frac{1}{5}\right)$ 有甚么关系？

延伸阅读

- **Geoffrey Grimmett & Dominic Welsh, *Probability: An Introduction***（Oxford University Press）。大学一年级教科书，涵盖容斥原理和配对问题。
- **William Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1**（Wiley），第 IV 章。经典之作，载有配对问题的许多变化。
- **Miklós Bóna, *A Walk Through Combinatorics***（World Scientific），关于容斥原理的部分。

9.9 生日问题

题目

STEP II 2007，第 13 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/07-s2-q13`

引子

一个房间里要有多少人，才会令其中两人同一天生日的机会大于一半？大多数人会猜 180 左右，即 365 的一半。答案却是 23。在一班三十人的课堂里，有人同一天生日的概率（probability）是 70%；在五十人的派对上，则是 97%。

这个结果令人意外，是因为我们把两个不同的问题混为一谈：“有没有人与*我*同一天生日？”（这确实需要大约 253 人）和“有没有*任何两个人*同一天生日？”。23 个人可以组成 $\binom{23}{2} = 253$ 对，而每一对都是一次出现相同生日的机会。这道题把这个直觉变得精确，而同样的数学亦决定了保护你银行户口的密码技术有多安全。

动笔之前

想一想 为甚么计算*没有*两位宾客同一天生日的概率会比较容易？请宾客排成一行：第二位避开第一位生日的机会是多少？第三位同时避开前两位的机会又是多少？

“至少有两人相同”包含许多互相重叠的情况（两人相同、三人相同、两对不同的人各自相同……）。它的补事件（complement）“全部不同”却是单一而简洁的事件。第二位宾客避开第一位生日的概率是 $\frac{364}{365}$；在此条件下，第三位同时避开前两位的概率是 $\frac{363}{365}$；如此类推。

逐步解题

阶段 1　一个有用的近似

想一想　$y = e^{-x}$ 在 $x = 0$ 处的切线（tangent line）是甚么？

由于 $e^{-x} = 1 - x + \frac{x^2}{2} - \cdots$，当 x 很小时，我们有 $e^{-x} \approx 1 - x$。取 $x = r/n$，得

$$\frac{n-r}{n} = 1 - \frac{r}{n} \approx e^{-r/n}. \qquad \blacksquare$$

误差约为 $\frac{r^2}{2n^2}$，当 r 远小于 n 时，这个误差很小。

阶段 2　有人同一天生日的概率

k 个生日全部不同的概率是

$$\frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdots \frac{365-(k-1)}{365} = \prod_{r=0}^{k-1} \frac{365-r}{365}.$$

对每个因子使用上述近似（approximation），再利用指数函数相乘时指数相加这个性质，

$$\prod_{r=0}^{k-1} e^{-r/365} = \exp\Big(-\frac{0+1+\cdots+(k-1)}{365}\Big) = \exp\Big(-\frac{k(k-1)}{730}\Big).$$

因此，至少有一对人同一天生日的概率约为 $1 - e^{-k(k-1)/730}$。 $\blacksquare$

想一想　看看指数：$\frac{k(k-1)}{730} = \binom{k}{2}\Big/365$。用文字表达，这说明了甚么？

它是宾客的配对数目，乘以某一对人同一天生日的概率 $\frac{1}{365}$：也就是生日相同的配对数目的期望值（expected number）。这个近似告诉我们，*没有*任何配对相同的机会是 $e^{-(\text{相同配对的期望数目})}$，这又一次透露出泊松分布（Poisson distribution）的影子，正如 Section 9.8 所见。

阶段 3　二十三

我们需要 $1 - e^{-k(k-1)/730} \geq \frac{1}{2}$，即 $e^{-k(k-1)/730} \leq \frac{1}{2}$，亦即

$$\frac{k(k-1)}{730} \geq \ln 2 \approx \frac{253}{365}, \qquad \text{即} \qquad k(k-1) \geq 506.$$

由于 $22 \times 21 = 462$ 而 $23 \times 22 = 506$，最小的值是 $k = 23$。 $\blacksquare$

（23 位宾客时的精确概率是 0.507，22 位宾客时是 0.476，所以这个近似给出了正确答案。）

阶段 4　与主人同一天生日

想一想 对每位宾客来说，不与主人同一天生日的概率是多少？在这方面，各位宾客之间是否互相独立（independent）？

每位宾客各自独立地以概率 $\frac{364}{365} \approx e^{-1/365}$ 避开主人的生日。因此 k 位宾客中无人与主人同一天生日的概率是 $\left(\frac{364}{365}\right)^k \approx e^{-k/365}$，而我们需要

$$e^{-k/365} \le \tfrac{1}{2}, \qquad \text{即} \qquad k \ge 365 \ln 2 \approx 253.$$

所以必须有 253 位宾客。 ■

这个对比正是重点所在：要与某一个指定的人生日相同，需要 253 位宾客；但要有某一对人生日相同，只需 23 人，因为 23 个人可以组成 253 对。

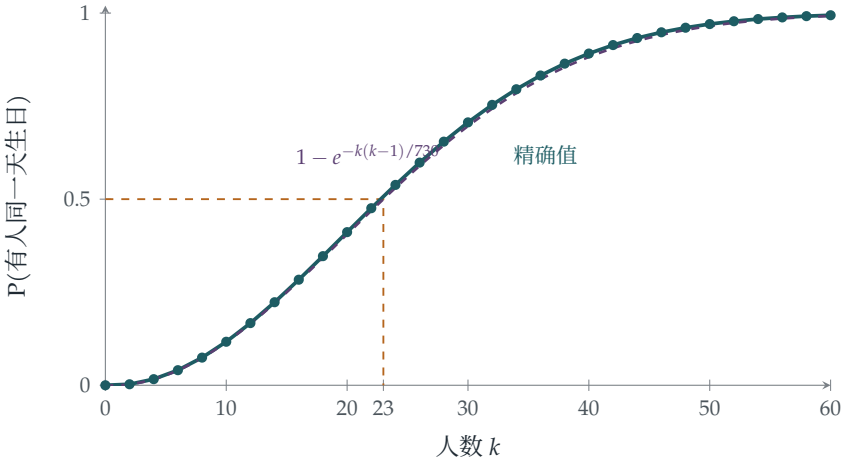


图 9.9.1 k 个人中至少有两人同一天生日的概率：精确值（蓝绿色）与近似值 $1-e^{-k(k-1)/730}$（紫色虚线）。两者非常接近，并且都在 $k=22$ 与 $k=23$ 之间越过 $\frac{1}{2}$。

回顾

- **利用补事件。**“全部不同”只是一个乘积；“至少有一对相同”却是一团乱麻。
- **先线性化，再取指数。**把每个因子 $1-x$ 换成 e^{-x}，乘积就变成了和。
- **数配对。**指数正是配对的数目乘以一对人相同的机会。$k(k-1)$ 中的平方，就是这个悖论的全部解释。

背后的数学

平方根定律

若把 365 天换成 n 个等可能的日子，同样的论证给出

$$\mathrm{P}(\text{有人相同，人数为}k) \approx 1 - e^{-k^2/(2n)},$$

所以只要 k 达到 $\sqrt{n}$ 左右，而不是 n，就很可能出现相同的情况。更准确地说，概率在 $k \approx 1.18\sqrt{n}$ 时达到 $\frac{1}{2}$，而在首次出现重复之前你需要遇到的人数，期望值约为 $\sqrt{\pi n/2}$（当 $n = 365$ 时，约为 24.6）。这条*平方根定律*（square-root law）正是巧合如此常见的原因：在任何一大批事件之中，总很可能有某一对在某方面吻合，尽管任何一个指定的吻合都不大可能发生。

生日攻击

电脑会计算文件的*杂凑值*（hash），以检查文件有否被窜改：那是一个指纹，可能的取值有（比方说）$n = 2^{128}$ 个。伪造者若想得到两份杂凑值相同的文件（一份无害的合约和一份欺诈的合约），并不需要试 2^{128} 份文件。根据生日问题的论证，为每份文件各自产生大约 $\sqrt{2^{128}} = 2^{64}$ 个变体，就很可能得到一对相同的杂凑值。这种*生日攻击*（birthday attack）正是现代杂凑函数采用 256 位元或以上的原因：有一半的位元因平方根定律而白白损失。类似的计算亦决定了随机识别码必须有多长，才能令两部电脑永远不会碰巧选中同一个。

谁最先提出?

生日问题通常归功于冯·米塞斯（Richard von Mises），他在 1939 年讨论过这个问题，不过它在此之前已经流传。据说 Harold Davenport 亦更早知道这个问题，只是觉得它太简单而没有发表。按 Warren Weaver 的说法，他在第二次世界大战期间喜欢在军官的晚宴上示范这个问题：他打赌在场二十二人之中有两人同一天生日，结果输了，直至一位女侍应透露她与其中一位军官同一天生日。

思维技巧

计算补事件；小心地近似

当“至少一个”很复杂时，就去求“一个也没有”的概率。当乘积难以处理时，把每个因子近似为指数函数，然后把指数相加。

这两个做法在大学里都是标准手法。第二个，即 $1 - x \approx e^{-x}$，是泊松近似理论的起点，也是渐近方法（asymptotic methods）的起点；后者用来估计那些复杂得

无法精确计算的量。

延伸练习

1 **首次重复。**试做 STEP II 2014，第 13 题（`step.maths.org/questions/14-s2-q13`）。这道题研究随机数产生器（random number generator）首次出现重复的值，其实是换了个面貌的生日问题，并把有关概率化成两个出人意表的恒等式（identities）。

2 **三人相同。**大约需要多少人，才会令某三人同一天生日的机会大于一半？（提示：同一天生日的三人组数目的期望值是 $\binom{k}{3}/365^2$。令它约等于 $\ln 2$。）真正的答案是 88。

3 **差之毫厘。**需要多少人，才会令两人的生日在同一天或相邻两天的机会达到一半？

4 **不均匀的生日。**现实中的生日并非均匀（uniform）分布。证明若各日的概率为 $p_1, \ldots, p_{365}$，则两个人同一天生日的机会是 $\sum p_i^2$，并且当所有 p_i 相等时这个值最小。（所以不均匀只会令相同的情况更容易出现。）

延伸阅读

- **William Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1**（Wiley），第 II 章。生日问题及其各种近似。
- **Persi Diaconis & Frederick Mosteller, "Methods for studying coincidences"**，*Journal of the American Statistical Association* 84 (1989)。解释巧合为何常见；非常易读。
- **David Hand, *The Improbability Principle***（Bantam）。平方根定律，以及不大可能的事情为何时刻都在发生的其他原因。

9.10 总能找赎

题目

STEP I 2011，第 12 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/11-s1-q12`

引子

这个规律实在叫人难以抗拒：$\frac{m}{m+1}$、$\frac{m-1}{m+1}$、$\frac{m-2}{m+1}$。若有 n 个人拿着 £2 硬币，答案想必就是 $\frac{m-n+1}{m+1}$。事实正是如此，而最巧妙的证明（proof）用上了组合数学（combinatorics）中数一数二优雅的想法：*反射原理*（reflection principle）。它把图形的一部分沿一条直线反射，从而将一个困难的计数问题（从不缺零钱找赎的队伍有多少种）变成一个容易的计数问题。

同一个问题有许多不同的面貌：点票过程中胜出者从未落后的选举（1887 年的伯特兰投票问题（Bertrand's ballot problem））、永不破产的赌徒、正确配对的括号。当 $m = n$ 时，数出来的就是*卡塔兰数*（Catalan numbers），这大概是组合数学中除二项式系数（binomial coefficients）以外最常出现的数列（sequence）。

动笔之前

想一想 每位拿 £1 的顾客会给你一枚硬币；每位拿 £2 的顾客会取走一枚（作为找赎）。你甚么时候会无法继续？

留意你手上 £1 硬币的数目。每来一位 £1 顾客，数目加 1；每来一位 £2 顾客，数目减 1。你无法继续的情况，恰好就是一位 £2 顾客来到而你手上没有硬币：数目会变成 -1。所以你能成功，当且仅当（if and only if）在队伍中的每一刻，到目前为止的 £1 顾客人数都不少于到目前为止的 £2 顾客人数。

逐步解题

阶段 1　一枚 £2 硬币

当 $n = 1$ 时，只有 £2 顾客排第一才会失败。£2 顾客排在 $m+1$ 个位置中任何一个的可能性均等，所以

$$\mathrm{P}(\text{成功}) = 1 - \frac{1}{m+1} = \frac{m}{m+1}. \qquad \blacksquare$$

阶段 2　两枚 £2 硬币

> **想一想**　第一个人必须拿着 £1 硬币。第二个人可以出现甚么情况？在每种情况下，你余下面对的是甚么局面？

第一个人拿着 £1 硬币的概率（probability）是 $\frac{m}{m+2}$；否则你立即失败。这时你手上有一枚硬币，而余下的队伍有 $m-1$ 个 £1 的人和两个 £2 的人。

- 若第二个人拿着 £1 硬币（概率为 $\frac{m-1}{m+1}$），你手上便有两枚硬币，无论两枚 £2 硬币何时出现都足够找赎。成功是必然的。
- 若第二个人拿着 £2 硬币（概率为 $\frac{2}{m+1}$），你手上又变回没有硬币，余下 $m-1$ 个 £1 的人和一个 £2 的人。根据第 (i) 部分，以 $m-1$ 代替 m，你成功的概率是 $\frac{m-1}{m}$。

所以

$$\mathrm{P}(\text{成功}) = \frac{m}{m+2}\left(\frac{m-1}{m+1} + \frac{2}{m+1}\cdot\frac{m-1}{m}\right) = \frac{m}{m+2}\cdot\frac{m-1}{m+1}\cdot\frac{m+2}{m} = \frac{m-1}{m+1}. \qquad \blacksquare$$

> **停一停，试一试**　试用同样的方法处理 $n = 3$。这是行得通的，但情况会愈分愈多。有没有方法可以一次过数出所有成功的队伍？

阶段 3　把队伍看成路径

> **想一想**　以你手上的硬币数目对已服务的人数作图。一条队伍看起来是甚么样子？一条成功的队伍又是甚么样子？

一条队伍就是一条路径，由 m 步上升（£1）和 n 步下降（£2）组成，从高度 0 出发，终于高度 $m-n$。这样的路径共有 $\binom{m+n}{n}$ 条，每条出现的可能性均等。一条队伍成功，恰好就是它的路径从不低于 0。所以我们需要数出坏路径，即那些触及高度 -1 的路径。

阶段 4　反射

想一想 取一条坏路径。找出它第一次触及 -1 的时刻，把路径在这一刻之前的部分沿直线 $y=-1$ 反射。新的路径从哪里开始，又在哪里结束?

反射后的路径从 -2 出发（即 0 对直线 $y=-1$ 的镜像），终点与原来的路径相同，即 $m-n$（图 9.10.1）。反过来说，每条由 -2 到 $m-n$ 的路径都必定在某处穿过 $y=-1$（因为 $m \geq n$，它起点在这条线之下，终点在这条线之上）；把它第一次到达这条线之前的部分反射，便得回一条由 0 出发的坏路径。所以由 0 出发的坏路径，与由 -2 到 $m-n$ 的所有路径恰好一一对应。

一条以 $m+n$ 步由 -2 走到 $m-n$ 的路径共上升了 $m-n+2$，所以它有 $m+1$ 步上升和 $n-1$ 步下降：这样的路径有 $\binom{m+n}{n-1}$ 条。因此

$$\mathrm{P}(\text{失败}) = \frac{\binom{m+n}{n-1}}{\binom{m+n}{n}} = \frac{n!\,m!}{(n-1)!\,(m+1)!} = \frac{n}{m+1},$$

而

$$\mathrm{P}(\text{成功}) = 1 - \frac{n}{m+1} = \frac{m-n+1}{m+1}.$$

当 $n=3$ 时，这就是 $\frac{m-2}{m+1}$。 ■

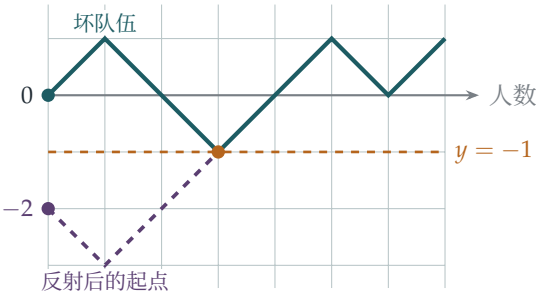


图 9.10.1 一条有 $m=4$ 枚 £1 硬币和 $n=3$ 枚 £2 硬币的队伍，画成一条路径（蓝绿色）。它在第三个人时首次到达 -1（琥珀色），这时你无法找赎。把路径在这一点之前的部分沿直线 $y=-1$ 反射，便得到一条由 -2 出发（紫色虚线）而终点相同的路径。坏队伍与由 -2 出发的路径之间的这种配对，就是反射原理。

回顾

- **找出合适的状态。**唯一要紧的是手上硬币的数目，这把队伍变成了一次游走。
- **以最初几步为条件。**第 (i) 和第 (ii) 部分的做法，是考虑最初几个人的情况，从而化约为较小的情形。
- **用双射（bijection）数出补集。**坏路径难以直接数出，但反射把它们与一个容易数算的集合配对起来。

背后的数学

卡塔兰数

当 $m = n$（£1 硬币刚好足够）时，成功队伍的数目是

$$\binom{2n}{n} - \binom{2n}{n-1} = \frac{1}{n+1}\binom{2n}{n},$$

即第 n 个卡塔兰数 C_n：$1, 1, 2, 5, 14, 42, 132, \ldots$。Richard Stanley 的著作 *Catalan Numbers* 列出了超过二百种由卡塔兰数计算数目的事物。其中包括：把 n 对括号正确排列的方法数目（“(())()”是允许的，“())(()”则不允许，理由正是上述的硬币论证，只要把“(”看成 £1、“)”看成 £2）、用互不相交的对角线把一个 $n+2$ 边形分割成三角形的方法数目，以及为 $n+1$ 个因子的乘积加上括号的方法数目。

随机游走

队伍就是随机游走（random walk），而反射原理是研究随机游走的基本工具之一。举例说，它可以证明由 0 出发的公平随机游走必定终会返回 0，也可以证明它在最初 $2n$ 步一直保持严格为正的概率是 $\frac{1}{2}\binom{2n}{n}2^{-2n} \approx \frac{1}{2\sqrt{\pi n}}$。在连续的形式下，把同样的反射论证应用于布朗运动（Brownian motion），便可得出其最大值的分布（distribution）。这是某些金融期权定价的基础，这类期权的回报取决于股价有否越过某个界限。

伯特兰投票问题

1887 年，约瑟夫·伯特兰（Joseph Bertrand）提出：在一场选举中，候选人 A 得 a 票，B 得 $b < a$ 票，那么 A 在整个点票过程中一直严格领先的概率是多少？答案是 $\frac{a-b}{a+b}$，同年由 Désiré André 证明，他的论证通常被视为反射原理（不过历史学家指出，André 实际的证明用的是一个稍有不同的双射）。我们的抽奖券问题是一个容许平手的变体：成功的意思是 £1 的数目从不低于 £2 的数目。其中一道练习会由我们的答案推导出伯特兰的答案。

思维技巧

双射计数

要数算一个集合，就找出它与一个你懂得数算的集合之间的一一对应（*one-to-one correspondence*）。

反射把坏队伍与由另一个起点出发、不受限制的路径配对起来。双射是组合数学中最令人满足的证明，因为它们解释了两个数为何相等，而不只是证明它们相等。在大学里，你会在组合数学中处处遇到双射证明，也会在研究随机游走和布朗运动时再次遇上反射原理。

延伸练习

1 **分情况处理三枚。**仿照第 (ii) 部分，以队伍中最初几个人为条件来证明第 (iii) 部分，看看反射原理省却了多少工夫。

2 **伯特兰投票问题。**在一场选举中，A 得 a 票，B 得 $b < a$ 票。证明 A 在整个点票过程中一直严格领先的概率是 $\frac{a-b}{a+b}$。（提示：第一票必须投给 A；此后，A 落后于 B 的票数绝不可多于…。）

3 **括号。**列出三对括号的 5 种正确排列，以及把五边形分割成三角形的 5 种方法。你能找出两者之间的一个双射吗？

4 **一张 £5 纸币。**一条队伍中有 m 个人拿着 £1 硬币，n 个人拿着 £5 纸币，而一张票的价钱是 £1。解释为何反射原理不再给出简单的答案。（这个推广困难得多。）

延伸阅读

- **William Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1**（Wiley），第 III 章。有关随机游走和反射原理的经典论述。
- **Richard Stanley, *Catalan Numbers***（Cambridge University Press）。超过二百种诠释，并附练习。
- **Marc Renault, "Lost (and found) in translation: André's actual method and its application to the generalized ballot problem"**，*American Mathematical Monthly* 115 (2008)。反射原理的历史。

9.11 集齐一套

题目

STEP II 1994，第 13 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/94-s2-q13`

引子

一本足球贴纸簿要集齐 600 张贴纸，贴纸每次随机发售一张。要贴满整本贴纸簿，你预计要买多少张贴纸？最初几百张很快到手，因为几乎每一张都是新的。但贴纸簿愈贴愈满，新贴纸就愈来愈罕有，单是最后一张，平均就要买 600 次。总数大约是 4200 张，是贴纸簿大小的七倍。

这就是*集券问题*（coupon collector's problem）。它的答案结合了两个想法：等候一次成功所需的时间，以及调和级数（harmonic series）$1+\frac{1}{2}+\frac{1}{3}+\cdots$，后者的增长与 $\ln n$ 相若。你在 Section 2.3 见过调和级数发散（diverge）；在这里，它那缓慢的增长正正就是集齐一套收藏品所要付出的代价。

动笔之前

想一想 如果每包各自独立地以概率 p 开出一张新卡，那么直至开出第一张新卡为止（连这一包计算在内），你开包数目的期望值（expectation）是多少？

开包数目 X 服从*几何分布*（geometric distribution）：$\mathrm{P}(X=k)=(1-p)^{k-1}p$。它的期望值是 $\frac{1}{p}$。（有一个快捷的看法：$\mathrm{E}(X)=1+(1-p)\mathrm{E}(X)$，因为你先开一包，然后有 $1-p$ 的概率回到起点。）如果有三分之一的机会开出新卡，你预计要等三包。

逐步解题

阶段 1　等候一张新卡

手上已有 r 张不同的卡时，一包开出新卡的概率是 $\frac{n-r}{n}$，与之前各包独立。因此，直至开出下一张新卡为止，所需包数的期望值是

$$\frac{1}{(n-r)/n} = \frac{n}{n-r}. \qquad \blacksquare$$

阶段 2　比较和与积分

> **想一想**　在区间 $r \le x \le r+1$ 上，$\frac{1}{x}$ 介乎哪两个值之间？这对曲线下的面积有甚么启示？

由于 $\frac{1}{x}$ 是递减的，在 $[r, r+1]$ 上有 $\frac{1}{r+1} \le \frac{1}{x} \le \frac{1}{r}$。在长度为 1 的区间上积分，得

$$\frac{1}{r+1} \le \int_r^{r+1} \frac{\mathrm{d}x}{x} \le \frac{1}{r}. \qquad \blacksquare$$

用图来看（图 9.11.1）：曲线下的面积夹在高度为 $\frac{1}{r+1}$ 的长方形与高度为 $\frac{1}{r}$ 的长方形之间。把这些不等式由 $r = 1$ 加到 $r = n-1$，各个积分便连接成 $\int_1^n \frac{\mathrm{d}x}{x} = \ln n$，于是

$$\frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \le \ln n \le 1 + \frac{1}{2} + \cdots + \frac{1}{n-1}. \qquad \blacksquare$$

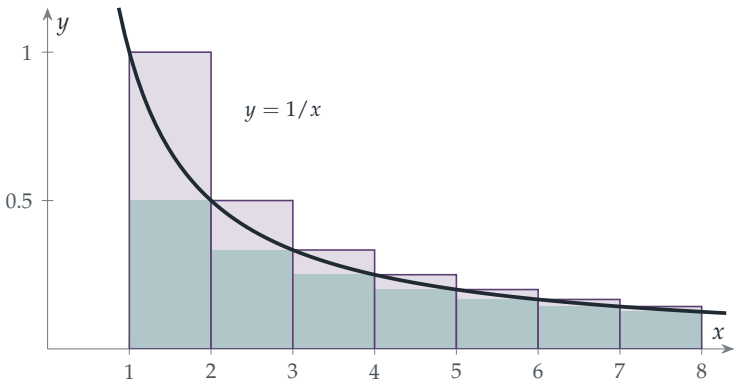


图 9.11.1　在每个区间 $[r, r+1]$ 上，$y = 1/x$ 之下的面积介乎蓝绿色长方形（高度 $\frac{1}{r+1}$）与较大的紫红色长方形（高度 $\frac{1}{r}$）之间。全部加起来，$\ln n$ 便介乎 $\frac{1}{2} + \cdots + \frac{1}{n}$ 与 $1 + \frac{1}{2} + \cdots + \frac{1}{n-1}$ 之间。曲线与紫红色长方形之间的空隙，加起来是一个有限的总和：欧拉常数（Euler's constant）$\gamma \approx 0.577$。

阶段 3　集齐全套

想一想 总包数是等候第一张卡、第二张新卡……各段等候时间之和。它的期望值是多少?

设 X_r 为由 r 张不同的卡增至 $r+1$ 张所需的包数。总数是 $X_0 + X_1 + \cdots + X_{n-1}$，而期望值可以相加，所以

$$\mathrm{E}(\text{总数}) = \sum_{r=0}^{n-1} \frac{n}{n-r} = n\Big(\frac{1}{n} + \frac{1}{n-1} + \cdots + 1\Big) = nH_n,$$

其中 $H_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n}$ 是第 n 个调和数（harmonic number）。由第 2 步，$\ln n + \frac{1}{n} \le H_n \le \ln n + 1$，所以当 n 很大时 H_n 约为 $\ln n$，而我预计大约要吃掉

$$n \ln n \text{ 包.} \qquad \blacksquare$$

回顾

- **把漫长的等候分成多个阶段。**总时间拆成 n 段等候时间，每一段都属于简单的几何分布。
- **期望值可以相加。**我们从头到尾都不需要总数的分布，只需要各部分的期望值。
- **比较和与积分。**那幅图把一个我们无法求值的和化成一个对数，误差至多为 1。

背后的数学

这个估计有多准确?

这个比较还可以再精确一些。图 9.11.1 中曲线与上方长方形之间的阴影空隙，总面积是有限的，所以 $H_n - \ln n$ 趋向一个极限，

$$\gamma = \lim_{n\to\infty} (H_n - \ln n) = 0.5772156649\ldots,$$

即欧拉常数。（没有人知道 γ 是否有理数。）因此包数的期望值是 $n\ln n + \gamma n + \frac{1}{2} + \cdots$。以 600 张贴纸的贴纸簿为例，$600 \times (\ln 600 + 0.577) \approx 4185$，这就是引子中所引述的数字。

时间都花到哪里去了? 各个阶段极不平均。集齐全套的前一半只需大约 $n\ln 2 \approx 0.69n$ 包，而单是最后一张卡，平均就要 n 包。大部分工夫都花在最后几张卡上，这正是收藏者要互相交换重复卡的原因。

集中现象

期望值只是个开始：总数的变动有多大？总数的方差（variance）是 $\sum nr/(n-r)^2$，约为 $\frac{\pi^2}{6}n^2$（Section 2.4 中的和 $\sum 1/k^2 = \pi^2/6$ 在这里出现了）。所以标准差（standard deviation）约为 $1.28n$，当 n 很大时远小于平均值 $n\ln n$：总数集中在其平均值附近。艾狄胥（Paul Erdős）和雷尼（Alfréd Rényi）在 1961 年证明了一个更精确的结果：总数减去 $n\ln n$ 再除以 n，会有一个极限分布（limiting distribution），即耿贝尔分布（Gumbel distribution），这个分布也描述了许多随机量之中最大的一个（Section 12.25）。

无处不在的集券问题

这个问题可追溯至棣莫弗（Abraham de Moivre，1712 年）和拉普拉斯（Laplace）。如今它以许多不同的面貌出现。要检查一个系统的每个组件，需要做多少次随机测试？在一个网络上进行随机漫步（random walk），要多久才会走遍每个节点？电脑的杂凑表（hash table）要探查多少次？每一种情况的答案都涉及 $n\ln n$。在生物学中，同样的计算可用来估计要有多少个定序“读段”（reads）才能覆盖整个基因组。

思维技巧

期望值的线性性质（linearity of expectation）

和的期望值等于期望值之和，不论各部分是否独立。把一个复杂的随机量拆成简单的部分。

这是概率论中最有用的事实之一。在这里，它把集券问题简化为 n 段几何等候；在 Section 9.8 中，它不用任何计算就给出了配对数目的期望值。到了大学，同一个想法推动着算法分析，以及组合数学中的概率方法（probabilistic method）。

延伸练习

1 **半套。**证明集齐 $\frac{n}{2}$ 张不同的卡（n 为偶数）所需包数的期望值是 $n(H_n - H_{n/2})$，并且这个值约为 $n\ln 2$。

2 **不同的卡。**证明开了 k 包之后，你手上不同的卡的数目的期望值是 $n\left(1-(1-\frac{1}{n})^k\right)$。（提示：对每一张卡，求出你拥有它的概率，然后相加。）大约要开多少包，才能得到全套的 95%？

3 **两位收藏者。**两个小孩收集同一套卡，并共用所开的各包，因此对其中一人来说是重复的卡，对另一人来说可能是新的。证明集齐两套所需包数的期望值约为 $n(\ln n + \ln\ln n)$。（这题较难；先说服自己这个数目远小于 $2n\ln n$。）

4 **更精确的界。**改用梯形法则（trapezium rule）的图像而不用长方形，证明 $H_n - \ln n$ 是递减的，而 $H_n - \ln n - \frac{1}{2n}$ 是递增的。由较小的 n 值推导出 $0.5 < \gamma < 0.7$。

延伸阅读

- **Geoffrey Grimmett & David Stirzaker, *One Thousand Exercises in Probability***（Oxford University Press）。集券问题的许多变化。
- **Julian Havil, *Gamma: Exploring Euler's Constant***（Princeton University Press）。关于 γ 和调和级数的一切，写得深入浅出。
- **Michael Mitzenmacher & Eli Upfal, *Probability and Computing***（Cambridge University Press）。集券问题及其在电脑科学中的应用。

9.12 随机子集

题目

STEP III 2019，第 12 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2019 年 STEP 试卷（zip 文件）内的 STEP III 试卷：

`www.ocr.org.uk/Images/696590-2019-past-papers.zip`

引子

完全随机地选出 $\{1,2,\ldots,100\}$ 的两个子集（subset）。它们没有任何共同元素的可能性有多大？直觉也许会告诉你：“相当大：大多数成对的集合都不相同”。事实上，这个概率是 $\left(\frac{3}{4}\right)^{100}$，约为 3×10^{-13}。随机集合很大（大约占全部元素的一半），所以它们几乎总会重叠。

这道题很短，当中唯一的想法却很有威力：不要把集合当作整体来想，而要每次只想一个元素。一个随机子集不过是 n 次独立（independent）的掷硬币，每个元素掷一次；于是，凡是关于随机子集的问题，都变成关于这些掷硬币结果的问题。

动笔之前

想一想 怎样用 n 条是非题的一串答案，来描述 $\{1,\ldots,n\}$ 的一个子集？

对每个元素 i，回答“i 在集合里吗？”这条问题。子集决定了答案，答案也决定了子集。每个元素有 2 个选择，所以一共有 2^n 个子集。均匀随机（uniformly at random）地选一个子集，就等于独立地掷一枚公平硬币来回答每条问题。

逐步解题

阶段 1 一个元素

按照掷硬币的描述，$\mathrm{P}(1 \in A_1) = \frac{1}{2}$。事实上，$2^n$ 个子集中恰好有一半包含 1：把每个不包含 1 的子集，与加入 1 之后所得的子集配成一对即可。

阶段 2　不相交的集合

> **想一想**　要使 $A_1 \cap A_2 = \varnothing$，每个元素 i 必须满足甚么条件？就单一个元素而言，这件事的概率是多少？为甚么可以把各个元素的概率相乘？

A_1 与 A_2 不相交（disjoint），当且仅当没有元素同时属于两者。对单一个元素 i 来说，有四种等可能的情况：两个集合都不在、只在 A_1、只在 A_2、两个都在（图 9.12.1 左）。只有最后一种是禁止的，所以 i 不同时属于两者的概率是 $\frac{3}{4}$。不同的元素由不同而且独立的掷硬币结果决定，所以

$$\mathrm{P}(A_1 \cap A_2 = \varnothing) = \left(\frac{3}{4}\right)^n. \qquad \blacksquare$$

对三个集合而言，元素 i 同时属于三者的概率是 $\frac{1}{8}$，所以 $\mathrm{P}(A_1 \cap A_2 \cap A_3 = \varnothing) = \left(\frac{7}{8}\right)^n$。一般而言，

$$\mathrm{P}(A_1 \cap \cdots \cap A_m = \varnothing) = \left(1 - \frac{1}{2^m}\right)^n. \qquad \blacksquare$$

阶段 3　子集

> **想一想**　要使 $A_1 \subseteq A_2$，一个元素的四种情况之中，哪一种是禁止的？

$A_1 \subseteq A_2$ 不成立，当且仅当有某个元素在 A_1 里而不在 A_2 里。同样是四种模式中有一种被禁止，所以

$$\mathrm{P}(A_1 \subseteq A_2) = \left(\frac{3}{4}\right)^n,$$

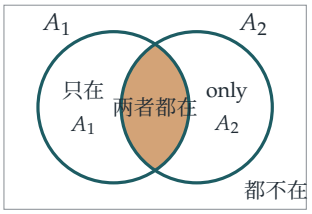
与不相交的情况相同。这并非巧合：$A_1 \subseteq A_2$ 当且仅当 A_1 与 A_2 的补集（complement）不相交，而随机集合的补集也是一个随机集合。

> **想一想**　要使 $A_1 \subseteq A_2 \subseteq A_3$，一个元素的 8 种归属模式（在或不在每个集合里）之中，哪些是容许的？

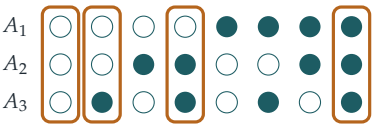
沿着 A_1, A_2, A_3 这条链（chain）走，一个元素一旦进入，就必须一直留在里面。所以容许的模式是“不在、不在、不在”、“不在、不在、在”、“不在、在、在”和“在、在、在”：8 种之中的 4 种（图 9.12.1 右）。因此 $\mathrm{P}(A_1 \subseteq A_2 \subseteq A_3) = \left(\frac{4}{8}\right)^n = \left(\frac{1}{2}\right)^n$。

对于由 m 个集合组成的链，一个容许的模式取决于该元素最先进入哪一个集合（$A_1, \ldots, A_m$ 其中之一，或者从不进入），所以 2^m 种模式之中有 $m+1$ 种是容许的：

$$\mathrm{P}(A_1 \subseteq A_2 \subseteq \cdots \subseteq A_m) = \left(\frac{m+1}{2^m}\right)^n. \qquad \blacksquare$$



每个元素：4 个等可能的位置



符合 $A_1 \subseteq A_2 \subseteq A_3$ 的模式：4 种（共 8 种）

图 9.12.1 左：S 的一个元素相对于 A_1 和 A_2 可以处于四个位置之一，每个位置的概率都是 $\frac{1}{4}$。不相交这个条件禁止“两者都在”（琥珀色）；$A_1 \subseteq A_2$ 则禁止“只在 A_1”。右：元素对 A_1、A_2、A_3 的八种归属模式（实心圆表示“在”）。链 $A_1 \subseteq A_2 \subseteq A_3$ 只容许加了方框的四种模式，即元素一旦进入便一直留在里面。

回顾

- **按元素分解。**每个集合层面的条件（不相交、子集、链），都是分别加在每个元素上的条件。
- **元素之间互相独立。**由于各个元素由独立的掷硬币结果决定，概率可以相乘。
- **数算模式。**每个答案都是（容许的模式数 ÷ 全部模式数）^{n}。

背后的数学

不用概率的计数

把每个概率乘以选择的总数，就得到一个计数。满足 $A_1 \subseteq A_2$ 的有序对 (A_1, A_2) 的数目是 $4^n \times \left(\frac{3}{4}\right)^n = 3^n$。你也可以直接数算这些有序对：先选定 A_2 的大小 k，然后以 $\binom{n}{k}$ 种方法选出 A_2 本身，再从它的 2^k 个子集中任选一个作为 A_1。所以

$$\sum_{k=0}^{n} \binom{n}{k} 2^k = 3^n,$$

这正是 $(1+2)^n$ 的二项式定理（binomial theorem）。用两种方法数算同一样东西，一次按元素、一次按大小，便证明了一条恒等式：这就是算两次（double counting）的方法，你在 Section 7.9 讨论柏拉图立体时已经见过。同样地，链 $A_1 \subseteq \cdots \subseteq A_m$ 的数目是 $(m+1)^n$。

子集的格

$\{1, \ldots, n\}$ 的所有子集按包含关系排序，形成一个称为布尔格（Boolean lattice）的结构，可以把它想像成一个 n 维立方体：每个子集是一个顶点，以它那一串是非题的答案作标

记，而沿着一条棱移动，就是加入或移走一个元素。链就是向上走的路径。有些深刻的结果描述了在给定条件下，子集族可以有多大。施佩纳定理（Sperner's theorem，1928 年）指出，在没有一个集合包含另一个集合的子集族之中，最大的一族有 $\binom{n}{\lfloor n/2 \rfloor}$ 个集合；艾狄胥—柯—拉多定理（Erdős–Ko–Rado theorem，1961 年）则说明，最多可以选出多少个 k 元集合，使得任意两个都有共同元素。两者都是组合数学（combinatorics）入门课程的标准内容，而把子集看成掷硬币的观点，给出了其中一些最漂亮的证明。

布尔的代数

布尔（George Boole）的 *The Laws of Thought*（《思维规律》，1854 年）把逻辑命题当作集合，把“与”、“或”、“非”当作交集、并集和补集。把一个集合编码成一串是非题的答案，正是电脑表示集合的方法：$\{1,\dots,64\}$ 的一个子集就是单独一个 64 位元的数，而交集就是“逐位元 AND”（bitwise AND）指令。本题的方法，即每次只看一个元素（一个位元），正是分析这类运算的方法。

思维技巧

每次只想一个元素

当对象是由每个元素各自的独立选择构成时，一个分别加在每个元素上的条件，其概率等于一个乘积。

这个方法把一个关于 2^n 个子集的问题，变成一个关于 4 种或 8 种模式的问题。到了大学，它会在组合数学（集合与多重集的生成函数（generating functions））、概率论（独立的坐标）和计算机科学（逐位元的推理）中再次出现。

延伸练习

1 **覆盖。**求 $\mathrm{P}(A_1 \cup A_2 = S)$ 和 $\mathrm{P}(A_1 \cup A_2 \cup \cdots \cup A_m = S)$。

2 **大小相同。**证明 $\mathrm{P}(|A_1| = |A_2|) = \binom{2n}{n}/4^n$。（提示：$|A_1| = |A_2|$ 当且仅当 $|A_1| + |S \setminus A_2| = n$。）证明当 n 很大时，这个概率约为 $1/\sqrt{\pi n}$。

3 **分成三份。**证明 S 的不相交子集所组成的有序对 (A, B) 的数目是 3^n，并求由不相交子集所组成的无序对 $\{A, B\}$ 的数目。

4 **小 n 的施佩纳定理。**对 $n = 3$ 和 $n = 4$，求 $\{1,\dots,n\}$ 的子集族中，没有一个集合包含另一个集合的最大一族。

延伸阅读

- **Miklós Bóna, *A Walk Through Combinatorics***（World Scientific），其中关于子集和偏序集（partially ordered sets）的章节。
- **Béla Bollobás, *Combinatorics: Set Systems, Hypergraphs, Families of Vectors and Combinatorial Probability***（Cambridge University Press）。一本简短而优雅的书，讲述子集族，由施佩纳定理说起。
- **Peter Cameron, *Combinatorics: Topics, Techniques, Algorithms***（Cambridge University Press）。一门入门课程，特别着重用两种方法计数。

9.13 做了手脚的骰子

题目

STEP I 1999，第 12 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/99-s1-q12`

引子

掷两颗公平骰子时，点数总和为 7 的可能性是总和为 2 的六倍。桌上游戏的设计者有时会希望情况并非如此：能否在骰子上做手脚（加上配重，使某些面较常出现），令 2 至 12 的每个总和都有相同的机会，即 $\frac{1}{11}$？那十二个概率（probability）可以任你选择，看来回旋的余地相当充裕。

答案是不能，而证明堪称一绝：它只用到十一个总和中的三个，以及不等式（inequality）$x+\frac{1}{x}\geq 2$。这是证明某件事不可能的典范；与证明某件事办得到相比，这往往更难，也更有趣。

动笔之前

想一想 对于 $x>0$，你会怎样证明 $x+\frac{1}{x}\geq 2$？等号何时成立？

两边同乘 $x>0$：不等式变成 $x^2+1\geq 2x$，即 $(x-1)^2\geq 0$，这显然成立。等号只在 $x=1$ 时成立。（另一做法是 $x+\frac{1}{x}-2=(\sqrt{x}-\frac{1}{\sqrt{x}})^2$。）这是算术平均—几何平均不等式（AM–GM inequality）最简单的形式：x 与 $\frac{1}{x}$ 的平均值不小于它们的几何平均（geometric mean）1。

逐步解题

阶段 1 两端的总和

想一想　掷出总和 2 有多少种方法？总和 12 呢？试以 q_i 和 r_j 表示它们的概率。

总和为 2 需要两颗骰子都掷出 1，总和为 12 则需要两颗都掷出 6，所以 $\mathrm{P}(2) = q_1 r_1$ 及 $\mathrm{P}(12) = q_6 r_6$。

为了引出矛盾，假设全部十一个总和的可能性相等，每个的概率都是 $\frac{1}{11}$。那么

$$q_1 r_1 = \frac{1}{11}, \qquad q_6 r_6 = \frac{1}{11}.$$

特别地，q_1、r_1、q_6、r_6 四者全为正数。

阶段 2　中间的总和

想一想　总和 7 可以由六种方法得出。其中哪两种只涉及 1 和 6 这两个面？试利用上面的等式写出它们的概率。

总和 7 包括结果 $(1,6)$ 和 $(6,1)$，所以

$$\mathrm{P}(7) \geq q_1 r_6 + q_6 r_1.$$

现在利用第 1 步的等式引入一个比。由于 $q_1 = \frac{1}{11 r_1}$ 及 $q_6 = \frac{1}{11 r_6}$，

$$q_1 r_6 + q_6 r_1 = \frac{1}{11}\left(\frac{r_6}{r_1} + \frac{r_1}{r_6}\right) \geq \frac{1}{11} \times 2 = \frac{2}{11},$$

这里用了 $x + \frac{1}{x} \geq 2$，其中 $x = r_6 / r_1 > 0$。因此 $\mathrm{P}(7) \geq \frac{2}{11}$，与 $\mathrm{P}(7) = \frac{1}{11}$ 矛盾。各个总和不可能全部等可能。 ■

停一停，试一试　看看这个证明为何行得通。要令 2 和 12 与 7 同样可能出现，1 和 6 就必须有相当高的概率，于是单凭 $(1,6)$ 和 $(6,1)$ 这两个“混合”结果，已足以令 7 的概率过高。图 9.13.1 显示了所涉及的方格。



图 9.13.1 两颗骰子的 36 个结果，各以其总和标示。证明只用到四个角落的方格。两个琥珀色方格的概率之积，等于两个深蓝绿色方格的概率之积，即 $q_1r_1q_6r_6 = q_1r_6 \cdot q_6r_1$；而根据算术平均—几何平均不等式，蓝绿色那一对之和至少是琥珀色那一对的几何平均的两倍。

阶段 3 同一颗骰子掷两次

想一想 如果把一颗骰子掷两次，两次点数相同的概率是多少？试以 q_j 表示。对公平骰子而言，这个概率是多少？

两次掷出相同点数的概率是 $q_1^2 + q_2^2 + \cdots + q_6^2$。对公平骰子而言，这是 $6 \times \frac{1}{36} = \frac{1}{6}$，已经远大于 $\frac{1}{36}$。

想一想 在骰子上做手脚，能否令 $\sum q_j^2$ 小于 $\frac{1}{6}$？你知道 $\sum q_j = 1$。

写成 $q_j = \frac{1}{6} + e_j$，其中各偏差 e_j 之和为零。那么

$$\sum q_j^2 = \sum\left(\frac{1}{36} + \frac{e_j}{3} + e_j^2\right) = \frac{1}{6} + \frac{1}{3}\sum e_j + \sum e_j^2 = \frac{1}{6} + \sum e_j^2 \geq \frac{1}{6}.$$

所以掷出重复点数的概率必定不小于 $\frac{1}{6}$，而且只有公平骰子才令等号成立。它当然不可能小于 $\frac{1}{36}$。 ■

回顾

- **反证法（proof by contradiction）。**我们假设所有总和等可能，然后推出不可能成立的结论。
- **只用所需的条件。**十一条等式之中，三条已经足够。选对等式正是关键的一步。
- **以不等式证明不可能。**第 (i) 部分的算术平均—几何平均不等式，以及第 (ii) 部分的“平方和非负”，都一举排除了一整类的可能性。

背后的数学

用多项式来证明

以下是一个完全不同的证明，用的是 Section 9.6 的生成函数（generating function）。设 $Q(x) = q_1x + \cdots + q_6x^6$ 及 $R(x) = r_1x + \cdots + r_6x^6$。假如所有总和的概率都是 $\frac{1}{11}$，则

$$Q(x)R(x) = \frac{x^2 + x^3 + \cdots + x^{12}}{11} = \frac{x^2}{11}\big(1 + x + \cdots + x^{10}\big).$$

现在 $Q(x)/x$ 是一个 5 次实系数多项式（polynomial），而每个奇数次的实多项式都有一个实根（它的图像由 $-\infty$ 走到 $+\infty$，或者反过来）。所以左边有一个非零的实根。但 $1 + x + \cdots + x^{10} = \frac{x^{11}-1}{x-1}$，它的根是 1 以外的 11 次单位根（roots of unity），当中没有一个是实数。矛盾。同样的论证表明，只要 $n - 1$ 是奇数，即 n 是偶数，无论怎样在两颗 n 面骰子上做手脚，都不能令各个总和等可能。（当 n 是奇数时，多项式的论证需要多花些工夫，但上面的算术平均—几何平均证明对每个 $n \geq 2$ 都适用。）

碰撞概率

第 (ii) 部分中的量 $\sum q_j^2$，即两次独立抽取结果相同的机会，称为一个分布的*碰撞概率*（collision probability）。第 (ii) 部分表明，均匀分布（uniform distribution）的碰撞概率最小。它量度一个分布有多集中：在生态学中，$\sum q_j^2$ 是辛普森多样性指数（Simpson's index of diversity），其中 q_j 是物种 j 在某个生境中所占的比例；在经济学中，它是量度市场集中度的赫芬达尔指数（Herfindahl index）；在密码学中，它量度一个密码分布有多容易被猜中。它也解释了 Section 9.9 结尾的那句话：各个生日的概率不相等，只会令生日相同的情况更有可能出现。

不可能性的证明

数学中一些最著名的结果，说的正是某件事办不到：没有分数的平方等于 2；没有尺规作图法能三等分任意角；没有根式公式能解所有五次方程；没有算法能判定每一个数学命题。每一个这样的结果都改变了所属的学科，因为要证明*没有*任何方法行得通，就必须了解所有方法构成的整个空间。灌铅骰子问题是一个缩影：要排除每一种做手脚的方式，我们需要一个每种方式都必须遵守的不变性质（算术平均—几何平均不等式）。

思维技巧

以反证法和不等式证明不可能

要证明没有任何参数的选取行得通，先假设有一组行得通，找出互相抵触的那几个条件，再用不等式证明它们不可能同时成立。

在大学里，反证法配合不等式是证明某些对象不存在的标准方法，见于分析、组合数学和最优化；在这些领域中，算术平均—几何平均不等式和柯西—施瓦兹不等式（Cauchy–Schwarz inequality）这类不等式为可达到的结果设下了界限。

延伸练习

1 **柯西—施瓦兹不等式。**利用柯西—施瓦兹不等式 $\left(\sum a_j b_j\right)^2 \le \sum a_j^2 \sum b_j^2$，取 $a_j = q_j$ 及 $b_j = 1$，为第 (ii) 部分给出一行的证明。

2 **两颗不同的骰子。**如果把两颗不同的灌铅骰子各掷一次，它们点数相同的概率能否小于 $\frac{1}{6}$？能否等于 0？

3 **面上数字不同时的均匀总和。**证明如果两颗骰子各面的数字另作安排，它们可以给出等可能的总和：试取一颗骰子的数字为 $0, 1, 2, 3, 4, 5$，另一颗为 $0, 6, 12, 18, 24, 30$。会出现哪些总和？为甚么这与题目并不矛盾？

4 **三颗骰子。**三颗各面为 1 至 6 的六面灌铅骰子，能否令 3 至 18 的所有总和等可能？（提示：利用总和 3、8、13 和 18，以及六个数的算术平均—几何平均不等式。为甚么多项式证明这次行不通？）

延伸阅读

- **J. Michael Steele, *The Cauchy–Schwarz Master Class***（Cambridge University Press）。一本出色的不等式入门书，由算术平均—几何平均不等式讲起。
- **Geoffrey Grimmett & Dominic Welsh, *Probability: An Introduction***（Oxford University Press）。讲述生成函数，以及它们在上述这类证明中的应用。
- **Ian Stewart, *Galois Theory***（CRC Press），开首数章。最著名的不可能性证明：五次方程。

我们学到了甚么

本章开首说过，计数是最古老的数学，而要数得巧妙，至今仍是最难的事情之一。我们提出了三个问题：怎样不逐一列举而数出答案？甚么时候某件事必然发生？计数与概率（probability）怎样互相帮助？以下是本章十三道题目所揭示的答案。

1. 怎样不逐一列举而数出答案?

本章逐步建立了一套工具。

- **递归关系（recurrences）。**在平面上多加一条直线（Section 9.1），或在围栏上多加一根柱子（Section 9.3），数目都只会以简单的方式改变，由此得出可以求解的递归关系。只要多记录一点资料，即最后一根柱子的颜色，这些递归关系便能自成一体。
- **生成函数（generating functions）。**把数列（sequence）包装成幂级数（power series），递归关系便化为代数运算（Section 9.4），既解释了为何 $1/89 = 0.01123\ldots$，也把数算组合变成多项式相乘：Section 9.5 的硬币、Section 9.6 的骰子，以及 Section 9.7 的等幂和，其中 $x = 1$ 处的重根（repeated root）包办了一切。
- **双射（bijections）。**反射原理（reflection principle）（Section 9.10）把不合格的队列与不受限制的路径一一配对，从而数出前者的数目；而取随机集合的补集（complement），则把子集变成互不相交的集合对（Section 9.12）。
- **容斥原理（inclusion–exclusion）。**先多算、后修正，便求得错排（derangements）的数目（Section 9.8）。
- **重复计数（double counting）。**以欧拉公式（Euler's formula）数算区域（Section 9.1），以及按大小数算子集对（Section 9.12），都是循不同途径得出相同答案，并顺道证明了一些恒等式。

2. 甚么时候某件事必然发生?

有好几节显示，单凭规模已足以迫出结构。根据鸽巢原理（pigeonhole principle），六个人当中必有三人互相认识，或有三人互不相识（Section 9.2），但五个人则不足够。二十三人当中有人同一天生日的机会已过半（Section 9.9），因为配对的数目随人数的平方增长。一个大集合的随机子集几乎必定互相重叠（Section 9.12）。还有一些事情，无论怎样尝试都不可能发生：不论怎样为两颗骰子灌铅，都无法令每个点数总和出现的机会均等

(Section 9.13)；而除了 Sicherman 的编号方式外，再没有其他骰子编号能重现普通骰子的点数总和分布（Section 9.6)。

3. 计数与概率怎样互相帮助?

在本章的后半部分, 概率是靠计数求得的: 错排用容斥原理, 生日问题用连乘积, 选票问题 (ballot problem) 用反射。但反过来, 概率也帮了计数的忙。期望值的线性性质 (linearity of expectation) 给出了集换券收集问题 (coupon collector) 的 nH_n (Section 9.11), 也给出了洗牌后配对数目的期望值, 完全毋须计数。把随机子集看成掷 n 次硬币（Section 9.12), 3^n 和 $(m+1)^n$ 这类数目便成了一行就能完成的计算。而艾狄胥（Erdős）的概率方法 (probabilistic method)（Section 9.2）则通过证明随机着色有正概率符合要求，证明了理想的着色确实存在。有两个常数凭空出现：e 出现在错排概率的极限中，而欧拉常数 γ 则出现在集齐一整套所需的代价中。

4. 计数的工具箱

方法	用于何处
多加一个时有何改变	直线（9.1)；围栏柱子（9.3)
鸽巢原理	拉姆齐（Ramsey)（9.2)
耦合递归关系	围栏柱子（9.3)
生成函数	9.4、9.5、9.6、9.7、9.13
从极端情况入手分类讨论	奇异骰子（9.6)；灌铅骰子（9.13)
容斥原理	错排（9.8)
补集与指数近似	生日（9.9)
反射原理	选票问题（9.10)
期望值的线性性质	集换券（9.11)；配对（9.8)
每次处理一个元素	随机子集（9.12)
以算术—几何平均不等式（AM–GM）导出矛盾	灌铅骰子（9.13)

下一步

- 错排的泊松极限（Poisson limit）会在 Section 12.11 再次出现，而随机变量（random variables）的生成函数则推动 Section 12.15 对家族树的研究。
- 耦合递归关系在 Section 12.16 会变成马可夫链（Markov chains)。
- 集换券收集问题中的冈贝尔分布（Gumbel distribution)，在 Section 12.25 用来描述众多随机量中的最大者。

如果你想在入读大学前再进一步

- **Miklós Bóna, *A Walk Through Combinatorics***：一本平易近人的入门教材，几乎涵盖本章的每一个概念。
- **Herbert Wilf, *generatingfunctionology***：生成函数的经典著作，可于网上免费阅读。
- **Ronald Graham, Donald Knuth & Oren Patashnik, *Concrete Mathematics***：求和、递归关系与生成函数，笔调风趣，习题丰富。
- **Geoffrey Grimmett & Dominic Welsh, *Probability: An Introduction***：大学一年级概率论中与计数有关的部分。

怀特黑德（Whitehead）曾称组合数学（combinatorics）为“拓扑学的贫民窟”。本章尝试说明，为何今天再没有人会这样说。每一个巧妙的计数背后，都有一个影响远超计数本身的想法：一个生成函数、一个双射、一个递归关系、一个不等式。这些想法把计数与代数、分析和概率连系起来；学会寻找这些想法，而不是再多找一个窍门，正是令计数成为数学的关键。

延伸练习答案

以下只是简短的答案和提示，并非完整解答：这里只给出结果和关键步骤，方便你核对自己的做法。请先认真做过每道练习，然后才翻看。

9.1 直线分割平面

1. 一个圆把平面分成 2 个区域。第 n 个圆与其余各圆相交于 $2(n-1)$ 点，这些点把它分成 $2(n-1)$ 段弧，每段弧把一个区域一分为二。所以 $R_n = R_{n-1} + 2(n-1)$，由此得 $n^2 - n + 2$。

2. 每条直线所增加的区域数目，是 1 加上该直线上的交点数目，所以总数是 $1 + n +$ (交点数目)。那 k 条平行线少了 $\binom{k}{2}$ 个交点：$1 + n + \binom{n}{2} - \binom{k}{2}$。

3. 第 n 个平面被其余平面截出 $n-1$ 条处于一般位置（general position）的直线，这些直线把它分成 $\frac{1}{2}(n^2 - n + 2)$ 块，每一块把空间中的一个区域一分为二。把这个数加到 $\frac{1}{6}\big((n-1)^3 + 5(n-1) + 6\big)$ 上，便得 $\frac{1}{6}(n^3 + 5n + 6)$。

4. 当 $n = 6$ 时有 31 个区域，而不是 32 个。在圆内，这个图有 $V = n + \binom{n}{4}$ 个顶点（vertex）（每个交点由四个点决定）和 $E = n + \binom{n}{2} + 2\binom{n}{4}$ 条边（edge）（圆上的各段弧，加上各条弦，而每个交点再多添两段）。欧拉（Euler）公式 $V - E + F = 2$ 把圆外的区域也算在内，所以区域的数目是 $F - 1 = 1 + \binom{n}{2} + \binom{n}{4}$。

9.2 小型版的拉姆齐定理

1. 双色角（bichromatic angle）最多有 $6 \times 6 = 36$ 个。每个非单色的三角形恰好含有两个双色角，所以这类三角形最多有 18 个。由于 $\binom{6}{3} = 20$，最少有 2 个三角形是单色的。

2. 每个点有 16 条边，所以其中最少有 6 条同色，设为红色，并设它们连到 $P_1, \ldots, P_6$ 各点。若 P_i 之间有任何一条边是红色，便有一个红色三角形。若没有，则各 P_i 构成一个以两种颜色著色的 K_6，其中必有单色三角形。

3. 点 v 有 9 条边，所以要么最少有 4 条是金色，要么最少有 6 条是银色。若有 4 个金色邻点：它们之间只要有一条金色边，便与 v 构成金色三角形，否则它们构成一个银色的 K_4。若有 6 个银色邻点：它们之中必有一个单色三角形；若是金色，证明完成，若是银色，它便与 v 构成一个银色的 K_4。

4. 给每一项标上 (i,d)：分别是以该项结尾的最长递增子序列（subsequence）和最长递减子序列的长度。对任意两项，较后的一项不是 i 较大，就是 d 较大，所以各个标记互不相同。假如没有长度为 3 的子序列，那么 5 项就只有 4 个可能的标记。两者皆无的例子：$2,1,4,3$。

9.3 为篱笆桩着色

1. 与之前一样，$r_{n+1}=s_n$（第 n 根柱不是红色，所以第 $n+1$ 根柱可以是红色），而 $r_n+s_n=(k-1)^{n-1}$，因为第一根之后的每根柱都有 $k-1$ 种选择。所以 $r_{n+1}+r_n=(k-1)^{n-1}$，而 $r_n=\frac{(k-1)^{n-1}+(k-1)(-1)^{n-1}}{k}$。至于围成一圈的情况：第 1 根柱的颜色有 k 种选择，然后是一排 $n+1$ 根柱，最后一根与第一根同色，由此得 $kr_{n+1}=(k-1)^n+(-1)^n(k-1)$。

2. $2,3,5,8,\ldots$：设共有 a_n 种方法,以白色结尾的一排可接在任何一排 $n-1$ 根柱之后,而以红色结尾的一排则接在以白色结尾的一排 $n-2$ 根柱之后,所以 $a_n=a_{n-1}+a_{n-2}$。答案是 F_{n+2}。

3. 中心有 4 种选择，然后是 n-圈的 3-著色：$4\bigl(2^n+2(-1)^n\bigr)$。

4. 当 $n=4$ 时：有 18 种著色。转四分之一圈不会保持任何一种著色不变，转半圈则保持其中 6 种（即相对的柱同色的那些）不变，所以有 $\frac14(18+0+6+0)=6$ 条项链。当 $n=5$ 时：有 30 种著色，没有任何旋转能保持其中一种不变，所以是 6。

9.4 把数列看成函数

1. 余下各项之和最多是 $F_7\sum_{r\ge7}2^{r-7}10^{-r}=\frac{13\times10^{-7}}{0.8}<2\times10^{-6}$。由于 $\sum F_r10^{-r}=\frac{10}{89}$，而首六项给出 0.112358，所以 $\frac{10}{89}$ 介乎 0.112358 与 0.112360 之间，于是 $\frac{1}{89}=0.011235\ldots$。若改以两位数为一组，则 $\sum F_r100^{-r-1}=\frac{1}{9899}$。

2. 把递推关系（recurrence）求和，得 $f(x)(1-7x+10x^2)=2-7x$，而 $\frac{2-7x}{(1-2x)(1-5x)}$ 可如题目所述分拆。至于 b_n：由 $10p+5q=40$ 和 $40p+10q=100$ 得 $p=1$，$q=6$；$t^2=t+6$ 的根是 3 和 -2，而 $b_n=4\cdot3^n+(-2)^n$。

3. $\sum nx^n=\frac{x}{(1-x)^2}$；在 $x=\frac{1}{100}$ 时它等于 $\frac{100}{9801}$，所以 $\frac{1}{9801}=\sum n\cdot100^{-n-1}$。这个规律在 98 这一组出错：100 那一组把 1 进位到 99 那一组，后者又再进位到 98 那一组，于是展开式是 $\ldots96\,97\,99\,00\,01\ldots$。

4. 乘以 $\frac{1}{1-x}=1+x+x^2+\cdots$ 之后，x^n 的系数变成 $u_0+\cdots+u_n$。平方数的生成函数（generating function）是 $\frac{x(1+x)}{(1-x)^3}$，所以其部分和的生成函数是 $\frac{x+x^2}{(1-x)^4}$，而 x^n 的系数是 $\binom{n+2}{3}+\binom{n+1}{3}=\frac{n(n+1)(2n+1)}{6}$。

9.5 凑零钱

1. x^{24} 的系数就是 $6a+6b+3c=24$ 的解的数目：对由 0 到 4 的每个 $s=a+b$，各有 $s+1$ 种选择，合共 15 种。一般而言，x^n 的系数是 0，除非 $n=3m$，这时它等于 $\frac{1}{2}(k+1)(k+2)$，其中 $k=\lfloor m/2\rfloor$。加上额外的因式后，系数变成部分和：x^{24}、x^{25} 和 x^{66} 的系数分别是 55、55 和 650。它们所数的，是用 1p 和 3p 硬币以及两种不同的 6p 硬币付款的方法数目。

2. 每个部分 k 要么用一次，要么完全不用。这个乘积可以逐项相消：$\prod\frac{1-x^{2k}}{1-x^k}$ 把分母中指数为偶数的因式全部约去，剩下 $\prod\frac{1}{1-x^{\text{奇数}}}$，也就是分拆（partition）成奇数部分的生成函数。当 $n=6$ 时：6、$5+1$、$4+2$、$3+2+1$，以及 $5+1$、$3+3$、$3+1+1+1$、$1+1+1+1+1+1$：两类各有四个。

3. 当 $n=0,1,2,\ldots$ 时，数值依次是 $1,1,2,2,3,4,5,6,7,8,10,\ldots$，而 $\frac{(n+4)^2}{20}$ 与它们的差距总不超过 $\frac{1}{2}$（已验算至 $n=60$；证明则要用到 $\frac{1}{(1-x)(1-x^2)(1-x^5)}$ 的部分分式（partial fractions））。

4. 最后一枚硬币是 1p 或 2p，所以 $c_n=c_{n-1}+c_{n-2}$：这就是斐波那契（Fibonacci）数，$c_n=F_{n+1}$。

9.6 奇怪的骰子

1. 每颗骰子都是 x 乘以 $1+x$、$1+x$、$1+x^2$、$1+x^2$ 这些因式中若干个的乘积，而且在 $x=1$ 时的值是 4，所以它要取其中两个因式。除了标准的一对之外，唯一的选择是 $x(1+x)^2$ 和 $x(1+x^2)^2$：各面是 $1,2,2,3$ 和 $1,3,3,5$。

2. 乘积是 $2x^3+x^4+x^5+4x^6+3x^7+8x^8+6x^9+5x^{10}+6x^{11}$，与表中数字吻合。

3. 可以：取 Sicherman 骰子对 $1,2,2,3,3,4$ 和 $1,3,4,5,6,8$，再加一颗标准骰子。

4. $(x+x^2)(x+\cdots+x^8)=x^2(1+x)^2(1+x^2)(1+x^4)$。硬币需要一个值为 2 的因式：除了标准的选择 $x(1+x)$ 之外，它可以是 $x(1+x^2)$（两面是 $1,3$，配上八面体（octahedron）骰子 $1,2,2,3,5,6,6,7$），也可以是 $x(1+x^4)$（两面是 $1,5$，配上 $1,2,2,3,3,4,4,5$）。

9.7 相等的幂和

1. $\sum(c+b_k)^3-\sum(c+a_k)^3=3c^2\left(\sum b_k-\sum a_k\right)+3c\left(\sum b_k^2-\sum a_k^2\right)+\sum b_k^3-\sum a_k^3$，而首两个括号都等于零。由 $\{1,4,6,7\}$、$\{2,3,5,8\}$ 出发，取 $c=8$：得 $\{1,4,6,7,10,11,13,16\}$ 和 $\{2,3,5,8,9,12,14,15\}$。

2. A 和 $B\cup\{0\}$ 都有 6 个元素，和为 33，平方和为 253。于是 $\sum_{a\in A}(n+a)^2=\sum_{b\in B\cup\{0\}}(n+b)^2$，这样便把 $\{n^2,\ldots,(n+11)^2\}$ 分成两组。

3. 用二项式定理（binomial theorem）展开 $(x+c)^r$：式中只涉及不高于 r 次的幂，而这些幂的和都相等。题目中的两个集合，在第二个加入 0 之后，合起来就是 $\{0,\dots,15\}$，再作平移便得任意 16 个连续整数。

4. 按 $ABABABAB$ 的次序，A 约得 1.676，B 约得 1.042。按 $ABBABAAB$ 的次序，A 约得 1.560，B 约得 1.158：仍不相等，但已接近得多。

9.8 没有人拿到自己的恤衫

1. $\mathrm{E}(C_1)=\frac{1}{m}$，$\mathrm{Var}(C_1)=\frac{1}{m}-\frac{1}{m^2}$，而 $\mathrm{Cov}(C_1,C_2)=\frac{1}{m(m-1)}-\frac{1}{m^2}=\frac{1}{m^2(m-1)}$。于是 $\mathrm{E}(N)=1$，而 $\mathrm{Var}(N)=m\left(\frac{1}{m}-\frac{1}{m^2}\right)+m(m-1)\frac{1}{m^2(m-1)}=1$。当 $m=4$ 时：$\mathrm{P}(N=0)=\frac{9}{24}=0.375$，相比之下 $\mathrm{e}^{-1}\approx 0.368$。

2. 第 1 张牌落在某个位置 i（有 $n-1$ 种选择）。若第 i 张牌落在位置 1，余下的就是 $n-2$ 张牌的错排（derangement）；若不是，则把位置 1 当作第 i 张牌的禁位，得到 $n-1$ 张牌的错排。验算：$D_3=2(1+0)=2$，$D_4=3(2+1)=9$。

3. 先选出 k 张留在原位的牌，再把其余的牌错排：$\binom{n}{k}\frac{D_{n-k}}{n!}=\frac{1}{k!}\cdot\frac{D_{n-k}}{(n-k)!}=\frac{d_{n-k}}{k!}$，它趋向 $\frac{\mathrm{e}^{-1}}{k!}$。

4. $30-(15+10+6)+(5+3+2)-1=8$，它等于 $30\cdot\frac{1}{2}\cdot\frac{2}{3}\cdot\frac{4}{5}$：容斥原理（inclusion–exclusion）的和式可以因式分解（这就是欧拉的 $\varphi(30)$）。

9.9 生日问题

1. $\mathrm{P}(X=r)=\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{r-2}{n}\right)\frac{r-1}{n}$，而这些概率之和为 1。至于期望值（expectation），$\mathrm{P}(Y=r)$ 在 $\mathrm{P}(Y\ge 1),\dots,\mathrm{P}(Y\ge r)$ 中各被计算一次。记 $Q_k=\mathrm{P}(X>k)=\prod_{j=1}^{k-1}\left(1-\frac{j}{n}\right)$。于是 $\mathrm{E}(X)=\sum(k+1)\frac{k}{n}Q_k$，同时 $\mathrm{E}(X)=1+\sum Q_k$；利用 $\sum\frac{k}{n}Q_k=1$ 并将两式相减，得 $\sum Q_k\left(1-\frac{k^2}{n}\right)=0$。

2. 由 $\binom{k}{3}\approx 365^2\ln 2$ 得 $k\approx 82$；这个近似忽略了各个三人组之间的相依关系，真正的答案是 88。

3. 14 人。（每一对人“接近”的概率是 $\frac{3}{365}$，而由 $\binom{k}{2}\frac{3}{365}\approx\ln 2$ 得 $k\approx 14$；精确计算的结果是：14 人时概率约为 0.537，13 人时约为 0.483。）

4. 两人同在第 i 天生日的概率是 p_i^2；把它们加起来。根据柯西–施瓦兹不等式（Cauchy–Schwarz），$1=\left(\sum p_i\right)^2\le 365\sum p_i^2$，等号仅在所有 p_i 都相等时成立。

9.10 总能找赎

1. 第一个人必须持有一枚 £1 硬币。若按其后几个人的情况取条件，便要分成许多情形（视乎每位 £2 顾客之后你手上有哪些硬币），每种情形最后都归结到第 (i) 和第

(ii) 部分的答案；把它们合起来，得 $\frac{m-2}{m+1}$。用反射原理（reflection principle）则一行便可完成。

2. 第一票必须投给 A。在以 A 开始的点票过程之中，坏的那些（即曾经出现平手的）可借着反射第一次平手之前的部分，与以 B 开始的点票过程一一对应。所以好的点票过程的数目是 $\binom{a+b-1}{a-1}-\binom{a+b-1}{a}$，除以 $\binom{a+b}{a}$ 便得 $\frac{a-b}{a+b}$。

3. (((())))、((()()))、((())())、(()(()))、(()()())。一个双射（bijection）：第一对括号把序列分成括号之内的部分和括号之后的部分；相应地，以某条固定的边为底的三角形把多边形分成两个较小的多边形。对两部分分别递归处理。

4. 每位 £5 顾客需要找回四枚 £1 硬币，所以路径每步上升 1 或下降 4。路径可以越过界线而不碰到它，于是反射再不能把坏的路径与一个简单的集合配对起来。

9.11 集齐一套

1. 各段等待时间的期望值之和是 $\frac{n}{n}+\frac{n}{n-1}+\cdots+\frac{n}{n/2+1}=n(H_n-H_{n/2})\approx n(\ln n-\ln\frac{n}{2})=n\ln 2$。

2. 每张卡缺少的概率是 $(1-\frac{1}{n})^k$；对 n 张卡求和。要达到 95%，需 $(1-\frac{1}{n})^k\approx \mathrm{e}^{-k/n}=0.05$，所以 $k\approx n\ln 20\approx 3n$。

3. 买了 t 包之后，某一张指定的卡出现少于两次的概率约为 $\mathrm{e}^{-t/n}(1+\frac{t}{n})$。当这类卡的期望数目 $n\mathrm{e}^{-t/n}(1+\frac{t}{n})$ 降至约 1 时，两个孩子便大致都集齐了；由此得 $\frac{t}{n}\approx \ln n+\ln\ln n$，远小于 $2\ln n$。

4. $\frac{1}{n+1}<\ln\frac{n+1}{n}$（$\frac{1}{x}$ 在 $[n,n+1]$ 上的曲线下面积），所以 $H_n-\ln n$ 递减。由凸性（convexity）可知梯形（trapezium）高估了面积：$\ln\frac{n+1}{n}<\frac{1}{2}\bigl(\frac{1}{n}+\frac{1}{n+1}\bigr)$，所以 $H_n-\ln n-\frac{1}{2n}$ 递增。取 $n=1$，得 $\gamma>\frac{1}{2}$；取 $n=4$，得 $\gamma<H_4-\ln 4\approx 0.697$。

9.12 随机子集

1. 每个元素必须属于最少一个集合：$\bigl(\frac{3}{4}\bigr)^n$，一般情况则是 $\bigl(1-2^{-m}\bigr)^n$。

2. $|S\setminus A_2|$ 与 $|A_2|$ 有相同的分布（distribution），并且与 A_1 独立（independent），所以所求概率就是 $|A_1|+|S\setminus A_2|=n$ 的概率。$(A_1,S\setminus A_2)$ 这一对集合对应于一个有 $2n$ 个元素的集合的随机子集，其大小为 n 的概率是 $\binom{2n}{n}/4^n$。由斯特灵（Stirling）公式（或沃利斯（Wallis）乘积，见 Section 4.4）得 $\approx\frac{1}{\sqrt{\pi n}}$。

3. 每个元素放进 A、放进 B，或两者都不放：3^n。满足 $A=B$ 的有序对只有 $(\varnothing,\varnothing)$，所以共有 $\frac{3^n-1}{2}+1=\frac{3^n+1}{2}$ 个无序对。

4. 当 $n=3$ 时，是 3 个集合（全部单元素集，或全部二元素集）。当 $n=4$ 时，是 6 个集合（全部二元素集）。一般情况是 $\binom{n}{\lfloor n/2\rfloor}$（Sperner 定理，Sperner's theorem）。

9.13 做了手脚的骰子

1. $1 = \left(\sum q_j\right)^2 \le 6\sum q_j^2$，所以 $\mathrm{P}(\text{相同}) = \sum q_j^2 \ge \frac{1}{6} > \frac{1}{36}$。

2. 两问的答案都是“可以”：所求概率是 $\sum q_j r_j$，而比方说，若一颗骰子永远掷出 1，另一颗永远掷出 2，它便等于 0。

3. 点数总和是 $0, 1, \ldots, 35$，每个恰好出现一次：是均匀分布。题目要求两颗骰子都标上 1 至 6，而这里的骰子并非如此。

4. 不可以。记 $\mathrm{P}(3) = q_1 r_1 s_1$ 和 $\mathrm{P}(18) = q_6 r_6 s_6$，两者都是 $\frac{1}{16}$。$\mathrm{P}(8) + \mathrm{P}(13)$ 包含 $q_1 r_1 s_6, \ldots, q_1 r_6 s_6$ 这六项（两个 1 和一个 6，或两个 6 和一个 1），它们的乘积是 $(\mathrm{P}(3)\mathrm{P}(18))^3$。由算术–几何平均不等式（AM–GM）得 $\mathrm{P}(8) + \mathrm{P}(13) \ge 6 \cdot \frac{1}{16} > \frac{2}{16}$。原来那个多项式证明照搬过来并不成立，因为 $1 + x + \cdots + x^{15}$ 确实有一个实根 -1。（不过它可以补救：三个实系数五次多项式需要三个实根，而 -1 只是单根。）

章

10

守恒定律、碰撞与引力

有些东西永远不变。知道是哪些，物理学就懂了一半。

本章内容

10.1	牛顿摆	STEP II 1990
10.2	叠球效应	STEP II 2010
10.3	火箭与枪炮	STEP III 2017
10.4	一个分子与一个活塞	STEP I 2002
10.5	有限时间内的无限次反弹	STEP III 2002
10.6	二体化为一体	STEP III 2004
10.7	球壳内的引力	STEP III 1999
10.8	排成三角形的三颗恒星	STEP I 1999
A	练习答案	第 717 页

甚么保持不变

> 有一个事实，或者你愿意的话，可称之为一条定律，支配着迄今所知的一切自然现象。这条定律至今未见任何例外。
>
> — 费曼（Richard Feynman）论能量守恒，*The Feynman Lectures on Physics*（1963）

学校里的力学（mechanics）通常离不开力：画一幅图，分解各力，再应用 $F = ma$。这套方法对付斜面上的木块绰绰有余，却对付不了历时仅千分之一秒的碰撞（collision），因为其间的力既巨大又未知；也对付不了一边飞行一边改变质量的火箭，或者三颗同时互相牵引的恒星。在这些情况下，出路是不再追问力，而改为追问甚么不能改变。

只要没有外力作用，动量（momentum）就守恒。在弹性碰撞（elastic collision）中能量守恒，而在其他碰撞中，能量的损失可以量度出来。质量、角动量（angular momentum）和质心（centre of mass）的运动也遵从类似的规律。这些定律不会告诉你一切，但它们永远成立，适用于复杂得根本无法仔细分析的系统，而且往往在动手计算之前，就已排除了大部分的可能性。

守恒定律为何重要

守恒定律（conservation law）是物理学中最可靠的工具。在称为 β 衰变（beta decay）的放射过程中，射出的粒子的能量加起来对不上数；泡利（Wolfgang Pauli）没有放弃能量守恒，反而在 1930 年预言有一种看不见的粒子带走了失去的能量，而中微子（neutrino）在二十六年后果然被探测到。工程师靠动量和能量的收支帐来设计防撞栏、安全气袋和火箭。天文学家靠观察质心来秤量恒星、寻找行星。到了 1918 年，诺特（Emmy Noether）解释了这些定律为何成立：每一条守恒定律都是大自然某种对称性（symmetry）的投影。动量守恒，是因为物理定律在任何地方都相同；能量守恒，是因为物理定律在任何时刻都相同。

引力（gravity）是本章的第二个主题。牛顿（Newton）正是用这种力把天上与地上统一起来，而它与守恒定律是天生的拍档：轨道、引力弹弓和塌缩的星云，全都要通过能量和动量，再加上一个关于球体的奇妙事实，才能理解。

你在大学会遇到甚么

大学一年级的动力学（dynamics）课程就从这里开始。你会遇到质点系统、质心与约化质量（reduced mass）、不同参考系（frame）中的碰撞、火箭之类的变质量系统，以及物

体在引力下的运动：开普勒（Kepler）定律、轨道、逃逸速度（escape velocity）。往后的课程会用拉格朗日（Lagrange）和哈密顿（Hamilton）的语言重新表述力学，届时诺特定理（Noether's theorem）会成为一个精确的命题，而守恒定律可以从单独一个函数的对称性中直接读出。再往前走，便是天体力学与三体问题（three-body problem）、分子运动论（kinetic theory）与统计力学（statistical mechanics）（Section 10.4 的活塞在那里会变成气体理论），以及爱因斯坦（Einstein）的引力理论。

贯穿本章的问题

三大问题

1. **守恒定律单凭自身能决定甚么？** 动量和能量何时能决定结果，何时只能排除某些可能？
2. **小步骤如何累积？** 一个简单的事件，例如一次碰撞或一次发射，重复多次甚至无限多次，会发生甚么事？
3. **如何驯服多个物体？** 哪些坐标的选择、哪些对称性，能把涉及几个相互作用物体的问题，化为我们解得到的问题？

本章由碰撞开始：牛顿摆（Newton's cradle）做得到甚么、做不到甚么（Section 10.1），以及一串球如何令速率倍增（Section 10.2）。接着探讨重复的过程：从货车上开炮与火箭方程（rocket equation）、分子撞向活塞而愈来愈快，以及在有限时间内弹跳无限多次的冰球。最后三节转向互相吸引或排斥的物体：质心与约化质量、球壳的引力，以及拉格朗日那个由恒星组成的旋转三角形。

本章的概念

§	课题	概念
10.1	牛顿摆	守恒定律作为约束条件
10.2	叠球	迭代（iteration）同一种碰撞；引力弹弓
10.3	火箭	变质量；火箭方程
10.4	分子运动论	绝热不变量（adiabatic invariant）
10.5	非弹性碰撞（inelastic collision）	有限时间内无限多次弹跳
10.6	二体问题（two-body problem）	质心与约化质量
10.7	牛顿引力	球壳定理（shell theorem）；自由落体式塌缩
10.8	三体问题	拉格朗日的等边三角形解

如何阅读本章

计算之前，先列出甚么是守恒的，以及在哪个参考系中守恒。然后问一问：换个位置看，问题会否简单一些？例如从质心看、从移动中的活塞看、从弹跳中的球看。本章不少题目在合适的参考系中几乎变得不费吹灰之力，而力学的大部分功夫，正在于找出这个参考系。

10.1 牛顿摆

题目

STEP II 1990，第 14 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/90-s2-q14`

引子

把牛顿摆（Newton's cradle）的一个球拉开，然后放手：另一端会有一个球飞出。拉开两个，就有两个飞出。为甚么？为甚么不是一个球以两倍速率飞出，或者三个球以三分之二的速率飞出？这两种情况的动量同样正确。

答案要从力学的两大守恒定律（conservation laws）说起。动量（momentum）与能量（energy）合起来所施加的限制，远比其中任何一个单独施加的严格得多，你想像得到的结果大多会被它们排除。本题把这两条定律用作约束条件（constraints）：不是用来计算会发生甚么，而是用来判断甚么可以发生。把守恒定律看成每一个可能结果都必须遵守的规则，这种想法贯穿整个物理学。

动笔之前

想一想 一粒质量为 m 的弹珠以速率 u 运动，撞上一粒质量为 M 的静止弹珠，而动能（kinetic energy）守恒。碰撞后两者的速度是多少？若 $M = m$ 又如何？

设碰撞后的速度为 p 和 q。由动量得 $mp + Mq = mu$，由能量得 $mp^2 + Mq^2 = mu^2$。把这两式写成 $m(u-p) = Mq$ 和 $m(u-p)(u+p) = Mq^2$，然后相除，得 $u + p = q$。于是

$$p = \frac{m-M}{m+M}\,u, \qquad q = \frac{2m}{m+M}\,u.$$

当 $M = m$ 时，第一粒弹珠停下，第二粒以速率 u 离开：两者只是交换了速度。这就是质量相等的弹珠能把运动沿着队列传递下去的原因。

逐步解题

阶段 1　只有最后两粒运动

想一想　假设碰撞后只有 A_n（速度 p）和 A_{n+1}（速度 q）运动。写出守恒定律。哪一粒弹珠的运动限制了 p 的正负号?

由动量和能量得

$$mp + \lambda m q = mu, \qquad \tfrac{1}{2}mp^2 + \tfrac{1}{2}\lambda m q^2 = \tfrac{1}{2}mu^2.$$

这正是质量为 m 的弹珠撞上质量为 λm 的弹珠时的方程，所以（除了 $p = u$、$q = 0$ 这个解之外，而该解需要 A_n 穿过 A_{n+1}）

$$p = \frac{1-\lambda}{1+\lambda}\,u, \qquad q = \frac{2}{1+\lambda}\,u.$$

若 $\lambda < 1$，两者皆为正，且 $p < q$：A_n 跟在 A_{n+1} 后面，但走得较慢，所以两者分开，一切并无矛盾。若 $\lambda > 1$，则 $p < 0$：A_n 向后运动。但 A_{n-1} 静止在它的正后方，并与它接触，而 A_n 不能穿过它。所以这种运动不可能发生。 ■

阶段 2　三粒弹珠

想一想　现在假设只有 A_{n-1}、A_n 和 A_{n+1} 运动，速度分别为 a、b、c。已知哪些弹珠静止、哪一粒在哪一粒后面，a、b、c 必须满足甚么不等式 (inequalities)?

弹珠 A_{n-2} 静止在 A_{n-1} 后面（它确实存在，因为 $n \geq 3$），所以 $a \geq 0$。而且没有弹珠可以越过前面的那一粒，所以 $a \leq b \leq c$。合起来就是 $0 \leq a \leq b \leq c$。守恒定律为

$$a + b + \lambda c = u, \qquad a^2 + b^2 + \lambda c^2 = u^2.$$

想一想　把动量方程平方，再与能量方程比较。剩下甚么?

平方后，$u^2 = a^2 + b^2 + \lambda^2 c^2 + 2ab + 2\lambda ac + 2\lambda bc$。减去能量方程，

$$0 = \lambda(\lambda - 1)c^2 + 2ab + 2\lambda c(a + b).$$

由于 $\lambda > 1$ 且 $a, b, c \geq 0$，右边每一项都非负，所以每一项都等于零。特别地，$c = 0$，于是 $0 \leq a \leq b \leq c$ 迫使 $a = b = 0$。但这样总动量就是 0，而不是 mu。矛盾：这种运动不可能发生。 ■

停一停，试一试 留意这个论证的结构。把动量方程平方会产生 $2ab$ 之类的交叉项（cross terms），而能量方程却没有这些项。两者之差量度了运动有多“分散”。当 $\lambda > 1$ 时，同一个论证可以排除只有 $A_k, \dots, A_{n+1}$ 以非负速度运动的任何结果，你看得出原因吗?

阶段 3 两粒可以运动的弹珠

想一想 向前运动的弹珠必须全部位于远端，而这种情况既然行不通，也许有一粒弹珠是向后运动的。哪一粒弹珠可以自由地向后运动?

只有 A_0 的后面没有任何东西。不妨试试：A_0 以速度 p 运动（可能是负数），A_{n+1} 以速度 q 运动，其余全部静止。守恒定律再次与质量为 m 和 λm 的两个物体碰撞时相同，所以

$$p = -\frac{\lambda - 1}{\lambda + 1}\,u, \qquad q = \frac{2}{\lambda + 1}\,u.$$

A_0 反弹，A_{n+1} 向前运动，中间的弹珠保持静止，一切并无矛盾。这就完全像是 A_0 直接撞上了 A_{n+1}，而中间那一列相等的弹珠充当了完美的信差。 ■

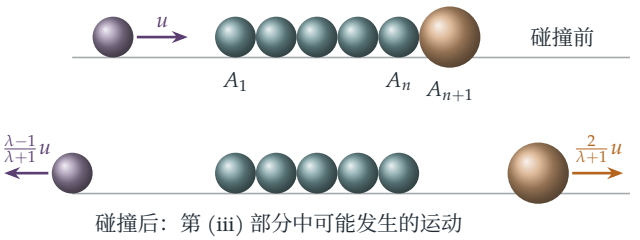


图 10.1.1 一列相等的弹珠（蓝绿色），末端是一粒较重的弹珠（琥珀色，$\lambda > 1$）。碰撞后，A_0 反弹，重弹珠离开，中间每一粒弹珠都保持静止，完全像是 A_0 直接撞上了 A_{n+1}。

回顾

- **以守恒定律作为约束条件。**我们从来没有仔细追踪每一次碰撞，只是问：动量、能量和几何条件容许哪些结果?
- **物理上的不等式十分重要。**“弹珠不能互相穿过”给出了不等式 $0 \le a \le b \le c$，而排除第 (i) 和第 (ii) 部分的，正是这些不等式，而不是守恒定律。
- **平方后相减。**把动量的平方与能量比较，会得出一个由交叉项组成的和，而这个和必须等于零。

背后的数学

为甚么只有一个球飞出

在由相等的球组成的牛顿摆中，假设有 k 个球以速率 w 一起离开。由动量得 $kw = u$，由能量得 $kw^2 = u^2$。两式相除，得 $w = u$ 及 $k = 1$。所以在离开的球一起运动的各种结果之中，只有“一个球以原来的速率离开”同时符合两条定律。两个球以一半速率离开，动量正确，但能量只有一半；一个球以两倍速率离开，则只有当它是半个球时动量才正确。

但问题并未就此解决。若有 n 个球，就有 n 个未知的速度，而守恒定律只有两条，所以同时符合两者的结果有无穷多个（试找出一个三个球以不同速率运动的结果）。实际出现哪一个，取决于各球在接触期间如何变形。1980 年代初，Frank Herrmann 和他的同事证明，我们熟悉的现象有赖于冲击沿着队列传递时*不会扩散开去*，而坚硬的钢球之间留有极小空隙时正是如此。若用软球，或者把球紧紧压在一起，牛顿摆的表现就没有那么干净利落；仔细的量度更显示，其他的球总会稍为移动。

守恒定律与对称

守恒定律从何而来？1918 年，诺特（Emmy Noether）证明了一条定理，它是现代物理学的基石之一：物理系统的每一种连续对称（continuous symmetry），都对应一个守恒量（conserved quantity）。如果物理定律在这里与在左边一米处相同（空间平移对称），动量就守恒。如果物理定律今天与明天相同（时间平移对称），能量就守恒。如果物理定律在每个方向都相同（旋转对称），角动量（angular momentum）就守恒。你会在大学二年级的拉格朗日力学（Lagrangian mechanics）课程中学到诺特定理（Noether's theorem）；本章自始至终都在运用它的推论。

惠更斯、雷恩与沃利斯

1668 年，英国皇家学会（Royal Society）征求物体碰撞问题的解答。沃利斯（John Wallis）处理了非弹性碰撞（inelastic collisions），而雷恩（Christopher Wren）和惠更斯（Christiaan Huygens）则各自独立地解决了弹性（elastic）的情况。惠更斯的关键洞见是转换参考系（frame）：在一艘随质心（centre of mass）运动的船上观察碰撞，这时两个相等的球只是以相反方向、原来的速率弹回，然后再变换回岸上的观点。他还注意到某一个量，即各质量乘以速率平方之和，在碰撞前后相同；不过要再过一个世纪，动能本身才被确认为一个守恒量。

思维技巧

利用守恒量排除不可能的情况

当一个过程复杂得无法追踪时，列出不会改变的量，看看它们容许哪些结果。

物理学家正是这样推断粒子加速器中的碰撞：当中相互作用的细节并不清楚，但动量、能量和电荷必须平衡。再配合一个不等式（这里是弹珠不能互相穿过），就可以证明某个结果不可能发生。

延伸练习

1 **同一个问题再现。**试做 STEP I 2006，第 11 题（`step.maths.org/questions/06-s1-q11`）：槽中有一列粒子，末端是较重的一粒。哪些最终状态是可能的？

2 **追踪各次碰撞。**把第 (iii) 部分的撞击看成一连串独立的两粒弹珠碰撞，由 A_0 撞上 A_1 开始。逐步追踪这一连串碰撞，并证明最后得到的正是第 (iii) 部分所求得的运动。

3 **两个球撞入。**在一个由五个相等的球组成的牛顿摆中，两个球同时释放，以速率 u 撞上其余三个。证明“两个球以速率 u 离开”与两条守恒定律并无矛盾，并找出另一个同样符合的结果。

4 **末端较轻。**若 $\lambda < 1$，追踪一连串的两粒弹珠碰撞。除了 A_n 和 A_{n+1} 之外，最后还有其他弹珠在运动吗？

延伸阅读

- **David Morin, *Introduction to Classical Mechanics***（Cambridge University Press），第 5 章。动量守恒与能量守恒，附有大量关于碰撞的习题。
- **Frank Herrmann & Peter Schmälzle, “Simple explanation of a well-known collision experiment”**，*American Journal of Physics* 49 (1981)。解释牛顿摆为何行得通。
- **Dwight Neuenschwander, *Emmy Noether's Wonderful Theorem***（Johns Hopkins University Press）。深入浅出地讲述对称与守恒之间的联系。

10.2 叠球效应

题目

STEP II 2010，第 10 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/10-s2-q10`

引子

把一个网球放在篮球上面，然后让两个球一起落下。篮球从地面反弹，网球随即向上飞射，升到原来下落高度的好几倍，有时甚至高得撞上天花板。这个实验请到户外去做。

题目的第 (i) 部分说明，单独一次碰撞有一个速率上限：无论撞过来的球有多重，一个静止的球被撞之后，速率都不可能超过来球速率的两倍。第 (ii) 部分则说明怎样打破这个上限：让运动沿着一串愈来愈轻的球传递下去，每传一步，速率都会再乘上一个倍数。把太空探测器送往外行星，靠的正是同一个想法；也有人提出，超新星（supernova）爆发的机制当中有一部分也是如此。

动笔之前

想一想 根据 Section 10.1：若质量为 M、以速率 u 运动的物体与质量为 m 的静止物体发生弹性碰撞（elastic collision），碰撞后 m 的速率是多少？

那一节的结果（把两个质量的角色对调）给出被撞粒子的速率为 $\frac{2M}{M+m}u$。还有另一个记法：在弹性碰撞中，相对速度（relative velocity）会反向。如果重球几乎没有减慢，轻球离开时的速率就必定比重球的速率大约多出 u，也就是大约 $2u$。

逐步解题

阶段 1　速率上限

取 $M = bm$，碰撞后 A 的速率为

$$v_A = \frac{2bm}{bm + m}\,u = \frac{2b}{b+1}\,u.$$

> **想一想** $\frac{2b}{b+1}$ 与 2 相比，哪个较大？当 $b \to \infty$ 时会怎样？

由于对任何 b 都有 $\frac{2b}{b+1} = 2 - \frac{2}{b+1} < 2$，所以 A 的速率总是小于 $2u$，无论 B 有多重。当 $b \to \infty$ 时，这个速率趋近 $2u$：一个无限重的球会把 A 弹开，就像 A 撞上一堵移动中的墙一样。 ■

阶段 2　沿着一串粒子传递

> **想一想** 在第 (ii) 部分，每个粒子的质量都是下一个粒子的 λ 倍。对于这串粒子中的每一次碰撞，第 (i) 部分告诉了我们甚么？

B_1（质量 $\lambda^n m$）撞上 B_2（质量 $\lambda^{n-1} m$），后者的质量是前者的 λ 分之一。在第 (i) 部分取 $b = \lambda$，可知 B_2 以速率 $\frac{2\lambda}{\lambda+1}u$ 离开。接着 B_2 撞上 B_3，后者的质量同样是前者的 λ 分之一，于是 B_3 以速率 $(\frac{2\lambda}{\lambda+1})^2 u$ 离开。经过 n 次这样的碰撞之后（最后一次是 B_n 撞上 A），

$$v_A = \left(\frac{2\lambda}{\lambda+1}\right)^n u.$$

（各个 B 之间其后再发生的碰撞，只会把 A 推得更快，不会使它减慢，因为从来没有东西从前方撞向 A。）

> **想一想** $\frac{2\lambda}{\lambda+1}$ 比 1 大还是比它小？

由于 $\lambda > 1$，故 $2\lambda > \lambda + 1$，因此比值 $r = \frac{2\lambda}{\lambda+1}$ 大于 1，而且当 $n \to \infty$ 时 $r^n \to \infty$。所以只要 n 取得够大，A 的速率要多快就有多快。 ■

阶段 3　快二十倍

取 $\lambda = 4$，则 $r = \frac{8}{5} = 1.6$，我们需要 $1.6^n > 20$。取以 10 为底的对数（logarithm），

$$\log_{10} 1.6 = \log_{10} 16 - 1 = 4\log_{10} 2 - 1 \approx 0.20412, \qquad \log_{10} 20 = 1 + \log_{10} 2 \approx 1.30103,$$

所以我们需要 $n > \frac{1.30103}{0.20412} \approx 6.37$。最小的值是 $n = 7$。（事实上 $1.6^6 \approx 16.8$，而 $1.6^7 \approx 26.8$。）

■

回顾

- **把单一结果反复套用。**第 (i) 部分的结果在第 (ii) 部分用了 n 次，把有界的增益变成无界的增益。

- **永远达不到的极限。**速率可以趋近上界 $2u$，却永远达不到，所以正确的写法是严格不等式（strict inequality）。
- **对数把乘方化为乘积。**“需要多少步”这个问题，变成了一道除法。

背后的数学

篮球与网球

设两个球从高度 H 一起落下，到达地面时的速率为 $U = \sqrt{2gH}$。假设一切碰撞都是完全弹性的，而且篮球比网球重得多。篮球从地面反弹，以速率 U 向上运动，随即遇上仍以速率 U 向下运动的网球。在篮球的参考系（frame）中，网球以 $2U$ 的速率迎面而来，撞上一个非常重的物体后反弹，再以 $2U$ 的速率离开。回到地面参考系，那就是以 $2U + U = 3U$ 的速率向上运动（图 10.2.1）。由于上升高度与抛射速率的平方成正比，网球会升到 $9H$ 的高度。

现实中的球并不是完全弹性的。若恢复系数（coefficient of restitution）为 e，质量比为 M/m，网球离开时的速率就是 $\frac{Me(2+e)-m}{M+m}U$。练习中的 STEP I 2014 Q10，就是由量度所得的高度反推质量比。

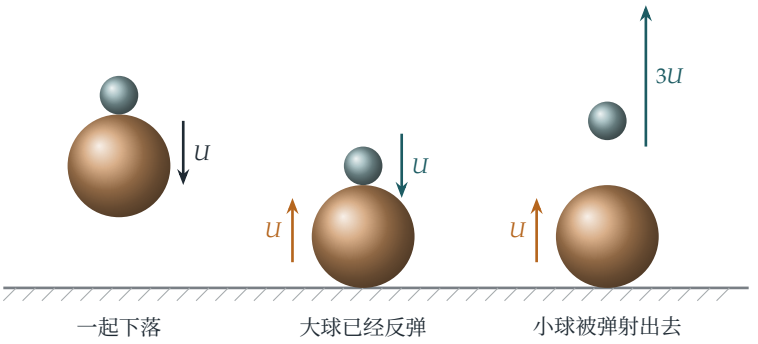


图 10.2.1 叠球效应：两个球都是完全弹性的，而且下面的球重得多。大球以速率 U 向上反弹，遇上以速率 U 落下的小球。相对于大球，小球以 $2U$ 的速率到达，又以 $2U$ 的速率离开，所以在地面参考系中它以 $3U$ 的速率离开，升到原来高度的九倍。

引力弹弓

太空探测器飞掠行星时并没有碰到行星，但引力的作用就像与一个非常重的物体发生一次非常柔和的弹性碰撞。在行星的参考系中，探测器绕过行星，离开时的速率与到达时相同。如果行星正以速率 V 朝探测器运动，那么在太阳的参考系中，探测器的速率最多可以增加 $2V$，与网球的情况完全一样。1977 年发射的两艘航行者号（*Voyager*）探测器，

利用了外行星一次罕见的排列，从木星飞往土星（航行者 2 号（*Voyager* 2）更继续飞往天王星和海王星），每次与行星相遇都获得加速。假如没有这些引力助推（gravity assist），前往海王星的旅程就要花上大约三十年，而不是十二年。能量来自行星，而行星因此减慢的幅度小得无法量度。

弹力球与超新星

1968 年，W. R. Mellen 在 *American Journal of Physics* 上描述了一种“弹力球反弹抛射体”（superball rebound projectile）：把几个由大至小的球叠起来，最顶的球可以用撞击速率的许多倍弹射出去。天体物理学家指出，这与核心塌缩超新星（core-collapse supernova）有一点粗略的相似之处：恒星向内坠落的外层物质撞上坚硬而致密的核心后反弹，冲击波（shock wave）把动量（momentum）向外带进密度愈来愈低的物质，把这些物质加速到极高的速率。叠球效应的示范常用来向学生解释这个想法，不过真实的超新星所涉及的物理比弹跳的球多得多，中微子（neutrino）就是其中之一。

思维技巧

把单一结果反复套用

如果一个过程只是把一个简单步骤重复多次，就先把这一步解决一次，然后把它复合起来。

每一步的增益因子都小于 2，但经过很多步之后，增益却变成无界，因为这些因子是相乘的。到了大学，这就是转移矩阵（transfer matrix）、多级放大器和多级火箭（Section 10.3）背后的想法：一条链的增益，等于链上各个环节的增益的乘积。

延伸练习

1 **真实的情况。**试做 STEP I 2014，第 10 题（`step.maths.org/questions/14-s1-q10`）：这是带有真实数据、反弹并不完美的叠球问题。

2 **非弹性的链。**试做 STEP I 2017，第 10 题（`step.maths.org/questions/17-s1-q10`）。这一题让碰撞沿着一串质量按几何级数增长的粒子传递下去。在甚么情况下，每个粒子都恰好参与两次碰撞？

3 **三个球。**把三个完全弹性的球叠起来，每个球都比它下面的球轻得多。证明最顶的球以 $7U$ 的速率离开，并升到 $49H$ 的高度。若有 k 个球，证明最顶的球以 $(2^k - 1)U$ 的速率离开。

4 **能量从何而来?** 在第 (ii) 部分，证明即使 A 的速率无界，总动能（kinetic energy）仍然守恒。那些重的粒子会怎样?

延伸阅读

- **David Morin, *Introduction to Classical Mechanics***（Cambridge University Press），第 5 章，当中包括叠球问题。
- **W. R. Mellen,“Superball rebound projectiles”**, *American Journal of Physics* 36 (1968)。篇幅短小，读来有趣。
- **Gary Flandro,“Fast reconnaissance missions to the outer solar system utilizing energy derived from the gravitational field of Jupiter”**, *Astronautica Acta* 12 (1966)。这篇论文提出了航行者号的“行星大巡游”（grand tour）。

10.3 火箭与枪炮

题目

STEP III 2017，第 11 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/17-s3-q11`

引子

太空中的火箭没有任何东西可以借力。它靠把质量向后抛出而前进，其余的就交给动量守恒（conservation of momentum）。一辆从车尾开炮的货车，就是一枚拆成一块块的火箭：每枚抛射体（projectile）都是一团喷出的废气。

那么，这些炮应该同时发射，还是逐一发射？这道题给出了明确的答案，但要看清楚它比较的是甚么。同时发射会令货车得到较大的速率，但也消耗较多能量，所以在两种情况下，这些炮所做的并不是同一件事。只要改变保持不变的量，答案就会逆转。把这一点弄清楚，便会引出太空飞行中最重要的方程——齐奥尔科夫斯基（Tsiolkovsky）的*火箭方程*（rocket equation），也会明白为甚么真正的火箭要分成多级。

动笔之前

想一想 某一门炮发射之前，货车以 V 行驶。发射之后，抛射体的速度是 $V-v$（即以速率 $v-V$ 向后运动）。抛射体*相对于发射前的货车*的速度是多少？

答案是 $(V-v)-V=-v$。所以每门炮都以速率 v 把抛射体向后抛出，这个速率是相对于货车在发射前一刻的速度而言的。这就是本题的固定规则；请牢记它，因为答案正是由它决定的。

逐步解题

阶段 1　同时发射

系统最初静止，动量为零。发射之后，货车（质量为 M）以 u 向前运动，而 n 枚抛射体则以速率 v 向后运动（因为发射前 $V=0$）。所以

$$Mu - nmv = 0, \qquad u = \frac{nmv}{M}.$$

所获得的动能（kinetic energy）为

$$K = \tfrac{1}{2}Mu^2 + \tfrac{1}{2}nmv^2 = \tfrac{1}{2}\frac{n^2m^2v^2}{M} + \tfrac{1}{2}nmv^2 = \tfrac{1}{2}nmv^2\Big(1 + \frac{nm}{M}\Big). \qquad \blacksquare$$

阶段 2　逐一发射

> **想一想**　在第 r 门炮发射前的一刻，货车连同余下的抛射体的质量是多少？速率又是多少？写出第 r 次发射的动量守恒式。

第 r 次发射之前，货车载有 $n-r+1$ 枚抛射体，所以它的质量是 $M+(n-r+1)m$，速率是 u_{r-1}。发射之后，载有 $n-r$ 枚抛射体的货车速率为 u_r，而射出的抛射体速度为 $u_{r-1}-v$。动量守恒：

$$\big(M+(n-r+1)m\big)u_{r-1} = \big(M+(n-r)m\big)u_r + m(u_{r-1}-v).$$

把含 mu_{r-1} 的项整理后，得 $\big(M+(n-r)m\big)(u_r-u_{r-1}) = mv$，这就是 $(*)$。 $\blacksquare$

把各个增量相加，

$$u_n = \sum_{r=1}^{n} \frac{mv}{M+(n-r)m} = \frac{mv}{M} + \frac{mv}{M+m} + \cdots + \frac{mv}{M+(n-1)m}.$$

第一项之后的每一项都小于 $\frac{mv}{M}$，而且 $n>1$，所以 $u_n < \frac{nmv}{M} = u$。 $\blacksquare$

> **停一停，试一试**　为甚么在这里逐一发射的效果较差？第一门炮发射时，它要加速的不只是货车，还有那 $n-1$ 枚未发射的抛射体，所以它带来的速率增幅，比起作用在空车上时要小。

阶段 3　能量的帐单

> **想一想**　写出第 r 次发射时动能的变化。$(*)$ 怎样帮助你化简含 $u_r^2 - u_{r-1}^2$ 的那一项？

动能的变化为

$$K_r - K_{r-1} = \tfrac{1}{2}\big(M + (n-r)m\big)u_r^2 + \tfrac{1}{2}m(u_{r-1} - v)^2 - \tfrac{1}{2}\big(M + (n-r+1)m\big)u_{r-1}^2.$$

最后两项合并成 $\tfrac{1}{2}m(v^2 - 2vu_{r-1}) - \tfrac{1}{2}\big(M + (n-r)m\big)u_{r-1}^2$。然后，把平方差因式分解，并利用 $(*)$，

$$\tfrac{1}{2}\big(M + (n-r)m\big)(u_r^2 - u_{r-1}^2) = \tfrac{1}{2}mv(u_r + u_{r-1}).$$

所以

$$K_r - K_{r-1} = \tfrac{1}{2}mv(u_r + u_{r-1}) + \tfrac{1}{2}mv^2 - mvu_{r-1} = \tfrac{1}{2}mv^2 + \tfrac{1}{2}mv(u_r - u_{r-1}). \qquad \blacksquare$$

由 $r = 1$ 加到 n，增量 $u_r - u_{r-1}$ 逐项相消（telescope），余下 u_n：

$$K_n = \tfrac{1}{2}nmv^2 + \tfrac{1}{2}mvu_n. \qquad \blacksquare$$

最后，$K = \tfrac{1}{2}nmv^2 + \tfrac{1}{2}nmv^2 \cdot \tfrac{nm}{M} = \tfrac{1}{2}nmv^2 + \tfrac{1}{2}mvu$，而 $u_n < u$，所以 $K_n < K$。 $\blacksquare$

回顾

- **每一步都用动量。**每次发射都是一个小型的二体问题：一边是货车连同所载的东西，另一边是一枚抛射体。
- **裂项求和。**$(*)$ 的增量和能量的增量都能干净利落地加起来。
- **仔细阅读结论。**同时发射得到较大的速率，但要付出较多能量。题目比较的是两个做功多少并不相同的过程。

背后的数学

应该让甚么保持不变？

真正的炮或火箭引擎，是以相对于*自身*的固定速率抛出抛射体的，而这个速率是在抛射体刚离开之后量度的。本题的参照对象却是货车在发射*之前*的速率。对一枚细小的抛射体来说，两者相差极微；但当 n 枚抛射体同时发射时，差别就很大：在第 (i) 部分，一大批沉重的抛射体以速率 v 被向后抛出，仿佛货车完全没有后座（recoil）一样，这需要大量能量。

假如改为每枚抛射体都以相对于发射*之后*的货车的速率 v 离开，结论就会逆转。同时发射得 $u = \frac{nmv}{M+nm}$；逐一发射得 $u_n = \sum_{k=1}^{n} \frac{mv}{M+km}$，后者较大，因为每一项都至少是 $\frac{mv}{M+nm}$。这时逐一发射胜出：每枚抛射体都是从一辆已在运动的货车上蹬出去的，而且越后发射的炮，要加速的质量越少。（请在练习中验证这一点。）

火箭方程

现在让抛射体变得极小而数目极多，并保持总质量 $\mu = nm$ 不变：这就是一枚连续燃烧燃料的火箭。u_n 的和变成一个积分（两种约定给出相同的极限）：

$$u_n = \sum_{k=0}^{n-1} \frac{mv}{M+km} \;\to\; \int_0^{\mu} \frac{v\,\mathrm{d}s}{M+s} = v \ln \frac{M+\mu}{M}.$$

这就是*齐奥尔科夫斯基火箭方程*：若火箭喷出的废气以相对于火箭的速率 v 离开，火箭所增加的速率为

$$\Delta u = v \ln \frac{m_{\text{初}}}{m_{\text{末}}}.$$

图 10.3.1 比较了这三个答案。麻烦就出在对数（logarithm）上：若排气速率约为 4.5 km/s（优良的化学燃料），要达到约 7.8 km/s 的轨道速率，质量比必须是 $e^{7.8/4.5} \approx 5.7$；一旦把空气阻力和重力计算在内，这个比值更接近 10 或以上。火箭几乎全是燃料。

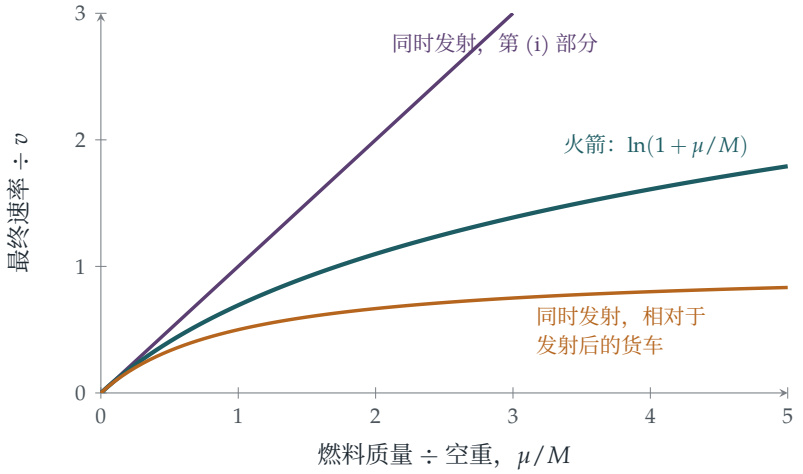


图 10.3.1　最终速率与燃料质量对空重之比的关系。按题目的约定，全部同时发射得 μ/M（紫红色）；但若排气速率是相对于后座中的货车量度，则只得 $\frac{\mu/M}{1+\mu/M}$（琥珀色）。连续发射的火箭介乎两者之间，为 $\ln(1+\mu/M)$（蓝绿色）。这幅图所表达的就是不等式 $\frac{x}{1+x} \le \ln(1+x) \le x$。

为甚么火箭要分级

对数也解释了分级（staging）的道理。火箭若在升空途中抛掉用完的燃料箱，就不必再加速这些无用的质量。各级所得的速率增量是*相加*的（就像那些炮一样），而每一级的增量则取决于该级质量的*比*，所以把火箭分成多级，就能达到总质量相同的单级火箭无法达到的速率。把太空人送上月球的土星五号（Saturn V）有三级；它的第一级在不足三分钟内烧掉约 2000 公吨燃料。

齐奥尔科夫斯基

康斯坦丁.齐奥尔科夫斯基(Konstantin Tsiolkovsky)是俄罗斯外省小城卡卢加(Kaluga)一位失聪的教师，他在 1903 年一篇题为“利用反作用装置探索外太空”的论文中发表了火箭方程。多级火箭、液态氢和液态氧燃料，以及太空站，他都一一推算过，比这些东西真正造得出来早了数十年。这条方程其实更早就在一些较不起眼的地方出现过(William Moore 在 1810 年已推导出其中一个版本)，但齐奥尔科夫斯基是第一个看出它对太空飞行意味着甚么的人。他写道：“地球是人类的摇篮，但人不能永远活在摇篮里。”

思维技巧

从离散过程到极限

连续的过程，往往可以理解为许多细小离散步骤的极限（*limit*）。先把离散的问题精确地解出来，再让步长缩小：和就变成积分。

在这里，n 门炮变成了一枚火箭，一个和变成了一个对数。到了大学，每逢推导物理定律，你都会见到同一手法：先分析一个微小的单元，再把结果加起来，成为一个积分或一条微分方程（differential equation）。

延伸练习

1 **另一种约定。**假设每枚抛射体都以相对于发射之后的货车的速率 v 离开，重做第 (i) 和第 (ii) 部分。证明这时逐一发射所得的最终速率较大。

2 **尘埃云中的彗星。**试做 STEP III 2009，第 11 题（`step.maths.org/questions/09-s3-q11`）。它和火箭一样是变质量（variable mass）问题，但彗星的质量是增加而不是减少。留意它的速率趋向甚么值。

3 **两级火箭。**一枚火箭的排气速率为 v，空重为 M，燃料质量为 $9M$。求它的最终速率。现在把它分成两级，每级的空重为 $\frac{1}{2}M$，燃料为 $4.5M$，第一级在燃料用尽后抛掉。比较两个最终速率。

4 **不等式。**在 $[0, x]$ 上把 $\frac{1}{1+s}$ 分别与 1 和 $\frac{1}{1+x}$ 比较，从而证明当 $x \geq 0$ 时，$\frac{x}{1+x} \leq \ln(1+x) \leq x$。

延伸阅读

- **Daniel Kleppner & Robert Kolenkow, *An Introduction to Mechanics***（Cambridge University Press），第 4 章。动量、变质量和火箭方程。
- **George Sutton & Oscar Biblarz, *Rocket Propulsion Elements***（Wiley）。工程学上的标准教科书；读完本节后，开首几章已可读懂。
- **Hermann Oberth, *Ways to Spaceflight***（NASA 译本，可在网上免费阅读）。一位先驱的亲身记述。

10.4 一个分子与一个活塞

题目

STEP I 2002，第 10 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/02-s1-q10`

引子

替单车轮胎打气时摸一摸气泵：它会发热。压缩气体会令气体变暖，柴油引擎（单靠压缩就能点燃燃料）、雪柜以至恒星的形成，背后都是这个原理。但为甚么压缩会令气体升温？温度就是分子的平均动能（kinetic energy），所以问题其实是：为甚么分子会加速？

这道题目只用一个分子就回答了这个问题。分子每次从推进中的活塞（piston）反弹，都会多得一点速率，就像网球碰上挥动中的球拍一样。把每次反弹准确地追踪下去，便会发现一条隐藏的定律：分子的速率乘以气缸的长度，几乎保持不变。这就是一个绝热不变量（adiabatic invariant），也是你在物理课遇到的气体定律在微观层面的来源。

动笔之前

想一想 一个以速率 w 运动的球，撞上一堵以速率 u 迎面移来的墙，并作弹性（elastic）反弹。反弹后球的速率是多少？（试在墙的参考系中思考。）

在墙的参考系（frame）中，球以 $w+u$ 接近；由于它是从一个极重的物体上弹性反弹，所以亦以 $w+u$ 离开。回到地面参考系，墙正以 u 朝着与反弹后的球相同的方向运动，所以球的速率是 $w+2u$。每次从推进中的活塞反弹，速率都增加 $2u$。（从气缸静止的一端反弹，则甚么也不会改变。）

逐步解题

阶段 1　反弹 n 次后的速率

质点（particle）起初的速率是 v。每次与活塞碰撞（collision），速率增加 $2u$，而与末端碰撞则不会改变速率。所以与活塞碰撞 n 次之后，

$$v_n = v + 2nu. \qquad \blacksquare$$

阶段 2　长度的递推关系

> **想一想**　在第 n 次与第 $(n+1)$ 次碰撞之间，质点由活塞走到末端再折返。它走了多远？活塞又移动了多远？

在时间 t_n 内，活塞移动了 ut_n，所以 $d_n - d_{n+1} = ut_n$。与此同时，质点以速率 v_n 由活塞走到末端（距离 d_n），再返回活塞（距离 d_{n+1}），所以 $d_n + d_{n+1} = v_n t_n$。把两条方程相除，便可消去 t_n：

$$\frac{d_n + d_{n+1}}{d_n - d_{n+1}} = \frac{v_n}{u}, \qquad \text{所以} \qquad d_{n+1}(v_n + u) = d_n(v_n - u).$$

代入 $v_n = v + 2nu$，得

$$d_{n+1} = \frac{v + (2n-1)u}{v + (2n+1)u}\, d_n. \qquad \blacksquare$$

这就是所求的递推关系（recurrence）。

阶段 3　裂项相消

> **想一想**　把 d_n 写成 d_1 乘以 $n-1$ 个这样的分数之积。哪些项会互相抵消？

$$d_n = d_1 \cdot \frac{v+u}{v+3u} \cdot \frac{v+3u}{v+5u} \cdots \frac{v+(2n-3)u}{v+(2n-1)u} = \frac{v+u}{v+(2n-1)u}\, d_1. \qquad \blacksquare$$

每个分子都与前一个分母抵消，只剩下第一个分子和最后一个分母。

阶段 4　$v = u$ 的情况

这时 $d_n = \frac{2u}{2nu} d_1 = \frac{d_1}{n}$，而

$$ut_n = d_n - d_{n+1} = d_1\Big(\frac{1}{n} - \frac{1}{n+1}\Big) = \frac{d_1}{n(n+1)}. \qquad \blacksquare$$

把这些加起来，$u(t_1 + t_2 + \cdots) = d_1\big(1 - \tfrac{1}{2} + \tfrac{1}{2} - \tfrac{1}{3} + \cdots\big) = d_1$：活塞在第一次碰撞后经过时间 d_1/u 到达气缸末端，这正是理所当然的结果。在活塞到达之前，质点反弹了无限多次（图 10.4.1）。

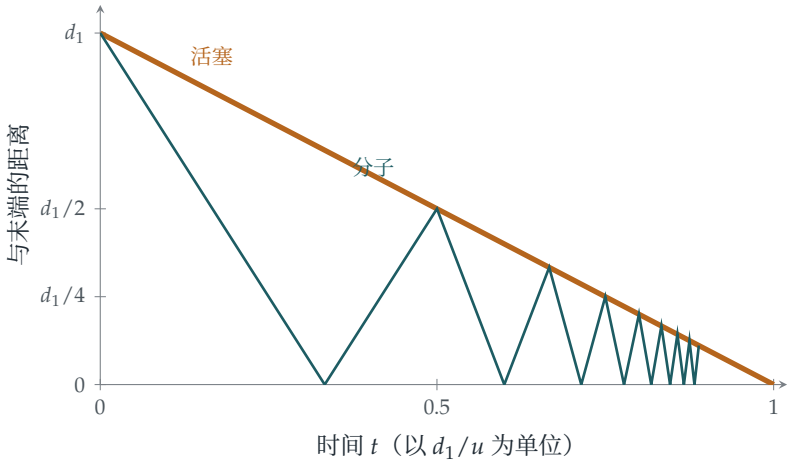


图 10.4.1 $v = u$ 情况下的时空图：与气缸末端的距离对时间。活塞（琥珀色）稳步推进；分子（蓝绿色）在活塞与末端之间来回反弹。它的锯齿形轨迹愈来愈陡（愈来愈快），也愈来愈短。第 n 次与活塞的碰撞发生在距离 d_1/n 处，所以分子的速率大致与 1/(长度) 成正比。

回顾

- **处理碰撞时转换参考系。**在活塞的参考系中，反弹再简单不过。
- **两条方程，消去未知量。**质点与活塞在同一段时间内所走的距离，给出了一个不含 t_n 的递推关系。
- **裂项相消（telescoping）。**那个乘积一下子化简了，正如 $\frac{1}{n(n+1)}$ 之和一样。

背后的数学

一个绝热不变量

把速率乘以长度：

$$v_n d_n = (v + 2nu) \cdot \frac{v+u}{v+(2n-1)u}\, d_1 \approx (v+u)d_1$$

这在 n 很大时成立，在 u 远小于 v 时亦然。所以

$$\text{速率} \times \text{长度} \approx \text{常数}.$$

如果活塞相对于分子移动得很慢，那么无论活塞移动多远，乘积 vd 都几乎不变。像这样在系统缓慢改变时几乎保持不变的量，称为*绝热不变量*。一般而言，往复运动的绝热不变量是 $\oint p\,\mathrm{d}q$，即动量（momentum）沿一个周期的积分；在这里它就是 $mv \times 2d$。

从一个分子到一团气体

由于 $v \propto 1/d$，动能 $\frac{1}{2}mv^2$ 与 $1/d^2$ 成正比。对于由许多沿同一方向运动的分子组成的气体，温度与平均动能成正比，所以 $T \propto 1/d^2$：长度减半，温度便变成四倍。在三维空间中，能量由三个运动方向分担，而同样的论证（费曼（Feynman）在他的 *Lectures* 中给出了这个论证）可推出单原子气体（monatomic gas）的绝热压缩定律，

$$TV^{2/3} = \text{常数}, \qquad \text{或等价地} \qquad pV^{5/3} = \text{常数}.$$

指数 $\frac{5}{3}$ 就是单原子气体著名的比热比（ratio of specific heats）γ。把单车气泵内的空气压缩至原来体积的五分之一，如果做得够快而且没有热量散失，其绝对温度会升高至原来的 $5^{0.4} \approx 1.9$ 倍（空气分子有两个原子，因而 $\gamma = 1.4$）：由 20°C 升至约 280°C。气泵之所以没有那么热，只是因为压缩得慢，热量又会散走。

费米的宇宙射线

1949 年，恩里科·费米（Enrico Fermi）提出，宇宙射线（cosmic rays）这些从太空而来、能量极高的粒子，其加速方式与本题的分子如出一辙：它们在星际气体云中运动的磁场上反复反弹，每次获得一点能量。现时认为，“费米加速”（Fermi acceleration）在超新星残骸的激波处发生，粒子在那里来回穿越激波，就像分子在活塞与墙壁之间往返一样。绝热不变量在量子力学的诞生中也扮演过一个角色：保罗·埃伦费斯特（Paul Ehrenfest）在 1916 年指出，被量子化的量必定是绝热不变量，因为缓慢的改变不可能令系统在量子态之间跃迁。

思维技巧

从物理过程建立递推关系

当一个过程不断重复时，把其中一个周期准确地描述出来，并写下状态怎样由一个周期变到下一个周期。

这里的状态是每次碰撞时的一对量（速率，长度）。这个递推关系可以精确地解出，而它的解揭示了一个近似守恒的量。在大学里，同样的方法（先考虑一个周期，然后得出递推关系或“回归映射”（return map））会用于轨道、振子，以及 Section 8.7 的迭代映射。

延伸练习

1 **不变量的精确形式。**证明 $v_n d_n - (v+u)d_1 = \frac{u(v+u)d_1}{v+(2n-1)u}$，并由此推出当 $n \to \infty$ 时 $v_n d_n \to (v+u)d_1$。

2 **把活塞拉出。**设活塞改为以速率 $u < v$ 向外抽出，把题目重做一次。证明质点会减速，并求 $v_n d_n$。（气体膨胀时会冷却。）

3 **能量。**在 $v = u$ 的情况下，随着活塞逼近末端，分子的速率会无止境地增长。求第 n 个周期内活塞所受的平均力（动量的改变除以 t_n），并证明它按 n^3 的速度增长。这说明了把活塞一直推到尽头所需的功有甚么特点?

4 **弹跳的球。**一个球在重力作用下于地板上作弹性弹跳，而地板被缓慢地升高。试猜出这个运动的一个绝热不变量。（提示：与 vd 相对应的量是平均速率乘以弹跳的高度，或等价地，能量乘以两次弹跳之间的时间。）

延伸阅读

- **Richard Feynman, Robert Leighton & Matthew Sands, *The Feynman Lectures on Physics*, Vol. I**（《费曼物理学讲义》第一卷，可在网上免费阅读），第 39 章“The kinetic theory of gases”（气体动力论）。同一个活塞论证，并推广至三维。
- **Stephen Blundell & Katherine Blundell, *Concepts in Thermal Physics***（Oxford University Press）。大学一年级程度的绝热过程与气体动力论。
- **Lev Landau & Evgeny Lifshitz, *Mechanics***（Butterworth-Heinemann），§49。对绝热不变量的精简论述，留待日后阅读。

10.5 有限时间内的无限次反弹

题目

STEP III 2002，第 11 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/02-s3-q11`

引子

埃利亚的芝诺（Zeno of Elea）曾经论证，你永远走不到一堵墙：你必须先走完一半的距离，然后再走完余下距离的一半，如此类推，永无止境。化解这个悖论的关键，在于无穷多步所需的总时间可以是有限的，也就是一个收敛（converge）的无穷级数（infinite series）。这道题目正是芝诺悖论（Zeno's paradox）的真实版本：一个冰球滑进楔形的角落，在两面墙之间反弹无穷多次，却在有限时间内到达角落。

大自然也会这样做。一个球在地板上弹跳，若恢复系数（coefficient of restitution）小于 1，它会在有限时间内弹跳无穷多次，然后静止下来。在颗粒物质（沙、谷物、雪）的电脑模拟中，这种“非弹性坍缩”（inelastic collapse）是个实实在在的难题：电脑要在有限时间内处理无穷多次碰撞。

动笔之前

想一想 当一个质点（particle）以恢复系数 e 撞向一面光滑的墙时，速度沿墙方向的分量会怎样？垂直于墙的分量又会怎样？

光滑的墙不会沿着墙面施力，所以速度平行于墙的分量保持不变。垂直分量则会反向，并乘以 e。因此，若质点以与墙成 α 的角度到达，并以 β 的角度离开，则

$$\tan\beta = e\tan\alpha,$$

而由于平行分量不变，（入射速率）$\cos\alpha$ =（离开速率）$\cos\beta$。

逐步解题

阶段 1　离开的角度

想一想　质点到达 M_1 时与墙 OA 成角 θ（在远离 O 的一侧量度）。它离开时与 OA 所成的角 φ 是多少？利用三角形 OM_1N_1，以 θ 表示 φ。

质点离开 M_1 时朝着 O 的一侧前进，与墙成角 φ，所以角 OM_1N_1 等于 φ。在三角形 OM_1N_1 中，位于 O 的角是 30°。在 N_1，角 M_1N_1B 等于 θ，而 O、N_1、B 在同一直线上，所以角 ON_1M_1 等于 $180^\circ-\theta$。三角形的内角和为 180°：

$$30^\circ+\varphi+(180^\circ-\theta)=180^\circ,\qquad \text{所以}\qquad \varphi=\theta-30^\circ.$$

（特别地，$\theta>30^\circ$。）于是由恢复定律 $\tan\varphi=e\tan\theta$ 可得

$$e=\frac{\tan(\theta-30^\circ)}{\tan\theta},$$

而由平行分量守恒 $u_1\cos\theta=v_1\cos\varphi$ 可得

$$\frac{v_1}{u_1}=\frac{\cos\theta}{\cos(\theta-30^\circ)}.$$ ■

阶段 2　规律重复出现

想一想　在 N_1，质点到达时与墙 OB 成角 θ（在远离 O 的一侧量度）。把这个情况与它到达 M_1 时的情况比较一下。接下来会发生甚么事？

N_1 的情况是 M_1 的情况的镜像：质点以角 θ 撞上一面墙，而这个角位于远离 O 的一侧。所以它以角 $\theta-30^\circ$ 离开，再以角 θ 撞上 OA，如此类推，永无止境。每一次反弹都是第一次反弹按比例缩放后的复制品。三角形 OM_kN_k、ON_kM_{k+1}、… 每一个都与前一个相似（图 10.5.1），而每次反弹时速率都乘以同一个因子。

阶段 3　两个比

想一想　由一次反弹到下一次反弹，与 O 的距离按甚么因子 ρ 缩小？速率按甚么因子 k 缩小？那么每一段路程所需的时间又按甚么因子缩小？

在三角形 OM_1N_1 中应用正弦定理（sine rule），可得

$$\rho=\frac{ON_1}{OM_1}=\frac{\sin\varphi}{\sin(180^\circ-\theta)}=\frac{\sin(\theta-30^\circ)}{\sin\theta}.$$

下一个三角形的所有长度都是这个三角形的 ρ 倍，所以路径的每一段都是前一段的 ρ 倍长。每次反弹时速率都乘以 $k=\frac{\cos\theta}{\cos(\theta-30^\circ)}$。时间等于长度除以速率，所以每一段所需的时间是前一段的

$$\frac{\rho}{k}=\frac{\sin(\theta-30^\circ)\cos(\theta-30^\circ)}{\sin\theta\cos\theta}=\frac{\sin(2\theta-60^\circ)}{\sin 2\theta}$$

倍。

阶段 4 一个等比级数

总时间是 $T_1(1+r+r^2+\cdots)$，其中 T_1 是第一段所需的时间，而 $r=\frac{\sin(2\theta-60^\circ)}{\sin 2\theta}$。这是一个等比级数（geometric series），若 $0<r<1$，总时间便是有限的。

想一想 $\sin(2\theta-60^\circ)<\sin 2\theta$ 在甚么时候成立？利用恒等式 $\sin X-\sin Y=2\cos\frac{1}{2}(X+Y)\sin\frac{1}{2}(X-Y)$。

$$\sin 2\theta-\sin(2\theta-60^\circ)=2\cos(2\theta-30^\circ)\sin 30^\circ=\cos(2\theta-30^\circ),$$

当 $2\theta-30^\circ<90^\circ$，即 $\theta<60^\circ$ 时，上式为正。另外，由于 $\theta>30^\circ$，故 $r>0$。所以当 $30^\circ<\theta<60^\circ$ 时，级数收敛：质点反弹无穷多次，它与 O 的距离每次按因子 $\rho<1$ 缩小，并在有限时间 $\frac{T_1}{1-r}$ 之后到达 O。 ■

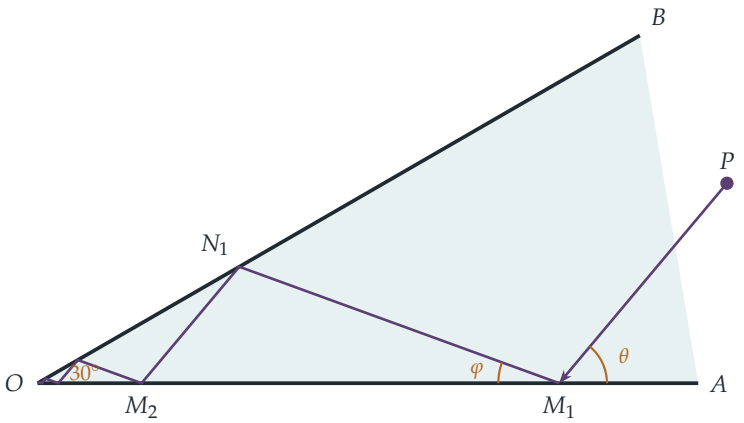


图 10.5.1 $\theta=50^\circ$ 时质点在楔形内的路径，此时 $\varphi=20^\circ$，$e=\tan 20^\circ/\tan 50^\circ\approx 0.31$。每一次反弹都是前一次的复制品，按 $\rho\approx 0.45$ 的比例缩小，而每一段所需的时间约为前一段的 0.65 倍。质点经过无穷多次反弹后，在有限时间内到达 O。

回顾

- **自相似性（self-similarity）。**一旦弄清楚一次反弹，其后每一次反弹都只是按比例缩放的复制品，所以整个无穷过程就归结为两个比。
- **正确地组合各个比。**距离按 ρ 缩小，速率按 k 缩小，所以时间按 ρ/k 缩小。
- **无穷多个事件，有限的总和。**公比小于 1 的等比级数收敛。

背后的数学

当 $\theta \geq 60^\circ$ 时

当 $\theta = 60^\circ$ 时，这个比恰好等于 1：每一段所需的时间都相同，质点虽然愈来愈接近 O，却永远到不了。（速率和每段的长度都按同一个因子缩小。）当 $\theta > 60^\circ$ 时，每一段所需的时间愈来愈长。在所有情况下，质点都会趋近 O，并反弹无穷多次；只有当 $\theta < 60^\circ$ 时，它才会在有限时间内到达。$e = \tan 30^\circ / \tan 60^\circ = \frac{1}{3}$ 这个值就是分界线。

弹跳的球与非弹性坍缩

同一现象最简单的例子：一个球掉落到地板上，反弹至高度 e^2H，然后是 e^4H，如此类推。两次弹跳之间的时间与每次弹跳时的速率成正比，而速率按因子 e 缩小，所以总时间是一个等比级数：设 $T_0 = \sqrt{2H/g}$ 为第一次下落的时间，则球在

$$T_0 + 2eT_0 + 2e^2T_0 + \cdots = T_0\,\frac{1+e}{1-e}.$$

之后停止弹跳。无穷多次弹跳，然后静止。1992 年，Sean McNamara 和 William Young 证明了，在模拟排成一直线的多个非弹性质点时，同样的事情也会发生：其中一组质点可以在有限时间内发生无穷多次碰撞，这就是“非弹性坍缩”。沙和颗粒流的模拟必须加入特别的规则来避免这种情况，而这个现象也有助解释为何颗粒气体会自发地聚集成团。

芝诺与收敛

芝诺的悖论约于公元前 450 年提出，困扰了哲学家两千年。数学上的解答，即无穷和可以有有限的值，要到十九世纪才变得严谨，当时柯西（Cauchy）、魏尔斯特拉斯（Weierstrass）等人以极限（limit）来定义收敛（这是第 2 章的主题）。这道题目令人惊讶之处，在于这个过程在物理上是无穷的：确确实实有无穷多次碰撞，发生在有限时间之内。研究多边形内的桌球运动和碰撞质点运动的数学家，把这种行为称为芝诺现象（Zeno phenomenon）；要证明由互相碰撞的硬球组成的系统在所有时刻都有明确定义的运动，这正是其中一个微妙之处。

思维技巧

为无穷过程求和

当一个过程以固定的缩放因子不断重复时，任何可加的量（时间、距离、能量）的总和都是一个等比级数。根据公比判断它是否收敛。

这里的关键是找出正确的比：不是距离之比，也不是速率之比，而是时间之比。在大学里，同样的推理支配着迭代法（iterative methods）的收敛性、分形（fractal）理论，以及对反复把问题减半的算法（algorithm）的分析。

延伸练习

1 **总距离。**证明对于 30° 与 90° 之间的每一个 θ，由 M_1 到 O 所走的总距离都是有限的，并以 OM_1 和 θ 表示这个总距离。

2 **其他楔形。**把 30° 换成角 α 的楔形，重做这道题目。证明 $\varphi = \theta - \alpha$，并求出质点在有限时间内到达 O 时 θ 须满足的条件。

3 **弹性墙。**若 $e = 1$，质点有可能到达 O 吗？（提示：把楔形沿它的墙反复反射，使路径变成一条直线。围绕 O 可以放下多少个 30° 的楔形？）

4 **弹跳的球。**一个球从 1 m 的高度掉下，$e = 0.8$。它会弹跳多久？它总共走了多远？（取 $g = 9.8\ \mathrm{m\,s^{-2}}$。）

延伸阅读

- **David Morin, *Introduction to Classical Mechanics***（Cambridge University Press），第 5 章。书中有多道关于弹跳的球和无穷多次连续碰撞的题目。
- **Sean McNamara & William Young, "Inelastic collapse and clumping in a one-dimensional granular medium"**（一维颗粒介质中的非弹性坍缩与聚团），*Physics of Fluids A* 4 (1992)。
- **Serge Tabachnikov, *Geometry and Billiards***（American Mathematical Society；可在网上免费阅读）。讲述桌球的数学，包括练习中用到的展开技巧。

10.6 二体化为一体

题目

STEP III 2004，第 11 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/04-s3-q11`

引子

两个物体互相吸引或排斥：地球与月球、两颗恒星、一个质子与一个电子。每个物体的运动都取决于另一个物体的位置，因此两条方程是耦合（coupled）的，看来非要同时求出两个运动不可。牛顿（Newton）的重大简化在于更换变量：不用两个位置，而改用质心（centre of mass）和间距（separation）。质心的运动仿佛完全不受力，而间距的运动则像一个质量经过修正的单一质点。一个难题就此化为两个易题。

本题就两个在直线上运动的质点，实际做一次这样的分离。过程中还有一个小小的奇迹：间距原来是 t 的二次式的平方根，而个中原因与题目本身的设置毫无关系，却完全源于平面上的运动。

动笔之前

想一想 写出每个质点的牛顿第二定律（Newton's second law）。每个力朝哪个方向作用？

这个力是排斥力，所以它把（左边的）P 推向左，把（右边的）Q 推向右：

$$2\ddot{x}_1 = -\frac{2}{z^3}, \qquad 1\cdot\ddot{x}_2 = +\frac{2}{z^3}.$$

这两条方程是耦合的，因为每一条都含有 $z = x_2 - x_1$。

逐步解题

阶段 1　间距

想一想　两条方程的哪一个组合只含有 z?

相减：$\ddot{z} = \ddot{x}_2 - \ddot{x}_1 = \frac{2}{z^3} + \frac{1}{z^3} = \frac{3}{z^3}$。 ■

间距的表现就像一个质量为 $\frac{2}{3}$ 的单一质点，受力 $\frac{2}{z^3}$ 推动。$\frac{2}{3} = \frac{2\times 1}{2+1}$ 这个数就是约化质量（reduced mass）$\frac{m_1 m_2}{m_1+m_2}$。

阶段 2　间距的能量

想一想　利用 $\ddot{z} = v\frac{\mathrm{d}v}{\mathrm{d}z}$，分离变量然后积分。$z$ 和 $v = \dot{z}$ 的初始值是多少?

由 $v\frac{\mathrm{d}v}{\mathrm{d}z} = \frac{3}{z^3}$，积分得 $\frac{1}{2}v^2 = -\frac{3}{2z^2} + C$。起初 $z = 1$ 而 $\dot{z} = \dot{x}_2 - \dot{x}_1 = 0 - 1 = -1$，所以 $\frac{1}{2} = -\frac{3}{2} + C$，得 $C = 2$:

$$v^2 = 4 - \frac{3}{z^2}.$$

（这就是间距的能量守恒（conservation of energy）：以每单位约化质量计，动能加上势能 $\frac{3}{2z^2}$ 是常数。）

阶段 3　出奇简单的解

想一想　要证明的答案是 $z^2 = 4t^2 - 2t + 1$。所以不妨考虑 $w = z^2$，而不是 z。以 w 表示，$\dot{w}^2$ 是甚么?

设 $w = z^2$，则 $\dot{w} = 2z\dot{z}$，所以

$$\dot{w}^2 = 4z^2\dot{z}^2 = 4z^2\left(4 - \frac{3}{z^2}\right) = 16w - 12.$$

求导得 $2\dot{w}\ddot{w} = 16\dot{w}$，所以（只要 $\dot{w} \neq 0$）$\ddot{w} = 8$。因此 w 是 t 的二次式：$w = 4t^2 + \alpha t + \beta$。起初 $w = 1$ 而 $\dot{w} = 2z\dot{z} = -2$，所以 $\beta = 1$ 及 $\alpha = -2$:

$$z = \sqrt{4t^2 - 2t + 1}. \qquad \blacksquare$$

（你可以直接验证：$\dot{w}^2 = (8t-2)^2 = 64t^2 - 32t + 4 = 16(4t^2 - 2t + 1) - 12$。）两个质点在 $t = \frac{1}{4}$ 时最接近，此时 $z = \frac{\sqrt{3}}{2}$，其后便永远愈离愈远。

阶段 4　质心

想一想　$2x_1 + x_2$ 满足甚么方程? 为甚么它是最自然要考虑的组合?

把两条运动方程相加，$2\ddot{x}_1 + \ddot{x}_2 = -\frac{2}{z^3} + \frac{2}{z^3} = 0$。两个力大小相等、方向相反（牛顿第三定律），所以总动量（momentum）$2\dot{x}_1 + \dot{x}_2$ 是常数。它的初始值是 $2\times 1 + 0 = 2$，而 $2x_1 + x_2$ 的初始值是 1，所以

$$2x_1 + x_2 = 1 + 2t.$$

质心位于 $\frac{2x_1+x_2}{3}$，以速率 $\frac{2}{3}$ 匀速移动。再结合 $x_2 - x_1 = z$，

$$x_1 = \frac{1 + 2t - \sqrt{4t^2 - 2t + 1}}{3}, \qquad x_2 = \frac{1 + 2t + 2\sqrt{4t^2 - 2t + 1}}{3}. \qquad \blacksquare$$

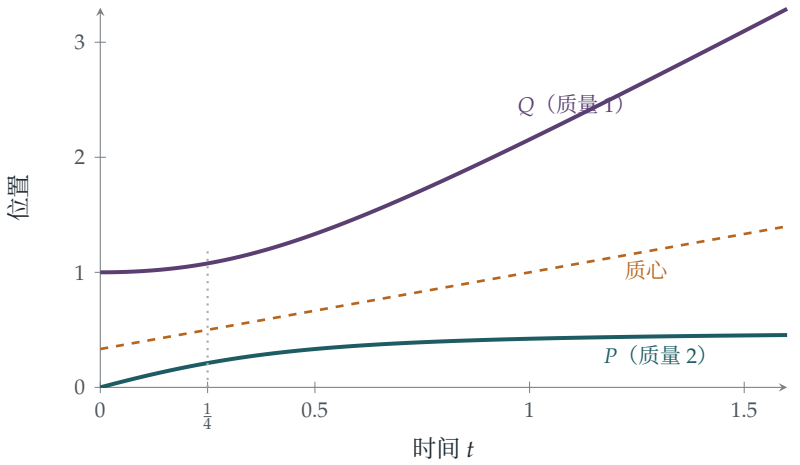


图 10.6.1 P（蓝绿色）和 Q（紫红色）的位置随时间的变化。P 向 Q 移动，受排斥力减慢并折返，逐渐趋近一个固定位置，而 Q 则被推开。两者在 $t = \frac{1}{4}$ 时最接近。质心（琥珀色虚线）自始至终以恒定速率移动。

回顾

- **更换变量以解除耦合。**两条方程之差只含有 z；加权之和只涉及质心。
- **能量积分。**写成 $\ddot{z} = v\frac{\mathrm{d}v}{\mathrm{d}z}$，便把二阶方程化为一阶方程：这就是能量守恒。
- **更换未知量。**z 的方程不好处理；z^2 的方程却简单不过。

背后的数学

为甚么 z^2 是二次式

对任何形如 $\ddot{z} = k/z^3$ 的力律，同样的计算都给出 $\frac{\mathrm{d}^2}{\mathrm{d}t^2}(z^2) = 2\dot{z}^2 + 2z\ddot{z} = 2\dot{z}^2 + \frac{2k}{z^2} =$ 常数，因为 $\frac{1}{2}\dot{z}^2 + \frac{k}{2z^2}$ 正是守恒的能量。所以 z^2 必定是 t 的二次式。为甚么？

隐藏的原因是这样的。设想一个质点在平面上自由运动，即沿直线以恒定速率前进。它与原点的距离 r 满足 $r^2 = (\text{一个二次式，变量为} t)$，因为 $r^2 = (x_0 + at)^2 + (y_0 + bt)^2$。而在极坐标（polar coordinates）中，它的径向运动方程是 $\ddot{r} - r\dot{\theta}^2 = 0$；由于角动量（angular momentum）$h = r^2\dot{\theta}$ 是常数，方程变成

$$\ddot{r} = \frac{h^2}{r^3}.$$

立方反比的排斥力，正正就是隐藏的角向运动所产生的离心（centrifugal）效应。在我们的问题中，$z^2 = 4t^2 - 2t + 1 = (2t - \tfrac{1}{2})^2 + \tfrac{3}{4}$ 是某一点与原点距离的平方，该点以速率 2 沿一条直线运动，而这条直线与原点的距离为 $\tfrac{\sqrt{3}}{2}$。一维上互相排斥的两个质点，就是二维自由运动的影子。

二体问题

只要两个物体之间的力仅取决于它们的间距，并沿两者的连线作用，本题的方法便适用，一维如是，三维亦如是。质心作匀速运动，而间距向量 $\mathbf{r}$ 服从 $\mu\ddot{\mathbf{r}} = \mathbf{F}(\mathbf{r})$，其中 $\mu = \frac{m_1 m_2}{m_1 + m_2}$。对引力而言，这就是开普勒问题（Kepler problem），由牛顿在《自然哲学的数学原理》（*Principia*）中解决：间距描出的轨迹是椭圆、抛物线或双曲线。地球和太阳都绕着它们的共同质心运行，而这个质心位于太阳内部；地球和月球则绕着距离地心约 4700 km 的一点运行，这一点在地球内部，但离地心颇远，所以地球每月都会摆动一次。天文学家正是靠侦测恒星的这种摆动，来寻找环绕其他恒星的行星。

可积系统

1969 年，Francesco Calogero 研究直线上多个质点以平方反比势（即立方反比力）互相排斥的情况，发现无论质点数目多少，这个系统竟然都可以精确求解。Jürgen Moser 在 1975 年解释了个中原因，他用上一个矩阵，其特征值（eigenvalues）在矩阵随时间演化时保持不变：这就是 Section 7.7 的 *Lax* 对（Lax pairs）。*Calogero–Moser* 系统如今是可积系统（integrable system）的核心例子，所谓可积系统，就是守恒量多得足以完全求解的系统；它把力学与随机矩阵、量子场论及孤立子理论联系起来。

思维技巧

选择能解除耦合的坐标

当两个量互相耦合时，找出不耦合的组合。和与差，配上适当的常数作加权，往往能把问题拆成互相独立的部分。

质心坐标与相对坐标是第一个例子。到了大学，同一个想法会以矩阵对角化（简正模态（normal modes），Section 11.2）、偏微分方程的分离变量法等形式出现，也见于量子力学。

延伸练习

1 **绳上的两个质量。**试做 STEP III 2023，第 9 题（载于 OCR 提供下载的 2023 年试卷，`www.ocr.org.uk/Images/707102-2023-past-papers.zip`）。题目考察以绳相连的两个质点，并利用质心把运动分离。留意其中一个质点何时向后移动。

2 **吸引力。**把力改为吸引力 $\frac{2}{z^3}$，重做本题。证明 z^2 仍然是 t 的二次式，并求两个质点何时相撞。

3 **隐藏的平面。**就题目中的运动，求出平面上与原点距离为 $z(t)$ 的那条直线运动轨迹，以及扫过的角 $\theta(t)$。验证 $z^2\dot{\theta}$ 是常数，且 $(z^2\dot{\theta})^2 = 3$。

4 **开普勒第三定律。**两颗质量分别为 m_1 和 m_2 的恒星相距 a，在引力作用下绕它们的质心作圆周运动。证明周期 T 满足 $T^2 = \frac{4\pi^2 a^3}{G(m_1+m_2)}$。

延伸阅读

- **John Taylor, *Classical Mechanics***（University Science Books），第 8 章。二体问题、约化质量与轨道。
- **Daniel Kleppner & Robert Kolenkow, *An Introduction to Mechanics***（Cambridge University Press），第 10 章，论有心力。
- **Francesco Calogero, *Classical Many-Body Problems Amenable to Exact Treatments***（Springer）。给有志挑战的读者：精确可解系统的世界。

10.7 球壳内的引力

题目

STEP III 1999，第 9 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/99-s3-q9`

引子

试想像一颗空心的行星：一层薄薄的岩石球壳（spherical shell），里面空无一物。假如你站在球壳内靠近壳壁的地方，也许会以为自己会被附近的岩石吸过去。事实上你会浮在原处：球壳内任何一处，球壳的引力都恰好是零。球壳较近的一侧距离较短，但较远的一侧面积较大，两种效应刚好完全抵销。

这就是**牛顿球壳定理**（*Newton's shell theorem*）。牛顿（Newton）花了多年才把它证明出来，而它对牛顿的引力理论至为重要：连同与它配对的另一个结果（在球体之外，引力就如全部质量都集中在球心一样），它说明了为何可以把地球和太阳当作点来处理。在练习和“背后的数学”一节中，我们会用这个定理计算一团气体云塌缩成恒星需要多少时间。

动笔之前

想一想 考虑圆环上两小段位于同一直径两端的部分。画出它们对质点（particle）所施的力。哪些分量（component）互相抵销，哪些相加？

这两小段与质点的距离相同。它们的引力沿轴方向的分量相等，因而相加；垂直于轴的分量大小相等、方向相反，因而抵销。把整个圆环这样逐对配起来，可知合力沿着轴，指向圆环的中心。

逐步解题

阶段 1　圆环

圆环上每一小段与质点的距离都是 $\sqrt{x^2+r^2}$，而连接两者的直线与轴成角 ψ，其中 $\cos\psi = \frac{x}{\sqrt{x^2+r^2}}$。质量为 δm 的一小段所施的引力为 $\frac{G\,\delta m}{x^2+r^2}$，其中沿轴方向的分量为 $\frac{G\,\delta m}{x^2+r^2}\cdot\frac{x}{\sqrt{x^2+r^2}}$。每一小段的距离和角度都相同，所以这些分量直接相加，得

$$\frac{Gmx}{(x^2+r^2)^{3/2}}. \qquad \blacksquare$$

阶段 2　把球壳切成圆环

> **想一想**　设球壳的中心为 O，质点位于轴上，与 O 相距 b。把球壳切成多个垂直于轴的薄圆环，并以圆环上的点在 O 处与轴所成的角 ϕ 来标记各圆环。圆环的半径、位置和质量分别是甚么？

角度为 ϕ 的圆环半径为 $a\sin\phi$，其中心沿轴与 O 相距 $a\cos\phi$，因此与质点相距 $a\cos\phi - b$。圆环的宽度为 $a\,\delta\phi$，所以面积为 $2\pi a\sin\phi\cdot a\,\delta\phi$，占球壳面积 $4\pi a^2$ 的 $\frac{1}{2}\sin\phi\,\delta\phi$。它的质量是 $\frac{1}{2}M\sin\phi\,\delta\phi$。根据第 1 阶段的结果，它沿着轴把质点拉向 $\phi = 0$ 的一端，力的大小为

$$\frac{G\cdot\frac{1}{2}M\sin\phi\,\delta\phi\,(a\cos\phi - b)}{\big((a\cos\phi-b)^2 + a^2\sin^2\phi\big)^{3/2}} = \frac{GM}{2}\cdot\frac{(a\cos\phi-b)\sin\phi\,\delta\phi}{(a^2-2ab\cos\phi+b^2)^{3/2}}.$$

（正负号会自动处理妥当：位于质点前方、满足 $a\cos\phi > b$ 的圆环把质点拉向一方，位于质点后方的圆环则把它拉向另一方。）

阶段 3　合适的变量

> **想一想**　分母含有 $s^2 = a^2 - 2ab\cos\phi + b^2$。$s$ 在几何上代表甚么？$\mathrm{d}s$ 以 $\mathrm{d}\phi$ 表示是甚么？

根据余弦定理（cosine rule），s 是质点到圆环的距离。求导可得 $2s\,\mathrm{d}s = 2ab\sin\phi\,\mathrm{d}\phi$，而由 $s^2 = a^2 - 2ab\cos\phi + b^2$ 解出 $a\cos\phi$，可得 $a\cos\phi - b = \frac{a^2-b^2-s^2}{2b}$。所以合力为

$$\frac{GM}{2}\int\frac{a^2-b^2-s^2}{2b\,s^3}\cdot\frac{s\,\mathrm{d}s}{ab} = \frac{GM}{4ab^2}\int_{a-b}^{a+b}\Big(\frac{a^2-b^2}{s^2}-1\Big)\mathrm{d}s,$$

其中 s 由 $a-b$（最近的点，$\phi=0$）变到 $a+b$（最远的点，$\phi=\pi$）。

阶段 4　结果为零

$$\int_{a-b}^{a+b}\Big(\frac{a^2-b^2}{s^2}-1\Big)\,\mathrm{d}s=(a^2-b^2)\Big(\frac{1}{a-b}-\frac{1}{a+b}\Big)-2b=(a^2-b^2)\cdot\frac{2b}{a^2-b^2}-2b=0.$$

球壳对质点不施任何力。 ■

> **停一停，试一试** 如果质点在球壳之外，即 $b>a$，情况会有甚么不同？s 的上下限变成 $b-a$ 和 $b+a$。重做最后一行，并证明所受的力为 $\frac{GM}{b^2}$，指向球心，就如全部质量都集中在球心一样。

回顾

- **先看对称性（symmetry）。**把相对的小段配对，未经任何计算便已看出力的方向。
- **由较简单的形状逐步建构。**点组成圆环；圆环组成球壳。
- **选择切合物理情境的变量。**把变量由角 ϕ 换成距离 s，一个棘手的积分便变得轻而易举。

背后的数学

牛顿的圆锥论证

牛顿本人的证明（《自然哲学的数学原理》（*Principia*）第一卷命题 70）完全没有用到积分。以质点为顶点，画一个狭窄的对顶圆锥，它在球壳上截出两小块曲面（图 10.7.1）。这两小块的面积与它们到质点的距离的平方 r_1^2 和 r_2^2 成正比（两小块与圆锥所成的角相等，因为一条弦在两端与圆所成的角相等）。所以它们的质量与 r_1^2 和 r_2^2 成正比，而它们的引力与质量除以距离平方成正比，因此大小相等、方向相反。每一个通过质点的对顶圆锥都互相抵销，所以合力为零。这种抵销之所以成立，正是因为平方反比定律（inverse-square law）：若换成距离的任何其他次方，球壳便会对质点产生推力或拉力。

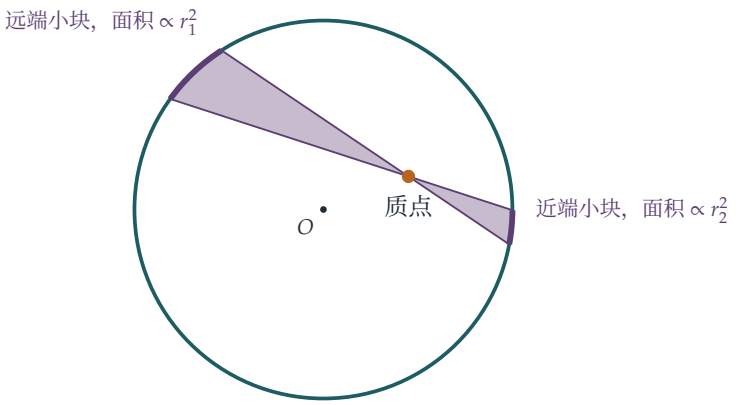


图 10.7.1　牛顿的论证。由质点出发的一个细窄对顶圆锥在球壳上截出两小块。远端的一块较大，面积之比等于距离平方之比，但它的引力也恰好按同一比例减弱。两个引力互相抵销，其余每一对亦然：在球壳之内，引力消失。

塌缩的云团

STEP III 1987 Q14 把这个定理反过来用。在均匀球体之外，位于球面的质点所受的引力为 GMm/r^2，就如全部质量都集中在球心一样。现在取一团静止的均匀球状尘埃云，半径为 R，质量为 M，让它在自身引力下塌缩。球面上的质点只感受到它以内的质量，而在所有物质向内坠落的过程中，这质量一直是 M，所以它向内的速率 $v = -\dot{r}$ 满足

$$v\frac{\mathrm{d}v}{\mathrm{d}r} = -\frac{GM}{r^2}, \qquad \text{所以} \qquad v^2 = 2GM\left(\frac{1}{r} - \frac{1}{R}\right).$$

代入 $r = R\cos^2\alpha$（其中一道练习），可得到达半径 r 所需的时间为 $\sqrt{\frac{R^3}{2GM}}\left(\alpha + \frac{1}{2}\sin 2\alpha\right)$。当 $\alpha = \frac{\pi}{2}$ 时，整团云塌缩成一点，所需时间就是**自由落体时间**（*free-fall time*）

$$t_{\mathrm{ff}} = \frac{\pi}{2}\sqrt{\frac{R^3}{2GM}} = \sqrt{\frac{3\pi}{32G\rho}},$$

其中 ρ 是密度。值得注意的是，云团的每一层球壳都在同一刻到达中心，因为这个时间只取决于密度（图 10.7.2）。对于一团典型的星际云，每立方米约有 10^{-17} kg，t_{ff} 约为 700 000 年：这决定了恒星形成的时间尺度。

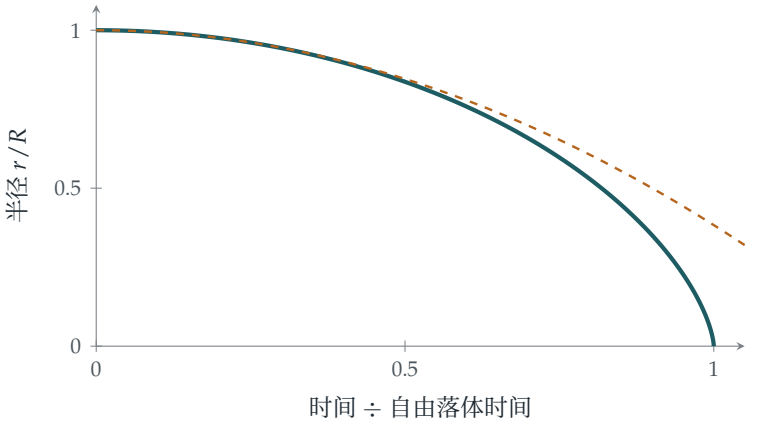


图 10.7.2 塌缩中的均匀尘埃云的半径随时间的变化（青绿色曲线）。它起初很慢，就像一块刚放手的石头（琥珀色虚线抛物线是恒定加速度的近似 $r = R - \frac{1}{2}\frac{GM}{R^2}t^2$），其后急剧加速，并在 $t = t_{\text{ff}}$ 时降至零。这条曲线是半条摆线（cycloid）。

牛顿的延迟与高斯定律

牛顿早在 1660 年代已掌握平方反比定律，但到 1687 年才发表他的引力理论。据一些历史学家所说，其中一个原因是他在证明球壳定理之前，无法说明为何可以把地球当作点质量（point mass）；一旦证明了，《自然哲学的数学原理》的其余部分便水到渠成。今天，这个定理只是高斯定律（*Gauss's law*）的一个简短推论。高斯定律指出，穿过任何闭合曲面向外流出的引力场总量，与曲面内的质量成正比。在球壳内取一个同心的球面。它没有包围任何质量，所以穿过它的总流量为零。单凭这一点并不足以令场为零（均匀场穿过任何闭合曲面的总流量都是零，从一边流入，从另一边流出），但这里对称性帮了忙：场必定笔直指向球心或背离球心，而且在球面上各处强度相同，所以总流量为零就意味着场为零。同一条定律也支配电场，这就是为甚么带电金属壳内没有电场，也是为甚么汽车能保护你免受雷击。

思维技巧

积分即求和

要求出一个连续物体的效应，可把它切成效应已知的小块，再把它们加起来。利用对称性来选择怎样切。

点组成圆环，圆环组成球壳，球壳组成实心球。在大学里，每一种场（引力场、电场、磁场）都是这样由场源计算出来的，由此更引出高斯（Gauss）定理和斯托克斯（Stokes）定理的向量微积分（vector calculus）。

延伸练习

1 **球壳之外。**完成“暂停思考”方框中的问题：证明在球壳之外，所受的力为 $\frac{GM}{b^2}$。由此推断，质量为 M 的均匀实心球对球外质点的引力，就如其质量集中在球心一样。

2 **塌缩时间。**试做 STEP III 1987，第 14 题（`step.maths.org/questions/87-s3-q14`）。这道题求一团尘埃云在自身引力下塌缩需要多少时间；用三角代换（trigonometric substitution）便可算出积分。

3 **穿越地球的隧道。**证明在均匀实心球之内，与球心相距 b 处所受的力为 $\frac{GMb}{a^3}$，与 b 成正比。由此推断，把一个球投进一条穿过地球中心的隧道，它会来回振荡，周期（period）为 $2\pi\sqrt{a^3/(GM)}$，约 84 分钟，与一颗贴着地面飞行的卫星的周期相同。

4 **另一种力的定律。**假如引力遵从立方反比定律，球壳内所受的力是否仍然为零？（重做第 3 阶段，以 $\frac{1}{s^3}$ 取代 $\frac{1}{s^2}$。）

延伸阅读

- **Daniel Kleppner & Robert Kolenkow, *An Introduction to Mechanics***（Cambridge University Press），关于万有引力的章节，当中包括球壳定理。
- **Subrahmanyan Chandrasekhar, *Newton's Principia for the Common Reader***（Oxford University Press）。由一位诺贝尔奖得主讲解牛顿的证明，包括命题 70 和 71。
- **Andrew Pontzen, *The Universe in a Box***（Riverhead）。讲述引力、塌缩和模拟如何解释宇宙的结构。

10.8 排成三角形的三颗恒星

题目

STEP I 1999，第 11 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/99-s1-q11`

引子

两个以引力互相吸引的天体，其运动可以完全解出，这是牛顿（Newton）证明了的，Section 10.6 亦已解释过：两者各自沿一条圆锥曲线（conic）绕共同的质心（centre of mass）运动，若两者互相束缚，轨道是椭圆；若能量足以互相逃逸，则是抛物线或双曲线。再加入第三个天体，一切就不同了。*三体问题*（three-body problem）没有能用常见函数表示的通解，而在 1890 年，庞加莱（Henri Poincaré）更发现它的解可以是混沌的，这是混沌（chaos）在数学中首次出现。

然而，有些特殊解却十分简单。1772 年，拉格朗日（Joseph-Louis Lagrange）发现，位于等边三角形（equilateral triangle）三个顶点的三个天体可以作刚性旋转，整个三角形像车轮一样转动。本题要找的，就是质量相等时的这个解。拉格朗日的三角形并非只是趣闻：太阳、木星和数以千计的“特洛伊”（Trojan）小行星正好构成这种位形，今天的太空望远镜亦停泊在类似的点上。

动笔之前

想一想 一颗星要以角速度（angular velocity）ω 沿半径为 R 的圆运动，它所受的总力必须是甚么？

总力的大小必须是 $m\omega^2 R$，方向指向圆心。因此我们要检查每颗星所受的总引力是否指向三角形的中心，并求出使其大小恰好相符的 ω。

逐步解题

阶段 1　几何

想一想　边长为 a 的等边三角形，每个顶点与中心相距多远？

中心位于每条中线（median）上，与顶点的距离是中线长度的三分之二。中线的长度为 $a\sin 60^\circ = \frac{\sqrt{3}}{2}a$，所以每个顶点到中心的距离是

$$R = \frac{2}{3}\cdot\frac{\sqrt{3}}{2}a = \frac{a}{\sqrt{3}}.$$

由对称性（symmetry）可知，三角形的中心就是三颗相同的星的质心。

阶段 2　一颗星所受的力

想一想　另外两颗星各以大小为 $\frac{\gamma m^2}{a^2}$ 的力拉它。这两个拉力之间的夹角是多少？它们的合力（resultant）是多少，又指向哪个方向？

两个拉力分别沿着三角形在该星处相交的两条边，所以它们之间的夹角是 60°。由对称性可知，合力沿着这个角的角平分线（bisector），而角平分线通过中心。每个拉力与角平分线成 30°，所以合力的大小是

$$2\cdot\frac{\gamma m^2}{a^2}\cos 30^\circ = \frac{\sqrt{3}\,\gamma m^2}{a^2},$$

方向指向中心（图 10.8.1）。

阶段 3　角速度

令合力等于 $m\omega^2 R$：

$$m\omega^2\frac{a}{\sqrt{3}} = \frac{\sqrt{3}\,\gamma m^2}{a^2}, \qquad \text{所以} \qquad \omega^2 = \frac{3\gamma m}{a^3}.$$

取这个角速度时，同样的计算（由对称性）适用于每一颗星，所以三颗星都绕中心作圆周运动，三角形保持刚性不变。 ■

阶段 4　中心的第四颗星

想一想　一颗质量为 λm 的星位于中心，与每个顶点的距离是 $\frac{a}{\sqrt{3}}$。它对每颗顶点上的星额外施加多大的力？质心会移动吗？中心的星有没有受到净力？

中心的星不会改变质心的位置，而由对称性可知，它所受的三个拉力互相抵消，所以它保持静止。它把每颗顶点上的星拉向中心，力的大小是 $\frac{\gamma\lambda m^2}{(a/\sqrt{3})^2} = \frac{3\gamma\lambda m^2}{a^2}$。现在顶点上的星所受指向中心的总力是 $(\sqrt{3}+3\lambda)\frac{\gamma m^2}{a^2}$，所以

$$m\omega^2\frac{a}{\sqrt{3}} = (\sqrt{3}+3\lambda)\frac{\gamma m^2}{a^2}, \qquad \omega^2 = \frac{(3+3\sqrt{3}\,\lambda)\gamma m}{a^3} = \frac{3\gamma m}{a^3}(1+\sqrt{3}\,\lambda). \qquad \blacksquare$$

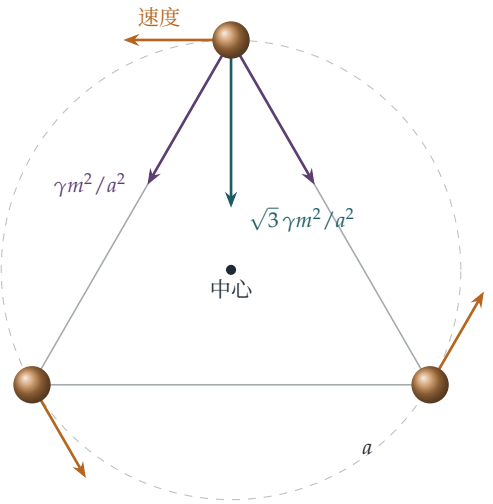


图 10.8.1 三颗相同的星位于等边三角形的三个顶点。顶部那颗星所受的两个拉力（紫色）合成一个指向中心的力（蓝绿色）。若每颗星以适当的速率（琥珀色）沿虚线圆运动，整个三角形便作刚性旋转。这就是拉格朗日对三体问题的等边三角形解。

回顾

- **猜一个对称的解。**我们并没有解出三体问题；我们只是猜了一种高度对称的运动，再检查牛顿定律是否容许它。
- **对称性决定方向。**合力必定沿着角平分线，所以只需计算它的大小。
- **检查每一个天体。**由对称性可知，验证了一颗星就等于验证了全部三颗；中心的星则需要另作（容易的）检查。

背后的数学

乔装的开普勒定律

把答案写成 $\omega^2 a^3 = \gamma(3m)$。对于相距 a、沿圆形轨道运行的两个天体，开普勒（Kepler）第三定律（Section 10.6 的一道练习）指出 $\omega^2 a^3 = \gamma(m_1 + m_2)$，右边出现的是总质量。三颗星遵从同一条定律，总质量为 $3m$。事实上，拉格朗日证明了等边三角形解对*任意*三个质量都存在，而不限于相等的质量：三角形绕质心旋转，并且 $\omega^2 a^3 = \gamma(m_1 + m_2 + m_3)$。欧拉（Euler）在几年前已找到第二族解，其中三个天体位于一条旋转的直线上。这五种位形（三种共线、两种三角形，不计标号的差别）是三个天体能保持形状不变的仅有情况。

拉格朗日点

取 m_1 为太阳，m_2 为一颗行星，m_3 为一颗微小的小行星。等边三角形解告诉我们，小行星可以停留在与太阳和行星构成等边三角形的第三个顶点上，在行星轨道上位于行星的前方或后方。这就是*拉格朗日点*（Lagrange points）L_4 和 L_5；欧拉的共线解则给出 L_1、L_2 和 L_3。木星有超过一万颗已知的“特洛伊”小行星，聚集在它的 L_4 和 L_5 点附近。詹姆斯·韦伯太空望远镜（James Webb Space Telescope）在日地系统的 L_2 点附近运行，位于地球以外 1.5 百万公里处，在那里它可以把太阳、地球和月球全部挡在遮阳罩之后。

稳定性

造星者的三角形能维持下去吗？对于相等的质量，答案是不能：最轻微的扰动也会不断增长，几颗星最终不是四散飞离，就是互相碰撞。Edward Routh 在 1875 年证明，拉格朗日的三角形只有在其中一个质量占绝对优势时才是稳定（stable）的，确切的条件是

$$27(m_1 m_2 + m_2 m_3 + m_3 m_1) < (m_1 + m_2 + m_3)^2.$$

对于太阳和木星，这个条件绰绰有余地成立，这就是特洛伊小行星得以长存的原因。对于三个相等的质量，$27 \times 3m^2 > 9m^2$，所以这个位形并不稳定。

庞加莱的错误与 8 字形轨道

1887 年，瑞典国王奥斯卡二世（Oscar II）悬赏征求多体问题的解。庞加莱在 1889 年凭一篇关于三体问题的长篇论文获奖，但论文付印期间，一位编辑发现了其中一处漏洞。在修正的过程中，庞加莱发现有些解的表现“复杂得惊人，我甚至无法尝试把它画出来”：这是人类对混沌的第一瞥。已印好的论文全数收回销毁，费用由庞加莱承担。新的特殊解至今仍不断被发现：1993 年 Cristopher Moore 以数值方法发现，而 2000 年 Alain Chenciner 与 Richard Montgomery 证明，三个相等的质量可以沿一条 8 字形曲线互相追逐。此后，人们借助电脑又找到了数以千计的周期轨道。

思维技巧

猜一个对称的解，然后验证

当一个问题难以在一般情况下解出时，不妨寻找对称性尽可能高的解。对称的猜想能把许多未知量减少到一两个，而验证它往往很容易。

这是数学和物理中一个强而有力的策略：对称的平衡态、稳态、行波和刚性旋转的位形，往往是最先找到的解，也是研究其余各种解的起点。接下来要问的，永远是这个对称解是否稳定。

延伸练习

1 **正方形上的四颗星。**四颗质量为 m 的星位于边长为 a 的正方形的四个顶点。证明它们可以绕中心作刚性旋转，其中 $\omega^2 = \frac{(4+\sqrt{2})\gamma m}{2a^3}$。

2 **两颗相同的星和一颗行星。**两颗质量为 M 的星相距 a，绕它们的质心运行。一颗质量可忽略的行星位于中心。它是否处于平衡？对于沿两星连线方向的微小位移，它是否稳定？垂直于连线的方向又如何？

3 **不相等的质量。**质量分别为 m_1、m_2、m_3 的三颗星位于边长为 a 的等边三角形的三个顶点，并绕它们的质心 G 旋转。证明 m_1 所受的引力为 $\frac{\gamma m_1(m_1+m_2+m_3)}{a^3}\overrightarrow{P_1G}$，其中 P_1 是它的位置。（提示：把各力写成向量，并利用质心的定义。）由此推导出拉格朗日的结果。

4 n **颗星。**n 颗相同的星位于一个正 n 边形的各个顶点，该正多边形内接于半径为 R 的圆。求刚性旋转的 ω^2，并把它表示成 $\frac{1}{\sin(k\pi/n)}$ 对 k 求和的形式。

延伸阅读

- **June Barrow-Green, *Poincaré and the Three Body Problem***（American Mathematical Society）。讲述那次悬赏以及庞加莱发现混沌的故事。
- **Richard Montgomery, "A new solution to the three-body problem"**, *Notices of the American Mathematical Society* 48 (2001)。由 8 字形轨道的其中一位发现者亲自讲解；可于网上免费阅读。
- **Florin Diacu & Philip Holmes, *Celestial Encounters***（Princeton University Press）。一部关于三体问题与混沌的通俗历史。

我们学到了甚么

本章开首说过，有些东西永远不变，而知道哪些东西不变，就已经掌握了半门物理学。我们提出了三个问题：守恒定律（conservation laws）单凭自身能决定甚么、细小的步骤如何累积起来，以及怎样驯服多个物体组成的系统。以下是本章八道题目给出的答案。

1. 守恒定律单凭自身能决定甚么？

两个物体沿直线作弹性碰撞（elastic collision）时，动量（momentum）和能量（energy）完全决定了碰撞的结果，而这一个结论在本章一再派上用场。物体再多一些，情况就不同了。在 Section 10.1 中，一排弹珠有许多种结果都同时符合这两条定律，要再加上一项物理事实（弹珠不能互相穿过），才能排除其中一些。把动量的平方减去能量，得出一个由交叉项组成的和，而这个和必须等于零；这个不等式（inequality）论证证明了某些运动根本不可能发生。至于牛顿摆（Newton's cradle）那种人所熟知的表现，原来取决于守恒定律以外的细节：冲击如何在球与球之间传递。

守恒定律也揭露了隐藏的假设。在 Section 10.3 中，所有炮同时发射所得的速率，比逐一发射为高，但这只是因为同时发射用了更多能量；一旦把该固定的量固定下来，结论就反过来了。

2. 细小的步骤如何累积起来？

有好几节都是把一个简单的步骤重复又重复。在 Section 10.2 中，一次碰撞最多只能令速率加倍，但经过一串愈来愈轻的球，增幅就变得没有上限：这正是篮球叠网球背后的原理，也是引力弹弓（gravitational slingshot）背后的原理。在 Section 10.3 中，许多次细小的发射化成了一个积分（integral），并得出齐奥尔科夫斯基（Tsiolkovsky）的对数（logarithm），由此解释了为何火箭的大部分都是燃料，以及为何火箭要分级。在 Section 10.4 中，分子每次从移动中的活塞反弹，速率都增加 $2u$，而一个精确的递推关系（recurrence）显示，速率与长度的乘积几乎保持不变：这是一个绝热不变量（adiabatic invariant），也是气体受压缩时升温的原因。而在 Section 10.5 中，无穷多次反弹只花了有限的时间，因为各段时间构成一个收敛（convergent）的等比级数（geometric series）。

3. 怎样驯服多个物体组成的系统?

方法是选取合适的坐标，并善用对称性（symmetry）。在 Section 10.6 中，质心（centre of mass）作匀速运动，而两物体的间距则表现得像一个带有约化质量（reduced mass）的单一质点；一个隐藏的平面甚至解释了为何间距的平方是时间的二次式。在 Section 10.7 中，对称性加上积分，证明了球壳（spherical shell）对内部不施加引力，对外部则有如一个质点；于是一团正在坍缩的尘埃云可以用单独一条方程来处理，而当中的自由下落时间（free-fall time）决定了恒星形成的步伐。而在 Section 10.8 中，一个对称的猜想解决了无法求解的三体问题（three-body problem）的一个特殊情况：拉格朗日（Lagrange）的旋转三角形，它在特洛伊小行星（Trojan asteroids）身上得到了体现。

4. 力学工具箱

方法	你在哪里用过
以守恒定律作为约束条件	牛顿摆（10.1）
转换参考系（frame）	碰撞（10.2）；活塞（10.4）
反复运用同一结果	叠球（10.2）
质量可变时的动量	火箭（10.3）
由求和到积分	火箭方程（10.3）；球壳（10.7）
递推关系与裂项相消（telescoping）	活塞（10.4）
等比级数	楔（10.5）
质心与约化质量	二体（10.6）
能量积分 $v\,\mathrm{d}v/\mathrm{d}z$	二体（10.6）；坍缩（10.7）
对称的猜想	三星（10.8）

下一步

- 第 11 章由守恒转到振荡（oscillation）与转动（rotation）：共振（resonance）、简正模式（normal modes）、真实的摆，以及在加速参考系和旋转参考系中的运动。
- Section 8.2 的能量方法和 Section 8.4 的极限环（limit cycles），是同一批思想在微分方程（differential equation）方面的体现。
- 庞加莱（Poincaré）在三体问题中发现的混沌（chaos），是 Section 8.7 的主题。

如果你想在入读大学之前再走远一点

- **Daniel Kleppner & Robert Kolenkow, *An Introduction to Mechanics***（力学导论）：经典的大学一年级教科书；要求很高，但收获甚丰。

- **David Morin, *Introduction to Classical Mechanics***（古典力学导论）：收录数以百计的题目，当中不少与本章的题目一样出人意表。
- **Richard Feynman, *The Character of Physical Law***（物理定律的本性）：为一般听众而设的讲座，内容涵盖守恒定律、对称性和引力。
- **Dwight Neuenschwander, *Emmy Noether's Wonderful Theorem***（诺特的奇妙定理）：为何有对称性就有守恒。

费曼（Feynman）说过，能量守恒至今未见任何例外。本章尝试展示守恒定律的威力有多大：排除不可能的结果、揭示隐藏的不变量，并把复杂得令人绝望的系统化为我们能够理解的系统。但本章也展示了它们的局限。守恒定律能够约束，却不一定能够决定；在它们止步之处，就要由力、对称性和参考系的细节接手。

延伸练习答案

以下只是简短的答案和提示，并非完整解答：只列出结果和关键步骤，方便你核对自己的做法。每道练习请先认真做过，然后才看答案。

10.1 牛顿摆

1. 所有碰撞（collision）结束后，运动中的质点后面不可能跟着一个较慢的质点。因此，若恰好只有一个质点在动，它必定是 A_n，而 $\lambda m v = m u$、$\lambda m v^2 = m u^2$ 在 $\lambda \neq 1$ 时互相矛盾。若只有 A_{n-1} 和 A_n 在动，则方程与质量为 m 和 λm 的两个质点相碰时的方程相同，得出 A_{n-1} 的速度为 $\frac{1-\lambda}{1+\lambda}u < 0$；但这样它便会撞上 A_{n-2}。有三个质点在动的情况，可完全仿照第 (ii) 部分排除。在动的两个质点是 A_0（以 $\frac{\lambda-1}{\lambda+1}u$ 向后运动）和 A_n（以 $\frac{2u}{\lambda+1}$ 向前运动）。

2. 质量相等的弹珠相碰时会交换速度，所以速率 u 一直传到 A_n。之后 A_n 撞上 A_{n+1}：A_{n+1} 以 $\frac{2u}{1+\lambda}$ 离开，而 A_n 则以 $\frac{\lambda-1}{\lambda+1}u$ 反弹。这个向后的速度沿着整列弹珠传回 A_0，使它以该速率向后离开。

3. “两球以 u 弹出”所带的动量（momentum）是 $2mu$，能量是 mu^2，与碰撞前相同。另一个解（之后不再有碰撞）是：最后三个球分别以 $\frac{u}{3}$、$\frac{u}{3}$ 和 $\frac{4u}{3}$ 离开（验算：总和为 $2u$，平方和为 $2u^2$）。真实的牛顿摆选择第一个解，是因为各球之间并非完全贴紧，所以碰撞是一对一对地先后发生的。

4. 速率一直传到 A_n，它再撞上较轻的 A_{n+1}：A_{n+1} 以 $\frac{2u}{1+\lambda}$ 离开，而 A_n 则以较慢的 $\frac{1-\lambda}{1+\lambda}u$ 跟在后面。其余质点全都不动。

10.2 叠球效应

1. 球以 $U = \sqrt{2gH}$ 到达地面，再以 eU 反弹，上升 e^2H。把小球放在上面时：大球以 eU 向上运动，遇上以 U 向下运动的小球。由动量守恒和恢复系数（restitution）的关系，可得小球向上的速率为 $w = U\frac{Me(2+e)-m}{M+m}$。小球球心由 $2R + r = 0.45$ 出发，升至 4.5，所以 $\frac{w^2}{2g} = 4.05$，而 $\frac{w}{U} = \sqrt{\frac{4.05}{1.8}} = 1.5$。代入 $e = \frac{2}{3}$，得 $\frac{16}{9}M - m = 1.5(M+m)$，于是 $M = 9m$。

2. 由动量守恒和恢复系数的关系，P_1 的速率为 $\frac{1+e}{1+\lambda}u$；每次碰撞的质量比都相同，情况重复出现，所以 P_n 以 $\left(\frac{1+e}{1+\lambda}\right)^n u$ 运动。速率为 s 的质点撞上下一个质点后，余下的

速率为 $\frac{1-\lambda e}{1+\lambda}s$，而下一个质点经过自己的下一次碰撞后，速率减至 $\frac{(1+e)(1-\lambda e)}{(1+\lambda)^2}s$。当 $\lambda e<1$ 时，前者只有在 $\frac{1-\lambda e}{1+\lambda}>\frac{(1+e)(1-\lambda e)}{(1+\lambda)^2}$ 时才会再次追上后者，即 $\lambda>e$；所以若 $e>\lambda$，每个质点只会被撞一次，也只会撞别的质点一次。若 $\lambda e=1$，每个质点撞上下一个之后便立即停下。

3. 最底的球以 U 反弹，遇上中间的球（正以 U 向下运动）：相对速率为 $2U$，所以中间的球以 $3U$ 离开。最上面的球与它相遇时相对速率为 $4U$，于是以 $7U$ 离开，上升 $49H$。一般而言 $u_{j+1}=2u_j+U$，所以 $u_j+U=2^jU$。

4. 弹性碰撞（elastic collision）中动能（kinetic energy）守恒。A 的能量是 $\frac12 m\big(\frac{2\lambda}{1+\lambda}\big)^{2n}u^2$，而最初的能量是 $\frac12\lambda^n mu^2$，又因 $\frac{4\lambda}{(1+\lambda)^2}<1$，所以 A 得到的能量永远不会多于原有的总能量。较重的质点继续向前运动，各自保留原来速率的 $\frac{\lambda-1}{\lambda+1}$ 倍，余下的能量便留在它们身上。

10.3 火箭与枪炮

1. 一次过全部发射：$Mu=nm(v-u)$，所以 $u=\frac{nmv}{M+nm}$。逐个发射：第 r 次发射给出 $\big(M+(n-r+1)m\big)(u_r-u_{r-1})=mv$，所以 $u_n=\sum_{j=1}^{n}\frac{mv}{M+jm}$，而除最后一项外，每一项都大于 $\frac{mv}{M+nm}$。因此 $u_n>u$。

2. 动量：$M(1+bx)v=MV$。于是 $(1+bx)\frac{\mathrm{d}x}{\mathrm{d}t}=V$，所以 $x+\frac12 bx^2=Vt$，而 $x=\frac{\sqrt{1+2bVt}-1}{b}$。有外力时，$\frac{\mathrm{d}}{\mathrm{d}t}\big(M(1+bx)v\big)=Mf$ 给出 $(1+bx)v=ft+V$。若速率为常数 c，则须有 $x=ct$ 及 $(1+bct)c=ft+V$，所以 $c=V$ 且 $bV^2=f$。否则 $x+\frac12 bx^2=\frac12 ft^2+Vt$，所以 $x\sim t\sqrt{f/b}$，而 $v\to\sqrt{f/b}$。

3. 单级：$v\ln 10\approx 2.30v$。两级：$v\ln\frac{10}{5.5}+v\ln\frac{5}{0.5}\approx 0.60v+2.30v=2.90v$。

4. 当 $0\le s\le x$ 时，$\frac{1}{1+x}\le\frac{1}{1+s}\le 1$；在 $[0,x]$ 上积分即可。

10.4 一个分子与一个活塞

1. 这个递推关系（recurrence）逐项相消后得 $d_n=\frac{(v+u)d_1}{v+(2n-1)u}$，所以 $v_nd_n=\frac{(v+2nu)(v+u)d_1}{v+(2n-1)u}$，它比 $(v+u)d_1$ 大 $\frac{u(v+u)d_1}{v+(2n-1)u}\to 0$。

2. 把 u 换成 $-u$：每次从后退中的活塞反弹，速率都减少 $2u$，所以 $v_n=v-2nu$（只要质点仍能追上活塞），$d_n=\frac{(v-u)d_1}{v-(2n-1)u}$，而 $v_nd_n=\frac{(v-2nu)(v-u)d_1}{v-(2n-1)u}$，同样接近 $(v-u)d_1$。

3. 当 $v=u$ 时，$v_n=(2n+1)u$。第 n 次碰撞时动量的改变量为 $m(v_{n-1}+v_n)=4nmu$，而 $t_n=\frac{d_1}{un(n+1)}$，所以平均力为 $\frac{4mu^2n^2(n+1)}{d_1}$，量级为 n^3。每个循环所作的功约为力 $\times\, ut_n$，大约是 $4nmu^2$，这些功加起来是无穷大：要把活塞完全推到底，需要无穷多的功，正如分子的能量 $\frac12 m(2n+1)^2u^2$ 会无止境地增长一样。

4. 不变量（invariant）是 $\oint p\,\mathrm{d}q=2\int_0^h m\sqrt{2g(h-z)}\,\mathrm{d}z\propto h^{3/2}$，其中 h 是弹跳离地

板的高度。所以当地板缓慢上升时，$h^{3/2}$（等价地说，即 v^3，或能量乘以周期）几乎保持不变。

10.5 有限时间内的无限次反弹

1. 由正弦定理(sine rule),$M_1N_1 = \frac{OM_1\sin 30^\circ}{\sin\theta}$,而相继各段路程按比率 $\rho = \frac{\sin(\theta-30^\circ)}{\sin\theta} < 1$ 缩短。总和为 $\frac{M_1N_1}{1-\rho} = \frac{OM_1}{2\big(\sin\theta-\sin(\theta-30^\circ)\big)} = \frac{OM_1}{4\sin 15^\circ\cos(\theta-15^\circ)}$。

2. 由三角形 OM_1N_1 的各角可得 $\varphi = \theta-\alpha$。时间之比为 $\frac{\rho}{k} = \frac{\sin(\theta-\alpha)\cos(\theta-\alpha)}{\sin\theta\cos\theta} = \frac{\sin(2\theta-2\alpha)}{\sin 2\theta}$，它小于 1 当且仅当 $\alpha < \theta < 45^\circ + \frac{1}{2}\alpha$。

3. 把楔形沿它的壁反射，可将路径展开成一条直线。围绕 O 刚好可放十二个楔形，而一条不通过 O 的直线最多穿过其中六个。所以质点只会反弹有限次，然后逃离，除非它正对着 O 射去。

4. 它首先下落 $t_0 = \sqrt{2/9.8} \approx 0.452\,\mathrm{s}$。每次弹跳历时 $2e^k t_0$，所以总时间为 $t_0\big(1+\frac{2e}{1-e}\big) = 9t_0 \approx 4.07\,\mathrm{s}$。总路程为 $1+2\sum_{k\ge1}e^{2k} = 1+\frac{2(0.64)}{0.36} \approx 4.56\,\mathrm{m}$。

10.6 二体化为一体

1. 绳子保持拉紧，相对速度与绳子垂直，大小恒为 u，所以 $\theta = \frac{ut}{r}$。由动量可得 $my + MY = mut$。利用 $A = G - \frac{M}{m+M}\overrightarrow{AB}$ 和 $\overrightarrow{AB} = r(\cos\theta, -\sin\theta)$，得 $y = \frac{1}{m+M}\big(mut + Mr\sin\frac{ut}{r}\big)$。于是 $\dot y = \frac{u}{m+M}\big(m + M\cos\frac{ut}{r}\big)$，它在 $\cos\frac{ut}{r} < -\frac{m}{M}$ 时为负，而这在 $M > m$ 时才有可能。

2. 现在 $\ddot z = -\frac{3}{z^3}$，所以 $\dot z^2 = \frac{3}{z^2} - 2$（利用 $z=1$、$\dot z = -1$）。设 $w = z^2$，则 $\ddot w = 2\dot z^2 + 2z\ddot z = -4$，所以 $z^2 = 1 - 2t - 2t^2$。两者在 $t = \frac{\sqrt{3}-1}{2} \approx 0.366$ 时相撞。

3. $z^2 = 4t^2 - 2t + 1 = |\mathbf{p} + \mathbf{q}t|^2$，其中 $\mathbf{p} = (1,0)$，$\mathbf{q} = (-1,\sqrt{3})$：这是通过 $(1,0)$、方向为 $(-1,\sqrt{3})$ 的直线，而 $\theta = \tan^{-1}\frac{\sqrt{3}t}{1-t}$。于是 $z^2\dot\theta = |\mathbf{p}\times\mathbf{q}| = \sqrt{3}$。这个斥力正好就是平面上自由运动的“离心”项 $\frac{h^2}{z^3}$，其角动量（angular momentum）为 $h = \sqrt{3}$。

4. 两颗星与质心(centre of mass)的距离分别为 $\frac{m_2a}{m_1+m_2}$ 和 $\frac{m_1a}{m_1+m_2}$。对第一颗星,$\frac{Gm_1m_2}{a^2} = m_1\omega^2\frac{m_2a}{m_1+m_2}$，所以 $\omega^2 = \frac{G(m_1+m_2)}{a^3}$，而 $T = \frac{2\pi}{\omega}$。

10.7 球壳内的引力

1. 在球壳外（$b > a$），与第 3 阶段相同的积分改为对 s 由 $b-a$ 积至 $b+a$，结果是 $\frac{GM}{b^2}$。实心球是多个同心球壳的并集，每个球壳的作用都如同其质量集中于球心。

2. 设 $r = R\cos^2\alpha$，则 $\frac{1}{r} - \frac{1}{R} = \frac{\tan^2\alpha}{R}$，所以速率为 $\sqrt{2GM/R}\tan\alpha$。于是 $\mathrm{d}t = \frac{2R\sin\alpha\cos\alpha\,\mathrm{d}\alpha}{\sqrt{2GM/R}\tan\alpha} = \sqrt{\frac{R^3}{2GM}}\,2\cos^2\alpha\,\mathrm{d}\alpha$；再积分即可。

3. 根据球壳定理（shell theorem），只有半径 b 以内的质量（即 $M\frac{b^3}{a^3}$）产生引力，结果是 $\frac{GMb}{a^3}$。这是简谐运动（simple harmonic motion），其中 $\omega^2 = \frac{GM}{a^3}$，周期为 $2\pi\sqrt{a^3/GM} \approx 84$ 分钟，这里取 $a = 6.37\times10^6\,\mathrm{m}$ 及 $GM = 3.99\times10^{14}\,\mathrm{m^3\,s^{-2}}$。在半径 a 处运行的卫星满足 $\omega^2 a = \frac{GM}{a^2}$：$\omega$ 相同。

4. 不会。换成 $\frac{1}{s^3}$ 后，被积函数不再化简为常数，球壳较近的一侧占了上风：质点会被拉向球壳最近的部分。只有平方反比定律（inverse-square law）才会完全抵消（球壳上一小块的面积按 s^2 增长，刚好抵消引力的减弱）。

10.8 排成三角形的三颗恒星

1. 每颗星与中心的距离为 $\frac{a}{\sqrt{2}}$。相邻两颗星的引力指向中心的分量合共为 $\frac{2\gamma m^2}{a^2}\cos45^\circ = \frac{\sqrt{2}\gamma m^2}{a^2}$，而对面那颗星的引力为 $\frac{\gamma m^2}{2a^2}$。令 $\frac{\gamma m^2}{a^2}\left(\sqrt{2}+\frac{1}{2}\right) = m\omega^2\frac{a}{\sqrt{2}}$，得 $\omega^2 = \frac{(4+\sqrt{2})\gamma m}{2a^3}$。

2. 在中心处各引力互相抵消，所以行星处于平衡（equilibrium）。沿着两颗星的连线方向，平衡是不稳定的：向其中一颗星移动距离 x，便会受到约为 $\frac{32\gamma Mx}{a^3}$ 的向外净力。在垂直于轨道平面的方向，平衡是稳定的，回复力约为 $\frac{16\gamma Mz}{a^3}$。（在轨道平面内垂直于连线的方向，引力有回复作用，但旋转会把这个方向的运动与不稳定的方向耦合起来。）

3. 作用于 m_1 的力为 $\frac{\gamma m_1}{a^3}\big(m_2(P_2-P_1)+m_3(P_3-P_1)\big)$。由于 $m_1P_1+m_2P_2+m_3P_3 = (m_1+m_2+m_3)G$，这个力等于 $\frac{\gamma m_1(m_1+m_2+m_3)}{a^3}(G-P_1)$。每颗星都被拉向 G，引力与它跟 G 的距离成正比，而且比例常数相同，所以三颗星可以像刚体一样一起旋转，其中 $\omega^2 = \frac{\gamma(m_1+m_2+m_3)}{a^3}$。

4. 相隔 k 个位置的那颗星，距离为 $2R\sin\frac{k\pi}{n}$，其引力指向中心的分量占 $\sin\frac{k\pi}{n}$。所以 $m\omega^2R = \frac{\gamma m^2}{4R^2}\sum_{k=1}^{n-1}\frac{1}{\sin(k\pi/n)}$，得 $\omega^2 = \frac{\gamma m}{4R^3}\sum_{k=1}^{n-1}\frac{1}{\sin(k\pi/n)}$。当 $n=3$、$R=\frac{a}{\sqrt{3}}$ 时，结果是 $\frac{3\gamma m}{a^3}$。

章

11

振动、转动与运动参考系

会摆动的东西、会旋转的东西，以及从移动中的位置所看到的东西。

本章内容

11.1	共振	STEP II 1991
11.2	简正模式	STEP III 2014
11.3	真实的摆	STEP III 1998
11.4	等时曲线	STEP III 1994
11.5	加速火车中的重力	STEP II 2021
11.6	后旋	STEP III 1997
11.7	杆中的辛普森法则	STEP II 1989
11.8	倾斜的平衡	STEP III 2025
11.9	安全抛物线	STEP II 2014
11.10	会朝你飞回来的球	STEP II 2019
11.11	外野手如何接球	STEP II 1998
11.12	橡皮绳上的蚂蚁	STEP II 2018
11.13	铲雪车问题	STEP III Specimen paper
11.14	是抛物线，不是悬链线	STEP I 1991
A	练习答案	第 815 页

看清运动

年轻理论物理学家的事业，就是在愈来愈抽象的层次上处理谐振子。

— 相传出自 Sidney Coleman

上一章利用守恒定律(conservation laws)来理解碰撞和引力。本章转向动力学(dynamics)入门课程的另外几个重要主题：会振动的东西、会转动的东西，以及从最有用的观点描述运动的艺术。

振动（oscillation）无处不在。单摆和弹簧是教科书上的例子，但同样的数学也描述了分子的振动、建筑物的摇晃、收音机里的电路，以及在量子力学中光本身的行为。原因在于，几乎任何处于稳定平衡（stable equilibrium）附近的东西，表现都像弹簧：把它拉开一点点，回复力（restoring force）便与位移成正比。因此，简谐振子（simple harmonic oscillator）是几乎所有物理系统的一阶近似，而物理学有很大一部分，就是要弄清楚当这个近似不太准确时会发生甚么事。

为甚么这很重要

工程师必须知道所建造的每一座结构的固有频率（natural frequency），使它们远离风、交通和地震的频率。化学家凭分子的振动来辨认分子。计时这门技艺，从惠更斯（Huygens）的摆钟到石英表和原子钟，就是制造完美振子的艺术。转动支配着一切，从桌球的旋转到人造卫星的稳定性皆然。至于哪些参考系（frame of reference）才算“自然”这个问题，则经由引力与加速度的等效性，引出了爱因斯坦（Einstein）的广义相对论。

本章最后有六道题目，归为一组，称为出人意表的轨迹：抛体、蚂蚁、外野手和铲雪车，它们的表现都与直觉所预料的不同。这些题目短小而优雅，每一道的核心都有一个真正的数学概念：包络线（envelope）、单调性（monotonicity）、反馈（feedback）、发散（divergence）、线性微分方程（linear differential equation），以及抛物线（parabola）与悬链线（catenary）之间的分别。

你在大学会遇到甚么

大学一年级的动力学课程涵盖简谐运动、阻尼（damped）谐运动和受迫（forced）谐运动、共振（resonance），以及耦合振子（coupled oscillators）及其简正模态（normal modes）（这是线性代数与力学交汇之处）。课程也会处理刚体（rigid body）、转动惯量

（moment of inertia）和滚动，以及在加速参考系和旋转参考系中的运动，当中会出现离心力（centrifugal force）和科里奥利力（Coriolis force）。往后的课程会把这一切发展成拉格朗日力学（Lagrangian mechanics）和哈密顿力学（Hamiltonian mechanics），坐标的选择在其中成为核心主题；又会发展成波动理论，而波就是在空间中散布开来的振动。非线性振动，例如大振幅的单摆，则通向椭圆函数（elliptic functions）和第 8 章的动力系统（dynamical systems）。

贯穿本章的问题

三大问题

1. **运动在甚么时候是简单的？**一个系统在甚么时候会以单一而固定的频率振动？若不是这样，又会发生甚么事？
2. **甚么才是合适的观点？**哪些坐标、哪些变量、哪个参考系能令问题变得容易？
3. **直觉可信吗？**日常对运动的直觉在哪里出错？数学又如何纠正它？

本章由振动开始：共振（Section 11.1）、简正模态、真实的单摆和等时曲线（tautochrone）。接着转向参考系和刚体：一列加速中的火车、一颗旋转的弹珠、一根表现得像三个质点的杆，以及一个可以按三种不同倾斜角度静止的立体。最后六节就是那些出人意表的轨迹。

本章的概念

§	课题	概念
11.1	受迫振动	共振与拍（beats）
11.2	耦合振子	简正模态就是特征向量（eigenvectors）
11.3	真实的单摆	椭圆积分（elliptic integrals）；算术几何平均（AGM）
11.4	等时曲线	以弧长（arc length）表示的简谐运动
11.5	运动参考系	假想力（fictitious forces）；等效原理（equivalence principle）
11.6	滚动与滑动	一个守恒的角动量（angular momentum）
11.7	转动惯量	动力等效（dynamical equivalence）；辛普森法则（Simpson's rule）
11.8	稳定性	通过质心（centre of mass）的法线
11.9	抛体	轨迹的包络线
11.10	抛体	距离不一定递增
11.11	接球	一条反馈法则
11.12	伸长的绳	分数与调和级数（harmonic series）
11.13	铲雪车	把时间看成距离的函数
11.14	悬索桥	是抛物线，不是悬链线

Section 11.6 和 Section 11.7 这两节用到转动动力学（rotational dynamics）和转动惯量，这些内容已于 2019 年从 STEP 考试范围中删去。收录它们，是因为这两道都是极好的题目；每一节都会给出所需的知识。

如何阅读本章

动手计算之前，先问问题目想要的是哪个变量。等时曲线用弧长，耦合弹簧用 $x \pm y$，蚂蚁用绳子的分数，铲雪车则把时间看成距离的函数：在每一种情况下，选对了变量，难题就会变成熟悉的问题。

11.1 共振

题目

STEP II 1991，第 12 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/91-s2-q12`

引子

推秋千上的小孩时，只要每一下都推在恰当的时刻，轻轻的推力也会累积成大幅的摆动；节奏不对的话，就几乎毫无效果。每个振动系统都有其*固有频率*（natural frequency），以这个频率去驱动它，所得的反应会大得与所花的力气完全不成比例。这就是*共振*（resonance）。歌手能够震碎酒杯，士兵过桥时要打乱步伐，收音机能够从数以千计的电台中选出其中一个，都是这个缘故。

这道题目摇动一条弹性绳的顶端，问绳子在甚么情况下会保持拉紧。答案含有因子 $\frac{1}{n-p}$，当驱动频率 p 趋近固有频率 n 时，这个因子会趋于无穷大：在接近共振时，轻轻一摇已足以令绳子松弛。

动笔之前

想一想 质点悬挂在平衡位置，绳的伸长量（extension）为 c。若绳的弹性模量（modulus）为 λ，则 λ、a、c、m 与 g 之间有甚么关系？当顶端向下移动了 $b\sin pt$，而质点向下移动了 x 时，伸长量是多少？

在平衡状态下，张力（tension）等于重量：$\frac{\lambda c}{a} = mg$。当顶端向下位移 $b\sin pt$，而质点向下位移 x 时，绳的长度增加了 $x - b\sin pt$，所以伸长量是 $c + x - b\sin pt$。

逐步解题

阶段 1　运动方程

取向下为正，作用于质点的力有它的重量 mg，以及向上的张力 $\frac{\lambda}{a}(c+x-b\sin pt)$：

$$m\ddot{x} = mg - \frac{\lambda}{a}(c+x-b\sin pt) = -\frac{\lambda}{a}(x-b\sin pt),$$

这里用了 $\frac{\lambda c}{a} = mg$。由于 $\frac{\lambda}{ma} = \frac{g}{c} = n^2$，上式即 $\ddot{x} = -n^2(x-b\sin pt)$。 ■

质点受到一个弹簧力，把它拉向移动中的点 $b\sin pt$。假如没有驱动，它会以角频率（angular frequency）n 振动，这就是它的固有频率。

阶段 2　求解

> ? **想一想**　方程是 $\ddot{x}+n^2x = n^2b\sin pt$。试用 $x = C\sin pt$ 这种形式的特解（particular solution），然后加上 $\ddot{x}+n^2x=0$ 的通解（general solution）。质点满足哪些初始条件？

代入 $x = C\sin pt$ 得 $(-p^2+n^2)C = n^2b$，所以 $C = \frac{n^2b}{n^2-p^2}$（这里需要 $p \neq n$）。余函数（complementary function）是 $D\cos nt + E\sin nt$。质点由平衡位置从静止开始运动：$x(0)=0$ 令 $D=0$，而 $\dot{x}(0)=0$ 给出 $Cp+En=0$，所以 $E = -\frac{nbp}{n^2-p^2}$。综合起来，

$$x(t) = \frac{bn}{n^2-p^2}(n\sin pt - p\sin nt). \qquad ■$$

阶段 3　保持拉紧

> ? **想一想**　绳子拉紧，当且仅当它的伸长量为正。试把伸长量 $c+x-b\sin pt$ 以 t 表示。

从 $x(t)$ 减去 $b\sin pt$，

$$\begin{aligned} x - b\sin pt &= \frac{b}{n^2-p^2}\left(n^2\sin pt - np\sin nt - (n^2-p^2)\sin pt\right) \\ &= \frac{bp}{n^2-p^2}(p\sin pt - n\sin nt). \end{aligned}$$

所以绳子自始至终保持拉紧，当且仅当

$$c + \frac{bp}{n^2-p^2}(p\sin pt - n\sin nt) \geq 0 \qquad \text{对所有} t \geq 0.$$

想一想 $p\sin pt - n\sin nt$ 最小可以负到甚么程度?

由于 $|\sin| \le 1$，我们有 $p\sin pt - n\sin nt \ge -(p+n)$。所以伸长量至少是

$$c - \frac{bp(n+p)}{n^2-p^2} = c - \frac{bp}{n-p},$$

当 $pb < (n-p)c$ 时，这个值为正。 ■

停一停，试一试 当 p 趋近 n 时，这个条件会变成怎样? 当 $p = 0.9n$ 时，摇动的振幅（amplitude）b 要有多小? 当 $p = 0.99n$ 时又如何?

回顾

- **由平衡位置量起。**利用平衡时的伸长量 c，重力便从方程中消去了。
- **特解加余函数。**频率为 p 的受迫响应，加上频率为 n 的自由振动，其中的常数由初始条件决定。
- **估界，不必求解。**我们没有准确求出伸长量的最小值；对正弦作一个粗略的估界，已足以得出一个充分条件（sufficient condition）。

背后的数学

共振极限

在解中令 $p \to n$。分子和分母都趋于零，计算极限（或使用洛必达法则（l'Hôpital's rule），对 p 求导）可得

$$x(t) \to \frac{b}{2}(\sin nt - nt\cos nt).$$

$nt\cos nt$ 这一项会无止境地增大：恰好以固有频率驱动时，振幅随时间线性增长，每一下推动都与运动同步，为它再添一分。在略低于共振的频率下，两个频率时而同步、时而错开，产生拍（beats）：振幅以缓慢的频率 $n-p$ 时涨时落（图 11.1.1）。两条结他弦的音高几乎调准时，听到的也是同样的拍。

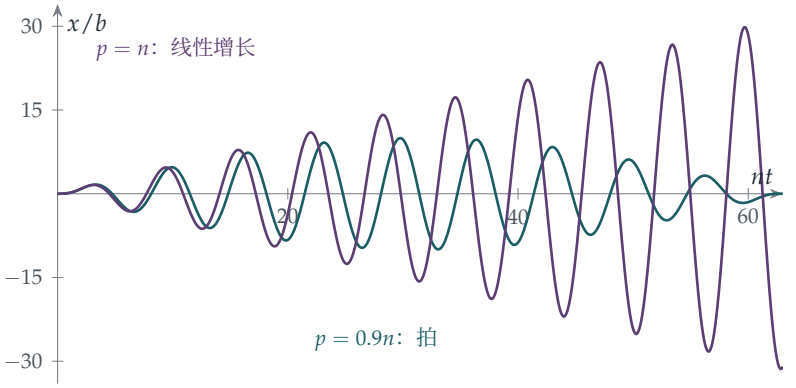


图 11.1.1　质点的响应（不考虑绳子松弛），以摇动振幅 b 为单位。当 $p = 0.9n$（蓝绿色）时，响应增大至约为 b 的十倍，然后再度消退，拍的周期为 $\frac{2\pi}{n-p}$。在恰好共振，即 $p = n$（紫红色）时，振幅无止境地增长。

阻尼与共振曲线

现实中的系统会因摩擦或空气阻力而损失能量，这就限制了振幅的增长。有阻尼（damping）时，系统对频率为 p 的驱动力的稳态响应，其振幅在固有频率附近达到峰值；峰愈尖锐，阻尼就愈小。工程师用品质因数（quality factor）Q 来描述这一点：粗略地说，它是系统在失去大部分能量之前所完成的振动次数。酒杯的 Q 数以千计，所以一个恰好等于其固有频率的持续音可以把它震碎。汽车的悬挂系统则刻意设计成 Q 值很低，以免路面颠簸令车身弹跳不止。

会动的桥

1831 年，一队士兵以整齐步伐操过曼彻斯特附近的布劳顿吊桥（Broughton Suspension Bridge），令桥身振动，最终倒塌；自此以后，军队过桥时都奉命打乱步伐。伦敦千禧桥（London Millennium Bridge）于 2000 年 6 月启用，人群一踏上桥，桥身便开始左右摇摆：行人不自觉地调整步伐去配合摇摆，同步地为它注入能量，结果该桥关闭了两年，以便加装阻尼器。1940 年塔科马海峡吊桥（Tacoma Narrows Bridge）那次著名的倒塌，常被说成是共振，但其实是另一种现象，即气动弹性颤振（aeroelastic flutter），当中风所施的力取决于桥本身的运动。要分析它，需要用到自激振动（self-excited oscillation）和极限环（limit cycle）的概念（Section 8.4）。

思维技巧

把受迫响应与自由响应叠加

对于受周期力驱动的线性方程，其解是驱动频率下的响应，加上固有频率下的自由振动，后者按初始条件选定。

这个方法之所以可行，全靠线性（linearity）：解可以相加。在大学里，同一原理应用于多个驱动频率之和，便引出傅里叶分析（Fourier analysis）、工程学中的传递函数（transfer function），以及物理学中的格林函数（Green's function）。

延伸练习

1 **下坠的升降机。**试做 STEP II 1996，第 11 题（`step.maths.org/questions/96-s2-q11`）：升降机内有一个质量系于弹性绳上，升降机先开始加速，其后停止加速。留意最终的振幅如何取决于加速持续的时间。

2 **共振之时。**试用特解 $x = Ct\cos nt$，直接解 $\ddot{x} + n^2 x = n^2 b \sin nt$，其中 $x(0) = \dot{x}(0) = 0$。验证所得结果与上文的极限一致。

3 **拍。**证明 $n\sin pt - p\sin nt$ 可写成两项之和，而当 p 接近 n 时，其中一项近似于 $2n\cos\frac{(n+p)t}{2}\sin\frac{(p-n)t}{2}$。由此解释拍的周期为何是 $\frac{2\pi}{n-p}$。

4 **有阻尼的共振。**对于 $\ddot{x} + 2k\dot{x} + n^2 x = F\cos pt$，其中 $k > 0$ 且很小，求稳态振幅，并证明当 $p^2 = n^2 - 2k^2$ 时振幅最大。

延伸阅读

- **A. P. French, *Vibrations and Waves***（MIT Introductory Physics Series）。第 4 章讲述受迫振动与共振，是这个课题的经典论述。
- **Steven Strogatz, *Sync***（Penguin）。讲述事物如何趋于同步，从萤火虫谈到千禧桥。
- **Henry Petroski, *To Engineer Is Human***（Vintage）。讲述工程上的失败个案，包括历史专栏中提到的几座桥。

11.2 简正模式

题目

STEP III 2014，第 10 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/14-s3-q10`

引子

一个挂在弹簧上的质点，会以简单而规律的方式上下弹动。把两个质点用弹簧连起来，运动看起来就复杂得多：每个质点都不规则地忽动忽停，能量在两者之间来回流转。然而在这复杂的表象背后，藏着两种极其简单的运动，称为**简正模态**（normal modes）：两个质点或一起同向运动，或恰好反向运动，而每个模态各以自己的频率（frequency）振荡。系统的任何运动，都是这两个模态的组合。

本题要找出夹在三条弹簧之间的两个质点的简正模态，然后证明一个令人惊讶的结论：若系统由静止开始，而其中一个质点有初始位移，则系统**永远不会**回到起始位置。原因在于两个频率之比是 $\sqrt{3}$，而 $\sqrt{3}$ 是无理数（irrational number）。

动笔之前

想一想 设 X 与 P 的距离为 $a+x$，Y 与 Q 的距离为 $a+y$，那么中间的弹簧 XY 有多长？当 x 和 y 皆为正数时，哪些弹簧被拉长，哪些被压缩？

三段长度之和为 $3a$，所以 $XY = 3a-(a+x)-(a+y) = a-x-y$。当 $x, y > 0$ 时，外侧两条弹簧被拉长（伸长量分别为 x 和 y），中间一条则被压缩（压缩量为 $x+y$）。

逐步解题

阶段 1　运动方程

沿着由 P 指向 Q 的方向量度 x。弹簧 PX 的伸长量（extension）为 x，所以它的张力（tension）$\frac{\lambda x}{a}$ 把 X 拉回 P 的方向。弹簧 XY 的伸长量为 $-(x+y)$，所以它的张力 $-\frac{\lambda(x+y)}{a}$ 把 X 拉向 Y（负的张力即是推力）。因此

$$m\ddot{x} = -\frac{\lambda x}{a} - \frac{\lambda(x+y)}{a} = -\frac{\lambda}{a}(2x+y). \qquad \blacksquare$$

根据对称性（把 X 与 Y、P 与 Q 的角色互换），

$$m\ddot{y} = -\frac{\lambda}{a}(2y+x).$$

阶段 2　解耦

想一想　每条方程都同时涉及 x 和 y。把两条方程怎样组合，才会只涉及 x 和 y 的某一个组合？

把两式相减及相加，得

$$m(\ddot{x}-\ddot{y}) = -\frac{\lambda}{a}(x-y), \qquad m(\ddot{x}+\ddot{y}) = -\frac{3\lambda}{a}(x+y).$$

两者都是简谐运动（simple harmonic motion）的方程：$x-y$ 以 $\omega^2 = \frac{\lambda}{ma}$ 振荡，而 $x+y$ 则以 $3\omega^2$ 振荡。

阶段 3　初始条件

起初 $XP = a$ 及 $YQ = \frac{1}{2}a$，所以 $x = 0$ 及 $y = -\frac{1}{2}a$，而且两个质点都处于静止。因此在 $t = 0$ 时，$x - y = \frac{1}{2}a$ 及 $x + y = -\frac{1}{2}a$，而两者的变化率皆为零，由此得

$$x - y = \tfrac{1}{2}a\cos\omega t, \qquad x + y = -\tfrac{1}{2}a\cos\sqrt{3}\,\omega t.$$

所以 $A = \frac{1}{2}a$ 及 $B = 0$。 $\blacksquare$

阶段 4　永不返回

想一想　把 y 写成 t 的函数。若 Y 要回到起始位置，便须 $y = -\frac{1}{2}a$。这时两个余弦必须各等于甚么？

两式相减，得 $y = \frac{1}{2}\big((x+y)-(x-y)\big) = -\frac{1}{4}a\big(\cos\sqrt{3}\,\omega t + \cos\omega t\big)$。要 $y = -\frac{1}{2}a$，便需要 $\cos\sqrt{3}\,\omega t + \cos\omega t = 2$；由于每个余弦最大只能是 1，两者必须都等于 1：

$$\omega t = 2\pi k, \qquad \sqrt{3}\,\omega t = 2\pi j$$

其中 k 和 j 为整数。当 $t > 0$ 时，$k \geq 1$，两式相除得 $\sqrt{3} = \frac{j}{k}$，但 $\sqrt{3}$ 是无理数，所以这是不可能的。因此 Y 永远不会回到起始位置。 $\blacksquare$

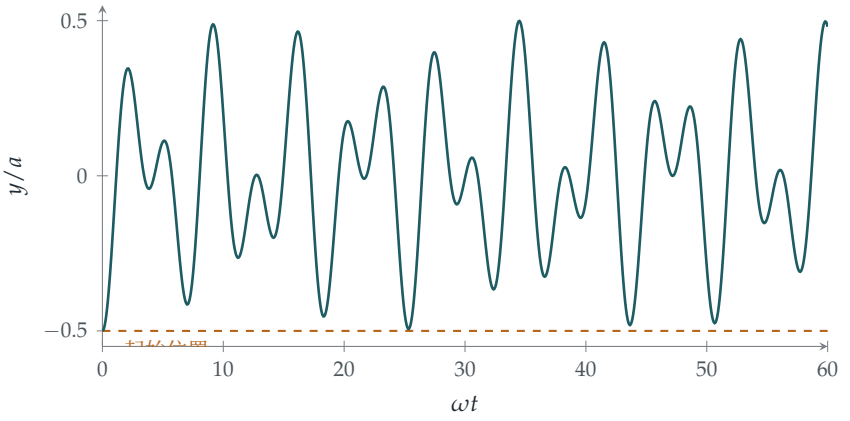


图 11.2.1 质点 Y 的位移 y。它是两个振荡之和，而两者频率之比为 $\sqrt{3}:1$，所以图形永不重复。Y 一次又一次地接近起始位置（虚线），例如在 $\omega t \approx 25$ 附近（该处 $\omega t \approx 8\pi$ 及 $\sqrt{3}\,\omega t \approx 14\pi$），但永远不会刚好到达。

回顾

- **找出能够解耦的组合。** $x-y$ 和 $x+y$ 各自满足一条属于自己的方程。
- **对称性会提示这些组合。** 把左右两边互换，系统保持不变，所以对称和反对称的组合是很自然的选择。
- **无理性会带来物理后果。** 运动之所以不是周期性的，是因为频率之比是无理数。

背后的数学

简正模态就是特征向量

把两条方程写成一条矩阵方程：

$$m\begin{pmatrix}\ddot{x}\\ \ddot{y}\end{pmatrix} = -\frac{\lambda}{a}\begin{pmatrix}2 & 1\\ 1 & 2\end{pmatrix}\begin{pmatrix}x\\ y\end{pmatrix}.$$

能够解耦的组合，正是这个矩阵的特征向量（eigenvectors）：$(1,-1)$，其特征值（eigenvalue）为 1；以及 $(1,1)$，其特征值为 3（Section 7.2）。在第一个模态中，$x=-y$：两个质点沿直线同向运动（记住 x 和 y 是从相反的两端量度的），中间的弹簧长度保持不变。在第二个模态中，$x=y$：两个质点反向运动，把中间的弹簧一压一拉，频率因而较高。频率的平方就是特征值乘以 $\frac{\lambda}{ma}$。

无论有多少个质点和弹簧，这个方法都适用。一个由 N 个耦合振子（coupled oscillators）组成的系统有 N 个简正模态，它们由一个对称矩阵的特征向量给出，而谱定理（spectral

theorem）保证这些模态必定存在，并且互相垂直。分子以简正模态振动，化学家正是凭分子所吸收的红外光来辨认它们。建筑物和桥梁也有简正模态，工程师必须令这些模态的频率远离风和地震的频率。当质点的数目趋向无限时，简正模态就变成振动弦线的谐波（harmonics），这正是傅里叶分析（Fourier analysis）的起点。

接近而不到达

虽然 Y 永远不会刚好返回，但它可以接近到任意程度。原因是克罗内克（Kronecker）的一条定理：若 α 是无理数，则各倍数 $\alpha, 2\alpha, 3\alpha, \ldots$ 取模 1 之后，可以任意接近 $[0,1]$ 内的每一个数。因此我们可以找到一些时刻，使 ωt 是 2π 的倍数，而 $\sqrt{3}\,\omega t$ 与另一个这样的倍数之差小于任何指定的容许误差。这种运动是*准周期*（quasi-periodic）的：它在一个环面（torus）上不断缠绕（每个模态对应一个角），却永远不会闭合。有理与无理频率比之间的这种区别，同样是关于太阳系稳定性的 KAM 定理的核心。

思维技巧

> **以变量代换把方程解耦**
>
> *当方程互相耦合时，设法找出各自独立演变的变量组合。在线性问题中，这些组合就是特征向量。*
>
> 这与 Section 10.6 中质心和相对坐标的想法如出一辙。到了大学，它就成为对角化（diagonalisation）：选取一组令矩阵成为对角矩阵的基（basis），一组耦合的方程便会分拆成各自独立的方程。

延伸练习

1 **能量。**证明总能量 $\frac{1}{2}m(\dot{x}^2+\dot{y}^2)+\frac{\lambda}{2a}\left(x^2+y^2+(x+y)^2\right)$ 是常数，并证明它可以分拆成两项之和，每个简正模态各占一项，而每一项本身都是常数。

2 **纯模态。**应该怎样让两个质点开始运动，才能使运动是周期性的，且周期为 $\frac{2\pi}{\omega}$？若周期为 $\frac{2\pi}{\sqrt{3}\,\omega}$ 又如何？

3 **较硬的中间弹簧。**若中间弹簧的弹性模量（modulus of elasticity）是 μ 而不是 λ，求两个简正模态的频率。当 μ/λ 取哪些值时，所有运动都是周期性的？

4 **三个质点。**三个质量相等的质点排成一线，由四条相同的弹簧连接在两个固定点之间。求对应的矩阵及其特征值 $(2-\sqrt{2},\,2,\,2+\sqrt{2})$，并描述三个简正模态。

延伸阅读

- **A. P. French, *Vibrations and Waves***（MIT Introductory Physics Series），第 5 章，讲述耦合振子与简正模态。
- **Gilbert Strang, *Introduction to Linear Algebra***（Wellesley-Cambridge），关于特征值的一章，当中以弹簧与质点作为主要例子。
- **Steven Strogatz, *Nonlinear Dynamics and Chaos***（CRC Press），第 8 章，讲述环面上的准周期运动。

11.3 真实的摆

题目

STEP III 1998，第 11 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/98-s3-q11`

引子

每个修读物理的学生都学过：单摆（pendulum）的周期（period）是 $2\pi\sqrt{l/g}$，与摆动的幅度无关。伽利略（Galileo）对此深信不疑，相传他曾在比萨大教堂以自己的脉搏为一盏摇晃的吊灯计时；摆钟之所以能够报时，靠的也正是这一点。然而这并不完全正确。这条公式源自近似式 $\sin\theta \approx \theta$，而摆动弧度大的单摆，所需的时间会较长。在 $90°$ 时，周期长了 18%；若单摆从几乎倒立的位置释放，所需的时间则要多长有多长。

本题先把精确的周期写成一个积分，再求出中等摆幅下的第一个修正项。这个积分无法用初等函数求出：它是一个*椭圆积分*（elliptic integral），而人们如何驯服这类积分的故事，由高斯（Gauss）的算术几何平均一直延伸到椭圆曲线的理论。

动笔之前

想一想 摆锤沿半径为 l 的圆周运动。以 θ 和 $\dot\theta$ 表示，它的速率和高度（相对于最低点）分别是甚么？

速率是 $l\dot\theta$，而距最低点的高度是 $l(1-\cos\theta)$。由于绳子不作功（其张力恒与运动方向垂直），能量守恒（conservation of energy）。

逐步解题

阶段 1 能量

在最大摆角 γ 处，摆锤瞬时静止。由能量守恒得

$$\tfrac{1}{2}ml^2\dot\theta^2 + mgl(1-\cos\theta) = mgl(1-\cos\gamma),$$

两边除以 ml，得 $\tfrac{1}{2}l\dot\theta^2 = g(\cos\theta-\cos\gamma)$。 ■

阶段 2　把周期写成积分

想一想　怎样用半角表示 $\cos\theta-\cos\gamma$？另外，由 $\theta=0$ 摆到 $\theta=\gamma$ 占了周期的多少？

利用 $\cos\theta = 1-2\sin^2\tfrac{\theta}{2}$，可得 $\cos\theta-\cos\gamma = 2(\sin^2\tfrac{\gamma}{2}-\sin^2\tfrac{\theta}{2})$，所以

$$\dot\theta = 2\sqrt{\frac{g}{l}}\left(\sin^2\frac{\gamma}{2}-\sin^2\frac{\theta}{2}\right)^{1/2}.$$

根据对称性，由 0 摆到 γ 需时四分之一个周期，所以

$$P = 4\int_0^{\gamma}\frac{\mathrm{d}\theta}{\dot\theta} = 2\sqrt{\frac{l}{g}}\int_0^{\gamma}\left(\sin^2\frac{\gamma}{2}-\sin^2\frac{\theta}{2}\right)^{-1/2}\mathrm{d}\theta. \qquad ■$$

阶段 3　消除奇点

想一想　被积函数（integrand）在 $\theta=\gamma$ 处变成无穷大。设 $k=\sin\tfrac{\gamma}{2}$，并作代换（substitution）$\sin\tfrac{\theta}{2} = k\sin\phi$。$0\le\theta\le\gamma$ 对应于 ϕ 的甚么范围？被积函数会变成怎样？

当 θ 由 0 变到 γ 时，ϕ 由 0 变到 $\tfrac{\pi}{2}$。求导得 $\tfrac{1}{2}\cos\tfrac{\theta}{2}\,\mathrm{d}\theta = k\cos\phi\,\mathrm{d}\phi$，所以

$$\mathrm{d}\theta = \frac{2k\cos\phi\,\mathrm{d}\phi}{\cos\frac{\theta}{2}} = \frac{2k\cos\phi\,\mathrm{d}\phi}{\sqrt{1-k^2\sin^2\phi}}.$$

另一方面，$\sin^2\tfrac{\gamma}{2}-\sin^2\tfrac{\theta}{2} = k^2(1-\sin^2\phi) = k^2\cos^2\phi$。那些棘手的因子互相抵销：

$$P = 2\sqrt{\frac{l}{g}}\int_0^{\pi/2}\frac{1}{k\cos\phi}\cdot\frac{2k\cos\phi}{\sqrt{1-k^2\sin^2\phi}}\,\mathrm{d}\phi = 4\sqrt{\frac{l}{g}}\int_0^{\pi/2}\frac{\mathrm{d}\phi}{\sqrt{1-k^2\sin^2\phi}}.$$

由于 $k<1$，新的被积函数性质十分良好。

阶段 4　第一个修正项

想一想 当 k 很小时，用二项式定理（binomial theorem）展开 $(1-k^2\sin^2\phi)^{-1/2}$。$\int_0^{\pi/2}\sin^2\phi\,\mathrm{d}\phi$ 等于甚么？

准确至一阶，$(1-k^2\sin^2\phi)^{-1/2}\approx 1+\frac{1}{2}k^2\sin^2\phi$，而 $\int_0^{\pi/2}\sin^2\phi\,\mathrm{d}\phi=\frac{\pi}{4}$。所以

$$P\approx 4\sqrt{\frac{l}{g}}\left(\frac{\pi}{2}+\frac{k^2}{2}\cdot\frac{\pi}{4}\right)=2\pi\sqrt{\frac{l}{g}}\left(1+\frac{k^2}{4}\right).$$

当 γ 很小时，$k=\sin\frac{\gamma}{2}\approx\frac{\gamma}{2}$，故 $k^2\approx\frac{\gamma^2}{4}$，于是

$$P\approx 2\pi\sqrt{\frac{l}{g}}\left(1+\frac{\gamma^2}{16}\right). \qquad \blacksquare$$

摆幅为 20°（$\gamma\approx 0.35$）时，修正量不足 1%；在 60° 时则约为 7%（图 11.3.1）。

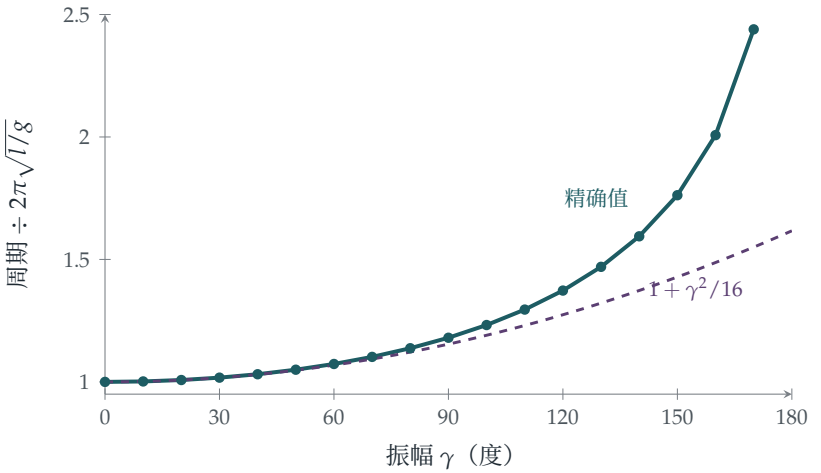


图 11.3.1 单摆的精确周期（蓝绿色）相对于小角度值的比，对振幅（amplitude）作图，并与近似式 $1+\gamma^2/16$（紫红色虚线）比较。在大约 60° 以内，这个近似极佳。当 $\gamma\to 180°$ 时，真正的周期趋向无穷大：一个倒立平衡的单摆要花无限长的时间才会倒下。

回顾

- **能量给出首次积分（first integral）。**二阶方程变成了一阶方程，而周期则变成了一个积分。
- **代换驯服奇点（singularity）。**被积函数在转向点处趋向无穷大；代换 $\sin\frac{\theta}{2}=k\sin\phi$ 消除了这个问题。

- **先求精确，再作近似。**我们先求出精确的公式，然后将它展开，这样同时可以知道近似的效果有多好。

背后的数学

椭圆积分

积分

$$K(k) = \int_0^{\pi/2} \frac{\mathrm{d}\phi}{\sqrt{1 - k^2 \sin^2 \phi}}$$

称为*第一类完全椭圆积分*（complete elliptic integral of the first kind），这个名称的由来，是因为与它相近的另一个积分给出椭圆的周界。精确的周期是 $P = 4\sqrt{l/g}\,K\bigl(\sin\tfrac{\gamma}{2}\bigr)$。把被积函数完全展开，再利用 Section 4.4 中的沃利斯积分（Wallis integrals）$\int_0^{\pi/2} \sin^{2n}\phi\,\mathrm{d}\phi$ 逐项积分，便得

$$P = 2\pi\sqrt{\frac{l}{g}}\Bigl(1 + \bigl(\tfrac{1}{2}\bigr)^2 k^2 + \bigl(\tfrac{1\cdot 3}{2\cdot 4}\bigr)^2 k^4 + \bigl(\tfrac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}\bigr)^2 k^6 + \cdots\Bigr).$$

$K(k)$ 无法以有限个熟悉的函数组合而成；它是一个名副其实的新函数。

高斯的非凡公式

计算 K 有一个好得多的方法。你在 Section 2.6 已见过*算术几何平均*（arithmetic–geometric mean）：由 a 和 b 开始，反复以它们的算术平均和几何平均取代它们。高斯在 1799 年发现

$$K(k) = \frac{\pi}{2\,\mathrm{AGM}\bigl(1, \sqrt{1 - k^2}\bigr)},$$

所以振幅为 γ 的单摆，其周期恰好是

$$P = \frac{2\pi\sqrt{l/g}}{\mathrm{AGM}\bigl(1, \cos\tfrac{\gamma}{2}\bigr)}.$$

AGM 收敛得极快：摆幅在大约 160° 以内时，只需三四步便可求得准确至十位有效数字的周期，图 11.3.1 中的数值正是这样算出来的。接近 180° 时，两个初始值 1 和 $\cos\frac{\gamma}{2}$ 相距甚远，需要多走一两步，但即使在 179°，五步也已足够。高斯在日记中写道，这项发现“开辟了一个全新的分析领域”，而他说得一点不错：把椭圆积分反演，引领阿贝尔（Abel）和雅可比（Jacobi）发现了*椭圆函数*（elliptic functions），而椭圆函数又引出了*椭圆曲线*（elliptic curves）。椭圆曲线如今是数论的核心（它们正是证明费马最后定理（Fermat's Last Theorem）的关键），也是保障互联网安全的密码学的核心。

伽利略的吊灯与惠更斯的时钟

伽利略大约在 1602 年提出，单摆的周期与振幅无关；对于他所研究的小幅摆动，这是个很好的近似。惠更斯（Christiaan Huygens）意识到这并不精确，并在 1659 年找到了一条曲线，沿着它摆动时周期确实与振幅无关：那就是摆线（cycloid），即 Section 11.4 的主题。他由 1656 年起制造的摆钟，把计时的误差由每天约一刻钟减至约十五秒。航海天文钟（marine chronometer）解决了在海上测定经度的难题，它后来完全舍弃了单摆，但直至 1920 年代，摆钟仍然是世上最准确的计时器。

思维技巧

由精确积分到级数近似

当一个量由无法求值的积分给出时，先把它变换成性质良好的形式，然后展开被积函数并逐项积分。

物理学家正是这样为理想化的结果求出修正项的，这也是*微扰理论*（perturbation theory）的起点。到了大学，你会学到逐项积分在甚么情况下才成立（一致收敛（uniform convergence）），以及怎样估计截断级数所造成的误差。

延伸练习

1. **系着的质点。**试做 STEP III 2015，第 9 题（`step.maths.org/questions/15-s3-q9`）：这个振子即使在小幅振动时，周期也取决于振幅。为甚么它会这样，而单摆却不会？

2. **下一项。**把二项式展开多取一项，证明 $P \approx 2\pi\sqrt{\frac{l}{g}}\left(1+\frac{\gamma^2}{16}+\frac{11\gamma^4}{3072}\right)$。

3. **徒手计算 AGM。**用三步 AGM 估算振幅为 90° 的单摆的周期，并与图中的数值 1.1803 比较。

4. **时钟。**一座摆钟经调校后，在振幅为 4° 时走时完全准确。若它的振幅增至 5°，它每天会慢多少秒？

延伸阅读

- **David Morin, *Introduction to Classical Mechanics***（Cambridge University Press），第 4 章，当中讨论了大振幅的单摆。

- **Jonathan Borwein & Peter Borwein, *Pi and the AGM***（Wiley）。高斯的发现及其应用，包括计算 π 的最快方法。
- **Dava Sobel, *Longitude***（Fourth Estate）。海上计时的故事，讲述单摆以外的发展。

11.4 等时曲线

题目

STEP III 1994，第 9 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/94-s3-q9`

引子

在一个碗内同时放开两粒弹珠，一粒靠近碗边，另一粒只比碗底高一点。哪一粒先到碗底？在普通的碗里，通常是起点较低的那粒胜出。不过有一种碗，无论弹珠从哪里出发，总是同时到达。这就是*等时曲线*（tautochrone，希腊文意为“相同时间”），它的横切面是一条*摆线*（cycloid），即滚动的轮子边缘上一点所描出的曲线。

惠更斯（Huygens）在 1659 年发现了这个性质，并用它设计出一种周期与摆幅无关的摆钟，修正了 Section 11.3 中发现的缺陷。本题只用寥寥数行便证明了等时性质，关键在于选对变量：用沿曲线量度的距离，而不是水平位置。

动笔之前

想一想 碗面是光滑的。有哪些力作用在质点上？其中哪些力有作功？

重力和法向反作用力（normal reaction）。法向反作用力垂直于碗面，因而垂直于运动方向，所以不作功。只有重力作功，故能量守恒。

逐步解题

阶段 1　沿曲线的能量

由于 s 是弧长（arc length），质点沿碗面的速率是 $\left|\frac{\mathrm{d}s}{\mathrm{d}t}\right|$。由能量守恒（conservation of

energy）得 $\frac{1}{2}m\dot{s}^2 + mgy =$ 常数，所以 $\dot{s}^2 + 2gy$ 是常数。 ■

阶段 2　以 s 表示的简谐运动

想一想　利用 $s = 2\sqrt{ky}$，把 y 换成 s 的函数。所得的方程描述的是哪一种运动？

由于 $y = \frac{s^2}{4k}$，能量方程变成

$$\dot{s}^2 + \frac{g}{2k}s^2 = \text{常数} = \frac{g}{2k}s_0^2,$$

其中 $s_0 = 2\sqrt{kh}$ 是起始弧长（质点由静止出发）。求导后得 $2\dot{s}\ddot{s} + \frac{g}{k}s\dot{s} = 0$，所以

$$\ddot{s} = -\frac{g}{2k}\,s.$$

这是以 s 为变量的简谐运动（simple harmonic motion），其中 $\omega^2 = \frac{g}{2k}$，与振幅（amplitude）无关：$s = s_0\cos\omega t$。

阶段 3　到达底部所需的时间

质点经过四分之一个周期后到达 V（该处 $s = 0$）：

$$\frac{\pi}{2\omega} = \frac{\pi}{2}\sqrt{\frac{2k}{g}} = \pi\sqrt{\frac{k}{2g}},$$

这与 h 无关。 ■

停一停，试一试　为甚么这样行得通？沿曲线的“回复力”是 g 乘以斜率，即 $g\frac{\mathrm{d}y}{\mathrm{d}s} = \frac{g}{2k}s$：恰好与沿曲线量度、离底部的距离成正比。这正是简谐运动的定义性质。起点较高的质点要走的路较长，但所受的推动也按比例加大。

阶段 4　两个质点

想一想　两个质点何时到达 V？到达时速率是多少？谁的动量较大？

两者从相反的两边，在同一时刻 $T = \pi\sqrt{k/(2g)}$ 到达 V。在 V 的速率是 ω 乘以 s 的振幅，与 $\sqrt{\text{高度}}$ 成正比。设较重的质点从高度 h 放开，在 V 的速率为 U；较轻的质点从 $\alpha^2 h$ 放开，速率为 αU。它们的动量（momentum）分别是 $\alpha m \cdot U$ 和 $m \cdot \alpha U$，方向相反：大小相等，方向相反。碰撞时的总动量为零。

想一想　总动量为零，而恢复系数（coefficient of restitution）为 α，碰撞刚结束时两者的速度是多少？

动量保持为零，所以碰撞后两个质点仍然朝相反方向运动，动量大小相等，而它们的相对速率是接近速率 $\alpha U + U$ 的 α 倍。若较轻的质点（质量 αm）此时以速率 w 运动，则较重的质点（质量 m）以速率 αw 运动，且 $w + \alpha w = \alpha(1+\alpha)U$，所以 $w = \alpha U$。每个质点反弹时，速率都乘以 α：质量为 αm 的质点以 αU 离开，质量为 m 的质点以 $\alpha^2 U$ 离开。

阶段 5 整个运动

每个质点现在沿自己的一边爬升，所达高度缩减为原来的 α^2 倍（因为高度与速率的平方成正比）：质量为 αm 的质点升至 $\alpha^2 h$，质量为 m 的质点升至 $\alpha^4 h$。由等时性质，两者在 $2T$ 之后恰好同一时刻回到 V，动量仍然大小相等、方向相反。同样的碰撞再次发生，只是所有速率都按 α 的比例缩小。因此：

- 两个质点总是在 V 碰撞，时刻为 $T, 3T, 5T, \ldots$；
- 第 n 次碰撞后，它们分别升至高度 $\alpha^{2n}h$ 和 $\alpha^{2n+2}h$；
- 运动按几何级数式衰减，但永不停止：碰撞有无限多次，并且永远持续下去。 ■

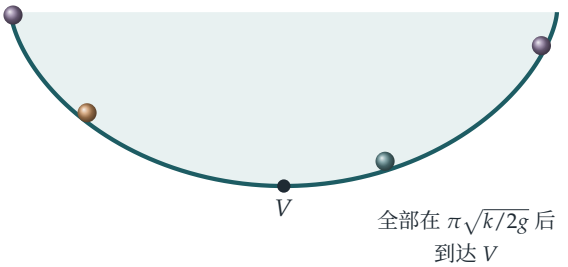


图 11.4.1 等时曲线：一条倒转的摆线，由沿直线滚动的轮子边缘上一点描出。弹珠从曲线上任何位置由静止放开，都在同一时刻到达最低点 V。对于由半径为 R 的轮子生成的摆线，弧长与高度满足 $s^2 = 8Ry$，所以 $k = 2R$。

回顾

- **选择自然的变量。**以弧长 s 表示时，运动恰好是简谐运动。以 x 或 y 表示，则一团糟。
- **能量给出方程。**对能量方程求导，便直接得出运动方程。
- **时间上的对称。**由于每一趟到达底部的行程所需时间都相同，两个质点永远保持同步，碰撞也就原样重复。

背后的数学

为甚么是摆线？

半径为 R 的轮子在一条直线下方滚动，描出摆线 $x = R(t + \sin t)$，$y = R(1 - \cos t)$。于是 $\frac{\mathrm{d}x}{\mathrm{d}t} = R(1 + \cos t)$ 及 $\frac{\mathrm{d}y}{\mathrm{d}t} = R \sin t$，所以沿曲线的速率是 $R\sqrt{2 + 2\cos t} = 2R \cos \frac{t}{2}$。积分后，由底部起计的弧长是 $s = 4R \sin \frac{t}{2}$，而 $y = 2R \sin^2 \frac{t}{2}$。因此 $s^2 = 8Ry$：摆线正是满足 $s = 2\sqrt{ky}$ 的曲线，其中 $k = 2R$。反过来说，要求回复力与 s 成正比，就迫使曲线必定是摆线。

最速降线

1696 年，约翰·伯努利（Johann Bernoulli）向欧洲的数学家提出挑战：一粒珠子从一点滑到较低的另一点（不在其正下方），沿哪一条曲线所需的时间最短？答案是最速降线（brachistochrone），它同样是一条摆线。据说牛顿（Newton）收到题目后，一个晚上就解决了；他匿名发表了自己的解法，但伯努利认出了作者，说是“从爪认出狮子”。这个问题是变分法（calculus of variations）的开山例子，你在 Section 4.7 已经见过：找出使某个由积分定义的量取最小值的曲线。

惠更斯的摆线钟

惠更斯知道单摆计时并不完美，因为它的周期取决于振幅。他在《*Horologium Oscillatorium*》（1673 年）中说明了怎样令摆锤沿摆线运动：把摆线挂在两块弯成摆线形状的“夹板”之间，摆线便会贴着夹板卷绕，使摆锤本身描出一条摆线（摆线的渐伸线（involute）是另一条摆线）。实际上，夹板因摩擦而引入的误差比它消除的还要多，钟表匠于是改为保持小幅摆动。然而惠更斯的分析开创了渐屈线（evolute）与渐伸线的理论，他这部著作也是十七世纪力学的伟大作品之一。

思维技巧

找出令运动变得简单的变量

沿曲线的运动，往往以沿曲线的距离来表示最为简单。要找出令方程变成线性的那个变量。

到了大学，这会发展成拉格朗日力学（Lagrangian mechanics）中广义坐标（generalised coordinates）的概念：任何能描述系统位形的变量都可以采用，选得好，就能令问题变得可解。

延伸练习

1 **不同的高度。**一个质点在等时曲线碗内从高度 h 放开。证明它在 V 的速率是 $\sqrt{2gh}$（由能量守恒，必然如此），并验证这与 ωs_0 一致。

2 **普通的碗。**在半径为 R 的半球形碗内，比较从很小的高度到达底部所需的时间与从碗边出发所需的时间。哪一个较长？（利用 Section 11.3 的结果。）

3 **质量相等。**把题目的第二部分重做一次，这次两个质点质量同为 m，分别从高度 h_1 和 h_2 放开，恢复系数为 e。它们是否仍然总是在 V 碰撞？

4 **摆线的弧长。**证明摆线的一拱长度为 $8R$，而拱下的面积为 $3\pi R^2$，即滚动圆面积的三倍。

延伸阅读

- **George Simmons, *Differential Equations with Applications and Historical Notes***（CRC Press）。书中包括等时曲线和最速降线，并附有相关历史。
- **E. A. Whitman, "Some historical notes on the cycloid"**，*American Mathematical Monthly* 50 (1943)。讲述摆线为何被称为"几何学家的海伦"。
- **Paul Nahin, *When Least Is Best***（Princeton University Press）。以历史故事的方式讲述最速降线及其他最优化问题。

11.5 加速火车中的重力

题目

STEP II 2021，第 10 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2021 年 STEP 试卷（zip 文件）内的 STEP II 试卷：

`www.ocr.org.uk/Images/696588-2021-past-papers.zip`

引子

火车驶离车站时，你会感到自己被压向椅背；桌上那杯咖啡的液面倾斜了；地板上的球滚向车尾。从车厢内看，一切就像重力倾斜了一样：除了向下，还稍微指向后方。爱因斯坦（Einstein）领悟到这不仅仅是比喻，并把这个领悟称为“我一生中最快乐的想法”：在局部范围内，重力与加速度是无法区分的。

这道题把一颗珠子穿在弯曲的金属线上，再放进一列正在加速的火车里。从火车上看，珠子的运动就像身处一个倾斜了的重力场（gravitational field）之中，倾斜的角度的正切为 a/g。金属线的底部不再是最低点，珠子会围绕一个新的平衡位置（equilibrium）振动。

动笔之前

想一想 火车以加速度 a 向西（即沿负 x 方向）加速。对车上的观察者来说，质量为 m 的物体看似受到怎样的“假想”力作用？

在火车的参考系(frame)中,只要加上一个假想力(fictitious force)$-m\times$(参考系的加速度)，牛顿第二定律（Newton's second law）便仍然成立。参考系在 x 方向的加速度为 $-a$，所以假想力是 $+ma$，指向东。连同重力一并考虑，珠子每单位质量所感受到的有效重力（effective gravity）为 $(a,-g)$：向东倾斜。

逐步解题

阶段 1　不能静止于原点

在原点处，金属线 $ky = x^2$ 是水平的。该处的法向反作用力（normal reaction）是铅垂的，因此无法抵消向东的假想力 ma。沿金属线方向的合力为 $ma \neq 0$，所以珠子不可能相对于火车保持静止。 ■

阶段 2　能量方程

> ? **想一想**　在火车的参考系中写出牛顿定律，当中要包括法向反作用力 $\mathbf{N}$。然后与速度 $(\dot{x}, \dot{y})$ 取点积。为甚么 $\mathbf{N}$ 会消失?

在火车的参考系中，

$$m(\ddot{x}, \ddot{y}) = \mathbf{N} + (ma, -mg).$$

金属线是光滑的，所以 $\mathbf{N}$ 垂直于金属线，因而也垂直于沿金属线方向的速度 $(\dot{x}, \dot{y})$。与速度取点积（dot product），$\mathbf{N}$ 便会消去：

$$m\dot{x}\ddot{x} + m\dot{y}\ddot{y} = ma\dot{x} - mg\dot{y}, \qquad \text{所以} \qquad \dot{x}(\ddot{x} - a) + \dot{y}(\ddot{y} + g) = 0. \quad ■$$

左边是 $\frac{1}{2}(\dot{x}^2 + \dot{y}^2) - ax + gy$ 的导数，因此这个量是常数。 ■

这就是火车参考系中的能量守恒（conservation of energy），其中每单位质量的势能（potential energy）$-ax + gy$ 已把假想力计算在内。

阶段 3　能升到多高?

> ? **想一想**　由于珠子从原点由静止出发，该常数为 0。在最高点，珠子瞬时静止。由此可得出甚么方程?

静止时，$\frac{1}{2}(\dot{x}^2 + \dot{y}^2) = 0$，所以 $ax = gy$。代入 $y = \frac{x^2}{k}$，得 $ax = \frac{gx^2}{k}$，所以（除了起点 $x = 0$ 之外）$x = \frac{ak}{g}$，而

$$b = \frac{a^2 k}{g^2}. \quad ■$$

珠子沿金属线向东摆上至点 $\left(\frac{ak}{g}, \frac{a^2 k}{g^2}\right)$，然后返回原点，如此往复不息。

阶段 4　最快的一点

> ? **想一想**　把速率的平方写成只含 x 的函数。它在哪里最大?

速率的平方是 $2(ax - gy) = 2\left(ax - \frac{gx^2}{k}\right)$，这是 x 的二次式，在 $x = \frac{ak}{2g}$ 处取得最大值，刚好在两个转向点（turning point）的正中间。该处速率的平方为 $2\left(\frac{a^2k}{2g} - \frac{a^2k}{4g}\right) = \frac{a^2k}{2g}$，所以最大速率是

$$a\sqrt{\frac{k}{2g}}. \qquad \blacksquare$$

想一想 求金属线在 $x = \frac{ak}{2g}$ 处的斜率。它与有效重力 $(a, -g)$ 的方向有甚么关系？

斜率是 $\frac{2x}{k} = \frac{a}{g}$。所以在这一点，金属线垂直于有效重力：它是金属线在倾斜重力场中的“最低”点，而珠子就围绕它振动（图 11.5.1）。

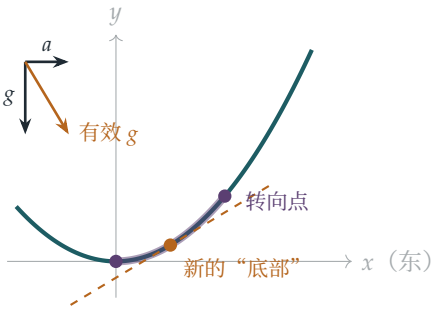


图 11.5.1 火车参考系中的金属线 $ky = x^2$（蓝绿色，取 $k = 1$ 及 $a/g = 0.6$）。有效重力（琥珀色箭头）向东倾斜，倾斜角的正切为 a/g。珠子在原点与 $x = ak/g$ 之间振动（紫红色），围绕着金属线与有效重力垂直的那一点，该点就是新的最低点。虚线是倾斜重力场中的一条“水平”线。

回顾

- **在方便的参考系中解题，并加上假想力。**在火车的参考系中，金属线是固定的，唯一的代价是多了一个恒力。
- **与速度取点积。**这样可以消去约束力（constraint force），并得出能量方程。
- **诠释答案。**转向点和最快的一点都有几何意义：有效重力倾斜了。

背后的数学

假想力

牛顿定律在惯性参考系（inertial frame）中成立，即没有加速的参考系。在加速参考系中，只要加上假想力，牛顿定律便可继续使用：对于加速度为 $\mathbf{A}$ 的参考系，要加上 $-m\mathbf{A}$；而在旋转参考系中，则要加上离心力（centrifugal force）和科里奥利力（Coriolis force）。它们之所以称为假想力，是因为没有任何物体施加这些力；它们纯粹是参考系造成的假象。但对身处其中的人来说，它们却是千真万确的。离心力使地球在两极变得扁平；科里奥利力决定了飓风旋转的方向（在北半球为逆时针），也令远程炮弹偏离原来的方向。

等效原理

恒定的假想力 $-m\mathbf{A}$ 与质量成正比，这一点与重力完全相同。在一个密封的箱子内，没有任何实验能够分辨箱子是静止在行星上，还是在虚空中加速：这就是*等效原理*（equivalence principle）。爱因斯坦以它作为广义相对论（general relativity）的出发点。既然光横越一个加速中的箱子时看来会弯曲，光在重力场中也必定会弯曲；既然加速火箭的前端和后端的时钟快慢不同，重力场中不同高度的时钟也必定如此。这两个预测都已得到证实，而每一颗 GPS 卫星都必须就第二个效应作出修正：单是这个效应，就令卫星上的时钟每天比地面的时钟快约 45 微秒。

爱因斯坦最快乐的想法

1907 年，爱因斯坦在伯尔尼的专利局工作，当时他正在撰写一篇关于狭义相对论（special relativity）的综述文章。据他后来忆述，就在这时，“我一生中最快乐的想法”出现了：一个从屋顶自由堕下的人，不会感觉到自己的重量。只要选取一个下堕中的参考系，重力就可以被“变换掉”。他再花了八年时间，并借助大量新的数学（由高斯（Gauss）和黎曼（Riemann）发展的弯曲空间微分几何），才把这个想法变成 1915 年的广义相对论。

思维技巧

在运动的参考系中寻找守恒量

在加速度恒定的参考系中，有一个恒定的假想力在起作用。它可以并入势能之中，而能量在该参考系中守恒。

把运动方程与速度取点积以消去约束力，这个技巧在任何参考系中都适用。到了大学，它会成为一个普遍的原理：若一个系统的方程并不显含时间，该系统便有一个守恒的能量。

延伸练习

1 **火车上的大炮。**试做 STEP I 2000，第 9 题（`step.maths.org/questions/00-s1-q9`）：在加速中的车厢内发射抛体（projectile），其中加速度的作用就如重力倾斜了一样。

2 **逆风。**试做 STEP II 2006，第 11 题（`step.maths.org/questions/06-s2-q11`）：抛体迎着恒定的风飞行，风可以在它落地之前把它吹回你这边。

3 **火车上的咖啡。**证明在加速中的火车上，一杯咖啡的液面是一个与水平面成 $\arctan(a/g)$ 角的倾斜平面。

4 **微小振动。**求珠子围绕 $x = \frac{ak}{2g}$ 作微小振动的周期。（提示：在新的底部附近，金属线近似于一个圆，其半径等于金属线在该处的曲率半径（radius of curvature）。）

延伸阅读

- **Daniel Kleppner & Robert Kolenkow, *An Introduction to Mechanics***（Cambridge University Press），当中讲述非惯性参考系与假想力的一章。
- **Albert Einstein, *Relativity: The Special and the General Theory***（有多个版本）。爱因斯坦亲自撰写的普及读物，当中包括加速箱子的讨论。
- **Pedro Ferreira, *The Perfect Theory***（Little, Brown）。一部广义相对论的传记。

11.6 后旋

题目

STEP III 1997，第 10 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/97-s3-q10`

引子

用手指按住一粒波子的后半部向下一压，波子便会一边向后旋转，一边向前弹出。它从你面前滑开，慢下来，停住，然后仿佛改变了主意似的滚回来。桌球（snooker）球手打“缩杆”（screw 或 draw）时用的正是同一效应：击打母球的下部，使它向后旋转，碰到目标球时停住，再退回来。

本题要求的是：后旋（backspin）恰好要多大，波子才会完全停定；又要多大，波子才会以初速的七分之一返回。计算时要追踪两个量，即线速度（linear velocity）和角速度（angular velocity），两者都由同一个摩擦力（frictional force）改变。这背后有一条漂亮的捷径：两者的某个组合完全不受摩擦力影响。

关于课程大纲的说明

转动动力学（rotational dynamics）已于 2019 年从 STEP 考试范围中删去。这里你只需要知道：质量为 m、半径为 a 的实心球，绕直径的转动惯量（moment of inertia）为 $I = \frac{2}{5}ma^2$；而力矩（torque）τ 按 $I\dot{\omega} = \tau$ 改变其角速度，这就是 $m\dot{v} = F$ 的转动版本。

动笔之前

想一想　波子以 v 向前运动，并以 ω 向后旋转。波子的最低点相对桌面移动得多快？摩擦力朝哪个方向作用？

最低点因波子本身的运动而有速度 v，再加上因后旋而来的 $a\omega$（后旋会把最低点带向前）。因此它以 $v+a\omega$ 向前滑动，而摩擦力向后作用，阻碍滑动。只有当 $v+a\omega=0$ 时，波子才会纯滚动（rolling without slipping），即只滚不滑。

逐步解题

阶段 1　摩擦力的作用

想一想　设摩擦力为 F，方向向后。它对 v 有甚么影响？它对球心的力矩为 Fa：这会增加还是减少后旋？

摩擦力使波子减速：$m\dot v=-F$，所以 $v=v_0-\frac{F}{m}t$。由于它在波子底部向后作用，它同时产生一个阻碍后旋的力矩 Fa：$I\dot\omega=-Fa$，于是代入 $I=\frac{2}{5}ma^2$ 得

$$\omega=\omega_0-\frac{5F}{2ma}t.$$

波子滑动期间，v 和 ω 都稳定地减少。

阶段 2　完全停定

想一想　波子要完全停下来，v 和 ω 必须在同一刻变成零。两者各在何时变成零？

$v=0$ 于 $t=\frac{mv_0}{F}$，而 $\omega=0$ 于 $t=\frac{2ma\omega_0}{5F}$。当 $v_0=\frac{2}{5}a\omega_0$ 时两者相等，即

$$\frac{v_0}{a\omega_0}=\frac{2}{5}.\qquad\blacksquare$$

在那一刻，波子既无线速度亦无角速度，而摩擦力（只在有滑动时才作用）随即消失：波子就此停在原地。

阶段 3　滚回来

想一想　若后旋较大，v 变成零时 ω 仍为正。最低点仍在向前滑动，所以摩擦力仍然向后作用。接下来会发生甚么事？滑动何时停止？

摩擦力继续把波子向后推，所以 v 变为负值，而 ω 继续减少。当 $v + a\omega = 0$ 时滑动停止：

$$v_0 + a\omega_0 - \left(\frac{F}{m} + \frac{5F}{2m}\right)t = 0, \qquad t^* = \frac{2m(v_0 + a\omega_0)}{7F}.$$

在那一刻，

$$v = v_0 - \frac{2(v_0 + a\omega_0)}{7} = \frac{5v_0 - 2a\omega_0}{7}.$$

此后波子只滚不滑，摩擦力不再作用，v 保持不变。要它以速率 $\frac{v_0}{7}$ 滚回来，我们需要 $5v_0 - 2a\omega_0 = -v_0$，所以

$$\frac{v_0}{a\omega_0} = \frac{1}{3}. \qquad \blacksquare$$

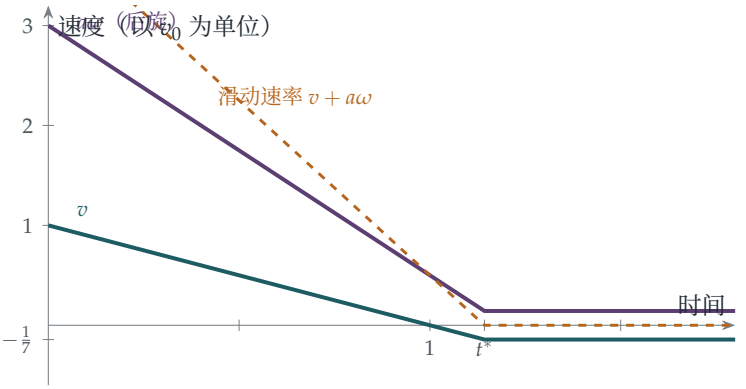


图 11.6.1 $v_0/(a\omega_0) = \frac{1}{3}$ 的情况，时间以 mv_0/F 为单位。摩擦力以恒定的速率减少向前的速度 v（蓝绿色）和后旋 $a\omega$（紫红色），两个速率之比为 2 : 5。波子在 $t = 1$ 时停止向前运动，但仍在滑动，所以被向后推，直至滑动速率（琥珀色）在 $t^* = \frac{8}{7}$ 时变成零。其后它以 $\frac{v_0}{7}$ 滚回来。

回顾

- **追踪两个量。**摩擦力同时改变 v 和 ω，两者的变率成固定比例。
- **知道力何时消失。**滑动摩擦力只在有滑动时才作用；波子一旦滚动，它便停止作用。
- **滑动速率是关键变量。**它线性地减少，而它变成零的一刻，正是由滑动转为滚动的一刻。

背后的数学

一个守恒量

要求最终速率，还有一个更快的方法。考虑波子对它与桌面接触点的角动量（angular momentum）。摩擦力正正作用于该点，所以对该点没有力矩，而法向反作用力（normal reaction）和重力都沿着通过该点的铅垂线作用。因此波子滑动期间，这个角动量守恒（conserved）。它等于 mva（来自球心的运动）减去 $I\omega$（后旋以负值计算）：

$$L = mav - \tfrac{2}{5}ma^2\omega.$$

当波子以速率 V 滚动时，$a\omega = -V$，角动量为 $maV + \tfrac{2}{5}maV = \tfrac{7}{5}maV$。令它等于初始值，

$$\tfrac{7}{5}maV = mav_0 - \tfrac{2}{5}ma^2\omega_0, \qquad V = \frac{5v_0 - 2a\omega_0}{7},$$

答案相同，却完全不必提及摩擦力的大小，甚至不必理会它是否恒定。最终速率单凭守恒便已确定；摩擦力只决定要花多少时间才达到这个速率。

桌球、保龄球与网球

同样的分析可以解释许多运动中的现象。桌球被击中正中央时，起初滑动，然后开始滚动，最终速率为初速的 $\frac{5}{7}$（在公式中取 $\omega_0 = 0$）。十瓶保龄球带侧旋掷出时，会沿球道滑行，到开始滚动时便“拐弯”。带上旋（topspin）的网球会急速下坠，不过这里的主要效应来自空气动力：旋转的球会使周围的空气偏转，这就是马格努斯效应（Magnus effect）。

科里奥利与桌球

科里奥利（Gaspard-Gustave Coriolis）如今因旋转参考系中的虚拟力（fictitious force）而为人所知（Section 11.5），他在 1835 年写了一整本关于桌球数学的书：*Théorie mathématique des effets du jeu de billard*。他证明了不少结果，其中包括：带旋转击出的球，在滑动期间沿抛物线运动，到滚动时才走直线；他亦精确地分析了本题的缩杆打法。研究这项运动的物理学家至今仍会引用此书。

思维技巧

追踪由同一个力改变的两个量

当一个力同时影响几个量时，写下它怎样改变每一个量，并找出一个它完全不会改变的组合。

这里摩擦力改变了 v 和 ω，却没有改变对接触点的角动量。在大学里，寻找这类组合是一个反复出现的主题：选择对哪一点取矩，使未知的力自动消去，正是选

择参考系的转动版本。

延伸练习

1 **上旋。**若波子起初带的是向前的旋转 ω_0 而非后旋，求它最终的滚动速率。$v_0/(a\omega_0)$ 取甚么值时它会加速?

2 **桌球。**一个球被击中正中央，以 v_0 开始滑动而不带旋转。证明它最终以 $\frac{5}{7}v_0$ 滚动，并求动能（kinetic energy）损失的比例。（滚动时的动能为 $\frac{1}{2}mV^2 + \frac{1}{2}I\omega^2$。）

3 **圆环。**对一个在自身所在平面内滚动的薄圆环重做本题，其 $I = ma^2$。现在 $v_0/(a\omega_0)$ 为何值时，它会以 $\frac{v_0}{7}$ 返回?

4 **缩杆打法。**桌球球杆必须击中球的甚么位置（在球心以上多高，以 a 表示），球才会从一开始便只滚不滑?（提示：考虑冲量（impulse）作用期间对接触点的角动量。）

延伸阅读

- **David Morin, *Introduction to Classical Mechanics***（Cambridge University Press），第 8 章，讲述角动量及带滑动的滚动。
- **Daniel Kleppner & Robert Kolenkow, *An Introduction to Mechanics***（Cambridge University Press），关于刚体运动的各章。
- **David Alciatore, *The Illustrated Principles of Pool and Billiards***（Sterling）。这项运动的物理学，附有大量图解。

11.7 杆中的辛普森法则

题目

STEP II 1989，第 13 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/89-s2-q13`

引子

一根均匀棒（uniform rod）受到推动和旋转时，会有甚么反应？原来就力学上的一切用途而言，一根棒的表现与三个质点（point mass）完全相同：两端各占其质量的六分之一，中点占三分之二。无论你对这根棒做甚么（推它，或让它绕任何轴旋转），都无法把它与这三个质点区分开来。

$\frac{1}{6}$、$\frac{2}{3}$、$\frac{1}{6}$ 这几个数也许似曾相识。它们正是*辛普森法则*（Simpson's rule）$\int_{-a}^{a} f \approx 2a\left(\frac{1}{6}f(-a)+\frac{2}{3}f(0)+\frac{1}{6}f(a)\right)$ 中的权（weight），这是你可能在高级程度（A level）课程中学过的数值积分（numerical integration）方法。这并非巧合：动力等效（dynamically equivalent）系统与数值积分法则，解决的其实是同一个问题。

关于课程范围的说明

转动惯量（moment of inertia）已于 2019 年从 STEP 考试范围中删去。一个物体对某轴的转动惯量是 $\sum m_i d_i^2$，即每份质量乘以它与该轴距离的平方之总和（对连续物体而言则是一个积分）。质量为 m、长度为 $2a$ 的均匀棒，对通过其中心的垂直轴的转动惯量为 $\frac{1}{3}ma^2$。*平行轴定理*（parallel axis theorem）指出：对任何轴的转动惯量，等于对通过质心（centre of mass）的平行轴的转动惯量加上 md^2，其中 d 是两轴之间的距离。

动笔之前

想一想　两个系统的质量相同、质心相同，而且对通过质心的每一条轴的转动惯量也相同。为甚么平行轴定理随即保证，它们对其他任何轴的转动惯量也相同？

对任何轴而言，两个系统的转动惯量都等于（对通过 O 的平行轴的转动惯量）$+md^2$。根据假设，第一项相同，而两个系统的 m 和 d 也相同。 ■

逐步解题

阶段 1　棒

想一想　在两端各放质量 p，在中点放质量 q。由对称性可知质心位置正确。总质量和对通过中心的垂直轴的转动惯量，分别要求甚么条件？

我们需要 $2p+q=m$，而对通过中点的垂直轴的转动惯量必须是 $\frac{1}{3}ma^2$：$2pa^2=\frac{1}{3}ma^2$。所以 $p=\frac{1}{6}m$，$q=\frac{2}{3}m$。

想一想　这是否对通过中心的每一条轴都成立，而不只是垂直的那一条？

考虑一条通过中心、与棒成 ψ 角的轴：棒上与中心相距 r 的点，与该轴的距离为 $r\sin\psi$。因此每一个转动惯量（不论是棒的还是质点的）都是垂直轴情况下数值的 $\sin^2\psi$ 倍，所以两者对所有 ψ 都相等。（不在包含棒的平面上的轴也是同样道理：只有各点与轴的距离才有关系，而对于一条通过 O 的直线上的点，这个距离与 r 成正比。）所以这根棒与质量 $\frac{1}{6}m$、$\frac{2}{3}m$、$\frac{1}{6}m$ 动力等效。 ■

阶段 2　长方形

想一想　把薄片（lamina）切成与其中一边平行的幼条，再把每一幼条换成它的三个等效质点。你会得到甚么？接下来又可以怎样做？

每一幼条都是一根棒，等效于把它质量的 $\frac{1}{6}$、$\frac{2}{3}$、$\frac{1}{6}$ 分别放在左边、中线和右边。把所有幼条加起来，薄片便等效于三根均匀棒（分别沿左边、中间和右边），质量为 $\frac{1}{6}M$、$\frac{2}{3}M$、$\frac{1}{6}M$。现在再把这三根棒各自换成比例相同的三个质点。每一点的质量是两个分数之积（图 11.7.1）：

$$\text{四角：}\ \tfrac{1}{6}\cdot\tfrac{1}{6}=\tfrac{1}{36},\qquad \text{各边中点：}\ \tfrac{1}{6}\cdot\tfrac{2}{3}=\tfrac{1}{9},\qquad \text{中心：}\ \tfrac{2}{3}\cdot\tfrac{2}{3}=\tfrac{4}{9},$$

全部再乘以 M。 ■

（严格来说，每一次替换都必须保持对通过整块薄片中心的每一条轴的转动惯量不变，而事实确是如此：每一次替换都保持了该部分的质量、质心，以及对通过该部分质心的各轴的转动惯量，其余的就交给平行轴定理。）

阶段 3 正方形对其对角线

对于边长为 $2a$ 的正方形，对角线通过两个角和中心，这些点没有任何贡献。另外两个角与对角线的距离为 $\sqrt{2}a$，而四个边的中点与它的距离为 $\frac{a}{\sqrt{2}}$。所以

$$I = 2\cdot\tfrac{1}{36}M\cdot 2a^2 + 4\cdot\tfrac{1}{9}M\cdot\tfrac{a^2}{2} = \tfrac{1}{9}Ma^2 + \tfrac{2}{9}Ma^2 = \tfrac{1}{3}Ma^2. \qquad ■$$

这与对通过中心、平行于一边的轴的转动惯量相同。（这并非偶然：对正方形而言，在其平面上通过中心的所有轴都给出相同的答案，这可由练习中的一个对称性论证得出。）

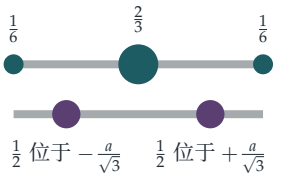
阶段 4 非均匀棒

> **想一想** 设棒的长度为 L，在距离其中一端 x 处的密度为 $\frac{2mx}{L^2}$。在 $x=0$、$\frac{L}{2}$、L 处分别放质量 p_0、p_1、p_2。令总质量、质心和对该端的转动惯量一一吻合，便得到三条方程。它们是甚么？

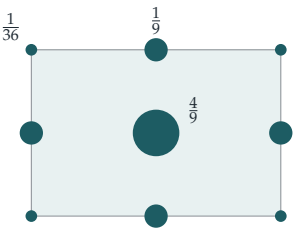
这根棒的总质量为 m，对较轻一端的一阶矩（first moment）为 $\int_0^L x\cdot\frac{2mx}{L^2}\,\mathrm{d}x = \frac{2}{3}mL$，二阶矩（second moment）为 $\int_0^L x^2\cdot\frac{2mx}{L^2}\,\mathrm{d}x = \frac{1}{2}mL^2$。令这些量吻合：

$$p_0+p_1+p_2 = m, \qquad \tfrac{1}{2}Lp_1 + Lp_2 = \tfrac{2}{3}mL, \qquad \tfrac{1}{4}L^2p_1 + L^2p_2 = \tfrac{1}{2}mL^2.$$

把第二式（除以 L）减去第三式（除以 L^2），得 $\frac{1}{4}p_1 = \frac{1}{6}m$，所以 $p_1 = \frac{2}{3}m$，继而 $p_2 = \frac{1}{3}m$，$p_0 = 0$。只需要两个质点：$\frac{2}{3}m$ 位于中点，$\frac{1}{3}m$ 位于较重的一端。 ■



两根等效的棒



薄片：$\frac{1}{6},\frac{2}{3},\frac{1}{6}$ 的乘积

图 11.7.1 左上：一根均匀棒与位于两端和中心的质量 $\frac{1}{6}$、$\frac{2}{3}$、$\frac{1}{6}$（以占其质量的分数表示）动力等效，这些正是辛普森法则的权。左下：它也等效于两份各占一半的质量，位于与中心相距 $\frac{a}{\sqrt{3}}$ 之处，即两点高斯积分的节点。右：一块长方形薄片等效于九个质量，图中圆点的面积与质量成正比。

回顾

- **令各阶矩吻合。**动力等效的意思，就是令质量分布的零阶矩、一阶矩和二阶矩一一吻合。
- **逐个维度建构。**长方形是由棒组成的棒，所以它的权是乘积。
- **数一数方程和未知数的数目。**要吻合的矩有三个，未知的质量也有三个；而有时解中恰好会出现一个零。

背后的数学

辛普森法则

令质量分布的各阶矩吻合，与设计一条数值积分法则是同一个问题。一条法则 $\int_{-a}^{a} f(x)\,\mathrm{d}x \;\approx\; w_0 f(-a) + w_1 f(0) + w_2 f(a)$ 若对 $f = 1, x, x^2$ 精确成立，就必须满足

$$w_0 + w_1 + w_2 = 2a, \qquad -a w_0 + a w_2 = 0, \qquad a^2 w_0 + a^2 w_2 = \tfrac{2}{3}a^3,$$

这正是棒所要满足的条件（其中 $f = 1, x, x^2$ 分别扮演质量、一阶矩和转动惯量的角色）。其解为 $w = 2a(\tfrac{1}{6}, \tfrac{2}{3}, \tfrac{1}{6})$：这就是辛普森法则。对称性还带来额外的收获：这条法则对 $f = x^3$ 也精确成立，因为两边都等于零，所以辛普森法则能精确地积分三次多项式，而它对光滑函数的误差与四阶导数成正比。因此，一根棒与它的三个质点不仅在质量、质心和转动惯量上吻合，连三阶矩也吻合。

高斯求积法与惯性张量

如果位置和质量都可以自由选择，两个质点便已足够：在 $\pm\frac{a}{\sqrt{3}}$ 两点各放一半质量。验算：$2 \cdot \frac{1}{2}m \cdot \frac{a^2}{3} = \frac{1}{3}ma^2$。这就是两点*高斯－勒让德法则*（Gauss–Legendre rule）$\int_{-a}^{a} f \approx a\big(f(-\frac{a}{\sqrt{3}}) + f(\frac{a}{\sqrt{3}})\big)$，它同样对三次多项式精确成立。高斯（Carl Friedrich Gauss）在 1814 年证明，只要巧妙地选取 n 个点，便能精确地积分次数为 $2n-1$ 的多项式；高斯求积法（Gaussian quadrature）如今已是数值分析（numerical analysis）的标准方法。

在力学中，一个物体对通过其中心的所有轴的转动惯量，都记录在一个 3×3 对称矩阵之中，称为*惯性张量*（inertia tensor）。两个物体若质量、质心和惯性张量都相同，对所有力的反应便完全一样，这就是为甚么工程师在模拟中可以用几个质点取代一个复杂的部件。惯性张量的特征向量（eigenvector）就是*主轴*（principal axes），即物体可以绕之稳定旋转的轴。

思维技巧

把系统换成等效的系统

如果两个系统在问题所涉及的每一个量上都一致，就把复杂的那个换成简单的那个。

这里只有三个矩有关系，所以一根连续的棒可以换成三个质点。在大学里，这个想法会以多种面貌出现：高斯求积法、电磁学和引力理论中的多极展开（multipole expansion）（在远处，电荷或质量分布只有头几个矩有影响），以及统计学中的矩估计法（method of moments）。

延伸练习

1 **辛普森法则。**用只有一个区间的辛普森法则估算 $\int_0^1 \frac{dx}{1+x^2}$，并与 $\frac{\pi}{4}$ 比较。验证这条法则对 $\int_{-1}^1 x^3\,dx$ 和 $\int_{-1}^1 x^2\,dx$ 精确成立，但对 $\int_{-1}^1 x^4\,dx$ 则不然。

2 **所有轴都一样。**证明正方形薄片对在其平面上通过中心的任何轴的转动惯量都是 $\frac{1}{3}Ma^2$。（提示：九个等效质点在旋转 90° 之下是对称的；这对一个关于 $\cos\psi$ 和 $\sin\psi$ 的二次式说明了甚么?）

3 **三角形。**证明一块质量为 M 的均匀三角形薄片，与位于其三边中点、质量各为 $\frac{1}{3}M$ 的三个质点动力等效。

4 **实心立方体。**为一个均匀实心立方体找出一个由 27 个质点组成的动力等效系统，并用它求出边长为 $2a$ 的立方体对其一条棱的转动惯量。

延伸阅读

- **Daniel Kleppner & Robert Kolenkow, *An Introduction to Mechanics***（Cambridge University Press），关于刚体和转动惯量的一章。
- **Endre Süli & David Mayers, *An Introduction to Numerical Analysis***（Cambridge University Press），关于数值积分的各章，包括辛普森法则和高斯求积法。
- **Philip Davis & Philip Rabinowitz, *Methods of Numerical Integration***（Dover）。经典参考书。

11.8 倾斜的平衡

题目

STEP III 2025，第 9 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2025 年 STEP 试卷（zip 文件）内的 STEP III 试卷：

`www.ocr.org.uk/Images/750486-2025-past-papers.zip`

引子

把一个橄球放在桌上，它会侧躺着停下来；把它竖起来，只要十分小心，它也可以这样平衡，但轻轻一碰就会倒下。每一个放在桌上的刚体（rigid body）都有一组平衡位置（equilibrium positions），其中有些是稳定（stable）的，有些是不稳定（unstable）的，而哪些位置存在，只取决于它的形状和质心（centre of mass）。一个物体可以有多少个平衡位置？最少又可以有多少个？

这道题构造了一个立体，形状像陀螺，上半部是圆顶，下半部是尖底，并证明它除了可以正立或倒立之外，还可以在三个不同的倾斜角度上静止。关键思路是一个简单的几何条件：物体能以其表面上某一点着地而静止，当且仅当表面在该点的法线（normal）通过质心。这背后有一项近年的发现：一个称为 Gömböc 的凸（convex）形体，它只有一个稳定位置和一个不稳定位置。

动笔之前

想一想 一个立体放在桌上，只以其曲面上的一点 X 与桌面接触。作用于它的力是它的重量（作用于质心 G）和在 X 的反作用力（reaction）。这两个力何时平衡？

桌面在 X 与曲面相切，而反作用力作用于 X，并垂直于桌面，即沿着曲面的法线。重量则沿铅垂方向作用，并通过 G。两个力只有在沿同一直线作用时才能平衡，所以通过 X 的铅垂线，也就是在 X 的法线，必须通过 G。因此：*物体能以 X 着地而静止，当且仅当在 X 的法线通过 G。*

逐步解题

阶段 1　质心

> **想一想**　把立体切成薄薄的圆柱壳（cylindrical shells），半径为 x，厚度为 δx，高度为 y。一个圆柱壳的体积是多少？它的质心在哪里？

由于 $y=\frac{r^n-x^n}{r^{n-1}}$，一个圆柱壳的体积是 $2\pi xy\,\delta x$，而它的质心在高度 $\frac{y}{2}$ 处。所以

$$V=\int_0^r 2\pi xy\,\mathrm{d}x=\frac{2\pi}{r^{n-1}}\Big(\frac{r^{n+2}}{2}-\frac{r^{n+2}}{n+2}\Big)=\frac{n\pi r^3}{n+2},$$

而对底面的矩（moment）是

$$\begin{aligned}\int_0^r 2\pi x\cdot\frac{y^2}{2}\,\mathrm{d}x&=\frac{\pi}{r^{2n-2}}\int_0^r x(r^n-x^n)^2\,\mathrm{d}x\\&=\pi r^4\Big(\frac{1}{2}-\frac{2}{n+2}+\frac{1}{2n+2}\Big)=\frac{n^2\pi r^4}{2(n+1)(n+2)}.\end{aligned}$$

（括号可以化简，是因为 $(n+1)(n+2)-4(n+1)+(n+2)=n^2$。）两者相除，得

$$\bar{y}=\frac{nr}{2(n+1)}.\qquad\blacksquare$$

当 $n=2$（抛物面，paraboloid）时，这是 $\frac{r}{3}$，正是大家熟悉的结果；当 $n\to\infty$ 时，立体变成圆柱，而 $\bar{y}\to\frac{r}{2}$。

阶段 2　法线与轴相交之处

曲线在 $x=rp$ 的斜率（gradient）是 $-\frac{nx^{n-1}}{r^{n-1}}=-np^{n-1}$，所以法线的斜率是 $\frac{1}{np^{n-1}}$。沿法线由 $\big(rp,r(1-p^n)\big)$ 走到 y-轴，x 减少了 rp，而 y 减少了 $\frac{rp}{np^{n-1}}=\frac{r}{np^{n-2}}$。所以

$$Y=r\Big(1-p^n-\frac{1}{np^{n-2}}\Big).\qquad\blacksquare$$

> **想一想**　当 $n=4$ 时，$Y=r\big(1-p^4-\frac{1}{4p^2}\big)$。它在哪里取最大值？当 $p\to0$ 时，以及在 $p=1$ 时，情况如何？

$\frac{\mathrm{d}Y}{\mathrm{d}p}=r\big(-4p^3+\frac{1}{2p^3}\big)=0$ 成立于 $p^6=\frac{1}{8}$，即 $p^2=\frac{1}{2}$。这时 $Y=r\big(1-\frac{1}{4}-\frac{1}{2}\big)=\frac{1}{4}r$。$\blacksquare$

当 $p\to0$ 时，$Y\to-\infty$（圆顶在顶部附近非常平坦，它的法线几乎是铅垂的，要在很低的地方才与轴相交）；在 $p=1$ 时，$Y=-\frac{1}{4}r$。所以当 p 由 0 增至 1 时，Y 由 $-\infty$ 上升至 $\frac{1}{4}r$，再回落至 $-\frac{1}{4}r$。

阶段 3　陀螺

想一想 第 (iii) 部分的立体，是一个圆顶（$n=4$）叠在一个倒转的抛物面（$n=2$）之上。它的质心在哪里？

由步骤 1，圆顶的体积是 $\frac{2\pi r^3}{3}$，质心在 $\frac{2r}{5}$；抛物面的体积是 $\frac{\pi r^3}{2}$，而由于它是倒转的，质心在 $-\frac{r}{3}$。所以

$$\bar{y} = \frac{\frac{2}{3}\cdot\frac{2}{5} - \frac{1}{2}\cdot\frac{1}{3}}{\frac{2}{3}+\frac{1}{2}}\, r = \frac{\frac{4}{15}-\frac{1}{6}}{\frac{7}{6}}\, r = \frac{3r}{35}.$$

G 位于 $\left(0, \frac{3r}{35}\right)$，比最阔之处略高。

阶段 4 数算平衡位置

想一想 倾斜的平衡位置需要有一点（不是 A 或 B），其法线通过 G。在曲面的每一部分上，有多少个点的法线与轴相交于高度 $\frac{3r}{35}$？

圆顶。它的法线与轴相交的高度 Y 由 $-\infty$ 上升至 $\frac{1}{4}r$，再下降至 $-\frac{1}{4}r$。由于 $-\frac{1}{4}r < \frac{3r}{35} < \frac{1}{4}r$，根据介值定理（intermediate value theorem），$\frac{3r}{35}$ 这个值在上升的一段取得一次，在下降的一段也取得一次：共两点。

尖底。由第 (ii) 部分取 $n=2$，再对 x-轴作反射，可知在参数为 p 的点上，法线与轴相交于 $-r\left(1-p^2-\frac{1}{2}\right) = r\left(p^2-\frac{1}{2}\right)$，这个值由 $-\frac{1}{2}r$ 递增至 $\frac{1}{2}r$。它恰好有一次等于 $\frac{3r}{35}$，就是在 $p^2 = \frac{41}{70}$ 时。

这三点各给出一个平衡位置，而三条法线与轴所成的角各不相同，所以 $\overrightarrow{AB}$ 与铅垂线成三个不同的非零角（图 11.8.1）。 ■

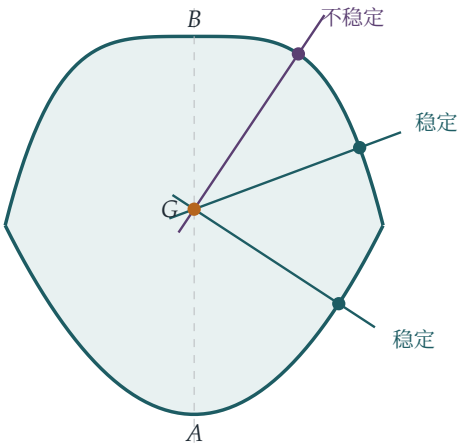


图 11.8.1 第 (iii) 部分的立体的截面，以及它的质心 G（琥珀色）。在三个标记的点上，曲面的法线都通过 G，所以立体可以用该点着地而静止。其中两个平衡位置是稳定的（曲面在该处局部地离 G 最近），一个是不稳定的（局部地最远）。

回顾

- **把物理条件化为几何。**平衡变成了“法线通过 G”。
- **不用求解也能证明存在。**我们从没有求出那些角。介值定理，加上函数 $Y(p)$ 的形状，已说明了它们有多少个。
- **重用一般结果。**第 (ii) 部分对一般的 n 所得的结果，一并处理了第 (iii) 部分中曲面的两个部分。

背后的数学

平衡位置即临界点

当立体以点 X 着地而静止时，G 离桌面的高度就是由 G 到在 X 的切平面（tangent plane）的距离；对凸体来说，当 GX 沿着法线时，这个距离就是 $|GX|$。一般而言，平衡位置就是由 G 到曲面的距离的临界点（*critical points*）：使 $|GX|$ 取驻值的点 X，正好就是法线通过 G 的点。稳定的平衡位置是这个距离的局部极小值（G 尽可能低），而不稳定的则是局部极大值或鞍点（saddle points）。

沿着截面由 A 走到 B，与 G 的距离依次是：在 A 较大（极大值：以尖端平衡，不稳定），在尖底上的倾斜点较小（极小值，稳定），在边缘较大，在圆顶上较低的倾斜点较小（稳定），在较高的倾斜点较大（不稳定），而在 B 较小（稳定：以圆顶的顶部着地）。正如图中所示，极小值和极大值必定交替出现。（事实上还有一个题目没有计算在内的平衡位置：立体可以不稳定地平衡在它尖锐的边缘上，曲面在那里没有唯一的法线。）

Gömböc

一个凸体最少可以有多少个平衡位置？任何物体都至少有一个稳定位置和一个不稳定位置：就是曲面上离 G 最近和最远的点。密度均匀的二维凸形状总是至少有四个（距离的两个极小值和两个极大值，交替出现）。1995 年，阿诺尔德（Vladimir Arnold）提出一个问题：密度均匀的三维凸体，能否只有两个平衡位置，即一个稳定、一个不稳定？2006 年，Gábor Domokos 和 Péter Várkonyi 构造出这样的一个形体，就是 *Gömböc*：无论以甚么姿势把它放下，它都会滚动摇晃，直至到达它唯一的稳定位置为止。Domokos 后来留意到，有些陆龟的壳很接近这个形状，这有助牠们在被翻转后自行翻身。

阿基米德与浮体

对物体稳定性的研究可以追溯至阿基米德（Archimedes）。他在《论浮体》（*On Floating Bodies*）中，根据形状和密度，分析了一个浮着的抛物面（船身的一个模型）何时能稳

定地直立，何时会翻侧。他只用几何便得出答案，当中涉及的正是本节所用的那种质心和法线的计算。造船工程师至今仍在运用他的想法，形式就是稳心（*metacentre*）：如果通过浮心（centre of buoyancy）的铅垂线与船的轴线的交点位于质心之上，船便是稳定的。

思维技巧

势能地形

平衡位置是势能（*potential energy*）的驻点（*stationary points*），而稳定的平衡位置是它的极小值。与其把驻点逐一解出来，不如数算它们的数目。

这里的势能与 G 的高度成正比，也就是由 G 到曲面的距离。在大学里，这个观点贯穿力学、化学（分子会落入能量的极小值）和拓扑学；在拓扑学中，莫尔斯理论（*Morse theory*）说明，曲面上任何函数的极小值、极大值和鞍点的数目，都受曲面的形状所限制。

延伸练习

1 **不倒翁玩具。**试做 STEP II 1996，第 9 题（`step.maths.org/questions/96-s2-q9`）：一个置于半球上的圆锥，何时会自行回正？（提示：相对于半球的球心，G 必须在哪里？）

2 **球体。**对于均匀实心球体，每一条法线都通过球心。这说明它的平衡位置有甚么特点？如果球体的质心稍为偏离球心，又会怎样？

3 **椭圆。**证明一块椭圆形圆盘在其自身所在的平面内沿一条直线滚动时，恰好有四个平衡位置，并指出哪些是稳定的。

4 **两个抛物面。**把第 (iii) 部分的圆顶换成抛物面 $ry = r^2 - x^2$，使立体对 x-轴对称。证明 G 位于原点，求出倾斜的平衡位置，并解释为何它们成对出现，且每对的倾斜角相等。

延伸阅读

► **Péter Várkonyi & Gábor Domokos, “Static equilibria of rigid bodies: dice, pebbles, and the Poincaré–Hopf theorem”**, *Journal of Nonlinear Science* 16 (2006)。提出 Gömböc 的

论文。

- **Daniel Kleppner & Robert Kolenkow, *An Introduction to Mechanics***（Cambridge University Press），论平衡与稳定性。
- **Chris Rorres, “Completing Book II of Archimedes’s *On Floating Bodies*”**, *Mathematical Intelligencer* 26 (2004)。阿基米德的稳定性分析，配以现代图解。

11.9 安全抛物线

题目

STEP II 2014，第 10 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/14-s2-q10`

引子

一座喷泉以相同的速率向四方八面喷水。水柱形成一个闪闪生辉的圆顶，外缘轮廓分明：在里面，每一点都有水经过；在外面，水一滴也到不了。这条边界是一条抛物线（parabola），称为*安全抛物线*（parabola of safety），因为炮口速率固定的大炮，无论以甚么角度发射，都打不中它外面的目标。

令人意外的是，这条边界本身并不是其中一条轨迹。它是整族轨迹的*包络线*（envelope），即一条与每条轨迹都相切一次的曲线，就像 Section 8.1 中直线 $y = 0$ 与那族抛物线相切一样。这道题目只需用到二次式的判别式（discriminant），便能把它求出来。

动笔之前

? **想一想** 用 t 表示 x 和 y，然后消去 t。怎样用 $\lambda = \tan\theta$ 表示 $\sec^2\theta$?

由 $x = ut\cos\theta$ 和 $y = ut\sin\theta - \frac{1}{2}gt^2$，代入 $t = \frac{x}{u\cos\theta}$：

$$y = x\tan\theta - \frac{gx^2}{2u^2\cos^2\theta} = \lambda x - \frac{gx^2}{2u^2}(1+\lambda^2),$$

其中用到 $\sec^2\theta = 1 + \tan^2\theta$。

逐步解题

阶段 1 固定 x，改变角度

想一想　对于固定的 $x>0$，把 y 看成 λ 的函数。它是哪一类函数？当 λ 取遍所有实数时，它会取哪些值？

对于固定的 x，y 是 λ 的二次函数，首项系数 $-\frac{gx^2}{2u^2}$ 为负：它是以 λ 为变量、开口向下的抛物线。它有极大值，而且当 $\lambda\to\pm\infty$ 时，$y\to-\infty$。由于它是连续的，所以它会取遍不超过极大值的每一个值。极大值出现在 $\frac{\mathrm{d}y}{\mathrm{d}\lambda}=x-\frac{gx^2}{u^2}\lambda=0$ 之处，即 $\lambda=\frac{u^2}{gx}$，此时

$$Y=\frac{u^2}{gx}\,x-\frac{gx^2}{2u^2}\left(1+\frac{u^4}{g^2x^2}\right)=\frac{u^2}{2g}-\frac{gx^2}{2u^2}.\qquad\blacksquare$$

（等价的说法是：点 (x,y) 可以到达，当且仅当二次方程 $\frac{gx^2}{2u^2}\lambda^2-x\lambda+\frac{gx^2}{2u^2}+y=0$ 有实根 λ，而判别式的条件给出 $y\le Y$。）

阶段 2　区域

曲线 $Y=\frac{u^2}{2g}-\frac{gx^2}{2u^2}$ 是一条开口向下的抛物线，顶点位于 $\left(0,\frac{u^2}{2g}\right)$，即垂直向上发射所达到的高度；它与地面相交于 $x=\pm\frac{u^2}{g}$，即最大射程。在它上面或下方（并在地面以上）的每一点都可以到达，而在它上方的点则全都到不了（图 11.9.1）。

想一想　经过严格位于曲线下方的一点，有多少条轨迹？经过曲线上的一点呢？

曲线下方的一点对应一个有两个实根的 λ 二次方程：两个抛射角，一高一低。曲线上的一点对应重根：恰好一条轨迹，并且在该点与曲线相切。

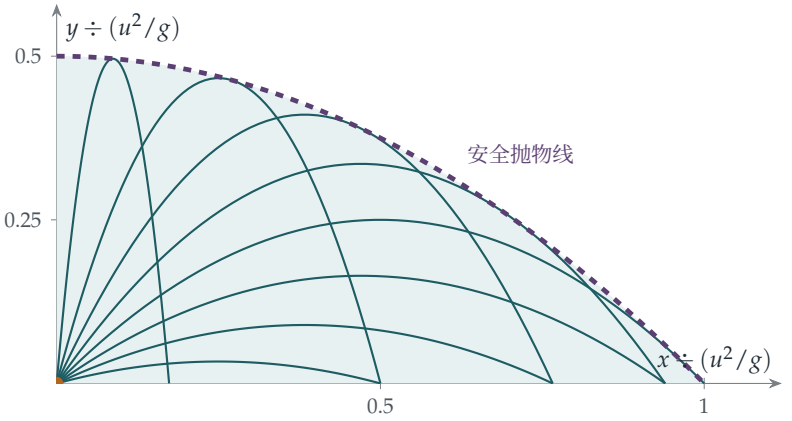


图 11.9.1　以固定速率 u 及角度 $15°,25°,\dots,85°$ 发射的轨迹（蓝绿色）。每条轨迹都与虚线所示的安全抛物线 $y=\frac{u^2}{2g}-\frac{gx^2}{2u^2}$ 恰好相切一次，而它下方的区域（阴影部分）正好是所有可到达的点所组成的集合。原点 O（琥珀色）是安全抛物线的焦点（focus）。

阶段 3　最远的点

想一想 距离 O 最远的可到达点（在地面上或地面以上）位于边界上。把距离的平方 x^2+Y^2 写成 x^2 的函数。它是递增还是递减?

设 $X=x^2$。则

$$x^2+Y^2=X+\left(\frac{u^2}{2g}-\frac{gX}{2u^2}\right)^2,\qquad \frac{\mathrm{d}}{\mathrm{d}X}(x^2+Y^2)=1-\frac{g}{u^2}\left(\frac{u^2}{2g}-\frac{gX}{2u^2}\right)=\frac{1}{2}+\frac{g^2X}{2u^4}>0.$$

因此，沿着边界向外移动时，距离不断增加，而最远的点就是边界与地面相交之处，即 $x=\frac{u^2}{g}$。最远距离为

$$\frac{u^2}{g},$$

由平常以 45° 发射的最大射程一击达到。 ■

回顾

- **把问题反过来问。**不问某条给定的轨迹会去到哪里，而问哪些轨迹会经过某个给定的点。
- **以参数为未知数的二次方程。**对于每一点，$\lambda=\tan\theta$ 所满足的方程都是二次方程，而它的判别式决定该点能否到达。
- **两个、一个或没有。**区域内的点有两个角度可以到达，边界上的点有一个，区域外的点则一个也没有。

背后的数学

包络线

单参数曲线族（one-parameter family of curves）$\mathrm{F}(x,y,\lambda)=0$ 往往有一条*包络线*：一条与族中每个成员都相切的曲线。求法是联立解 $\mathrm{F}=0$ 和 $\frac{\partial \mathrm{F}}{\partial\lambda}=0$，这正是我们对 λ 求 y 的极大值时所做的事。包络线在科学中随处可见。一杯咖啡表面上那条明亮的光曲线，就是反射光线的包络线，称为*焦散线*（caustic）。超音速飞机的音爆，由它所发出声波的包络面传送，那就是*马赫锥*（Mach cone）。而在 Section 8.1 中，一个微分方程（differential equation）的一族解的包络线，本身也是一个解，即所谓的“奇异”解。

焦点

安全抛物线有一个奇妙的性质：它的焦点就是发射点 O。（抛物线 $y=c-\frac{x^2}{4c}$ 的焦点位于原点；这里 $c=\frac{u^2}{2g}$。）每条轨迹也都是抛物线，而所有轨迹的焦点都落在以 O 为圆心、

半径为 $\frac{u^2}{2g}$ 的圆上。这可以用能量来解释：抛射体在高度 y 时的速率是 $\sqrt{u^2-2gy}$，在准线（directrix）的高度 $\frac{u^2}{2g}$ 处为零；而以同一条直线为准线并且经过 O 的抛物线，其焦点都在那个圆上。

托里拆利的抛物线

伽利略（Galileo）在 1609 年前后证明了抛射体沿抛物线运动，并且在水平地面上，最大射程出现在 45°。他的学生托里拆利（Evangelista Torricelli）以发明气压计而较为人熟悉，他在 1644 年的著作 *De motu gravium* 中发现了安全抛物线。对炮手来说，它回答了一个实际问题：究竟哪些目标在射程之内？托里拆利还证明了，经过给定一点的轨迹成对出现，两者的角度关于该点的方向与铅垂线之间的角平分线对称，这就是本题中的高射和低射。

思维技巧

曲线族的包络线

要找出一族曲线所覆盖区域的边界，可固定一点，然后问：参数取哪些值时，会有曲线经过这一点？边界就是解的数目发生变化的地方。

同一个想法，即随着点的移动来数算解的数目，是光学中焦散线研究的基础，也是动力系统（dynamical systems）中分歧（bifurcation）理论的基础，更是突变理论（catastrophe theory）的基础；突变理论把方程的解的数目可以怎样变化加以分类。

延伸练习

1 **爆开的烟花。**试做 STEP II 2023，第 10 题（载于 OCR 提供下载的 2023 年试卷，`www.ocr.org.uk/Images/707102-2023-past-papers.zip`）。这道题把爆炸产生的粒子看成一个球面，其球心作自由下落，并找出一颗沿着安全抛物面（paraboloid of safety）运动的粒子。

2 **向斜坡上发射。**求速率为 u 的抛射体沿倾角为 β 的斜面向上发射时的最大射程。（提示：直线 $y = x\tan\beta$ 与安全抛物线在哪里相交？）证明最佳角度平分斜坡与铅垂线之间的夹角。

3 **两个角度。**证明能到达点 P 的两个角度，关于 OP 与铅垂线夹角的平分线对称。

4 **焦点。**证明抛射角为 θ 的轨迹，其焦点位于 $\left(\frac{u^2}{2g}\sin 2\theta, -\frac{u^2}{2g}\cos 2\theta\right)$，并由此证明这些焦点都落在以 O 为圆心的一个圆上。

延伸阅读

- **David Morin, *Introduction to Classical Mechanics***（Cambridge University Press），第 3 章，当中讨论了轨迹的包络线。
- **J. W. Bruce, P. J. Giblin & C. G. Gibson, "On caustics of plane curves"**, *American Mathematical Monthly* 88 (1981)。几何学中的包络线与焦散线。
- **E. H. Lockwood, *A Book of Curves***（Cambridge University Press）。包络线、焦散线及许多经典曲线，并附作图方法。

11.10 会朝你飞回来的球

题目

STEP II 2019，第 9 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2019 年 STEP 试卷（zip 文件）内的 STEP II 试卷：

`www.ocr.org.uk/Images/696590-2019-past-papers.zip`

引子

把球向前抛出，它在落地之前会一直离你而去。这是理所当然的。把球垂直向上抛，它却会回到你手上。在这两种情况之间，必定有些事情起了变化：必定存在某些角度，令球先远离你，然后向你靠近，到落地之前又再远离。本题要找出的正是那个确切的分界线：$\sin\alpha = \frac{2\sqrt{2}}{3}$，即大约 70.5°。

这个结果是数学中一种常见情况的好例子：人人都以为某个量是单调（monotonic）的，其实不然。它同时展示了一个在大学里经常用到的标准技巧：要判断一个距离是否正在增加，就对它的平方求导（differentiate），这样便可避开平方根。

动笔之前

想一想 为甚么处理 OP^2 比处理 OP 容易？“OP 递增”与“OP^2 递增”是同一回事吗？

$OP^2 = x^2 + y^2$ 是 t 的多项式（polynomial），而 OP 则涉及平方根。由于 $OP \geq 0$，而平方运算在非负数上是递增的，所以 OP 递增的时候，正正就是 OP^2 递增的时候。

逐步解题

阶段 1 距离的平方

由 $x = ut\cos\alpha$ 及 $y = ut\sin\alpha - \frac{1}{2}gt^2$，可得

$$OP^2 = u^2t^2\cos^2\alpha + u^2t^2\sin^2\alpha - ugt^3\sin\alpha + \tfrac{1}{4}g^2t^4 = u^2t^2 - ugt^3\sin\alpha + \tfrac{1}{4}g^2t^4.$$

想一想 对 t 求导。抽出因子 t 之后，剩下的是甚么？它在甚么情况下对所有 t 皆为正？

$$\frac{\mathrm{d}}{\mathrm{d}t}OP^2 = 2u^2t - 3ugt^2\sin\alpha + g^2t^3 = t(g^2t^2 - 3ug\sin\alpha\, t + 2u^2).$$

当 $t > 0$ 时，其正负号与二次式（quadratic）$g^2t^2 - 3ug\sin\alpha\, t + 2u^2$ 的正负号相同。这个二次式的首项系数为正，所以只要它的判别式（discriminant）为负，它对每一个 t 都是正的：

$$9u^2g^2\sin^2\alpha - 8u^2g^2 < 0 \quad\Longleftrightarrow\quad \sin^2\alpha < \frac{8}{9} \quad\Longleftrightarrow\quad \sin\alpha < \frac{2\sqrt{2}}{3}.$$

在这情况下，OP 在整个飞行过程中都在递增。 ■

阶段 2　较陡的抛掷

想一想 若 $\sin\alpha > \frac{2\sqrt{2}}{3}$，这个二次式有两个实根。它们是正数吗？较小的那个根是否出现在质点落地之前？

两根之和为 $\frac{3u\sin\alpha}{g} > 0$，两根之积为 $\frac{2u^2}{g^2} > 0$，所以两根皆为正数；而在两根之间，这个二次式为负，因而 $\frac{\mathrm{d}}{\mathrm{d}t}OP^2$ 也为负。质点在 $T = \frac{2u\sin\alpha}{g}$ 时落地。较小的根是

$$t_1 = \frac{u}{2g}(3\sin\alpha - \sqrt{9\sin^2\alpha - 8}) < \frac{u}{2g}\cdot 3\sin\alpha < \frac{2u\sin\alpha}{g} = T.$$

所以紧接 t_1 之后（而且在落地之前），OP 正在递减。 ■

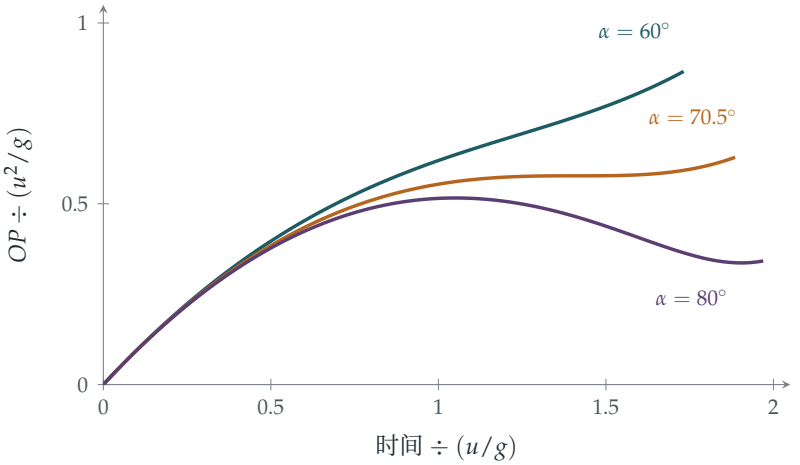


图 11.10.1 三个不同角度下，球与抛球者的距离随时间的变化，直至落地为止。在 60°（蓝绿色）时，距离一直增加。在临界角（critical angle）$\arcsin\frac{2\sqrt{2}}{3} \approx 70.5°$（琥珀色）时，距离有一瞬间停顿。在 80°（紫红色）时，球先远离，然后有一段时间向抛球者靠近，到落地前一刻又再远离。

阶段 3 移动的观察者

想一想 Q 沿地面朝相反方向以速率 v 离开 O。从 Q 量度，水平距离是 $x + vt$。这对计算有甚么影响?

间距 QP 的水平分量为 $(u\cos\alpha + v)t$，垂直分量为 $ut\sin\alpha - \frac{1}{2}gt^2$。所以

$$QP^2 = \big((u\cos\alpha + v)^2 + u^2\sin^2\alpha\big)t^2 - ugt^3\sin\alpha + \tfrac{1}{4}g^2t^4,$$

与之前的式子相同，只是第一项中的 u^2 换成了 $w^2 = (u\cos\alpha + v)^2 + u^2\sin^2\alpha$。其导数（derivative）为 $t\big(g^2t^2 - 3ug\sin\alpha\, t + 2w^2\big)$，而它对所有 $t > 0$ 皆为正的条件是

$$9u^2\sin^2\alpha < 8w^2 \quad\Longleftrightarrow\quad u^2\sin^2\alpha < 8(u\cos\alpha + v)^2,$$

这是在两边消去 $8u^2\sin^2\alpha$ 之后所得的结果。两边取平方根（$u\sin\alpha$ 与 $u\cos\alpha + v$ 皆为正数），即得 $u\sin\alpha < 2\sqrt{2}(u\cos\alpha + v)$，整理后就是

$$2\sqrt{2}\,v > (\sin\alpha - 2\sqrt{2}\cos\alpha)u. \qquad \blacksquare$$

（这个条件足以令 QP 在整个飞行过程中递增，因为此时导数对所有 $t > 0$ 皆为正。）

回顾

- **对平方求导。**距离的平方是多项式，而它的导数与距离本身的导数正负号相同。

- **恒为正的二次式。**“在所有时间都递增”这个条件，化成了一个关于判别式的条件。
- **把新问题化为旧问题。**移动的观察者只改变了一个系数，所以之前的分析可以直接套用。

背后的数学

几何观点

OP 递减的时候，正正就是 P 的速度指向“内侧”的时候，即速度与向量 $\overrightarrow{OP}$ 成钝角（obtuse angle），也就是 $\overrightarrow{OP}\cdot\mathbf{v}<0$ 的时候。对于陡峭的抛掷，质点在接近飞行最高点时，位于抛球者的高处而略为偏前，几乎是水平地向外移动，但已开始下坠；相对于 $\overrightarrow{OP}$ 的方向而言，一旦速度中向下的部分盖过向外的部分，质点便会靠近。对于较平的抛掷，质点总是在前方足够远的位置，以致它的速度从不指回 O。临界情况就是速度恰好只在一个瞬间与 $\overrightarrow{OP}$ 垂直，亦即图 11.10.1 中距离停顿的那一刻。

单调性并非必然

许多“显然”会增加的量，其实并不如此。月球与地球的距离会来回振荡；在弯曲的路上，即使其中一辆车一直领先，两车之间的距离仍可以减少；猎物大量繁殖的时候，捕食者的数量却可以下降。要证明一个量是单调的，通常就是要证明某个导数的正负号固定不变，而这里所用的技巧（化为多项式，再检查它的根或判别式）正是标准的做法。在分析学（analysis）中，单调性往往是证明收敛（convergence）的关键：有界的单调数列必定收敛，正如你在第 2 章所见。

思维技巧

藉导数检查单调性

要判断一个量是否一直递增，可对它（或对它的某个方便处理的函数，例如它的平方）求导，然后问：导数会否变号？

在大学里，这类关于正负号的论证会用来证明不等式、找出极大值和极小值，以及研究微分方程（differential equation）解的稳定性；在后者中，一个一直递减的量（如 Section 8.2 所示）便证明了系统会趋于平稳。

延伸练习

1 **有多近？** 对于 $\alpha = 80°$，求 OP 取得局部极大值（local maximum）和局部极小值（local minimum）的时间，以及该两刻的距离，以 u^2/g 为单位。

2 **在斜面上。** 把球沿一个与水平成 β 角的斜面向上抛出，并落在斜面上，就这个情况重做第 (i) 部分。临界角（从斜面量起）会改变吗？

3 **考虑空气阻力。** 假设球还受到一个与其速度成正比的阻力（resistance）。不作计算，你预期临界角会大于还是小于 70.5°？为甚么？

4 **从高处抛出。** 若球从悬崖顶抛出，并落在远低于崖顶的地方，证明无论以甚么角度抛出，OP 最终都会递增，并判断是否存在某些角度，令它在某个阶段递减。

延伸阅读

- **James Walker, *The Flying Circus of Physics***（Wiley）。数以百计出人意表的日常物理谜题。
- **David Morin, *Introduction to Classical Mechanics***（Cambridge University Press），第 3 章，讲述抛体运动，并附有数道有违直觉的题目。
- **J. M. Steele, *The Cauchy–Schwarz Master Class***（Cambridge University Press）。证明一样东西恒大于另一样东西的艺术。

11.11 外野手如何接球

题目

STEP II 1998，第 11 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/98-s2-q11`

引子

球被高高击向外野。出色的外野手会立刻起步，并且刚好在球落下的一刻赶到。这是怎样做到的？没有人能在不足一秒之内心算出抛体运动（projectile motion）的方程，何况外野手根本不知道球的速率和角度。然而外野手，以至接飞碟的狗，都能轻易做到。

1968 年，物理学家 Seville Chapman 提出了一个答案，而本题正是要证明它。假如外野手留意着球的仰角（angle of elevation）θ，并且移动自己的位置，使 $\tan\theta$ 以恒定的速率增加，她就会刚好到达球的落点。她毋须懂得任何物理，只需能察觉球看起来是升得比稳定上升快，还是比稳定上升慢。一个困难的预测问题，就这样换成了一条简单的反馈（feedback）法则。

动笔之前

想一想 设击球手位于原点，外野手在距离 D 处。若球位于 (x, y)，$\tan\theta$ 是甚么？

球在外野手眼睛上方高度 y 处，并在她前方水平距离 $D - x$ 处（球正朝她飞来）。所以 $\tan\theta = \frac{y}{D-x}$。

逐步解题

阶段 1　位置恰到好处的外野手

想一想　设球的水平速率为 U，初始垂直速率为 W，并在时间 T 落下（回到眼睛的高度）。试以 U、W、T、g 和 t 表示 x、y 和 D。你能把 y 因式分解吗？

我们有 $x = Ut$ 和 $y = Wt - \frac{1}{2}gt^2$，而球在 $T = \frac{2W}{g}$ 时回到眼睛的高度，外野手正在该处：$D = UT$。由于 $W = \frac{1}{2}gT$，

$$y = \tfrac{1}{2}gt(T - t), \qquad D - x = U(T - t).$$

因子 $T - t$ 可以约去：

$$\tan\theta = \frac{\frac{1}{2}gt(T-t)}{U(T-t)} = \frac{g}{2U}\,t.$$

所以 $\frac{\mathrm{d}}{\mathrm{d}t}\tan\theta = \frac{g}{2U}$，是一个常数。 ■

阶段 2　奔跑中的外野手

想一想　现在外野手从距离 D_0 处起步，以速率 v 跑向击球手，所以她的位置是 $D_0 - vt$。她在时间 T 接到球。D_0 是甚么？把刚才的计算重做一次。

要在时间 T 接到球，她必须身处球的落点：$D_0 - vT = UT$，所以 $D_0 = (U + v)T$。她与球的水平距离是 $D_0 - vt - Ut = (U + v)(T - t)$，而该因子同样可以约去：

$$\tan\theta = \frac{\frac{1}{2}gt(T-t)}{(U+v)(T-t)} = \frac{g}{2(U+v)}\,t.$$

所以只要她以适当的速率奔跑，$\tan\theta$ 就会以恒定的速率增加。

想一想　题目问的是逆命题（converse）：假如她奔跑时令 $\frac{\mathrm{d}}{\mathrm{d}t}\tan\theta$ 保持不变，她会接到球吗？

假设她以速率 v 奔跑，而 $\tan\theta = \frac{\frac{1}{2}gt(T-t)}{D_0-(U+v)t}$ 以恒定的速率 c 增加，所以 $\tan\theta = ct$（在 $t = 0$ 时它等于 0）。于是

$$\tfrac{1}{2}g(T - t) = c(D_0 - (U + v)t) \quad \text{对所有} t.$$

比较 t 的系数和常数项：$\frac{1}{2}g = c(U + v)$ 及 $\frac{1}{2}gT = cD_0$。两式相除，得 $D_0 = (U + v)T$：这正是在时间 T 身处落点的条件。所以她应选择 v 使 $\tan\theta$ 保持稳定上升，这样她就能接到球。 ■

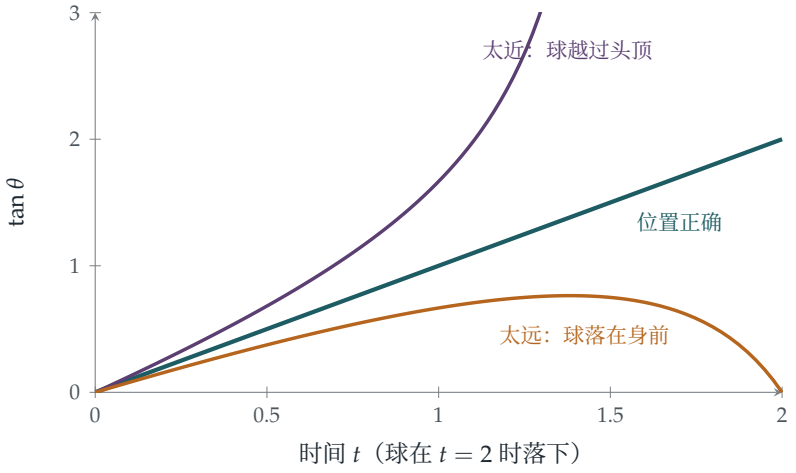


图 11.11.1 站着不动的外野手所见的情况，球的落点在距离 2 处（取 $g = 2$，水平速率为 1）。站在正确的位置时，$\tan\theta$ 沿直线上升（蓝绿色）。站得太近，$\tan\theta$ 会加速上升，球越过她的头顶（紫红色）；站得太后，它升得愈来愈慢，然后下降（琥珀色）。法则是：移动自己的位置，使球的影像稳定上升。

回顾

- **找出公因子。**高度和水平距离都含有 $T - t$，把它约去后，$\tan\theta$ 就变成线性。
- **证明逆命题。**要说明这条法则行得通（即它确实令人接到球），需要反方向的论证，也就是比较系数。
- **容易察觉的不变量（invariant）。**$\tan\theta$ 是否稳定上升，凭肉眼就能判断；落点则不能。

背后的数学

光学加速度抵消

Chapman 的法则现称为*光学加速度抵消*（optical acceleration cancellation）：外野手移动自己的位置，使“光学加速度” $\frac{d^2}{dt^2}\tan\theta$ 保持为零。假如球的影像加速上升，球就会越过她的头顶，所以她要后退；假如影像减速，球就会落在她身前，所以她要向前跑（图 11.11.1）。这个策略没有告诉她球会落在哪里，却能把她带到那里。这是一条反馈控制律（feedback control law），与导引导弹所用的同出一辙：导弹采用*比例导引*（proportional navigation），使指向目标的视线角保持不变，那正是同一个想法的横向版本。

外野手实际怎样做

真实的外野手是否采用 Chapman 的法则？1990 年代有人把摄影机装在棒球外野手的头上做实验，结果并不一致。Michael McBeath 及其同事于 1995 年在 *Science* 上提出，对于打向一侧的球，外野手会沿弯曲的路线奔跑，使球的影像在视野中沿直线移动（即“线性光学轨迹”，linear optical trajectory），这同样能把他们带到接球的位置。其后的研究显示，这两种策略以及两者的混合都有人使用，而且外野手往往并非以恒定速率奔跑。2004 年的一项研究发现，接飞碟的狗所走的路线，与同样的光学法则吻合。无论细节如何，这个令人惊叹的结论依然成立：接球并不需要预测，只需要正确的反馈。

Chapman 的论文

Seville Chapman 的“Catching a baseball”于 1968 年刊登在 *American Journal of Physics*，全文只有三页，内容基本上就是本题的计算。他指出这个策略只在没有空气阻力时才成立，而真实棒球的轨迹会因阻力和旋转而变形，令情况变得复杂。尽管如此，大脑利用视野中简单的不变量来解决复杂的控制问题，这个想法在心理学和机械人学中已变得十分重要。

思维技巧

找出容易观察的不变量

当某个量难以预测时，不妨找一个与它相关、可以直接观察的量，而这个量恰好在一切顺利时表现得简单。

在这里，落点难以预测，但 $\tan\theta$ 的变化率却容易监察。在大学里，这就是反馈控制的想法，也是选择一个可观察量、让它包含你所需信息的想法。

延伸练习

1. **向后跑。**假如球会落在外野手起步位置的后方，证明她应该背离击球手奔跑，而同一条法则依然适用。

2. **不是 θ 本身。**证明对于位置恰到好处的外野手，$\frac{\mathrm{d}\theta}{\mathrm{d}t}$ 并非常数。为甚么 $\tan\theta$ 才是应该留意的量？

3. **眼睛的高度。**假如球是从外野手眼睛高度以下 h 的位置击出，对于位置恰当的外野手，$\tan\theta$ 是否仍然是 t 的线性函数？（假设她在眼睛的高度接球。）

4 **空气阻力。**在线性空气阻力之下，水平速度会按指数衰减。对于位置恰当的外野手，$\tan\theta$ 是否仍然是 t 的线性函数？

延伸阅读

- **Seville Chapman，“Catching a baseball”**，*American Journal of Physics* 36 (1968)。
- **Michael McBeath、Dennis Shaffer 及 Mary Kaiser，“How baseball outfielders determine where to run to catch fly balls”**，*Science* 268 (1995)。
- **Dennis Shaffer 等，“How dogs navigate to catch Frisbees”**，*Psychological Science* 15 (2004)。

11.12 橡皮绳上的蚂蚁

题目

STEP II 2018，第 10 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/18-s2-q10`

引子

一只蚂蚁沿着一条长一公里的橡皮绳出发，每秒走一厘米。每过一秒，绳子都会均匀地再拉长一公里。绳的远端以每秒一千米的速度飞驰而去，比蚂蚁快十万倍。蚂蚁最终能走到绳的尽头吗？

令人惊讶的是，答案是能。绳子拉伸时会带着蚂蚁一起前进；虽然蚂蚁每次走过的路程占绳长的比例愈来愈小，但这些比例加起来就像调和级数（harmonic series），而调和级数是发散（diverge）的（Section 2.3）。蚂蚁终会到达，只是所需的时间远远超过宇宙的年龄。本题精确地解出这个问题的连续版本，并发现所需时间随两个速率之比呈指数增长。

动笔之前

想一想 绳子是均匀拉伸的：每一小段都按相同的倍数伸长。在时间 t，绳长为 $a+ut$。若距离钉子 x 的一点位于全长的 $\frac{x}{a+ut}$ 处，这一点移动得有多快？

绳上每一点所占全长的比例 f 保持不变，所以它与钉子的距离是 $f(a+ut)$，速率是 $fu=\frac{ux}{a+ut}$。靠近钉子的点几乎不动；远端则以 u 的速率移动。

逐步解题

阶段 1 蚂蚁的方程

蚂蚁相对于绳子的速率是 v，还要加上牠脚下那段绳子的速率：

$$\frac{\mathrm{d}x}{\mathrm{d}t} = v + \frac{ux}{a+ut}.$$

> **想一想** 用商法则（quotient rule）对 $\frac{x}{a+ut}$ 求导，并代入 $\dot{x}$ 的方程。

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{x}{a+ut}\right) = \frac{\dot{x}}{a+ut} - \frac{ux}{(a+ut)^2} = \frac{1}{a+ut}\left(v + \frac{ux}{a+ut}\right) - \frac{ux}{(a+ut)^2} = \frac{v}{a+ut}. \qquad \blacksquare$$

蚂蚁已走过的绳长比例以 $\frac{v}{a+ut}$ 的速率增加：即牠自身的速率除以当时的绳长。拉伸的影响完全消失了，因为拉伸不会改变任何一点所占的比例。

阶段 2　到达尽头

> **想一想** 由 0 积分至 t。比例何时达到 1?

蚂蚁从钉子出发，所以

$$\frac{x}{a+ut} = \int_0^t \frac{v\,\mathrm{d}s}{a+us} = \frac{v}{u}\ln\frac{a+ut}{a}.$$

当比例等于 1 时，蚂蚁到达尽头：$\ln\frac{a+uT}{a} = \frac{u}{v} = k$，所以 $a+uT = ae^k$，即

$$uT = a(e^k - 1). \qquad \blacksquare$$

对数可以无限增大，所以无论绳子拉得多快，蚂蚁总会到达。

阶段 3　回程

> **想一想** 往回走时，蚂蚁相对于绳子的速度是 $-v$。现在牠所占比例的变化率是多少？牠在时间 T 由比例 1 出发。

以 $-v$ 重复同样的计算，得 $\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{x}{a+ut}\right) = -\frac{v}{a+ut}$，所以

$$\frac{x}{a+ut} = 1 - \frac{v}{u}\ln\frac{a+ut}{a+uT}.$$

当 $a+ut = (a+uT)e^k$ 时，比例降至 0。由于 $a+uT = ae^k$，回程所需时间为

$$t - T = \frac{(a+uT)(e^k-1)}{u} = \frac{ae^k \cdot uT/a}{u} = Te^k. \qquad \blacksquare$$

回程所需的时间是去程的 e^k 倍：蚂蚁掉头的时候，绳子已是开始时的 e^k 倍长。

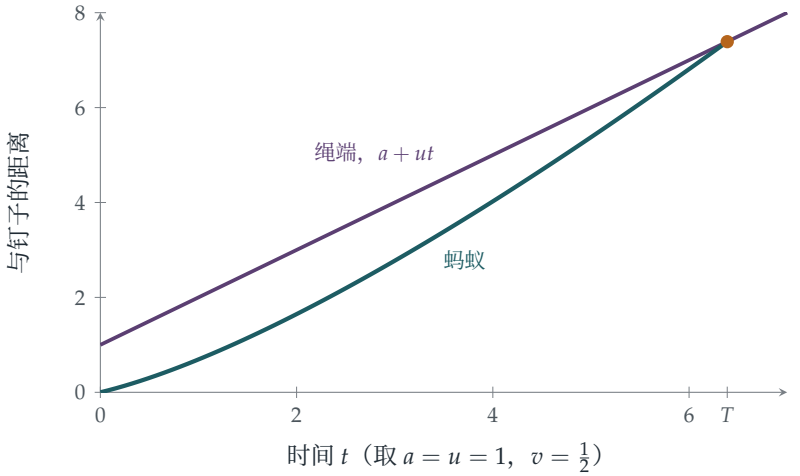


图 11.12.1 蚂蚁（蓝绿色）沿着绳子行走，绳端（紫红色）移动的速率是蚂蚁步行速率的两倍（$k=2$）。蚂蚁被拉伸带着前进，牠与钉子的距离增长得愈来愈快。牠在 $T=e^2-1\approx 6.39$ 时到达尽头。

回顾

- **选对变量。**已走过的绳长比例是最自然的量：拉伸不会改变它。
- **乔装了的积分因子（integrating factor）。**乘以 $\frac{1}{a+ut}$ 之后，一条线性微分方程（linear differential equation）就变成了一个恰好的导数。
- **发散胜过速度。**比例像对数一样增长，而对数是无界的，所以蚂蚁总会到达。

背后的数学

调和级数的版本

这个谜题通常是以离散的拉伸来叙述的。一条绳长 1 公里；蚂蚁在第一秒走 1 厘米；然后绳子拉长 1 公里；然后蚂蚁再走一厘米，如此类推。在第 n 秒，蚂蚁在长 n 公里的绳上走了 1 厘米：所占比例为 $\frac{1}{100\,000\,n}$。拉伸不会改变牠所占的比例，所以 N 秒之后，牠已走过绳子的

$$\frac{1}{100\,000}\left(1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{N}\right)\approx\frac{\ln N+0.577}{100\,000}$$

。当 $H_N\approx 100\,000$ 时牠便到达，那大约是 $e^{100\,000}$ 秒之后，这个数有超过 43 000 位数字。（作为比较，宇宙的年龄约为 4×10^{17} 秒。）我们的连续版答案与此吻合：$k=\frac{u}{v}=100\,000$，而 $T\approx\frac{a}{u}e^k$。

膨胀宇宙中的光

绳上的蚂蚁，正是光穿越膨胀宇宙的一个模型。遥远的星系被空间本身的膨胀带着远离我们，而它们发出的光则以速率 c 在膨胀的空间之中向我们传播，就像蚂蚁沿着拉伸中的绳子行走一样。遥远星系的光最终能否到达我们这里，取决于膨胀率如何随时间变化。在一个以稳定速率膨胀的宇宙中，就像那条绳子一样，光最终总会到达。而在一个加速膨胀的宇宙中（我们的宇宙看来正是如此），有些星系的光永远不会到达我们这里：这就是*宇宙事件视界*（cosmic event horizon）。对蚂蚁来说，相应的情况是一条按指数拉伸的绳子，绳端的速率与绳长成正比地增长（可在练习中试试）。

加德纳的蚂蚁

这个谜题似乎是法国数学家 Denys Wilquin 在 1972 年发明的，当时说的是橡皮筋上的一条虫；其后马丁·加德纳（Martin Gardner）在他的 *Scientific American* 专栏和著作 *Time Travel and Other Mathematical Bewilderments* 中介绍，令它广为人知。它是最著名的例子之一：结果违反直觉，却又容易证明；它也是人们说明调和级数发散时最爱用的例子。

思维技巧

找出不受复杂因素影响的量

当问题中某个部分改变了一切（这里是拉伸），就去找一个它不会改变的量。

拉伸会改变每一段距离，却不改变任何比例。改用比例来处理，拉伸就从问题中消失了。在大学里，这个想法会以宇宙学和流体力学中的*共动坐标*（comoving coordinates）的形式出现，更一般地说，就是寻找在某个变换下的不变量（invariant）。

延伸练习

1 **中途。**蚂蚁何时走到绳子的一半？那是在时间 $\frac{T}{2}$ 之前还是之后？

2 **指数拉伸。**假设绳长改为 $ae^{\lambda t}$。证明蚂蚁仅当 $v > \lambda a$ 时才能到达尽头。（这就是宇宙事件视界。）

3 **加速的绳子。**若绳长为 $a + \frac{1}{2}bt^2$，蚂蚁是否总能到达？

4 **离散的谜题。**在离散版本中，蚂蚁需要多少秒才能走过绳子的 1%？请使用 $H_N \approx \ln N + 0.577$。

延伸阅读

- **Martin Gardner, *Time Travel and Other Mathematical Bewilderments***（W. H. Freeman）。书中收录了橡皮绳上的蚂蚁。
- **Julian Havil, *Gamma: Exploring Euler's Constant***（Princeton University Press）。调和级数，以及这只蚂蚁。
- **Barbara Ryden, *Introduction to Cosmology***（Cambridge University Press）。膨胀的空间、视界与共动坐标，属大学本科程度。

11.13 铲雪车问题

题目

STEP III Specimen paper，第 11 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/spec-s3-q11`

引子

雪不停地均匀落下。一辆铲雪车（snowplough）沿着一条路出发，前方的积雪愈来愈深，车速也愈来愈慢。稍后，第二辆铲雪车沿着已经清理过的路面跟上。它遇到的只是第一辆车驶过之后才落下的雪，积雪较浅，所以走得较快，最终会追上前车。那是甚么时候？

这是一道著名问题的变体，原题由 Ralph Agnew 于 1942 年发表，后来成为微分方程（differential equations）中最广为人知的趣题之一。它的妙处在于，第二辆车的速率取决于*第一辆车何时驶过每一点*，所以第 (i) 部分的解会代入第 (ii) 部分的方程之中。而这条方程“倒过来”写，把时间看成距离的函数，才最为巧妙。

动笔之前

想一想 第一辆铲雪车在时间 t 遇到的积雪深度为 $h + kt$。它的速率是多少？求 x 作为 t 的函数较容易，还是求 t 作为 x 的函数较容易？

它的速率是 $\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{k}{\alpha(h+kt)}$。这条方程可以干净利落地分离变量（separation of variables）。把它倒过来，$\frac{\mathrm{d}t}{\mathrm{d}x} = \frac{\alpha(h+kt)}{k}$，这是一条以 t 为 x 的函数的简单线性（linear）方程。对第一辆车来说，两种写法都行得通；对第二辆车来说，倒过来的形式却必不可少。

逐步解题

阶段 1 第一辆铲雪车

由 $\frac{\mathrm{d}t}{\mathrm{d}x} = \alpha\left(\frac{h}{k} + t\right)$ 可知，$\frac{h}{k} + t$ 这个量随 x 按指数增长（exponential growth）：$\frac{h}{k} + t = Ce^{\alpha x}$。当 $x = 0$ 时 $t = 0$，所以 $C = \frac{h}{k}$，于是

$$t_1(x) = \frac{h}{k}\left(e^{\alpha x} - 1\right). \qquad \blacksquare$$

第一辆车按指数规律减慢：走过相等的距离，所需的时间按指数递增。

阶段 2　第二辆铲雪车的方程

想一想　第二辆铲雪车在时间 t 到达位置 y。那里的积雪有多深？（第一辆车刚驶过之后，深度是多少？）

第一辆车在时间 $t_1(y)$ 清除了位置 y 上的所有积雪。自此之后，雪以速率 k 累积，所以在时间 t 的深度为 $z = k\big(t - t_1(y)\big)$。因此第二辆车的速率是

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \frac{k}{\alpha k\big(t - t_1(y)\big)} = \frac{1}{\alpha\big(t - t_1(y)\big)},$$

倒过来便得

$$\frac{1}{\alpha}\frac{\mathrm{d}t}{\mathrm{d}y} = t - \frac{h}{k}\left(e^{\alpha y} - 1\right). \qquad \blacksquare$$

阶段 3　求解

想一想　这条方程对 t 是线性的：$\frac{\mathrm{d}t}{\mathrm{d}y} - \alpha t = -\frac{\alpha h}{k}\left(e^{\alpha y} - 1\right)$。应该用哪个积分因子（integrating factor）？初始条件是甚么？

乘以 $e^{-\alpha y}$，

$$\frac{\mathrm{d}}{\mathrm{d}y}\left(te^{-\alpha y}\right) = -\frac{\alpha h}{k}\left(1 - e^{-\alpha y}\right), \qquad te^{-\alpha y} = -\frac{\alpha h}{k}y - \frac{h}{k}e^{-\alpha y} + C.$$

第二辆车在 $t = T$ 时由 $y = 0$ 出发，所以 $T = -\frac{h}{k} + C$，得 $C = T + \frac{h}{k}$。再乘回 $e^{\alpha y}$ 并整理，

$$t_2(y) = \frac{h}{k}\left(e^{\alpha y} - 1\right) + \left(T - \frac{\alpha h y}{k}\right)e^{\alpha y}. \qquad \blacksquare$$

（题目要求验证，方法是：求导后代入；各项合并后正好是 $\alpha\big(t - t_1(y)\big)$。）

阶段 4　相撞

想一想　两辆铲雪车在同一时间到达同一点时便会相撞。试比较 $t_2(y)$ 与 $t_1(y)$。

两者之差为 $t_2(y) - t_1(y) = \left(T - \frac{\alpha h y}{k}\right)e^{\alpha y}$，即两辆车在位置 y 的时间差距。这个差距缩减至零时，

$$y = \frac{kT}{\alpha h}. \qquad \blacksquare$$

在这一点，第二辆车追上第一辆车：它前方的积雪深度为零，而在这个模型中，它的速率变成无限大（图 11.13.1）。

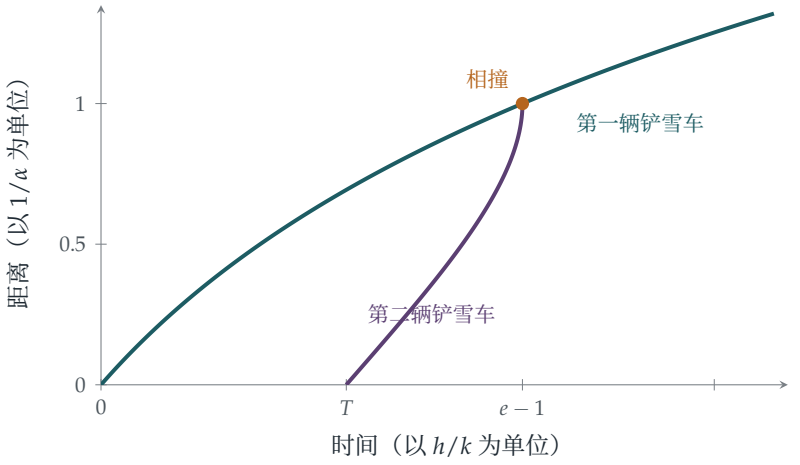


图 11.13.1 两辆铲雪车的距离对时间图，其中 $T = \frac{h}{k}$（即第二辆车在积雪深度倍增之时出发）。第一辆车（蓝绿色）随积雪加深而减慢。第二辆车（紫红色）较迟出发，路面已清理了一部分，它一边拉近距离一边加速，在距离 $\frac{kT}{\alpha h} = \frac{1}{\alpha}$、时间 $(e-1)\frac{h}{k}$ 处追上第一辆车。

回顾

- **把方程倒过来。**把 t 看成位置的函数，两条方程都是线性的。
- **前一个解代入后一条方程。**第一辆车的时间表 $t_1(y)$ 出现在第二辆车的方程之中。
- **作比较，不必解两次。**相撞的条件来自差 $t_2 - t_1$，而这个差有一个简单的因子。

背后的数学

Agnew 的铲雪车

Ralph Palmer Agnew 的版本出自他 1942 年的教科书 *Differential Equations*：某天早上开始下起大雪，雪势均匀不变；一辆铲雪车在正午出发，第一个小时走了 2 英里，第二个小时走了 1 英里；雪是甚么时候开始下的？铲雪车的速率与积雪深度成反比，而深度则与开始下雪以来所经过的时间成正比。答案（见下面的一道练习）涉及黄金比（golden ratio）：雪大约在正午前 37 分钟开始下。其后才陆续出现像本题这样有两辆铲雪车的版本。

会传递的信息

第二辆车的运动取决于路面上记录下来的历史：每一点最后一次被清理的时间。这是具有记忆（memory）的系统的一个简单例子，在这类系统中，某处的状态取决于该处先前发生过的事。交通流、冰川的运动和山火的蔓延都有这种特性，而它们背后的数学，即一阶偏微分方程（first-order partial differential equations）及其特征线（characteristics）的理论，是大学二年级的标准课题。把时间写成位置的函数（“前沿何时到达这里?”），正是处理这类问题时所用的到达时间（arrival times）方法。

思维技巧

选择自变量（independent variable）

把变量的角色对调，微分方程可能会简单得多。如果变化率取决于某处发生某事的时间，就把时间写成位置的函数。

本题的两条方程写成 $t(y)$ 的方程后都变成线性。在大学里，自由选择哪个变量作自变量（并在方便时更换）是经常用到的手法，由流体力学中的速度图法（hodograph method），到几何中以弧长（arc length）作参数，比比皆是。

延伸练习

1 **Agnew 的问题。**雪在正午前某个未知的时间开始以稳定的速率落下。一辆铲雪车的速率与积雪深度成反比，它在正午出发，第一个小时走了 2 英里，第二个小时走了 1 英里。证明若雪在正午前 t_0 小时开始下，则 $\left(1+\frac{2}{t_0}\right)^2=\left(1+\frac{1}{t_0}\right)^3$，由此推出 $\frac{1}{t_0}$ 是黄金比，并求雪何时开始下。

2 **验证。**用求导验证 $t_2(y)$ 满足第 (ii) 部分的微分方程。

3 **时间差距。**证明若 $kT>h$，时间差距 $t_2(y)-t_1(y)$ 先增加，然后减少至零，并求它在何处最大。若 $kT\le h$，又会怎样？试解释其意义。

4 **三辆铲雪车。**第三辆铲雪车在时间 $2T$ 出发。写出它的运动方程。它追上第二辆车，是在第二辆车追上第一辆车之前还是之后?

延伸阅读

► **Ralph Palmer Agnew, *Differential Equations***（McGraw-Hill, 1942）。铲雪车问题的原始

出处。

- **George Simmons, *Differential Equations with Applications and Historical Notes***（CRC Press），书中讨论了铲雪车问题和许多其他问题。
- **Richard Haberman, *Mathematical Models***（SIAM）。交通流与特征线，程度浅易。

11.14 是抛物线，不是悬链线

题目

STEP I 1991，第 10 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/91-s1-q10`

引子

用手拿着一条链的两端，它便会垂成一条称为*悬链线*（catenary）的曲线（此名源自拉丁文的“链”字）。伽利略（Galileo）以为它是抛物线（parabola），其实不然，少年时代的惠更斯（Huygens）在 1646 年已证明了这一点。不过，若像悬索桥（suspension bridge）那样，用等距排列的吊杆把沉重的路面挂在链上，链*确实*会形成抛物线。分别在于重量如何分布：悬链线的重量沿弯曲的链身分布，而桥的重量则沿水平方向均匀分布。

本题只需在每个节点分解力（resolving forces），便能证明这条抛物线。关键在于张力（tension）的水平分量在整条链上处处相同，所以每一段的斜率都与它所承托的重量成正比。结果是一个优美的规律：相邻节点之间的落差按奇数 $1, 3, 5, 7, \ldots$ 递增，而这些奇数的部分和（partial sums）正是完全平方数。

动笔之前

想一想 在每个节点 A_i 上有三个力作用：相邻两段链的张力，以及吊杆上所挂的重量 w。重量无法影响张力的哪一个分量？

重量是铅垂的，所以在每个节点上，两个张力的水平分量必须互相平衡。因此，张力的水平分量（设为 H）在链的每一段都相同。

逐步解题

阶段 1　每个节点上的方程

设线段 $A_{i-1}A_i$ 与水平线所成的俯角为 ϕ_i（由 A_{i-1} 向中间方向量度），则其张力 T_i 的水平分量为 $T_i\cos\phi_i$，铅垂分量为 $T_i\sin\phi_i$。对 $1\le i\le n$，在 A_i 分解力：

$$T_i\cos\phi_i = T_{i+1}\cos\phi_{i+1} = H,\qquad T_i\sin\phi_i - T_{i+1}\sin\phi_{i+1} = w.$$

（外侧的一段向上向外拉，铅垂分量较大；内侧的一段向内并稍微向下拉；两者之差承托着吊杆。）

> **想一想**　在中间的节点 A_{n+1} 会怎样？试利用桥的对称性。

由对称性，在 A_{n+1} 相接的两段互为镜像，所以各自承托该吊杆的一半重量：$T_{n+1}\sin\phi_{n+1} = \tfrac{1}{2}w$。

阶段 2　由重量求斜率

由中间向外逐段计算，铅垂分量为

$$T_{n+1}\sin\phi_{n+1}=\tfrac{1}{2}w,\quad T_n\sin\phi_n=\tfrac{3}{2}w,\quad \dots,\quad T_1\sin\phi_1=\tfrac{2n+1}{2}w.$$

把各式除以 H，得斜率为 $\tan\phi_i = \frac{(2(n+1-i)+1)w}{2H}$。线段 $A_{i-1}A_i$ 的斜率同时等于 $\frac{y_{i-1}-y_i}{h}$，所以

$$\frac{y_n-y_{n+1}}{h}=\frac{w}{2H},\qquad \frac{y_{n-1}-y_n}{h}=\frac{3w}{2H},\qquad \dots,\qquad \frac{y_0-y_1}{h}=\frac{(2n+1)w}{2H}.$$

取倒数后再乘以相应的奇数，可知 $\frac{h}{y_n-y_{n+1}}$、$\frac{3h}{y_{n-1}-y_n}$、…、$\frac{(2n+1)h}{y_0-y_1}$ 各项均等于 $\frac{2H}{w}$。 ■

阶段 3　抛物线

> **想一想**　设 j 为由中间向外数的节点序数，使得 A_{n+1-j} 与 A_{n+1} 的水平距离为 jh。它比 A_{n+1} 高出多少？

把各段落差相加，并设 $c=\frac{wh}{2H}$：

$$y_{n+1-j}-y_{n+1} = c(1+3+5+\cdots+(2j-1)) = cj^2.$$

若 x 由中间量起，则 $x=jh$，所以该点高出最低节点 $\frac{c}{h^2}x^2=\frac{w}{2Hh}x^2$。由对称性，另一边亦然，而两个支承点（$j=n+1$）也符合同一公式。因此所有点都位于抛物线

$$y = y_{n+1}+\frac{w}{2Hh}(x-x_{n+1})^2.$$ ■

上。

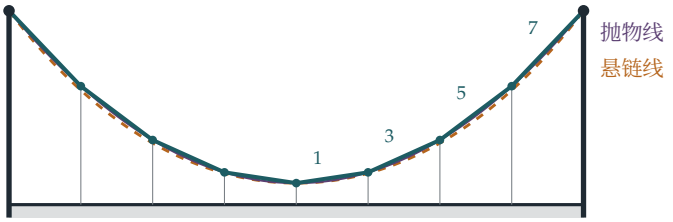


图 11.14.1 一条在相等水平间距上承受相等负载的链（蓝绿色）。它的节点位于一条抛物线上（紫红色），而由中间向外，相邻节点之间的落差之比为 $1:3:5:7$。一条只受自身重量、两端和最低点相同的自由链，则会垂成悬链线（琥珀色虚线），其底部较平坦，近支承点处较陡。

回顾

- **找出不变的量。**只有铅垂负载时，水平张力处处相同。
- **由中间开始累加。**每一段都承托它与中间之间的全部重量，所以铅垂分量每次增加 w。
- **奇数之和是平方数。**$1+3+\cdots+(2j-1)=j^2$ 把一列斜率变成了一条抛物线。

背后的数学

抛物线还是悬链线？

把上述论证改为连续的版本，便能同时解释这两条曲线。对于一条水平张力恒为 H 的链 $y(x)$，张力的铅垂分量为 Hy'，而在一小段上，它的变化等于该段所承受的负载：$H\,\mathrm{d}y' = \mathrm{d}W$。

- 若负载沿*水平方向*均匀分布（沉重的路面，每单位路面长度的重量为 ρ），则 $Hy''=\rho$，所以 y 是二次函数：一条抛物线。
- 若负载是链本身的重量，*沿链身*均匀分布（每单位链长的重量为 ρ），则 $Hy''=\rho\sqrt{1+y'^2}$，其解为悬链线 $y=c\cosh\frac{x}{c}$，其中 $c=H/\rho$。

在底部附近，链几乎是水平的，两者差不多一样；它们在支承点附近才有分别，因为那里的链较陡，同一段水平距离内包含较多的链。真实的悬索桥，桥面远比缆索重，所以形状十分接近抛物线。

倒转的链与拱

链处于纯张力状态；把它的形状上下倒转，便得到一个处于纯压力（compression）状态的拱，完全没有弯曲的倾向。胡克（Robert Hooke）在 1675 年以一个字谜（anagram）的形式发表了这个发现，解开后的意思是：“柔软的绳怎样悬垂，倒转过来，坚硬的拱便怎样屹立。”雷恩（Christopher Wren）与胡克把这个想法用于圣保罗大教堂（St Paul's Cathedral）的圆顶，而高迪（Antoni Gaudí）设计巴塞隆拿的圣家堂（Sagrada Família）时，则从天花板垂下挂着重物的绳子，再把照片倒转来看。圣路易斯的拱门（Gateway Arch）就是一条（加权的）倒置悬链线。

悬链线的挑战

1690 年，雅各布·伯努利（Jakob Bernoulli）向数学家提出挑战：求出悬垂的链的形状。不出一年，他的弟弟约翰（Johann）、莱布尼兹（Leibniz）和惠更斯都发表了解答。约翰·伯努利后来得意地忆述，出题的哥哥自己反而解不出来。悬链线是新兴的微积分（calculus）最早的胜利之一，并引出了双曲函数（hyperbolic functions）cosh 和 sinh；到了十八世纪，Vincenzo Riccati 和兰伯特（Johann Heinrich Lambert）对这些函数作了有系统的研究。

思维技巧

在每个节点分解力

对于由节点和直杆件组成的结构，在每个节点上令各力平衡。不变的分量和逐步累积的负载，往往会揭示结构的形状。

到了大学，这会发展成工程学中桁架（trusses）和框架（frameworks）的分析，而在连续极限下，则成为缆索、梁和薄膜的微分方程（differential equations）。由离散的链过渡到连续的链，与 Section 10.3 中由求和过渡到积分是同一步。

延伸练习

1 **偶数条吊杆。**把题目改为 $2n$ 条吊杆再做一次，此时链的最低处是中间的一段水平线段。各节点是否仍然位于一条抛物线上？

2 **张力。**以 H、w 和 n 表示链中的最大张力。它出现在哪里？

3 **悬链线。**验证 $y = c\cosh\frac{x}{c}$ 满足 $y'' = \frac{1}{c}\sqrt{1+y'^2}$。证明悬链线由 $x = 0$ 到 $x = X$ 的长度为 $c\sinh\frac{X}{c}$。

4 **伽利略的错误。**一条长度为 $2L$ 的链悬挂在相距 $2d$ 的两个支承点之间。当 L 只比 d 稍大时，证明悬链线的垂度（sag）与抛物线的垂度近似相同。为甚么伽利略的实验没有显示出两者的分别？

延伸阅读

- **E. H. Lockwood, *A Book of Curves***（Cambridge University Press）。悬链线、抛物线及其相关曲线。
- **Jacques Heyman, *The Stone Skeleton***（Cambridge University Press）。由胡克的链说起，讲述砖石拱和圆顶如何屹立不倒。
- **David Billington, *The Tower and the Bridge***（Princeton University Press）。桥梁的艺术与工程。

我们学到了甚么

这一章由会摆动的物体、会旋转的物体，以及从运动中的位置所看到的景象说起。我们提出了三个问题：运动在甚么时候才算简单？怎样才是合适的观点？直觉是否可靠？以下是本章十四道题目所揭示的答案。

1. 运动在甚么时候才算简单？

简谐运动（simple harmonic motion）只有单一频率（frequency），而且频率与振幅（amplitude）无关，是最理想的情况。以固有频率（natural frequency）驱动它，便会产生共振（resonance）（Section 11.1）：稍为偏离共振时出现拍（beats），恰好处于共振时则无止境地增长。两个耦合振子（coupled oscillators）原来只是两个各自独立的简单振子乔装而成（Section 11.2），但两者频率之比是无理数 $\sqrt{3}$，所以合成的运动永不重复。真实的摆一点也不简单（Section 11.3）：它的周期（period）随振幅增大，当中涉及一个椭圆积分（elliptic integral），而高斯（Gauss）的算术几何平均（arithmetic–geometric mean）能以惊人的速度把它算出来。至于等时曲线（tautochrone）（Section 11.4），则说明了只要把轨道弯成摆线（cycloid），运动便可以再次变得完全简单。

2. 怎样才是合适的观点？

几乎每一道题目，都是靠选对变量或选对参考系（frame）而解决的。

- **坐标。**弧长（arc length）令等时曲线变得简单；组合 $x \pm y$ 把两个弹簧解耦（decouple）；改用占绳长的比例，便消去了蚂蚁问题中绳子的伸长（Section 11.12）；把时间写成距离的函数，令铲雪车的方程变成线性（Section 11.13）。
- **参考系。**在加速的火车中（Section 11.5），一个虚拟力（fictitious force）使重力倾斜，珠子于是围绕新的最低点振动。爱因斯坦（Einstein）把这个观察发展成广义相对论（general relativity）。
- **取矩的参考点。**对于旋转的弹珠（Section 11.6），由于摩擦力正正作用于接触点，所以对接触点的角动量（angular momentum）守恒，毋须知道摩擦力的大小也可求得最终速率。
- **等效系统。**一根杆的表现与三个质点完全相同（Section 11.7），各质点的权重正是辛普森法则（Simpson's rule）中的权重；而一个立体的平衡位置，就是其法线（normal）

穿过质心（centre of mass）之处（Section 11.8）。

3. 直觉是否可靠?

往往并不可靠。大炮射程所及区域的边界并不是一条弹道，而是所有弹道的包络线（envelope）（Section 11.9）。把球抛得够陡，它在落地之前会朝你飞回来（Section 11.10）。外野手毋须预测球的落点，只要令球的影像保持稳定上升，便可以把球接住（Section 11.11）。一只蚂蚁沿着绳子慢慢爬行，即使绳子伸长的速度比牠快十万倍，牠仍然能到达尽头（Section 11.12）。悬垂的链并不是抛物线，但承托着桥面的链却是（Section 11.14）。在每一个例子中，一段简短的计算都以精确的陈述取代了自信却错误的直觉，而且通常还附带一个解释。

4. 处理振动与参考系的工具箱

方法	在哪里用过
特解加余函数（particular plus complementary solution）	共振（11.1）
以特征向量（eigenvector）解耦	简正模式（normal modes）（11.2）
能量积分与代换	摆（11.3）
自然变量	等时曲线（11.4）；蚂蚁（11.12）
虚拟力	火车（11.5）
对某一点的角动量	回旋（11.6）
配对各阶矩（matching moments）	辛普森杆（11.7）
数算驻点（stationary point）	倾斜的平衡（11.8）
判别式（discriminant）与包络线	安全抛物线（11.9）
导数的正负号	飞回来（11.10）
比较系数	外野手（11.11）
积分因子（integrating factor）	蚂蚁（11.12）；铲雪车（11.13）
在接点处分解力	桥（11.14）

下一步

- 接下来的两章转向概率与统计，由 Section 12.1 开始。
- 简正模式就是特征向量（Section 7.2），而 Section 8.6 的量子谐振子（quantum harmonic oscillator），就是本章的振子在量子力学中的版本。
- 极限环（limit cycle）（Section 8.4）是能够自我维持的振动，而混沌（chaos）（Section 8.7）则是非线性振子受驱动时可能出现的现象。

如果你想在入读大学之前再进一步

- **A. P. French, *Vibrations and Waves***：讲解振动、共振与简正模式的经典入门书。
- **David Morin, *Introduction to Classical Mechanics***：转动、参考系，以及许多出人意表的问题。
- **Steven Strogatz, *Sync***：由萤火虫到桥梁，振子如何变得步伐一致。
- **George Simmons, *Differential Equations with Applications and Historical Notes***：等时曲线、最速降线（brachistochrone）与铲雪车问题，连同它们的历史。

Coleman 曾打趣说，物理学家的一生，就是以愈来愈抽象的方式处理谐振子。本章尝试说明个中原因：振子是最简单的运动，而只要选对观点，几乎任何事物看起来都像一个振子。学会寻找这个观点，不论是一个变量、一个参考系，还是一个组合，正是力学所教的最有用的技巧。

延伸练习答案

以下只是简短的答案和提示，并非完整解答：只列出结果和关键步骤，让你核对自己的做法。请先认真做完每道练习，然后才翻看。

11.1 共振

1. 在升降机的参考系（frame）中，重力实际上相当于 $g-f$，而弦线的劲度（stiffness）是 $\frac{mg}{a}$（在 mg 的作用下它伸长 a）。因此 $\ddot{y}+\frac{g}{a}y=g-f$，开始时 $y=a$、$\dot{y}=0$。新的平衡位置（equilibrium）是 $a-\frac{af}{g}$，所以 $y=a-\frac{af}{g}(1-\cos\omega t)$：振幅（amplitude）为 $\frac{af}{g}$，而最小伸长量 $a(1-\frac{2f}{g})$ 是正数。升降机停止加速时，平衡位置回到 a，此时质点的位移是 $-\frac{af}{g}(1-\cos\omega T)$，速度是 $-\frac{af}{g}\omega\sin\omega T$。新的振幅是 $\frac{af}{g}\sqrt{(1-\cos\omega T)^2+\sin^2\omega T}=\frac{2af}{g}\left|\sin\frac{1}{2}\omega T\right|$。

2. 由 $x=Ct\cos nt$ 得 $\ddot{x}+n^2x=-2Cn\sin nt$，所以 $C=-\frac{nb}{2}$。再加上 $\frac{b}{2}\sin nt$ 以符合初始条件：$x=\frac{b}{2}(\sin nt-nt\cos nt)$。对题目中的公式就 p 使用洛必达法则（L'Hôpital's rule），当 $p\to n$ 时会得到相同的结果。

3. $n\sin pt-p\sin nt=n(\sin pt-\sin nt)+(n-p)\sin nt=2n\cos\frac{(n+p)t}{2}\sin\frac{(p-n)t}{2}+(n-p)\sin nt$。第二项很小。第一项是一个快速振荡，被包在缓慢变化的包络（envelope）$\left|\sin\frac{(n-p)t}{2}\right|$ 之内，而这个包络每隔 $\frac{2\pi}{n-p}$ 重复一次。

4. 振幅是 $\frac{F}{\sqrt{(n^2-p^2)^2+4k^2p^2}}$。把分母对 p^2 求极小：$-2(n^2-p^2)+4k^2=0$，所以 $p^2=n^2-2k^2$。

11.2 简正模式

1. 求导后，能量的变化率是 $\dot{x}\left(m\ddot{x}+\frac{\lambda}{a}(2x+y)\right)+\dot{y}\left(m\ddot{y}+\frac{\lambda}{a}(x+2y)\right)=0$。设 $u=x-y$ 及 $v=x+y$，能量便是 $\left(\frac{m}{4}\dot{u}^2+\frac{\lambda}{4a}u^2\right)+\left(\frac{m}{4}\dot{v}^2+\frac{3\lambda}{4a}v^2\right)$，每个括号都是一个简谐振子（simple harmonic oscillator）的能量，所以各自都是常数。

2. 周期 $\frac{2\pi}{\omega}$：以 $x=-y$ 及 $\dot{x}=-\dot{y}$ 开始（使得 $v\equiv0$）。周期 $\frac{2\pi}{\sqrt{3}\omega}$：以 $x=y$ 及 $\dot{x}=\dot{y}$ 开始。

3. 作用于 X 的力是 $-\frac{\lambda}{a}x-\frac{\mu}{a}(x+y)$。模式（mode）$x-y$ 仍然有 $\omega^2=\frac{\lambda}{ma}$，而 $x+y$ 现在则是 $\frac{\lambda+2\mu}{ma}$。当且仅当频率之比 $\sqrt{1+2\mu/\lambda}$ 是有理数时，所有运动都是周期性

的。

4. 矩阵是 $\frac{k}{m}\begin{pmatrix}2&-1&0\\-1&2&-1\\0&-1&2\end{pmatrix}$，特征向量（eigenvector）为 $(1,\sqrt{2},1)$、$(1,0,-1)$ 和 $(1,-\sqrt{2},1)$，分别对应 $2-\sqrt{2}$、2 和 $2+\sqrt{2}$。三个模式是：三者一同摆动，中间的摆幅最大；外侧两个反向摆动，中间的静止不动；以及中间的与外侧两个反向摆动。

11.3 真实的摆

1. 能量：$\frac{1}{2}m\dot{x}^2+\frac{\lambda}{2a}\left(\sqrt{x^2+a^2}-a\right)^2=\frac{1}{2}mv^2$。位移最大时 $\sqrt{x^2+a^2}=a+\frac{v}{k}$，所以 $x_{\max}=\sqrt{\frac{2av}{k}+\frac{v^2}{k^2}}$。当 v 很小时，伸长量约为 $\frac{x^2}{2a}$，所以 $\dot{x}^2\approx v^2-\frac{k^2x^4}{4a^2}$。设 $x=Xu$ 及 $X=\sqrt{2av/k}$，周期 $4\int_0^X\frac{\mathrm{d}x}{\dot{x}}$ 便化为 $\frac{4X}{v}\int_0^1\frac{\mathrm{d}u}{\sqrt{1-u^4}}$。回复力（restoring force）与 x^3 成正比，而不是与 x 成正比（在 $x=0$ 时弦线刚好处于自然长度），所以不存在简谐极限。

2. $(1-m\sin^2\phi)^{-1/2}\approx 1+\frac{m}{2}\sin^2\phi+\frac{3m^2}{8}\sin^4\phi$，其中 $m=\sin^2\frac{\gamma}{2}\approx\frac{\gamma^2}{4}-\frac{\gamma^4}{48}$。对 ϕ 取平均得 $1+\frac{m}{4}+\frac{9m^2}{64}$，而 γ^4 项是 $-\frac{1}{192}+\frac{9}{1024}=\frac{11}{3072}$。

3. 由 1 和 $\cos 45^\circ=0.70711$ 开始：$(0.85355,0.84090)$，然后 $(0.84722,0.84720)$，然后 0.847213。周期因子是 $\frac{1}{0.847213}=1.18034$。

4. 周期因子的变化是 $\frac{1}{16}(\gamma_5^2-\gamma_4^2)\approx 1.71\times10^{-4}$（以弧度计），所以时钟每天大约慢 15 秒。

11.4 等时曲线

1. 由于 $y=\frac{s^2}{4k}$，$\ddot{s}=-g\frac{\mathrm{d}y}{\mathrm{d}s}=-\frac{g}{2k}s$：这是 $\omega^2=\frac{g}{2k}$ 的简谐运动（simple harmonic motion）。由 $s_0=2\sqrt{kh}$ 出发，在 V 处的速率是 $\omega s_0=\sqrt{2gh}$。

2. 这个碗相当于一个长度为 R 的摆。从很小的高度出发，到达底部所需的时间是 $\frac{\pi}{2}\sqrt{R/g}$。从碗边（$\gamma=90^\circ$）出发，所需时间是这个值的 1.180 倍，约为 $1.85\sqrt{R/g}$。

3. 会。两者都在四分之一周期后到达 V，并在该处碰撞。无论它们碰撞后以甚么速度离开，各自都以 s 作简谐运动，而且周期相同，所以各自恰好在半个周期后回到 V，并再次在 V 碰撞。（碰撞后静止于 V 的质点则会一直留在该处。）

4. 设 $x=R(\theta-\sin\theta)$ 及 $y=R(1-\cos\theta)$，则 $\mathrm{d}s=2R\sin\frac{\theta}{2}\,\mathrm{d}\theta$，所以一拱的长度是 $\int_0^{2\pi}2R\sin\frac{\theta}{2}\,\mathrm{d}\theta=8R$。面积是 $\int y\,\mathrm{d}x=R^2\int_0^{2\pi}(1-\cos\theta)^2\,\mathrm{d}\theta=3\pi R^2$。

11.5 加速火车中的重力

1. 在车厢的参考系中，每单位质量受到一个向后（向南）的力 a，方向与发射方向相同。飞行时间为 $\frac{2V\sin\theta}{g}$，所以射程是 $\frac{V^2}{g}\big(\sin 2\theta+\frac{a}{g}(1-\cos 2\theta)\big)$。当 $\tan 2\theta=-\frac{g}{a}$ 时它的导数为零，所以 $2\theta\approx\frac{\pi}{2}+\frac{a}{g}$。极大值是 $\frac{V^2}{g}\big(\sqrt{1+a^2/g^2}+\frac{a}{g}\big)\approx\frac{V^2}{g}+\frac{V^2a}{g^2}$。

2. $x=ut\cos\theta-\frac{1}{2}ft^2$，在 $t=\frac{2u\sin\theta}{g}$ 时着地，着地点为 $x=\frac{2u^2\sin\theta}{g^2}(g\cos\theta-f\sin\theta)$。若 $\tan\theta>\frac{g}{2f}$，水平速度 $u\cos\theta-ft$ 会在着地前变为负数，所以 $\alpha=\tan^{-1}\frac{g}{2f}$。当 $f=g$ 及 $\theta=45^\circ$ 时，加速度刚好与初速度方向相反：抛体（projectile）沿直线 $y=x$ 笔直向上，停下，然后沿原路直接返回 O。

3. 有效重力（effective gravity）是向下的 g 加上向后的 a。自由液面与它垂直，所以倾斜的角度是 $\tan^{-1}\frac{a}{g}$。

4. 有效重力的大小是 $\sqrt{g^2+a^2}$，在 $x=\frac{ak}{2g}$ 处斜率是 $\frac{a}{g}$，曲率半径（radius of curvature）是 $\rho=\frac{k}{2}\big(1+\frac{a^2}{g^2}\big)^{3/2}$。周期是 $2\pi\sqrt{\rho/\sqrt{g^2+a^2}}=2\pi\sqrt{\frac{k(g^2+a^2)}{2g^3}}$，当 $a=0$ 时简化为 $2\pi\sqrt{\frac{k}{2g}}$。

11.6 后旋

1. 摩擦力作用于接触点，所以对该点的角动量（angular momentum）守恒：$maV+\frac{2}{5}ma^2\frac{V}{a}=mav_0+\frac{2}{5}ma^2\omega_0$。因此 $V=\frac{5v_0+2a\omega_0}{7}$，当 $\frac{v_0}{a\omega_0}<1$ 时它大于 v_0。

2. $V=\frac{5}{7}v_0$。最终能量是 $\frac{1}{2}m\big(\frac{5}{7}v_0\big)^2\big(1+\frac{2}{5}\big)=\frac{5}{7}\cdot\frac{1}{2}mv_0^2$，所以损失了 $\frac{2}{7}$ 的能量。

3. 以 $I=ma^2$ 重复同一论证，得 $V=\frac{1}{2}(v_0-a\omega_0)$。令 $V=-\frac{v_0}{7}$，得 $\frac{v_0}{a\omega_0}=\frac{7}{9}$。

4. 在球心上方高度 h 处施加冲量（impulse）J，得 $mv=J$ 及 $\frac{2}{5}ma^2\omega=Jh$。滚动需要 $v=a\omega$，所以 $h=\frac{2}{5}a$。

11.7 杆中的辛普森法则

1. $\frac{1}{6}\big(1+4\cdot\frac{4}{5}+\frac{1}{2}\big)=\frac{47}{60}\approx0.7833$，而 $\frac{\pi}{4}\approx0.7854$。对于 $\int_{-1}^{1}$，法则 $\frac{1}{3}\big(\mathrm{f}(-1)+4\mathrm{f}(0)+\mathrm{f}(1)\big)$ 对 x^3 给出 0，对 x^2 给出 $\frac{2}{3}$（两者都是准确值），但对 x^4 给出 $\frac{2}{3}$，而它的积分其实是 $\frac{2}{5}$。

2. 对于与基准方向成角 ψ 的轴，转动惯量（moment of inertia）是一个二次型（quadratic form）$A\cos^2\psi+B\sin^2\psi+2C\sin\psi\cos\psi$。在 $\psi\mapsto\psi+90^\circ$ 之下不变，迫使 $A=B$ 及 $C=0$，所以对每一条轴都相同：$\frac{1}{3}Ma^2$，即对平行于一边的轴的值。

3. 在每边中点各放置 $\frac{1}{3}$ 的面积，便能准确地在三角形上积分任何二次式（可在直角三角形上验证 1、x、x^2、xy）。所以这三个质点与该薄片（lamina）有相同的质量、质心（centre of mass）和二阶矩（second moments）。

4. 在每个方向上，取杆在 $-a$、0、a 处的权重 $\frac{1}{6}$、$\frac{2}{3}$、$\frac{1}{6}$ 的乘积：顶点处为 $\frac{M}{216}$，棱的中点处为 $\frac{M}{54}$，面心处为 $\frac{2M}{27}$，中心处为 $\frac{8M}{27}$。对通过中心且平行于一条棱的轴，$I = M\left(\frac{a^2}{3} + \frac{a^2}{3}\right) = \frac{2}{3}Ma^2$。根据平行轴定理（parallel axis theorem），对一条棱的转动惯量是 $\frac{2}{3}Ma^2 + M(\sqrt{2}a)^2 = \frac{8}{3}Ma^2$。

11.8 倾斜的平衡

1. 由公共面起向圆锥方向量度，G 位于 $\frac{-\frac{2}{3}\cdot\frac{3a}{8} + \frac{\lambda}{3}\cdot\frac{\lambda a}{4}}{\frac{2}{3}+\frac{\lambda}{3}} = \frac{a(\lambda^2-3)}{4(2+\lambda)}$。在半球上，接触点处的法线必定通过该面的中心，所以当 G 低于该中心，即 $\lambda < \sqrt{3}$ 时，玩具会自行回正。侧卧时，玩具沿圆锥的一条母线（generator）及在边缘处与桌面接触；若通过 G 的铅垂线落在顶点与边缘之间，它便会停留在那里。这需要 G 的高度至少为 $-\frac{a}{\lambda}$，化简后即 $\lambda^3 + \lambda + 8 > 0$：恒成立。

2. 均匀球体上的每一点都是平衡位置（随遇平衡）。若质心稍微偏离球心，便只有两个：G 在最低处（稳定）和 G 在最高处（不稳定）。

3. 平衡出现在法线通过中心之处：即两轴的四个端点。靠在短轴的一端是稳定的（G 尽可能低）；靠在长轴的一端则不稳定。

4. 由对称性，G 在原点。当 $n = 2$ 时，$Y = r\left(\frac{1}{2} - p^2\right)$，而倾斜平衡需要法线通过 G，所以 $Y = 0$ 及 $p = \frac{1}{\sqrt{2}}$。该处的斜率是 $-\sqrt{2}$，所以 $\overrightarrow{AB}$ 倾斜的角度是 $\tan^{-1}\sqrt{2} \approx 54.7°$，若靠在另一个抛物面上则是 125.3°。两者互为关于 x 轴的反射，这就是倾斜角成对出现的原因。

11.9 安全抛物线

1. 每个质点位于 $ut\,\hat{\mathbf{n}} - \frac{1}{2}gt^2\mathbf{j}$，即在以 $-\frac{1}{2}gt^2\mathbf{j}$ 为球心、半径为 ut 的球面上。从 $(0, \frac{u^2}{2g})$ 水平射出的质点沿 $y = \frac{u^2}{2g} - \frac{gx^2}{2u^2}$ 运动：这正是安全抛物线（parabola of safety）。以角 α 射出的质点在时间 $\frac{u}{g\sin\alpha}$ 于 $x = \frac{u^2}{g\tan\alpha}$ 处与它相切；水平射出的质点需时 $\frac{x}{u}$ 到达该处，所以它必须在时间 $\frac{u(1-\cos\alpha)}{g\sin\alpha} = \frac{u}{g}\tan\frac{\alpha}{2}$ 发射（这需要 $\alpha \geq 45°$，使切点在地面之上）。

2. 该直线与安全抛物线相交于 $R\sin\beta = \frac{u^2}{2g} - \frac{gR^2\cos^2\beta}{2u^2}$ 之处，得 $R = \frac{u^2}{g(1+\sin\beta)}$。在极大值处，练习 3 的两个角重合，所以最佳角度是角平分线的方向，即 $45° + \frac{1}{2}\beta$。

3. 设 $\lambda = \tan\theta$，两个值满足 $\lambda_1 + \lambda_2 = \frac{2u^2}{gx}$ 及 $\lambda_1\lambda_2 = 1 + \frac{2u^2y}{gx^2}$。所以 $\tan(\theta_1 + \theta_2) = -\frac{x}{y}$ 及 $\theta_1 + \theta_2 = 90° + \varphi$，其中 φ 是 OP 的仰角。OP 与铅垂线的角平分线位于 $\frac{1}{2}(90° + \varphi)$。

4. 顶点位于 $\left(\frac{u^2}{2g}\sin 2\theta, \frac{u^2}{2g}\sin^2\theta\right)$，焦距（focal length）是 $\frac{u^2\cos^2\theta}{2g}$，所以焦点（focus）的高度是 $\frac{u^2}{2g}(\sin^2\theta - \cos^2\theta)$。无论 θ 是多少，它与 O 的距离都是 $\frac{u^2}{2g}$。

11.10 会朝你飞回来的球

1. 取单位使 $u = g = 1$，则 $\frac{\mathrm{d}}{\mathrm{d}t}OP^2 = t(t^2 - 3t\sin\alpha + 2)$，其根为 $t \approx 1.050$（局部极大，$OP \approx 0.516\frac{u^2}{g}$）和 $t \approx 1.904$（局部极小，$OP \approx 0.336\frac{u^2}{g}$）。球在 $t \approx 1.970$ 着地。
2. 沿斜面及垂直于斜面方向分解，$OP^2 = u^2t^2 - ugt^3\sin(\alpha' + \beta) + \frac{1}{4}g^2t^4$，其中 α' 由斜面起量度。这就是平地的公式，只是 $\alpha = \alpha' + \beta$。所以条件 $\sin(\alpha' + \beta) > \frac{2\sqrt{2}}{3}$ 所涉及的是与水平所成的角；若由斜面起量度，临界角便降至 $70.5° - \beta$（前提是球尚未着地）。
3. 较小。阻力（drag）抵消水平运动，并使下降变得更陡，所以球落回时更接近铅垂方向，令 OP 更容易缩短。
4. OP^2 的公式不涉及着地点，所以与之前一样，当且仅当 $\sin\alpha > \frac{2\sqrt{2}}{3}$ 时，OP 会有一段时间递减（而飞行时间较长，意味着递减阶段必定在着地之前出现）。最终下落占主导，OP 便会递增。

11.11 外野手如何接球

1. 若她站在 D_0 处，并以速度 v 背离击球手（batsman）跑动，球便在她前方 $D_0 + vt - Ut$ 的距离处。要接到球，她需要 $D_0 + vT = UT$，这时距离是 $(U - v)(T - t)$，所以 $\tan\theta = \frac{gt}{2(U-v)}$，同样是线性的。若球落在她身后，则 $UT > D_0$ 及 $v > 0$：她向后跑。
2. $\theta = \tan^{-1}\frac{gt}{2U}$，所以 $\dot\theta = \frac{g/2U}{1+(gt/2U)^2}$，它是递减的。$\tan\theta$ 之所以是线性的，是因为高度 $\frac{1}{2}gt(T - t)$ 和距离 $U(T - t)$ 有公因式 $T - t$。
3. 会。相对于她眼睛的高度是一个在接球时刻为零的二次式，所以可因式分解为 $(T - t)(c + dt)$；除以 $U(T - t)$ 后剩下一个线性函数（它起初是负数：她最初要向下看）。
4. 不会。有阻力时，水平距离是 $\frac{U}{k}(\mathrm{e}^{-kt} - \mathrm{e}^{-kT})$，而高度并不等于它乘以 t 的一个线性函数，所以 $\tan\theta$ 不再是线性的。

11.12 橡皮绳上的蚂蚁

1. $\frac{x}{a+ut} = \frac{1}{k}\ln\frac{a+ut}{a}$，它在 $t = \frac{a}{u}(\mathrm{e}^{k/2} - 1)$ 时等于 $\frac{1}{2}$。由于 $\frac{T}{2} = \frac{a}{2u}(\mathrm{e}^{k/2} - 1)(\mathrm{e}^{k/2} + 1)$ 而 $\frac{1}{2}(\mathrm{e}^{k/2} + 1) > 1$，这个时刻在 $\frac{T}{2}$ 之前。
2. $\frac{\mathrm{d}}{\mathrm{d}t}\frac{x}{L} = \frac{v}{L} = \frac{v}{a}\mathrm{e}^{-\lambda t}$，所以 $\frac{x}{L} = \frac{v}{a\lambda}(1 - \mathrm{e}^{-\lambda t})$，只有当 $v > \lambda a$ 时才会达到 1。
3. $\frac{x}{L} = v\int_0^t \frac{\mathrm{d}s}{a+\frac{1}{2}bs^2}$，它趋向 $\frac{v\pi}{\sqrt{2ab}}$。所以只有当 $v > \frac{\sqrt{2ab}}{\pi}$ 时，蚂蚁才会到达。
4. 这需要 $\frac{H_N}{100\,000} = \frac{1}{100}$，所以 $H_N = 1000$，而 $N \approx \mathrm{e}^{999.4} \approx 10^{434}$ 秒。

11.13 铲雪车问题

1. 正午后时间 t 的积雪深度与 $t+t_0$ 成正比，所以铲雪车走过的距离是 $c\ln\frac{t+t_0}{t_0}$。条件 $x(1)=2$ 及 $x(2)=3$ 给出 $\left(1+\frac{1}{t_0}\right)^3=\left(1+\frac{2}{t_0}\right)^2$。设 $u=\frac{1}{t_0}$，这便化简为 $u^2-u-1=0$，所以 $u=\frac{1+\sqrt{5}}{2}$ 及 $t_0\approx 0.618$ 小时：大约在 11:23 开始下雪。

2. $\frac{\mathrm{d}t}{\mathrm{d}y}=\frac{\alpha h}{k}\mathrm{e}^{\alpha y}-\frac{\alpha h}{k}\mathrm{e}^{\alpha y}+\alpha\left(T-\frac{\alpha h y}{k}\right)\mathrm{e}^{\alpha y}$，而 $\alpha\left(t-(\mathrm{e}^{\alpha y}-1)\frac{h}{k}\right)$ 与它相同。

3. 时间差距是 $\left(T-\frac{\alpha h y}{k}\right)\mathrm{e}^{\alpha y}$，它的导数在 $y=\frac{kT-h}{\alpha h}$ 处为零。若 $kT>h$，差距会一直增大到该处，然后缩小，在 $y=\frac{kT}{\alpha h}$ 处减至 0。若 $kT\le h$，导数 $\alpha\mathrm{e}^{\alpha y}\left(T-\frac{h}{k}-\frac{\alpha h y}{k}\right)$ 对所有 $y>0$ 都是负数，所以差距从一开始便缩小。第二辆铲雪车面对的积雪深度是 $k\times$（差距）；只要这比第一辆铲雪车在同一地点遇到的积雪深，它便会落后得更远。

4. $\frac{1}{\alpha}\frac{\mathrm{d}t_3}{\mathrm{d}y}=t_3-t_2(y)$。写成 $t_3=t_2+g$，得 $g=\mathrm{e}^{\alpha y}\left(T-\alpha T y+\frac{\alpha^2 h y^2}{2k}\right)$。若 $kT\ge 2h$，它在 $y=\frac{k}{\alpha h}\left(T-\sqrt{T^2-2hT/k}\right)$ 处为零，早于第二辆铲雪车追上第一辆之处；若 $kT<2h$，则第二辆会先追上第一辆。

11.14 是抛物线，不是悬链线

1. 是。水平张力 H 仍然是常数，中间一段的斜率为 0，其后斜率每次递增 $\frac{w}{H}$。以水平一段的中点为起点量度 x，第 k 个接点位于 $x=(k-\frac{1}{2})h$，比最低的接点高出 $\frac{wh}{2H}k(k-1)$，即 $\frac{w}{2Hh}\left(x^2-\frac{h^2}{4}\right)$：是一条抛物线。

2. 张力是 $\sqrt{H^2+V^2}$，在垂直分量 V 最大之处最大：即在两端的一段，该处 V 承托一半的负载，即 $\frac{(2n+1)w}{2}$。所以 $T_{\max}=\sqrt{H^2+\frac{1}{4}(2n+1)^2w^2}$。

3. $y'=\sinh\frac{x}{c}$ 及 $y''=\frac{1}{c}\cosh\frac{x}{c}=\frac{1}{c}\sqrt{1+\sinh^2\frac{x}{c}}$。长度是 $\int_0^X\cosh\frac{x}{c}\,\mathrm{d}x=c\sinh\frac{X}{c}$。

4. 对于垂度（sag）为 s 的平缓曲线，悬链线（catenary）和抛物线的长度都约为 $2d+\frac{4s^2}{3d}$，所以两者都给出 $s\approx\sqrt{\frac{3}{2}d(L-d)}$；它们只在 s^4 阶才有分别。对于一条几乎拉紧的链，这个差别远小于伽利略（Galileo）所能量度的任何数值。

章

12

概率

不确定的事情，可以精确地推理。直觉呢，原来不行。

本章内容

12.1	精确地取条件	STEP I 2007
12.2	基本比率谬误	STEP II 1998
12.3	“随机”是甚么意思?	STEP III 2022
12.4	两两独立，却非相互独立	STEP III 2021
12.5	最后一位乘客的座位	STEP II 2021
12.6	多一枚硬币	STEP I 2016
12.7	淘汰赛中谁会遇上谁	STEP II 样本试卷
12.8	Penney 游戏	STEP III 2001
12.9	用不公平的硬币作公平的决定	STEP I 2019
12.10	把绳头系起来	STEP III 2008
12.11	邮票、杯子与 e	STEP I 1996
12.12	何时停止寻找	STEP I 2001
12.13	三人决斗	STEP II 1995
12.14	混合策略	STEP I 1989
12.15	灭绝还是爆发	STEP II 2003
12.16	顺向与逆向的马可夫链	STEP III 1996
12.17	巴拿赫的火柴盒	STEP II 1992
12.18	无记忆性	STEP III 2021
12.19	平面上的随机点	STEP II 2012
12.20	蒲丰投针	STEP I 1993
12.21	折断一根棒	STEP III 1990
12.22	没有平均值的分布	STEP III 1995
12.23	由概率得出斯特灵公式	STEP II 2016
12.24	中心极限定理	STEP III 2008
12.25	众多之中的最大值	STEP III 2025
A	练习答案	第 969 页

机会的推理

概率论不过是化为计算的常识而已。

— 拉普拉斯（Pierre-Simon Laplace），*Théorie analytique des probabilités*（1812）

拉普拉斯这句名言只说对了一半。概率（probability）确实是精确化了的常识；但正如本章一再显示的，任由常识自行其是，就会出错。人们以为，罕见疾病的检测结果呈阳性，就表示你很可能患了病；以为若第一位登机的乘客坐错了座位，最后一位乘客就很难坐到自己的座位；以为一串掷硬币的结果，不可能胜过另一串同样长度的结果；以为多次量度的平均值总会稳定下来。这些全都是错的，而每一次只需简短的计算，便能揭示真相，更可贵的是，还能揭示当中的原因。

概率始于十七世纪有关赌博的问题，见于帕斯卡（Blaise Pascal）与费马（Pierre de Fermat）之间的书信。直到 1933 年，柯尔莫哥洛夫（Andrey Kolmogorov）订立了它的公理（axiom），概率才成为严格数学的一个分支。在这期间以及其后，它已发展成数学中应用最广的部分之一。

概率为何重要

凡是要处理不确定性的领域，都用到概率：医学（诊断检测与临床试验）、法律（证据的分量）、金融（风险与保险）、工程（可靠性与洪水）、电脑科学（随机化算法、密码学、机器学习）、物理（量子力学与统计力学）以及生物学（遗传学与流行病）。现代人工智能在很大程度上就是应用概率：语言模型所做的，是预测下一个字词出现的概率。而能够正确地就风险、机会和证据进行推理，是任何受过教育的人所能拥有的最宝贵技能之一。

你在大学会遇到甚么

大学一年级的概率课程涵盖公理、条件概率（conditional probability）与贝叶斯定理（Bayes' theorem）、独立性（independence）、离散与连续随机变量（random variable）、期望值（expectation）与方差（variance）、母函数（generating function），以及两大极限定理（limit theorem）：大数定律（law of large numbers）和中心极限定理（central limit theorem）。其后的课程会研究随机过程（random process），包括马可夫链（Markov chain）、泊松过程（Poisson process）、随机漫步（random walk）和分支过程（branching process），也会研究测度论概率（measure-theoretic probability，它为极限和连续分布奠定稳固的基础）和鞅（martingale），并把概率与统计学、分析、组合数学和物理联系起来。

贯穿本章的问题

三个大问题

1. **问题究竟在问甚么？** 随机试验（random experiment）是甚么？以甚么为条件？“随机”又是甚么意思？
2. **巧妙的切入点在哪里？** 甚么时候可以用对称性（symmetry）、指示变量（indicator）、取条件（conditioning）或精心选定的状态，来取代冗长的计算？
3. **长远而言会怎样？** 随机过程随时间如何演变？许多随机量的和、平均值和最大值又是甚么模样？

本章分为五部分。*清晰思考*（Section 12.1 至 Section 12.4）处理各种陷阱：取条件、基础比率（base rate）、“随机”的含义，以及独立的含义。*对称性与巧妙论证*（Section 12.5 至 Section 12.9）收集了一些解法既短且美的问题。*期望值、博弈与停止*（Section 12.10 至 Section 12.14）转向平均值与决策：何时停止、向谁开枪、怎样随机化。*随机过程*（Section 12.15 至 Section 12.19）追踪随机性在时间和空间中的表现。而*极限与连续概率*（Section 12.20 至 Section 12.25）则以几何概率（geometric probability）和几个重要的极限定理作结。

本章的思想

§	课题	思想
12.1	取条件	“指定的一个”与“至少一个”之别
12.2	贝叶斯定理	基础比率与自然频率（natural frequency）
12.3	贝特朗悖论（Bertrand's paradox）	“随机”需要定义
12.4	独立性	两两独立（pairwise independent）不等于相互独立（mutually independent）
12.5	飞机座位	两个座位之间的对称性
12.6	多一枚硬币	互补而对称的事件
12.7	锦标赛	数算比赛场数与配对
12.8	彭尼博弈（Penney's game）	首步分析（first-step analysis）；非传递性（non-transitivity）
12.9	公平硬币	从偏差中提取公平
12.10	随机环圈	无记忆的指示变量之和
12.11	杯子与 e	阶乘（factorial）与指数级数（exponential series）
12.12	最优停止（optimal stopping）	$\frac{1}{e}$ 法则
12.13	三人决斗（truel）	策略上的弱势
12.14	博弈论（game theory）	随机化策略；极小化极大（minimax）
12.15	分支过程	灭绝作为不动点（fixed point）
12.16	马可夫链	平稳分布（stationary distribution）；逆转
12.17	巴拿赫（Banach）的火柴盒	负二项分布（negative binomial distribution）
12.18	指数分布（exponential distribution）	无记忆性（memorylessness）与独立性
12.19	空间随机性	最近邻；瑞利（Rayleigh）分布
12.20	蒲丰投针（Buffon's needle）	几何概率与 π
12.21	随机三角形	以面积表示概率
12.22	柯西分布（Cauchy distribution）	没有平均值的分布
12.23	斯特灵公式（Stirling's formula）	由正态曲线（normal curve）得出阶乘
12.24	中心极限定理	矩母函数（moment generating function）
12.25	极值（extreme value）	冈贝尔分布（Gumbel distribution）

其中一节，即 Section 12.24，用到矩母函数，而矩母函数已于 2019 年从 STEP 考试范围中删去；该节会解释所需的内容。

怎样阅读本章

当一道概率题令人摸不着头脑时，不妨就一个细小的情况写出样本空间（sample space），然后逐一计数。当答案简单得令人起疑时，就去找出解释它的对称性。而当你的直觉与

计算结果不一致时，请相信计算，然后弄清楚直觉为何出错：真正的学习正在于此。

12.1 精确地取条件

题目

STEP I 2007，第 13 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/07-s1-q13`

引子

“已知”（given that）是概率中最危险的字眼。这道题最后四个部分的说法听起来几乎一模一样：编号为 3 的圆片最先被取出；编号为 3 的圆片被取出；最先取出的是一块有编号的圆片；取出的圆片中有一块有编号。其中三个的答案都是 $\frac{1}{2}$，第四个却不是。

这个分别，正正就是著名的“男孩还是女孩”谜题背后的分别（一个家庭有两个孩子，其中至少一个是男孩；两个都是男孩的机会有多大?）。这道谜题难倒了一代又一代的学生，也难倒了不少专业数学家。当中的教训是：条件（condition）是一个精确的陈述，说明哪些结果仍然可能出现；而信息是怎样得来的，与信息本身同样重要。

动笔之前

? **想一想** 从十一块圆片中取出四块。对于“四块都有编号”这个事件（event）而言，抽取的次序有没有影响？四块圆片组成的等可能（equally likely）组合共有多少个?

每一个由四块圆片组成的组合被抽中的可能性都相同，所以我们可以不理会次序：共有 $\binom{11}{4} = 330$ 个等可能的组合。

逐步解题

阶段 1 没有条件

由四块有编号圆片组成的组合有 $\binom{5}{4} = 5$ 个，所以

$$\mathrm{P}(\text{四块都有编号}) = \frac{5}{330} = \frac{1}{66}.$$

阶段 2 3 号最先取出

想一想 编号为 3 的圆片最先被抽出之后，余下的三次抽取是怎样的?

其余三块圆片是从余下十块（四块有编号、六块空白）中随机抽出的三块，全部 $\binom{10}{3} = 120$ 个组合都是等可能的。所以

$$\text{(ii)}\quad \frac{\binom{4}{3}}{120} = \frac{4}{120} = \frac{1}{30}, \qquad \text{(iii)}\quad \frac{\binom{4}{1}\binom{6}{2}}{120} = \frac{4 \times 15}{120} = \frac{1}{2},$$

因为在 (iii) 中，其余三块之中必须恰好有一块有编号。

阶段 3 3 号被取出（在某一次）

想一想 编号为 3 的圆片是在哪一次抽出的，有没有关系？已知 3 号在四块之中，对其余三块你能说些甚么?

根据对称性（symmetry），已知这四块圆片中包含 3 号，其余三块就是从余下十块中随机抽出的三块，与第 2 阶段完全一样。所以 (iv) 的答案也是 $\frac{1}{2}$。

阶段 4 最先取出的是有编号的圆片

已知第一块圆片有编号（不论是哪个号码），其余三块同样是从余下十块圆片中随机抽出的三块，当中四块有编号、六块空白。所以 (v) 的答案也是 $\frac{1}{2}$。

阶段 5 取出的圆片中有一块有编号

想一想 现在的条件是“四块之中至少有一块有编号”。这与之前任何一个条件相同吗？330 个组合之中有多少个符合这个条件?

这个条件有所不同：它没有指明任何一块特定的圆片。只有 $\binom{6}{4} = 15$ 个组合完全没有编号圆片，所以有 315 个符合条件。其中 $\binom{5}{2}\binom{6}{2} = 10 \times 15 = 150$ 个恰好包含两块有编号的圆片。所以

$$\text{(vi)}\quad \frac{150}{315} = \frac{10}{21}. \qquad \blacksquare$$

停一停，试一试 为甚么 (vi) 与 (v) 不同？在 (v) 中，我们知道哪一块圆片有编号（第一块），而其余三块不受限制。在 (vi) 中，我们只知道有某一块圆片有编号。含有三块或四块编号圆片的组合，可以用更多方式满足“某一块圆片有编号”，所以它们不会像在 (v) 中那样比重偏低。

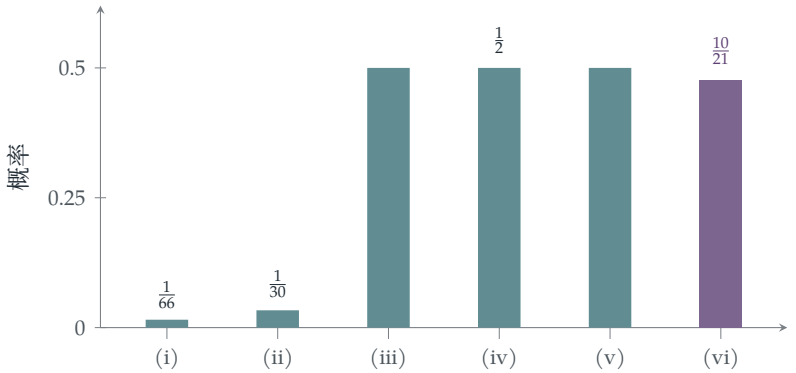


图 12.1.1 六个答案。第 (iii)、(iv) 和 (v) 部分的条件都是某一块特定的圆片有编号，答案是 $\frac{1}{2}$。第 (vi) 部分的条件只是有某一块圆片有编号，答案是 $\frac{10}{21}$。

回顾

- **条件会收窄样本空间（sample space）。** $P(A \mid B)$ 是 B 的结果之中同时属于 A 的比例。
- **对称性可以省功夫。** 3 号是最先抽出，还是在某一次抽出，并无分别，因为所有位置的地位都相同。
- **“特定的一个”不等于“至少一个”。** 听起来相似的条件，选出来的结果集合可以大不相同。

背后的数学

男孩还是女孩的谜题

马丁·加德纳（Martin Gardner）在 1959 年提出这个问题：史密斯先生有两个孩子，其中至少一个是男孩；两个都是男孩的概率是多少？列出等可能的家庭组合 BB、BG、GB、GG，剔除 GG 后剩下三个，其中一个是 BB：答案是 $\frac{1}{3}$。但如果我们知道年长的孩子是男孩，答案就是 $\frac{1}{2}$。这正是我们这道题第 (vi) 和第 (v) 部分的缩影。加德纳后来承认，他最初提出的问题含糊不清：答案取决于我们是怎样得知有一个男孩的。如果我们随机遇上其中一个孩子，而他是男孩，答案是 $\frac{1}{2}$；如果我们问“你有没有至少一个儿子？”而对

方回答有，答案就是 $\frac{1}{3}$。概率取决于产生信息的过程，而不仅是信息本身。

蒙提霍尔问题（Monty Hall）

概率中争议最多的谜题，背后也是同一个教训。在一个游戏节目中，三道门的其中一道后面有一辆汽车。你选了 1 号门。主持人知道汽车在哪里，他必定会打开另外两道门的其中一道，让你看到门后是一只山羊，并且必定会让你选择是否换门；当汽车就在你所选的门后，以致他有选择余地时，他会在另外两道门之间随机挑选。他打开了 3 号门，并让你改选 2 号门。你应该换吗？应该：换门后胜出的概率是 $\frac{2}{3}$。主持人选哪道门并非随机；这取决于汽车的位置，因而带有信息。确切的规则十分重要。如果主持人只要可以就一定打开 3 号门，那么看到 3 号门打开，1 号门和 2 号门就变得同样可能，换门后胜出的概率只有 $\frac{1}{2}$；然而，一个事先决定必定换门的参赛者，仍然有 $\frac{2}{3}$ 的机会胜出，因为换门胜出，恰好就是第一次选错的时候。Marilyn vos Savant 在 1990 年于杂志专栏中给出这个答案时，收到数以千计的来信指她错了，其中有些更是出自数学家之手。用电脑模拟一次，或者仔细列出各个等可能的情况，就可以解决争议。

贝叶斯（Bayes）与拉普拉斯（Laplace）

条件概率(conditional probability)的系统化基础,是由托马斯·贝叶斯(Thomas Bayes)和皮埃尔–西蒙·拉普拉斯（Pierre-Simon Laplace）奠定的：贝叶斯的论文在他死后于 1763 年发表，而拉普拉斯则在 1774 年以一般形式写出了现称为贝叶斯定理（Bayes' theorem）的法则。安德烈·柯尔莫哥洛夫（Andrey Kolmogorov）在 1933 年提出的公理（axioms），是现今每一个大学课程的起点，当中定义 $\mathrm{P}(A \mid B) = \mathrm{P}(A \cap B)/\mathrm{P}(B)$，而概率论的其余部分都建基于这个定义之上。下一节 Section 12.2 会展示贝叶斯定理在实际应用中的作用。

思维技巧

准确把握条件

计算条件概率之前，先确切写下哪些结果符合条件。问一问信息是怎样得来的。

在大学里，条件概率和条件期望（conditional expectation）是概率与统计中最重要的工具之一，也是最常出错的地方。把样本空间写出来，至少就一个小规模的情况写出来，是最好的防范方法。

延伸练习

1 **推广。**设有 b 块空白圆片和 n 块有编号的圆片，从中抽出 k 块，证明 (iv) 和 (v) 的答案永远相等，并求出 (vi) 在甚么情况下与它们相等。

2 **星期二出生的男孩。**一个家庭有两个孩子，其中至少一个是在星期二出生的男孩。证明两个都是男孩的概率是 $\frac{13}{27}$。（列出 14×14 个等可能的情况，或者用巧妙的方法计数。）

3 **蒙提霍尔问题。**列出各个结果，验证 $\frac{2}{3}$ 这个答案。如果主持人不知道汽车在哪里，随机打开另外两道门的其中一道，而门后刚好是一只山羊，那又如何?

4 **两张 A。**一手 13 张的桥牌中至少有一张 A。这手牌中至少有两张 A 的概率是多少?试与已知这手牌中有葵扇 A 时的概率比较。

延伸阅读

- **Jason Rosenhouse, *The Monty Hall Problem***（Oxford University Press）。讲述这道谜题、它的历史，以及它的众多变化。
- **Geoffrey Grimmett & Dominic Welsh, *Probability: An Introduction***（Oxford University Press），有关条件概率的各章。
- **Joseph Blitzstein & Jessica Hwang, *Introduction to Probability***（CRC Press；可于网上免费阅读）。对条件概率的处理十分出色，并载有许多悖论。

12.2 基本比率谬误

题目

STEP II 1998，第 12 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/98-s2-q12`

引子

某种疾病的检测，对患者的准确度达 90%，而且甚少误报。Frankie 的检测结果呈阳性。Frankie 应该有多担心？大多数人都会答“大约 90%”，而在许多研究中，大多数医生也是这样回答。在这道题目中，正确答案是 $\frac{1}{11}$，约为 9%，足足小了十倍。

错处在于忽略了*基础比率*（base rate）：这种疾病很罕见，而健康的人为数众多，所以即使假阳性（false positive）率很低，误报的数目仍会远远多于真正的病例。这种*基础比率谬误*（base-rate fallacy）在医学、法律（在法律上称为检控官谬误（prosecutor's fallacy））、保安检查和机器学习中都举足轻重。贝叶斯定理（Bayes' theorem）正是对症良药。

动笔之前

想一想 试想像有 1000 人。当中有多少人患病？这些患者之中有多少人检测呈阳性？健康的人之中又有多少人检测呈阳性？

在 1000 人当中，有 10 人患病，其中 9 人检测呈阳性。在 990 个健康的人当中，有 $\frac{1}{11}$ 检测呈阳性，即 90 个假阳性。因此共有 99 人检测呈阳性，其中只有 9 人真正患病。

逐步解题

阶段 1 贝叶斯定理

想一想　以 D 表示“Frankie 患有此病”，以 $+$ 表示“AL 检测呈阳性”。你已知 $\mathrm{P}(+\mid D)$，想求 $\mathrm{P}(D\mid +)$。两者有甚么关系?

根据条件概率（conditional probability）的定义，$\mathrm{P}(D\mid +)=\frac{\mathrm{P}(D\cap +)}{\mathrm{P}(+)}$，而 $\mathrm{P}(+)$ 可按 Frankie 是否患病拆成两部分：

$$\begin{aligned}\mathrm{P}(D\mid +)&=\frac{\mathrm{P}(+\mid D)\mathrm{P}(D)}{\mathrm{P}(+\mid D)\mathrm{P}(D)+\mathrm{P}(+\mid \bar{D})\mathrm{P}(\bar{D})}\\&=\frac{0.9\times 0.01}{0.9\times 0.01+\frac{1}{11}\times 0.99}=\frac{0.009}{0.009+0.09}=\frac{1}{11}.\end{aligned}$$ ■

这正是那一千人当中的 $\frac{9}{99}$。

阶段 2　第二次检测

想一想　只有 AL 检测呈阳性的人才会接受第二次检测。在这群人之中，患者占多大比例？把这个比例当作新的“基础比率”。

在 AL 检测呈阳性的人当中，有 $\frac{1}{11}$ 是患者，$\frac{10}{11}$ 不是。以这两个数作为先验概率（prior probability），再次应用贝叶斯定理，得

$$\mathrm{P}(D\mid \text{两次均呈阳性})=\frac{0.9\times\frac{1}{11}}{0.9\times\frac{1}{11}+p\times\frac{10}{11}}=\frac{0.9}{0.9+10p}.$$

令上式等于 $\frac{1}{2}$，得 $10p=0.9$，所以 $p=0.09$。 ■

（这里假设在已知 Frankie 是否患病的条件下，两项检测的误差互相独立（independent），题目亦隐含这个假设。）

停一停，试一试　当 $p=0.09$ 时，STEP 的假阳性率与 AL 的相若，但第二次阳性结果却令概率由 $\frac{1}{11}$ 升至 $\frac{1}{2}$。为甚么质素相同的检测，第二次的影响会大得多?

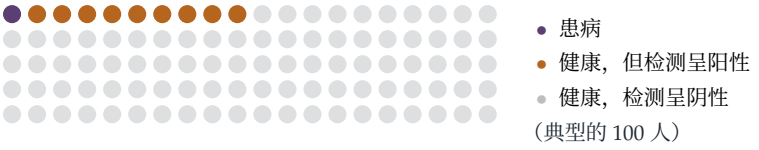


图 12.2.1　典型的一百人。其中一人患病（紫色），十次检测中有九次会呈阳性。在 99 个健康的人当中，约有 9 人会被误判为阳性（琥珀色）。在检测呈阳性的人当中，健康的人与患者的数目之比约为十比一。

回顾

- **自然频率（natural frequencies）。**想像 1000 人，答案便一目了然，也容易验算。
- **先验乘以似然（likelihood）。**后验概率（posterior probability）既取决于检测的准确度，也取决于基础比率。
- **分阶段更新。**昨天的后验就是今天的先验：第二次检测是由 $\frac{1}{11}$ 开始，而不是 $\frac{1}{100}$。

背后的数学

胜算与证据

以胜算（odds）来表达，贝叶斯定理最为清晰。D 的胜算是 $\mathrm{P}(D) : \mathrm{P}(\bar{D})$。一项证据 E 会把胜算乘以似然比（likelihood ratio）$\frac{\mathrm{P}(E|D)}{\mathrm{P}(E|\bar{D})}$：

$$\text{后验胜算} = \text{先验胜算} \times \text{似然比}.$$

这里的先验胜算是 $1:99$，而 AL 的似然比是 $0.9 \div \frac{1}{11} = 9.9$，于是后验胜算为 $9.9:99 = 1:10$，即概率为 $\frac{1}{11}$。第二项独立检测的似然比若为 10（$p = 0.09$ 时的 STEP 正是如此），便会把胜算带到 $1:1$。取对数之后，证据就可以相加：每项独立检测所贡献的，是其似然比的对数。图灵（Alan Turing）在布莱切利园（Bletchley Park）破解恩尼格玛（Enigma）密码电文时，用的正是这个方法，并以“ban”和“deciban”作为量度证据的单位。

检控官谬误

1999 年，事务律师 Sally Clark 被裁定谋杀两名年幼儿子罪名成立，定罪的部分依据是一项证据，指同一家庭发生两宗婴儿猝死的机会是“7300 万分之一”。姑且不论这个数字本身有错（它假设两宗死亡互相独立，忽略了遗传和环境因素），这个推论还把 $\mathrm{P}(\text{证据} \mid \text{无辜})$ 与 $\mathrm{P}(\text{无辜} \mid \text{证据})$ 混为一谈。母亲连杀两子同样极为罕见，真正应该比较的是这两种解释。皇家统计学会（Royal Statistical Society）发表公开信指出这个错误，她的定罪于 2003 年获撤销。当中的数学与本题完全相同。

Gigerenzer 与自然频率

在 1990 年代，心理学家 Gerd Gigerenzer 指出，医生若获得以条件概率表达的检测统计数字，遇到本题这类问题时大多错得离谱；但同样的资料若以自然频率表达（“在 1000 名女性当中，有 10 人患乳癌；其中 8 人检测呈阳性…”），大多数医生都能答对。他的研究改变了许多国家向病人传达医疗风险的方式。

思维技巧

用贝叶斯定理把条件概率倒转过来

$\mathrm{P}(A \mid B)$ 与 $\mathrm{P}(B \mid A)$ 是两个不同的数。要由其中一个求出另一个，你需要基础比率 $\mathrm{P}(A)$。

贝叶斯推理，即根据证据更新信念，是现代统计学和机器学习的基础，由垃圾邮件过滤器以至医学诊断都用得上。到了大学，你会遇到它的进阶形式——贝叶斯推断（Bayesian inference），届时处理的不再是单一的概率，而是先验分布和后验分布。

延伸练习

1 **雪柜里的尸体。**试做 STEP II 1987，第 16 题（`step.maths.org/questions/87-s2-q16`）：贝叶斯定理应用于两名可靠程度不同的证人，是 STEP 中较令人难忘的题目情境之一。

2 **要多罕见？**某项检测的灵敏度（sensitivity）为 0.9，假阳性率为 $\frac{1}{11}$。疾病要有多普遍，阳性结果才表示患病的机会高于一半？

3 **重复检测。**假如重复进行 AL 检测，而在已知受测者真实状况的条件下各次结果互相独立，那么两次 AL 检测均呈阳性后，患病的概率是多少？三次之后呢？

4 **筛查。**某项筛查计划为一百万人检测某种病症，其患病率（prevalence）为 10 000 分之 1，所用检测的灵敏度为 99%，假阳性率为 1%。有多少人会被告知可能患有此病症？当中又有多少人真的患病？

延伸阅读

- **Gerd Gigerenzer, *Reckoning with Risk***（Penguin）。自然频率、医学检测与法庭。
- **David Spiegelhalter, *The Art of Statistics***（Pelican）。内容包括医学筛查和 Sally Clark 案。
- **Sharon Bertsch McGrayne, *The Theory That Would Not Die***（Yale University Press）。贝叶斯定理的历史，包括图灵如何运用这个定理。

12.3 “随机”是甚么意思?

题目

STEP III 2022，第 12 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2022 年 STEP 试卷（zip 文件）内的 STEP III 试卷：

`www.ocr.org.uk/Images/696587-2022-past-papers.zip`

引子

“在圆上随机选一条弦（chord），它的期望长度（expected length）是多少?”这条问题听起来好像有一个答案。其实没有，除非你先说清楚“随机”是甚么意思。在圆周上随机取两点再连起来，期望长度是 $\frac{4a}{\pi} \approx 1.27a$。在圆内随机取一点，再随机取一个方向作过该点的弦，答案却是 $\frac{16a}{3\pi} \approx 1.70a$。两种做法都是选取“一条随机弦”的合理方法。

这就是贝特朗悖论（Bertrand's paradox），由约瑟夫·贝特朗（Joseph Bertrand）于 1889 年发表，用意是提出警告。只有在指明随机试验（random experiment）之后，概率（probability）才有明确的定义。这个悖论促使数学家在二十世纪为概率论建立严格的基础：在这套基础之中，每一个概率都定义在一个精确界定的结果空间之上。

动笔之前

想一想 若角 AOB 为 θ，弦 AB 的长度是多少?（由 O 作垂线。）

由 O 所作的垂线平分该弦，也平分该角，所以 $AB = 2a\sin\frac{\theta}{2}$。

逐步解题

阶段 1　圆周上的两点

$$\mathrm{E}(AB) = \frac{1}{2\pi}\int_0^{2\pi} 2a\sin\frac{\theta}{2}\,\mathrm{d}\theta = \frac{a}{\pi}\Big[-2\cos\frac{\theta}{2}\Big]_0^{2\pi} = \frac{a}{\pi}\cdot 4 = \frac{4a}{\pi}. \qquad \blacksquare$$

阶段 2　圆内的一个随机点

想一想　对于 $0 \le r \le a$，$\mathrm{P}(R \le r)$ 是多少？对它求导，便得到密度。

$\mathrm{P}(R \le r)$ 是半径为 r 的圆盘面积除以整个圆盘的面积，即 $\frac{r^2}{a^2}$。所以 R 在 $[0,a]$ 上的密度（density）是 $\frac{2r}{a^2}$。（点较有可能离圆心较远，因为那里的面积较大。）

想一想　一条过 C 的弦与 OC 成角 t。这条弦与 O 相距多远？

由 O 到该直线的距离是 $R\sin t$，所以弦长为 $2\sqrt{a^2 - R^2\sin^2 t}$。

阶段 3　对 C 取平均

设 $s = \sin t$，

$$\begin{aligned} L(t) &= \int_0^a 2\sqrt{a^2 - r^2 s^2}\cdot\frac{2r}{a^2}\,\mathrm{d}r = \frac{4}{a^2}\Big[-\frac{(a^2 - r^2s^2)^{3/2}}{3s^2}\Big]_0^a \\ &= \frac{4}{a^2}\cdot\frac{a^3 - a^3(1-s^2)^{3/2}}{3s^2} = \frac{4a(1-\cos^3 t)}{3\sin^2 t}. \qquad \blacksquare \end{aligned}$$

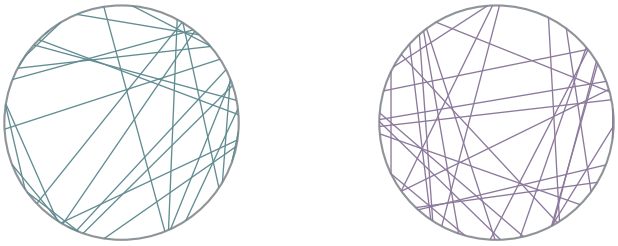
想一想　把 $1 - \cos^3 t$ 和 $\sin^2 t = 1 - \cos^2 t$ 因式分解。有甚么可以约去？

设 $c = \cos t$，则 $\frac{1-c^3}{1-c^2} = \frac{(1-c)(1+c+c^2)}{(1-c)(1+c)} = \frac{1+c+c^2}{1+c} = c + \frac{1}{1+c}$，而 $1 + \cos t = 2\cos^2\frac{t}{2}$。所以

$$L(t) = \frac{4a}{3}\Big(\cos t + \tfrac{1}{2}\sec^2\tfrac{t}{2}\Big). \qquad \blacksquare$$

阶段 4　方向也随机

$$\begin{aligned} \mathrm{E}\big(L(T)\big) &= \frac{2}{\pi}\int_0^{\pi/2}\frac{4a}{3}\Big(\cos t + \tfrac{1}{2}\sec^2\tfrac{t}{2}\Big)\,\mathrm{d}t \\ &= \frac{8a}{3\pi}\Big[\sin t + \tan\tfrac{t}{2}\Big]_0^{\pi/2} = \frac{8a}{3\pi}(1+1) = \frac{16a}{3\pi}. \qquad \blacksquare \end{aligned}$$



两个随机端点
平均长度 $\frac{4a}{\pi} \approx 1.27a$

圆盘内的随机中点
平均长度 $\frac{4a}{3} \approx 1.33a$

图 12.3.1 以两种不同方法选出的二十八条“随机弦”。左：圆周上两个独立的均匀分布点。右：以圆盘内一个均匀随机的点作为弦的中点。右边的方法在边缘附近产生的短弦较少。不同的方法给出不同的分布，也给出不同的答案。

回顾

- **指明试验。**“一条随机弦”在每一部分都化成了一个精确的步骤，而每个步骤给出不同的答案。
- **由面积求密度。**$\mathrm{P}(R \le r)$ 直接来自几何；求导便得到密度。
- **分阶段取平均。**先对 C 的位置取平均（t 固定），再对 t 取平均。

背后的数学

贝特朗的三个答案

贝特朗原本的问题是：一条随机弦比内接等边三角形的边长 $\sqrt{3}a$ 更长的概率是多少？他给出了三种方法和三个答案。

- 随机端点：固定一个端点；若另一个端点落在圆周中间的三分之一之内，弦便是长的。概率为 $\frac{1}{3}$。
- 随机取一条半径，再在其上随机取一点作为中点：若弦与 O 的距离小于 $\frac{a}{2}$，弦便是长的。概率为 $\frac{1}{2}$。
- 在圆盘内任意位置随机取中点：若中点与 O 的距离在 $\frac{a}{2}$ 之内，弦便是长的。概率为 $\frac{1}{4}$。

每一种方法都是弦的集合上一个不同的概率分布（probability distribution），当中并无矛盾：它们回答的是不同的问题。这里的教训是，“均匀随机”只有相对于某个选定的参数化（parametrisation）才有意义，而对于连续的集合，通常并没有唯一自然的参数化。

有没有正确答案?

1973 年，物理学家 Edwin Jaynes 主张第二种方法才是“正确”的，理由是把圆稍为移动或缩放，答案都不应改变：一条随机弦应该就像一根抛到地上所画圆圈上的饮管。只有与圆心距离均匀分布的那个分布才具有这种不变性（invariance），而抛饮管的实验也证实了 $\frac{1}{2}$。同一个想法，即正确的分布就是尊重情境中各种对称性的那一个，正是积分几何（integral geometry）的基础；积分几何是研究随机直线的数学，它也解释了 Section 12.20 中的布丰投针（Buffon's needle）。

从悖论到公理

贝特朗在 *Calcul des probabilités*（1889）中发表了他的弦悖论，部分原因是要嘲讽他眼中对“几何概率”的草率运用。庞加莱（Henri Poincaré）也讨论过它。1900 年，希尔伯特（David Hilbert）把概率的公理化（axiomatisation）列入他为新世纪提出的著名问题之中，而柯尔莫哥洛夫（Andrey Kolmogorov）在 1933 年解决了这个问题：概率是结果空间上的一个测度（measure），任何关于“随机”对象的问题，都必须先说明是哪一个空间、哪一个测度。贝特朗悖论正是说明为何必须如此的标准例子。

思维技巧

界定随机试验

计算概率之前，先准确描述随机对象是怎样产生的。不同的描述可以给出不同的答案。

在大学里，每一个概率计算都由一个样本空间（sample space）和一个概率测度（probability measure）出发。在统计学和物理学中，选择合适的随机模型（往往借助对称性）是工作的一大部分。

延伸练习

1 **贝特朗的概率。**验证上文的三个概率 $\frac{1}{3}$、$\frac{1}{2}$、$\frac{1}{4}$。

2 **另一个平均值。**证明若弦的中点在圆盘内均匀选取（右边的图），期望长度是 $\frac{4a}{3}$。（提示：这就是 $L(\frac{\pi}{2})$。为甚么?）

3 **随机距离。**若弦与 O 的距离在 $[0, a]$ 上均匀分布，证明期望长度是 $\frac{\pi a}{2}$。

4 **三种方法，三个分布。**对三种方法中的每一种，求弦长 ℓ 在 $[0, 2a]$ 上的概率密度（probability density）。

延伸阅读

- **Edwin Jaynes, “The well-posed problem”**, *Foundations of Physics* 3 (1973)。不变性论证，易读而又发人深省。
- **Gábor Székely, *Paradoxes in Probability Theory and Mathematical Statistics*** (Reidel)。贝特朗悖论及其他许多悖论。
- **Daniel Klain & Gian-Carlo Rota, *Introduction to Geometric Probability*** (Cambridge University Press)。积分几何，留待日后研读。

12.4 两两独立，却非相互独立

题目

STEP III 2021，第 12 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2021 年 STEP 试卷（zip 文件）内的 STEP III 试卷：

`www.ocr.org.uk/Images/696588-2021-past-papers.zip`

引子

玩家 1 和玩家 2 掷出相同的点数。这能告诉你玩家 2 和玩家 3 是否掷出相同点数吗？出人意料的是，不能：这两个事件是独立（independent）的。玩家 1 和玩家 3 的情况也一样。但如果 1 与 2 相同，而 2 又与 3 相同，那么 1 必定与 3 相同。三个事件，每两个之间都独立，合起来却不独立。

两两独立（pairwise independence）比完全独立弱，而这个差别至关重要。就方差（variance）而言，两两独立已经足够，本题正是利用了这一点。但在许多其他用途上，这并不足够。电脑科学家会刻意使用两两独立的随机数，因为产生它们的成本远低于产生完全独立的随机数，而对许多算法来说已经够用。

动笔之前

想一想 $\mathrm{P}(X_{12}=1)$ 是多少？$\mathrm{P}(X_{12}=1 \text{ 及 } X_{23}=1)$ 又是多少？

无论玩家 1 掷出甚么，玩家 2 与之相同的概率都是 $\frac{1}{6}$。$X_{12}=1$ 和 $X_{23}=1$ 同时成立，即三人掷出相同的点数：概率为 $\frac{6}{216}=\frac{1}{36}$。

逐步解题

阶段 1 两两独立

由于 $\mathrm{P}(X_{12}=1, X_{23}=1)=\frac{1}{36}=\frac{1}{6}\cdot\frac{1}{6}=\mathrm{P}(X_{12}=1)\mathrm{P}(X_{23}=1)$，而每个变量只取两个值，所以 X_{12} 与 X_{23} 独立。（对于只取两个值的变量，联合分布表中的一格便决定了其余各格；练习 1 会把这一点说清楚。）没有共同玩家的两对，例如 X_{12} 与 X_{34}，显然是独立的。因此所有 X_{ij} 都两两独立。

阶段 2 得分的平均值与方差

得分为 $S=\sum_{i<j}X_{ij}$，是 $\binom{n}{2}$ 个变量之和，每个变量的平均值为 $\frac{1}{6}$，方差为 $\frac{1}{6}\cdot\frac{5}{6}=\frac{5}{36}$。期望值（expectation）总是可以相加的，所以 $\mathrm{E}(S)=\frac{1}{6}\binom{n}{2}$。

> **想一想** 和的方差在甚么时候等于方差之和？

和的方差等于方差之和，再加上协方差（covariance）之和的两倍，而独立的一对变量协方差为零。两两独立已经足够：

$$\mathrm{E}(S)=\frac{n(n-1)}{12},\qquad \mathrm{Var}(S)=\frac{5}{36}\binom{n}{2}=\frac{5n(n-1)}{72}.$$

阶段 3 方差公式

若各 Y_i 的平均值为零，则 $\mathrm{Var}(\sum Y_i)=\mathrm{E}((\sum Y_i)^2)$，把平方展开后，每个 Y_i^2 出现一次，每个乘积 Y_iY_j（其中 $i<j$）出现两次。取期望值便得到公式。 ■

阶段 4 奇偶游戏

> **想一想** $\mathrm{E}(Z_{12})$ 是多少？$Z_{12}Z_{23}$ 可以取哪些值？它在甚么时候不等于零？

$Z_{12}=1$ 的概率为 $\frac{3}{36}=\frac{1}{12}$（两人都掷出 2、都掷出 4 或都掷出 6），而取值 -1 的概率为 $\frac{1}{12}$；所以 $\mathrm{E}(Z_{12})=0$ 而 $\mathrm{E}(Z_{12}^2)=\frac{1}{6}$。

只有当三位玩家都掷出相同点数时，才有 $Z_{12}Z_{23}\neq 0$，而此时 $Z_{12}=Z_{23}$（同为 $+1$ 或同为 -1），所以乘积为 1。因此 $\mathrm{E}(Z_{12}Z_{23})=\frac{1}{36}$，这并不等于 $\mathrm{E}(Z_{12})\mathrm{E}(Z_{23})=0$：它们并不独立。 ■

（更直接的看法是，$\mathrm{P}(Z_{12}=1, Z_{23}=-1)=0$，而每个事件单独发生的概率都是 $\frac{1}{12}$。）

阶段 5 方差

总得分 $T=\sum Z_{ij}$ 的平均值为 0。根据第 (ii) 部分，$\mathrm{Var}(T)$ 等于 $\binom{n}{2}\cdot\frac{1}{6}$，再加上 $\mathrm{E}(Z_{ij}Z_{kl})$ 对所有“两对玩家”的组合求和后的两倍。没有共同玩家的两对，贡献为 0（它们独立，而且各自的平均值为零）。恰好有一位共同玩家的两对，每组贡献 $\frac{1}{36}$，而这样的组合共有

$n\binom{n-1}{2}$ 个（先选共同的玩家，再选另外两人）。所以

$$\begin{aligned}\operatorname{Var}(T) &= \frac{n(n-1)}{12} + 2\cdot\frac{1}{36}\cdot\frac{n(n-1)(n-2)}{2} \\ &= \frac{n(n-1)}{36}\big(3+(n-2)\big) = \frac{n(n^2-1)}{36}.\end{aligned}$$

■

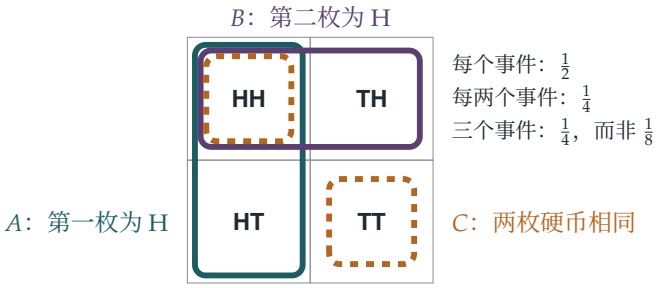


图 12.4.1 伯恩斯坦（Bernstein）的例子：两枚公平硬币。事件“第一枚硬币是正面”（蓝绿色）、“第二枚硬币是正面”（紫红色）和“两枚硬币结果相同”（琥珀色虚线）的概率各为 $\frac{1}{2}$，而每两个事件都相交于概率为 $\frac{1}{4}$ 的单一方格，所以它们两两独立。但三个事件全部相交于 HH，其概率为 $\frac{1}{4}$，而非 $\frac{1}{8}$：其中任何两个事件都决定了第三个。

回顾

- **指示变量（indicator variables）。**把得分写成指示变量之和，一切便化为对一对对玩家的考虑。
- **方差只需要看成对的变量。**和的方差只取决于两两之间的协方差，所以两两独立已足以令方差相加。
- **独立性要验证，不要假设。**两个看似相同的游戏，却有不同的相依结构。

背后的数学

多个事件的独立性

如果对每一个子族，交集的概率都等于各概率之积，则称事件 $A_1,\dots,A_n$ 独立。对三个事件来说，这意味着四个条件：三个是关于每一对的，另一个是 $\mathrm{P}(A\cap B\cap C)=\mathrm{P}(A)\mathrm{P}(B)\mathrm{P}(C)$。伯恩斯坦（Sergei Bernstein）在 1920 年代提出图 12.4.1 中的例子，说明两两之间的条件并不能推出三者的条件。（反过来，单凭三者的条件也不能推出两两之间的条件。）题目中“点数相同”的事件 X_{12}、X_{23}、X_{13} 与伯恩斯坦的硬币如出一辙。

不相关不等于独立

协方差为零，即 $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$，是比独立更弱的条件，称为不相关（uncorrelated）。STEP III 2000 Q14（见练习）说明，对于只取两个值的变量，这两个概念是一致的，但取三个值时则不然。一个简单的例子：设 X 在 $\{-1, 0, 1\}$ 上均匀分布，并设 $Y = X^2$。那么 $\mathrm{E}(XY) = \mathrm{E}(X^3) = 0 = \mathrm{E}(X)\mathrm{E}(Y)$，然而 Y 完全由 X 决定。协方差只量度线性关联。

廉价的随机性

对电脑来说，产生随机数的成本很高。1979 年，Larry Carter 和 Mark Wegman 证明，对许多算法而言，例如把数据储存在伪随机位置的*杂凑*（hashing），两两独立的随机选择与完全独立的同样有效。两两独立的数值可以用极低的成本产生：例如，取随机的 a 和 b，则 $a + bk \pmod{p}$ 对 $k = 0, 1, \ldots, p-1$ 给出 p 个数值，其中任何两个都是独立的。本题的方差计算正是它行得通的原因：切比雪夫（Chebyshev）不等式只需要方差，而方差只需要看成对的变量。

思维技巧

分解为指示变量

要求一个计数的平均值和方差，可把它写成指示变量之和，每一个可能被数到的对象对应一个指示变量。求平均值只需要考虑单个指示变量；求方差则需要考虑成对的指示变量。

这是概率论中最有用的技巧之一，早在 Section 9.8 已经出现过。在大学里，随机图、算法和统计检验的分析都以它为基础。

延伸练习

1 **两个值。**试做 STEP III 2000，第 14 题（`step.maths.org/questions/00-s3-q14`）。该题证明，对于各自只取两个值的随机变量，$\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$ 必然推出独立，并要求给出一个取三个值的反例。

2 **人数更多的队伍。**在第 (i) 部分中取 $n = 7$，求得分的平均值和标准差，并把例子中的得分 7 与它们比较。

3 **再谈生日。**利用指示变量，求 k 个人当中生日相同的人数对的平均值和方差（Section 9.9）。这些指示变量是否两两独立？

4 **三者成立但两两不独立。**找出三个事件，每个的概率都是 $\frac{1}{2}$，满足 $\mathrm{P}(A \cap B \cap C) = \frac{1}{8}$，但并非两两独立。

延伸阅读

- **Geoffrey Grimmett & David Stirzaker, *Probability and Random Processes***（Oxford University Press），开首几章讨论独立性。
- **Michael Mitzenmacher & Eli Upfal, *Probability and Computing***（Cambridge University Press）。两两独立及其在算法中的应用。
- **Joseph Blitzstein & Jessica Hwang, *Introduction to Probability***（CRC Press；可于网上免费阅读），关于指示变量与协方差。

12.5 最后一位乘客的座位

题目

STEP II 2021，第 11 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2021 年 STEP 试卷（zip 文件）内的 STEP II 试卷：

`www.ocr.org.uk/Images/696588-2021-past-papers.zip`

引子

一百名乘客登上一班满座的飞机。第一位乘客遗失了登机证，于是随便找个座位坐下。其后每位乘客，若自己的座位仍然空着就坐自己的座位，否则就在余下的空位中随机选一个。最后一位乘客坐到自己座位的概率（probability）是多少？混乱或会蔓延至整架飞机，所以你可能预期这个概率很小；又或者大部分乘客根本不受影响，概率应该接近 1。

无论乘客人数是多少，答案都恰好是 $\frac{1}{2}$。本题以递推关系（recurrence）和归纳法（induction）证明这个结果。不过，还有一个利用对称性（symmetry）的漂亮论证，一句话就能说完；它解释了为甚么答案是 $\frac{1}{2}$，更可让你立即写出任何一位乘客的答案。

动笔之前

停一停，试一试 列出所有可能发生的情况，求出 P_2 和 P_3。

若有两个座位，T_1 会坐 S_1（这时 T_2 坐到 S_2）或 S_2，两者的概率各为 $\frac{1}{2}$：$P_2 = \frac{1}{2}$。若有三个座位：如果 T_1 坐 S_1（概率为 $\frac{1}{3}$），人人都没有问题；如果 T_1 坐 S_3，T_3 就坐不到自己的座位；如果 T_1 坐 S_2，则 T_2 要在 S_1 和 S_3 之间随机选择，而 T_3 只有在 T_2 选中 S_1 时才没有问题。所以 $P_3 = \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{2}$。

逐步解题

阶段 1 递推关系

想一想 假设 T_1 坐了 S_k，其中 $2 \le k \le n-1$。乘客 $T_2, \dots, T_{k-1}$ 会怎样？当 T_k 到达时，哪些座位是空的？这个局面与一列较小的列车相比又如何？

乘客 $T_2, \dots, T_{k-1}$ 发现自己的座位空着，便坐了下来。当 T_k 到达时，空着的座位是 S_1 和 $S_{k+1}, \dots, S_n$：共有 $n-k+1$ 个。T_k 从中随机选一个，此后每位乘客若自己的座位空着就坐自己的座位，否则就随机选一个。这正是有 $n-k+1$ 个座位的原来问题，T_k 扮演第一位乘客的角色，而 S_1 则扮演 T_k“自己的”座位的角色（即一旦被选中，一切便回复正常的那个座位）。因此，T_n 坐到 S_n 的条件概率（conditional probability）是 P_{n-k+1}。

现在按 T_1 所坐的座位分情况考虑，每个座位的概率都是 $\frac{1}{n}$：坐 S_1 必定成功，坐 S_n 必定失败，而坐 S_k 则给出 P_{n-k+1}。所以

$$P_n = \frac{1}{n}\Big(1 + 0 + \sum_{k=2}^{n-1} P_{n-k+1}\Big) = \frac{1}{n}\Big(1 + \sum_{r=2}^{n-1} P_r\Big). \qquad \blacksquare$$

阶段 2 答案

想一想 $P_2 = P_3 = \frac{1}{2}$。猜出 P_n，并用强归纳法（strong induction）证明。

断言：对所有 $n \ge 2$，$P_n = \frac{1}{2}$。当 $n = 2$ 时断言成立。若 $P_r = \frac{1}{2}$ 对 $2 \le r \le n-1$ 成立，则

$$P_n = \frac{1}{n}\Big(1 + \frac{n-2}{2}\Big) = \frac{1}{n} \cdot \frac{n}{2} = \frac{1}{2}. \qquad \blacksquare$$

阶段 3 倒数第二位乘客

想一想 试找一个更巧妙的论证。到最后，留给 T_n 的是哪个座位？有可能是 S_2 吗？

最后留给 T_n 的座位不是 S_1 就是 S_n。它不可能是 S_k（其中 $2 \le k \le n-1$），因为 T_k 早就会把它坐了。而在此之前的每一次随机选择，对 S_1 和 S_n 都是对称的：每当有被占了座位的乘客要选位，S_1 和 S_n 被选中的机会均等。两者之中先被选中的那个决定了结局，所以两者同样可能成为最后剩下的座位：$P_n = \frac{1}{2}$。

同一个论证适用于任何一位乘客 T_k，$k \ge 2$。只有当某位较早的乘客占了 T_k 的座位，他们才会失去自己的座位。座位被占的连锁反应会一直持续，直至有人选中 S_1 为止；在 $S_1, S_k, S_{k+1}, \dots, S_n$ 这些座位之中，T_k 到达之前的选位者对它们一视同仁，所以最先被选

中的座位同样可能是这 $n-k+2$ 个座位中的任何一个，而 T_k 只有在最先被选中的是 S_k 时才会失去座位。所以

$$\mathrm{P}(T_k \text{ 坐在} S_k) = \frac{n-k+1}{n-k+2}.$$

当 $k = n-1$（其中 $n \geq 3$）时，这给出 $Q_n = \frac{2}{3}$；而 $Q_2 = \frac{1}{2}$，因为这时 $T_{n-1} = T_1$，他们坐在 S_1 的概率是 $\frac{1}{2}$。 ■

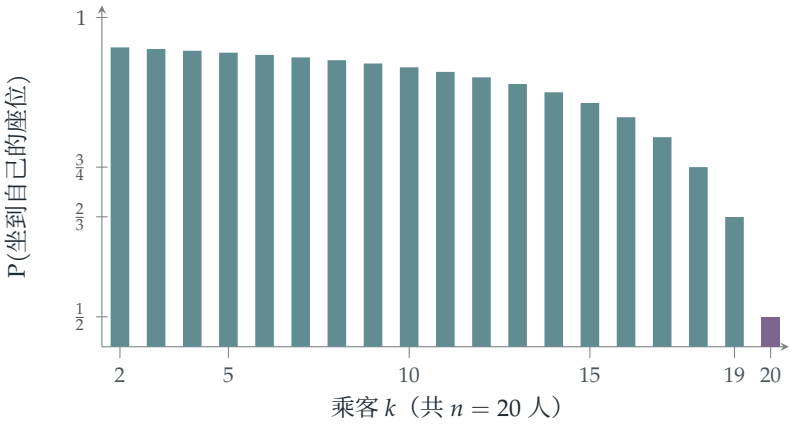


图 12.5.1 乘客 k 坐到自己座位的概率，列车共有 $n = 20$ 个座位。当 $2 \leq k \leq n$ 时，概率是 $\frac{n-k+1}{n-k+2}$：由 T_2 的 $\frac{19}{20}$ 一直下降至 T_{19} 的 $\frac{2}{3}$，以及最后一位乘客的 $\frac{1}{2}$（紫红色）。

回顾

- **化为较小的同类问题。**在 T_1 坐了 S_k 之后，余下的问题就是座位较少的同一个问题。
- **强归纳法。**递推关系牵涉所有较小的情况，所以归纳时要假设结果对所有这些情况都成立。
- **找出对称性。**座位 S_1 和 S_n 自始至终可以互换，这样不用任何代数运算就解释了 $\frac{1}{2}$ 的由来。

背后的数学

对称论证

第二个证明体现了概率论中一个威力强大的想法：若两个结果可由随机过程的某种对称性互相对应，它们的概率就相同。这里的过程在 S_1 和 S_n 其中一个被选中之前，并不区分这两个座位，所以两者成为最后剩下的座位的可能性必定相等。对称论证往往只凭一

个观察，就能取代冗长的计算。接下来的几节有很多这样的例子：Section 12.6 中多掷一次硬币，Section 12.7 中的淘汰赛抽签。

有多少位乘客被迫换位?

被迫换位的乘客（即坐不到自己座位的乘客）的数目同样有很好的规律。T_1 除非选中 S_1，否则就算是换了位；而 T_k（其中 $k \geq 2$）被迫换位的概率是 $\frac{1}{n-k+2}$。把这些概率相加，被迫换位的乘客数目的期望值（expected number）是

$$\left(1-\frac{1}{n}\right)+\left(\frac{1}{n}+\frac{1}{n-1}+\cdots+\frac{1}{2}\right)=H_n-\frac{1}{n},$$

大约是 $\ln n$：即使飞机上有 300 个座位，这场混乱平均也只影响约六位乘客。调和数（harmonic numbers）又再出现了（Section 9.11）。

思维技巧

寻找对称性

如果随机过程无法分辨两个结果，它们的可能性就相等。找出它们最初变得可以分辨的那一刻。

在大学，对称论证在概率论中随处可见，由反射原理（reflection principle）（Section 9.10）到可交换性（exchangeability）皆是；可交换性是指一列随机变量（random variables）无论以任何次序排列都有相同的分布，这个性质是不少现代统计学的基础。

延伸练习

1 **两位健忘的乘客。**如果最先的两位乘客都随机选位，最后一位乘客坐到自己座位的概率是多少?

2 **被迫换位的乘客。**列出所有可能情况，验证 $n=3$ 时被迫换位的乘客数目的期望值。

3 **迂回的路线。**利用递推关系，证明 $\sum_{r=2}^{n-1} P_r = \frac{n-2}{2}$ 与对称论证互相吻合。

4 **最后一个座位。**证明在最后一位乘客坐不到自己座位的条件下，最后一位乘客坐到 S_1 的概率是 1。为甚么?

延伸阅读

- **Peter Winkler, *Mathematical Puzzles: A Connoisseur's Collection***（A K Peters）。收录了飞机座位谜题，以及许多其他可用对称性解决的谜题。
- **Frederick Mosteller, *Fifty Challenging Problems in Probability***（Dover）。一本经典的题集。
- **Joseph Blitzstein & Jessica Hwang, *Introduction to Probability***（CRC Press；可于网上免费阅览）。

12.6 多一枚硬币

题目

STEP I 2016，第 12 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/16-s1-q12`

引子

Bob 比 Alice 多掷一枚硬币。他掷得的正面（heads）比 Alice 多的概率（probability）是多少？他占了优势，但多掷一次不保证多一个正面，而且平手算他输。用硬算的方法做出头两个情况，两次的答案都是 $\frac{1}{2}$。这是巧合吗？

不是巧合，而且有一个只需一句话的证明：只要留意到 Bob 掷得的正面或反面（tails）必有其一比 Alice 多，但不会两者皆多。这道题目是一个典范，示范了计算如何提示出规律，而巧妙的对称性（symmetry）又如何解释这个规律。

动笔之前

想一想 Alice 掷两次：她掷得正面的次数的分布（distribution）是甚么？掷三次的 Bob 又如何？

Alice 掷得 $0, 1, 2$ 个正面的概率分别是 $\frac{1}{4}, \frac{2}{4}, \frac{1}{4}$；Bob 掷得 $0, 1, 2, 3$ 个正面的概率分别是 $\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8}$。

逐步解题

阶段 1 硬算

第 (i) 部分，按 Alice 掷得的正面数目逐项相加：

$$\begin{aligned}\mathrm{P}(B>A) &= \tfrac{1}{4}\mathrm{P}(B\ge 1)+\tfrac{2}{4}\mathrm{P}(B\ge 2)+\tfrac{1}{4}\mathrm{P}(B=3)\\ &= \tfrac{1}{4}\cdot\tfrac{7}{8}+\tfrac{1}{2}\cdot\tfrac{4}{8}+\tfrac{1}{4}\cdot\tfrac{1}{8}=\frac{7+8+1}{32}=\frac{1}{2}.\end{aligned}$$

第 (ii) 部分是同样的计算，只是项数较多；数一数 $8\times 16=128$ 个等可能的结果，其中 64 个是 Bob 领先（图 12.6.1），所以答案同样是 $\frac{1}{2}$。

阶段 2 拆开 Bob 的投掷

想一想 把 Bob 的 $n+1$ 次投掷看成先掷 n 次，再额外掷一次。掷完头 n 次之后，他可能领先 Alice、与她平手，或者落后。每种情况下会发生甚么事?

二人各掷 n 次之后，Bob 领先的概率是 p_2，平手的概率是 p_1，落后的概率是 p_2（由 Alice 与 Bob 之间的对称性可知）。如果他领先，他会保持领先。如果平手，他额外的一掷有 $\frac{1}{2}$ 的概率令他领先。如果他落后，多一个正面最多只能追成平手。所以

$$\mathrm{P}(\text{Bob 的正面较多})=p_2+\tfrac{1}{2}p_1.$$

由于这三种情况涵盖了所有可能，$2p_2+p_1=1$，因此 $p_2+\frac{1}{2}p_1=\frac{1}{2}$。多了一枚硬币，Bob 掷得较多正面的概率对每个 n 都刚好是 $\frac{1}{2}$。 ■

停一停，试一试 以下就是那个一句话的证明。Bob 掷 $n+1$ 次，所以他的正面和反面数目加起来是 $n+1$，比 Alice 的多一。试证明 Bob 的正面比 Alice 多，当且仅当他的反面不比 Alice 多。由于正面与反面是对称的，这两个事件的概率相等，所以各为 $\frac{1}{2}$。

事实上，记 $H_A+T_A=n$ 及 $H_B+T_B=n+1$：若 $H_B>H_A$，则 $H_B\ge H_A+1$，所以 $T_B=n+1-H_B\le n-H_A=T_A$；反过来，若 $T_B\le T_A$，则 $H_B\ge H_A+1$。所以“正面较多”与“反面较多”是互补（complementary）的，而把正面与反面对调（由于硬币是公平的，这不会改变概率）就会把两者互换。

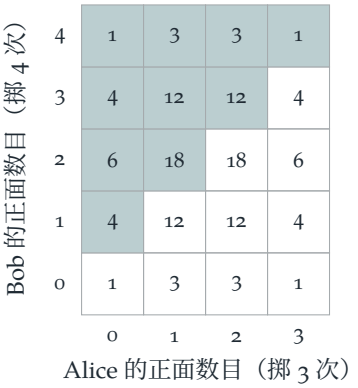


图 12.6.1 第 (ii) 部分：在 128 个等可能的结果之中，每种正面数目组合所占的个数。Bob 在有底色的格子中领先，这些格子刚好包含一半的结果。

回顾

- **计算小的情况。**两次计算都提示答案是 $\frac{1}{2}$。
- **按自然的划分来取条件。**把 Bob 额外的一掷分开，问题便化为一场势均力敌的比试再加一枚硬币。
- **找出互补的一对。**“正面较多”与“反面较多”互为补事件，并且在一个对称之下互换。

背后的数学

对称与补事件

那个漂亮的证明结合了两个事实：在一个对称之下互换的两个事件概率相等，而两个互补事件的概率之和为 1。若两者同时成立，则每个事件的概率都是 $\frac{1}{2}$。同样的推理在概率论中随处可见。在航班座位问题（Section 12.5）中，S_1 与 S_n 在一个对称之下互换，而最后剩下的恰好是其中一个。在六贯棋（Hex）中，由棋盘的对称性，加上恰有一方必胜这个事实，可知先手一方占优，不过在大棋盘上无人知道必胜策略是甚么。

平手的可能性有多大？

当二人投掷次数相同时，平手的概率 p_1 就很重要。Alice 与 Bob 平手，即 Alice 的正面数目等于 Bob 的正面数目；利用对称性，把 Bob 的正面换成他的反面，这个概率就等于 Alice 的正面数目加上 Bob 的正面数目等于 n 的概率，亦即在 $2n$ 次投掷中共有 n 个正面。

所以

$$p_1 = \binom{2n}{n}\frac{1}{4^n} \approx \frac{1}{\sqrt{\pi n}},$$

这里用了斯特灵近似（Stirling's approximation）（Section 12.23）。平手会变得罕见，但变得很慢：各掷 100 次，平手的机会仍然约有 5.6%。

思维技巧

先计算，再猜想，然后解释

当计算得出一个简单得可疑的答案时，就要找出原因。简单的答案通常有简单的解释。

在大学里，这个习惯正是理解与计算的分野。这里的第一个证明确认了答案；第二个证明则解释了答案，而且可以立即推广（试做练习 2）。

延伸练习

1 **多掷两枚硬币。**若 Bob 掷 $n+2$ 枚硬币而 Alice 掷 n 枚，证明 $\mathrm{P}(\text{Bob 的正面较多}) = \frac{1}{2} + \frac{1}{2}\mathrm{P}(H' = H_A)$，其中 H' 是 Bob 头 $n+1$ 次投掷中的正面数目。

2 **不公平的硬币。**若二人所用的硬币掷出正面的概率都是 p，多掷一次时答案是否仍然是 $\frac{1}{2}$？那个漂亮的证明中哪一步不再成立？

3 **平手。**用正文中的技巧证明 $p_1 = \binom{2n}{n}/4^n$，并求它在 $n = 1, 2, 3$ 时的值。

4 **骰子。**Alice 掷 n 粒骰子，Bob 掷 $n+1$ 粒。Bob 掷得较多六点的概率是否等于 $\frac{1}{2}$？

延伸阅读

- **Peter Winkler, *Mathematical Mind-Benders***（A K Peters）。收录大量可用对称性解决的概率谜题。
- **Gabor Székely, *Paradoxes in Probability Theory and Mathematical Statistics***（Reidel）。
- **Frederick Mosteller, *Fifty Challenging Problems in Probability***（Dover）。

12.7 淘汰赛中谁会遇上谁

题目

STEP II Specimen paper，第 16 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/spec-s2-q16`

引子

温布顿网球锦标赛（Wimbledon）有 128 名选手参加抽签。随机选出其中两人：他们在赛事中某个阶段相遇的机会有多大？你也许会尝试把每一轮相遇的概率（probability）加起来，而每一项都是“双方晋级”与“被抽在一起”这些机会的乘积。这样做是可行的，但有一个好得多的方法。数一数比赛的场数：每场比赛都由一对选手对赛，而比赛恰好有 127 场。根据对称性（symmetry），每一对选手成为其中任何一场的对赛双方的可能性都相同。因此所求的机会就是 127 除以选手配对的总数。

这道题是“数对了东西，概率问题便迎刃而解”的绝佳例子。答案 $\frac{1}{2^{n-1}}$ 只需一行便可得出。

动笔之前

想一想 有 2^n 名选手的淘汰赛（knockout tournament）共要进行多少场比赛？（每场比赛淘汰一名选手。）

每场比赛恰好淘汰一名选手，而除了一人之外所有选手都被淘汰，所以共有 $2^n - 1$ 场比赛。

逐步解题

阶段 1 第一轮

把选出的两名选手称为 X 和 Y。在第一轮，X 的对手是其余 2^n-1 名选手中任何一人的可能性都相同，所以

$$\mathrm{P}(\text{相遇于轮次1}) = \frac{1}{2^n-1}.$$

阶段 2 关键的对称性

> **想一想** 每一轮的抽签都是随机的，而每场比赛都相当于掷一次公平的硬币。这个过程对待某一对选手的方式，会否与对待另一对有所不同？

不会。由于抽签是随机的，而且每场比赛的胜负机会都是 50–50，把选手重新标号并不会改变任何概率。所以任何一场指定的比赛（例如决赛）由 $\binom{2^n}{2}$ 对选手中任何一对对赛的可能性都相同。

阶段 3 决赛

决赛是一场特定的比赛，所以它由 X 和 Y 对赛的概率是

$$\frac{1}{\binom{2^n}{2}} = \frac{2}{2^n(2^n-1)}.$$

（验算：两人都必须打进决赛，这需要他们分处不同的半区，并且各自胜出 $n-1$ 场比赛。）

阶段 4 在任何阶段

> **想一想** X 和 Y 最多只能相遇一次。这一点怎样让你可以把各场比赛的概率相加？

设 I_m 在第 m 场比赛由 X 和 Y 对赛时取值 1。在 2^n-1 场比赛中，每一场都有 $\mathrm{P}(I_m=1) = \binom{2^n}{2}^{-1}$，而这些事件是互斥（mutually exclusive）的（他们不可能相遇两次），所以

$$\mathrm{P}(\text{相遇}) = \frac{2^n-1}{\binom{2^n}{2}} = \frac{2(2^n-1)}{2^n(2^n-1)} = \frac{1}{2^{n-1}}. \qquad \blacksquare$$

以温布顿为例，$n=7$，两名指定选手相遇的概率是 $\frac{1}{64}$。作为验算，在 2^n-1 场比赛中，第一轮占了 2^{n-1} 场，由此得出 $\frac{2^{n-1}}{\binom{2^n}{2}} = \frac{1}{2^n-1}$，与第 1 阶段的结果一致。

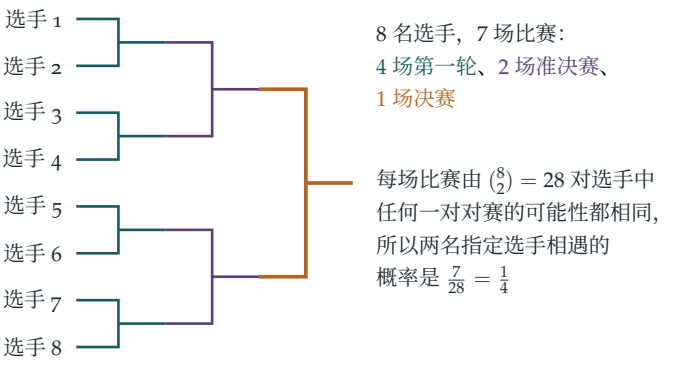


图 12.7.1 $2^3 = 8$ 名选手的淘汰赛（图中画成固定的赛程表，但题目中每一轮都会重新随机抽签）。比赛共有 7 场。根据对称性，每一场由 28 对选手中任何一对对赛的可能性都相同，所以选出的两名选手相遇的概率是 $\frac{7}{28} = \frac{1}{4} = \frac{1}{2^{n-1}}$。

回顾

- **数比赛的场数。**比赛的场数 $2^n - 1$ 是关键的事实。
- **各对选手之间的对称性。**每一对选手成为任何一场指定比赛的对赛双方的可能性都相同。
- **对各场比赛的线性性质。**由于两名选手最多相遇一次，相遇的概率就是相遇次数的期望值，也就是对各场比赛求和。

背后的数学

期望次数

第 4 阶段的论证其实是一个期望值（expectation）的计算：X 和 Y 相遇的次数是若干个指示变量（indicator）之和，每场比赛对应一个；而它的期望值（由于它只取 0 或 1，所以等于相遇的概率）就是各个指示变量的期望值之和。这就是期望值的线性性质（linearity of expectation），在 Section 9.11 和 Section 12.4 中都曾用到。它最大的好处是完全不必理会各轮之间错综复杂的相依关系。

设有种子的赛事

现实中的赛事设有种子（seeded）：最强的选手会被编排在适当的位置，使他们要到后段轮次才可能相遇。而且现实中的比赛也不是 50–50 的。这两点都破坏了对称性，于是计算两名指定选手相遇的机会便成了一项货真价实的运算。为体育运动建立模型的统计学

家正是运用这类计算来预测赛事，当中每场比赛的概率都是根据选手的评分估计出来的。对称的情况则是一条基线，用来衡量各种现实因素的影响。

思维技巧

数对的东西

只要选对了对象，概率往往可以用（有利对象的数目）÷（等可能对象的数目）求得。在这里，对象就是比赛和选手配对。

选择数甚么，正是组合概率（combinatorial probability）的艺术。在大学里，同一个想法会以算两次（double counting）的形式出现（Section 7.9），也会出现在概率方法（probabilistic method）之中：在那里，某种结构的期望数目可以用来证明它存在。

延伸练习

1 **逐轮计算。**直接证明 X 和 Y 在第 k 轮相遇的概率是 $\frac{2^{n-k}}{\binom{2^n}{2}}$，并验证这些概率加起来等于 $\frac{1}{2^{n-1}}$。

2 **固定的赛程表。**假如赛程表改为预先固定（如图所示），但所有选手都被随机编入其中，答案会否改变？

3 **一名强手。**假设选手 X 每场比赛都胜出。现在 X 与 Y 相遇的概率是多少？

4 **联赛。**在一个有 $2n$ 支球队的联赛中，每两队都对赛一次，共有多少场比赛？相应的“相遇概率”是甚么？

延伸阅读

- **Frederick Mosteller, *Fifty Challenging Problems in Probability***（Dover），书中收录了有关淘汰赛的问题。
- **David Spiegelhalter, *The Art of Statistics***（Pelican），讲述建立模型与预测。
- **Joseph Blitzstein & Jessica Hwang, *Introduction to Probability***（CRC Press；可于网上免费阅读），讲述指示变量与对称性。

12.8 Penney 游戏

题目

STEP III 2001，第 13 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/01-s3-q13`

引子

你和朋友各自选定一个由三次掷硬币结果组成的序列，然后不断掷硬币，直至其中一个序列出现为止。在任何一个时刻，每个三次投掷的序列出现的机会都相同，那么无论选哪一个，理应同样好吧？完全不是。无论朋友选哪个序列，你都可以选出一个胜过它的序列，胜算（odds）有时更高达 7 比 1。

这就是 *Penney 游戏*（Penney's game），由 Walter Penney 于 1969 年发明。本题要证明 A 胜 B，B 胜 C，C 胜 D，而 D 又胜 A：这是一个循环，就像包剪揼一样。“胜过”这种关系并不具传递性（transitive）。长远而言，后选的一方总能胜出。

动笔之前

想一想 假设至今的投掷结果以 TT 结尾，而且仍未有人胜出。THH 有没有可能比 TTH 先出现？

从 TT 出发，之后每掷出一个 T，最后两次投掷仍然是 TT，而第一个 H 一出现就凑成 TTH。要 THH 出现，必须先有 TH，而 TH 只能由一个 T 后接一个 H 而来，但这个 H 已经凑成了 TTH。所以 TT 一旦出现，A 必胜。

逐步解题

阶段 1 状态

想一想 两个序列都不可能在两次投掷内完成。掷了两次之后,从每个状态(state)出发，下一次投掷会带来甚么结果?

由于游戏只取决于最后两次投掷，我们可以追踪状态（即最后两次投掷的结果）：

- 从 TT 出发：A 胜（上面已证），所以 $p_{\mathrm{TT}} = 1$。
- 从 HT 出发：下一次投掷会得出 TH 或 TT，所以 $p_{\mathrm{HT}} = \frac{1}{2}p_{\mathrm{TH}} + \frac{1}{2}p_{\mathrm{TT}}$。
- 从 TH 出发：掷出 H 便凑成 THH（B 胜）；掷出 T 则得出 HT。所以 $p_{\mathrm{TH}} = \frac{1}{2}\cdot 0 + \frac{1}{2}p_{\mathrm{HT}}$。
- 从 HH 出发：掷出 H 仍是 HH；掷出 T 则得出 HT。所以 $p_{\mathrm{HH}} = \frac{1}{2}p_{\mathrm{HH}} + \frac{1}{2}p_{\mathrm{HT}}$，由此得 $p_{\mathrm{HH}} = p_{\mathrm{HT}}$。

阶段 2 求解

把 $p_{\mathrm{TH}} = \frac{1}{2}p_{\mathrm{HT}}$ 代入 HT 的方程：$p_{\mathrm{HT}} = \frac{1}{4}p_{\mathrm{HT}} + \frac{1}{2}$，所以 $p_{\mathrm{HT}} = \frac{2}{3}$、$p_{\mathrm{HH}} = \frac{2}{3}$ 及 $p_{\mathrm{TH}} = \frac{1}{3}$。四个起始状态的概率（probability）各为 $\frac{1}{4}$，所以

$$\mathrm{P}(A\text{ 胜}) = \tfrac{1}{4}\left(\tfrac{2}{3} + \tfrac{2}{3} + \tfrac{1}{3} + 1\right) = \tfrac{2}{3}. \qquad \blacksquare$$

阶段 3 B 对 C

想一想 C 想要 HHT，B 想要 THH。C 在甚么情况下胜出?（留意第一个 HH 之前紧接着的必定是甚么。）

如果头两次投掷是 HH，那么其后第一个 T 便凑成 HHT，THH 来不及出现：C 胜。否则，在 HH 首次出现之前的某处必有一个 T，因此第一个 HH 的前一次投掷必定是 T：THH 就在那一刻出现，早于 HHT（HHT 还需要在 HH 之后再掷出一个 T）。所以只有当游戏以 HH 开始时 C 才会胜出：

$$\mathrm{P}(B\text{ 胜}) = \tfrac{3}{4}. \qquad \blacksquare$$

阶段 4 循环

用同样的方法（或“背后的数学”一节的捷径）可以证明 C（HHT）胜 D（HTT）的概率是 $\frac{2}{3}$，而 D（HTT）胜 A（TTH）的概率是 $\frac{3}{4}$；后者正是第 3 阶段把正面与反面对调后的镜像。所以

$$A\text{ 胜}B, \qquad B\text{ 胜}C, \qquad C\text{ 胜}D, \qquad D\text{ 胜}A. \qquad \blacksquare$$

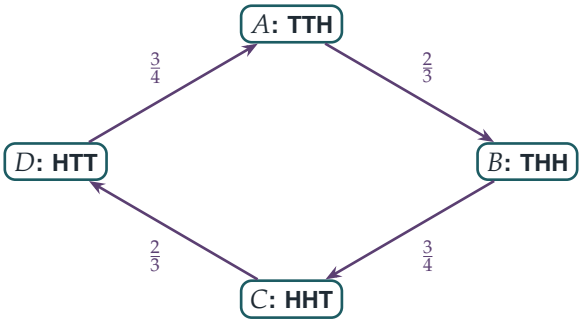


图 12.8.1 Penney 循环。由 X 指向 Y 的箭头表示 X 的序列胜过 Y 的序列，箭头旁为相应的概率。每个序列都被另一个序列击败，所以后选的一方总能挑出胜算较高的序列。

回顾

- **选对状态。**只有最后两次投掷才有影响，所以四个状态、四条方程便概括了整个游戏。
- **首步分析（first-step analysis）。**每条方程都是以下一次投掷为条件：从某状态出发的概率，就是下一次投掷所到达的各个状态的概率的平均值。
- **找出决定性的事件。**在第 3 阶段，“第一个 HH 之前是 T”这一点无须列方程便决定了胜负。

背后的数学

为甚么有些序列较好

在任何一个位置，所有三次投掷的序列出现的机会都相同，而每个序列的期望等待时间（expected waiting time）也不难计算。但各序列的等待时间并不相同：TTH 平均要等 8 次投掷，TTT 却要等 14 次，因为与自身重叠的序列（例如 TTT，只差一步的 TTH 会令已累积的两次投掷前功尽废）出现得较慢。而在两个序列的竞赛中，关键在于每个序列与对方如何重叠。THH 胜 HHT，是因为 HHT 每次出现（除了在最开头之外），前面都有一个 T，这表示 THH 刚刚已经出现。后选一方的窍门总是一样：取对手序列的头两个符号，再在前面加上与对手中间符号相反的那个符号。

康威公式（Conway's formula）

约翰·康威发现了一条计算胜算的绝妙公式。对于长度为 n 的两个序列 X 和 Y，定义相关数（correlation）XY 为一个 n 位的二进制数：由左至右，若 X 中由第 i 个符号开始的部

分与 Y 开头相同长度的部分相同，则第 i 位为 1。那么 Y 比 X 先出现的胜算是

$$(XX - XY) : (YY - YX).$$

取 $X = \text{HHT}$ 及 $Y = \text{THH}$：$XX = 100_2 = 4$（只有整个 HHT 与自身吻合）；$XY = 001_2 = 1$（只有 HHT 末尾的 T 与 THH 的开头吻合）；$YY = 100_2 = 4$；而 $YX = 011_2 = 3$（THH 末尾的 HH 和 H 都与 HHT 的开头吻合）。所以 THH 先出现的胜算是 $(4-1) : (4-3) = 3 : 1$，即第 3 阶段的 $\frac{3}{4}$。康威公式适用于任何长度、任何字母表的序列，其证明用到鞅（martingale）的理论：想像有一间赌场，赌客就序列的出现下注，再利用公平游戏的期望利润为零这一事实。

Penney 与加德纳（Gardner）

Walter Penney 于 1969 年在 *Journal of Recreational Mathematics* 的一篇短文中介绍了这个游戏。马丁·加德纳于 1974 年在他的 *Scientific American* 专栏中介绍了它，使它广为人知，文中还包括康威公式，这条公式是康威发现的，但他一直没有发表。这类非传递关系（埃弗龙骰子（Efron's dice）是另一个例子：骰子 A 胜 B，B 胜 C，C 胜 D，而 D 胜 A）一再令人惊讶，在投票理论中也很重要，因为多数人的偏好可以形成循环：这就是孔多塞悖论（Condorcet's paradox）。

思维技巧

首步分析

若一个过程在各状态之间转移，要求出与它有关的概率或期望值（expectation），可以第一步为条件。这样每个状态都给出一条线性方程。

这是处理马可夫链（Markov chain）的基本技巧，你会在 Section 12.16 以及大学概率课程中一再遇到。同样的方法可以求出先到达某一状态而非另一状态的概率、被吸收前的期望时间，以及赌博问题中的破产概率。

延伸练习

1 **四人同场。**试做 STEP II 2015，第 12 题（`step.maths.org/questions/15-s2-q12`）：四名玩家同时参与的 Penney 游戏。

2 **等待时间。**证明直至 HH 首次出现的期望投掷次数是 6，而直至 HT 首次出现的则是 4。（使用首步分析。）

3 **最佳对策。**对于八个长度为三的序列，逐一找出以最大概率胜过它的序列。哪个序列最容易被击败?

4 **康威公式。**用康威公式验证 HHT 胜 HTT 的概率是 $\frac{2}{3}$。

延伸阅读

- **Martin Gardner, *Time Travel and Other Mathematical Bewilderments***（W. H. Freeman）。书中包括 Penney 游戏和康威公式。
- **Ronald Graham, Donald Knuth & Oren Patashnik, *Concrete Mathematics***（Addison-Wesley），第 8.4 节。以生成函数（generating function）处理 Penney 游戏。
- **David Williams, *Probability with Martingales***（Cambridge University Press）。以鞅给出的证明，留待日后再读。

12.9 用不公平的硬币作公平的决定

题目

STEP I 2019，第 11 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2019 年 STEP 试卷（zip 文件）内的 STEP I 试卷：

`www.ocr.org.uk/Images/696590-2019-past-papers.zip`

引子

你要作出一个公平（fair）的决定，但手上唯一的硬币是弯的，而且你不知道它偏差（bias）有多大。这样还能公平地作出选择吗？冯纽曼（John von Neumann）在 1951 年给出的答案是：掷两次。若结果是 HT，就选酒店；若是 TH，就选茶室；若是 HH 或 TT，就再掷两次。由于 HT 和 TH 的概率（probability）都是 pq，所以不论 p 是多少，这个决定都完全公平。

本题的程序刚好相反：它在出现 HH 和 TT 时作决定，这样做很快，却不公平（选酒店的概率是 $\frac{p^2}{p^2+q^2}$）。本题要证明的是，不论硬币如何，这个程序都不会拖得很久。比较这两个程序，可以看到产生随机性时的一个基本取舍（trade-off）：速度与公平之间的取舍。

动笔之前

想一想 在一轮之中，作出决定的概率是多少？没有决定的概率又是多少？$2pq$ 最大可以取甚么值？

若两枚硬币结果相同，便作出决定，概率是 p^2+q^2；否则他们继续下去，概率是 $2pq$。由于 $pq = p(1-p) \le \frac{1}{4}$（当 $p = \frac{1}{2}$ 时等号成立），所以 $2pq \le \frac{1}{2}$。

逐步解题

阶段 1 在第 n 轮作出决定

各轮之间互相独立（independent）。要在第 n 轮作出决定，须先有 $n-1$ 轮结果不同，然后一轮结果相同：

$$\mathrm{P}(\text{作出决定的轮次为}n) = (2pq)^{n-1}(p^2+q^2). \qquad \blacksquare$$

阶段 2 在第 n 轮或之前作出决定

? **想一想** 在首 n 轮中没有作出决定的概率是多少?

n 轮之内没有决定的概率是 $(2pq)^n \le \left(\frac{1}{2}\right)^n$，所以不论 p 是多少，在第 n 轮或之前作出决定的概率至少是 $1-\frac{1}{2^n}$。 $\blacksquare$

阶段 3 三个人

? **想一想** 第一轮出现两个正面和一个反面之后，第三个人把反面翻成正面。第二轮要发生甚么事才会有决定？如果第一轮是两个反面和一个正面呢?

第一轮作出决定的概率是 p^3+q^3。否则，有两枚硬币结果相同，另一枚不同。若多数是正面（概率为 $3p^2q$），那枚不同的硬币会被翻成正面，而第二轮要作出决定，就需要另外两人都掷出正面：概率是 p^2（此时已不可能全是反面）。同理，若多数是反面（概率为 $3pq^2$），要作出决定就需要两个反面：q^2。所以

$$\mathrm{P}(\text{作出决定的轮次不迟于}2) = p^3+q^3+3p^2q\cdot p^2+3pq^2\cdot q^2.$$

? **想一想** 把所有东西用 $s=pq$ 表示。（利用 $p+q=1$。）

我们有 $p^3+q^3 = (p+q)^3-3pq(p+q) = 1-3s$，以及 $3p^4q+3pq^4 = 3s(p^3+q^3) = 3s(1-3s)$。所以

$$\mathrm{P} = 1-3s+3s-9s^2 = 1-9s^2 \ge 1-\frac{9}{16} = \frac{7}{16},$$

因为 $s \le \frac{1}{4}$。 $\blacksquare$

回顾

- **补事件（complement）。**“在第 n 轮或之前至少发生一次”最容易的算法是 1−（“n 轮之内从未发生”）。
- **为最坏情况定界。**未知的 p 是靠不等式（inequality）$pq \le \frac{1}{4}$ 来处理的。
- **对称函数（symmetric functions）。**对 p 和 q 对称的式子，用 $p+q=1$ 和 pq 来表示时便会简化。

背后的数学

冯纽曼的妙计

题目中的程序很快，但有偏差。冯纽曼的程序则是公平的：掷两次，出现 HT 或 TH 就作决定，出现 HH 或 TT 就重复。每次尝试成功的概率是 $2pq$，所以平均需要 $\frac{1}{2pq}$ 次尝试；当硬币偏差很大时，这会很慢（图 12.9.1）。题目的程序每轮作出决定的概率是 $p^2+q^2 \geq \frac{1}{2}$，但当 $p > \frac{1}{2}$ 时，它的结果偏向酒店。

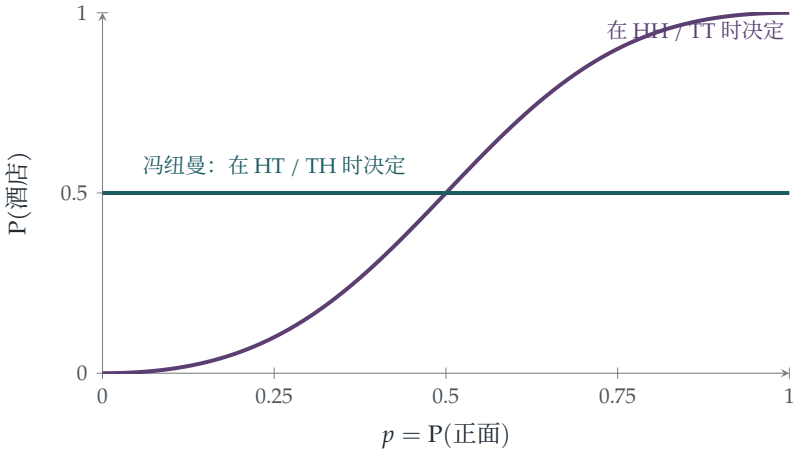


图 12.9.1 选中酒店的概率，对应硬币的偏差。在两枚硬币结果相同时作决定（紫红色）很快，但偏向较可能出现的一面。冯纽曼的程序（蓝绿色）对每一个 p 都完全公平，代价是当硬币偏差很大时需时较长。

提取随机性

冯纽曼的妙计是第一个*随机性提取器*（randomness extractor）：一种把不完美的随机性变成完美随机性的程序。它浪费很多（每两次投掷最多只得到一个公平的位元，而且往往一个也得不到）；更好的方法会把被弃掉的 HH 和 TT 重新利用，从而提取硬币所含的几乎全部随机性，而这随机性的多少是用硬币的*熵*（entropy）来量度的，即每次投掷 $-p\log_2 p - q\log_2 q$ 个位元。提取器用于电脑的硬件随机数产生器：这些产生器收集不可预测但有偏差的物理杂讯，并必须把它变成无偏差的位元，供密码学（cryptography）使用。

冯纽曼与随机数

冯纽曼在 1951 年的论文“Various techniques used in connection with random digits”中描述了这个妙计。当时他和同事正在为设计核武器而进行一些最早期的电脑模拟（即

“蒙地卡罗方法”（Monte Carlo method））。在同一篇论文中，他就用算术方法产生随机数说了一句名言：“任何人若考虑用算术方法来产生随机数字，当然就是身处罪恶之中。”

思维技巧

利用对称性消去未知量

当某个参数未知时，找出那些不论参数取何值、概率都相等的事件，并以它们作为决定的依据。

HT 和 TH 有相同的概率 pq，所以比较这两者就把 p 完全消去了。在大学里，同样的想法在统计学中以枢轴量（pivotal quantities）的形式出现，也出现于实验设计之中：在那里，随机化（randomisation）和对称性会把未知的效应抵消掉。

延伸练习

1 **轮数的期望值。**求第 (i) 部分的程序所需轮数的期望值（expected number），以及冯纽曼的程序所需轮数的期望值。对哪些 p 而言，冯纽曼的程序较慢?

2 **公平的三选一。**利用一枚有偏差的硬币，设计一个程序，从三个选项中选出一个，使每个选项的概率都恰好是 $\frac{1}{3}$。

3 **用公平硬币造出有偏差的骰子。**利用一枚公平的硬币，设计一个程序，模拟一枚掷出正面的概率恰好是 $\frac{1}{3}$ 的硬币。所需投掷次数的期望值是多少?

4 **再看三个人。**在第 (ii) 部分中，他们去酒店的概率是多少？请以 p 的函数表示。当 $p = \frac{1}{2}$ 时，这个程序公平吗?

延伸阅读

- **John von Neumann，“Various techniques used in connection with random digits”**，*National Bureau of Standards Applied Mathematics Series* 12 (1951)。篇幅短小而著名。
- **Donald Knuth，*The Art of Computer Programming*, Vol. 2**（Addison-Wesley），第 3 章，论随机数。
- **Michael Mitzenmacher & Eli Upfal，*Probability and Computing***（Cambridge University

Press），论提取随机性与熵。

12.10 把绳头系起来

题目

STEP III 2008，第 13 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/08-s3-q13`

引子

把一百条绳子放进袋中，只让绳头露在外面。随机选两个绳头绑在一起，如此重复，直至再没有松开的绳头为止。最后你会得到一堆纠缠的绳圈。究竟有多少个？一百条绳子原则上可以结成一百个小环，也可以结成一个巨大的圈。出人意表的是，绳环数目的期望值（expected number）只有约 3.3；即使有四万条绳子，也不过约 6。

只要看出一个关键之处，问题便几乎不费吹灰之力：每打一个结之后，不论刚才发生了甚么事，箱中都只是少了一条绳子。这个过程会“忘记”自己的历史，于是绳环的数目就变成一连串独立（independent）掷硬币结果之和，只是每次的概率（probability）各有不同。

动笔之前

想一想 把两个绳头绑在一起之后，还剩下多少条绳子（结成环的不再算作一“条”）？这是否取决于有没有结成环？

如果两个绳头属于同一条绳子，这条绳子便成为一个环，余下 $n-1$ 条。如果两个绳头属于不同的绳子，这两条绳子便接成较长的一条，同样余下 $n-1$ 条。无论是哪一种情况，我们都剩下 $n-1$ 条绳子，每条各有两个绳头。

逐步解题

阶段 1　第一个结

绳头共有 $2n$ 个。先选定一个；与它相绑的绳头等可能地是其余 $2n-1$ 个绳头中的任何一个，而当中恰好有一个属于同一条绳子。因此第一个结结成环的概率是 $\frac{1}{2n-1}$，而第一步所产生的绳环数目的期望值就是 $\frac{1}{2n-1}$。

阶段 2　所有的结

> ? **想一想**　根据上面的观察，打第二个结之前的情况是怎样的？第三个呢？设 I_k 为指示变量，表示在剩下 k 条绳子时所打的结是否结成环。各个 $\mathrm{E}(I_k)$ 是多少？

每次打结之前，箱中有 k 条绳子，其中 $k = n, n-1, \dots, 1$，而这个结结成环的概率是 $\frac{1}{2k-1}$。（最后一个结在只剩一条绳子时打，必定结成环。）绳环的总数是 $R = I_n + I_{n-1} + \cdots + I_1$，所以

$$\mathrm{E}(R) = \sum_{k=1}^{n} \frac{1}{2k-1} = 1 + \frac{1}{3} + \frac{1}{5} + \cdots + \frac{1}{2n-1}. \qquad \blacksquare$$

阶段 3　方差

> ? **想一想**　各个指示变量（indicator）I_k 是否独立？（某一步结成环的概率，是否取决于之前发生过的事？）

在每一步，不论之前发生过甚么事，结成环的概率都是 $\frac{1}{2k-1}$，因为当时的格局永远是“k 条绳子”。所以各个 I_k 互相独立，方差（variance）可以相加：

$$\mathrm{Var}(R) = \sum_{k=1}^{n} \frac{1}{2k-1}\left(1 - \frac{1}{2k-1}\right) = \sum_{k=1}^{n} \frac{2k-2}{(2k-1)^2}. \qquad \blacksquare$$

阶段 4　四万条绳子

> ? **想一想**　试用调和数（harmonic numbers）表示直至 $2n-1$ 为止所有奇数的倒数之和。

把直至 $2n$ 为止所有正整数的倒数之和，减去其中偶数的倒数之和，便得

$$\sum_{k=1}^{n} \frac{1}{2k-1} = H_{2n} - \tfrac{1}{2}H_n \approx \ln(2n) - \tfrac{1}{2}\ln n = \tfrac{1}{2}\ln(4n).$$

当 $n = 40\,000$ 时，$4n = 160\,000 = 20^4$，所以 $\mathrm{E}(R) \approx \frac{1}{2} \cdot 4 \ln 20 \approx 6$。$\blacksquare$（准确值是 6.28；两者之差来自欧拉常数（Euler's constant），因为 $H_n \approx \ln n + 0.577$。）

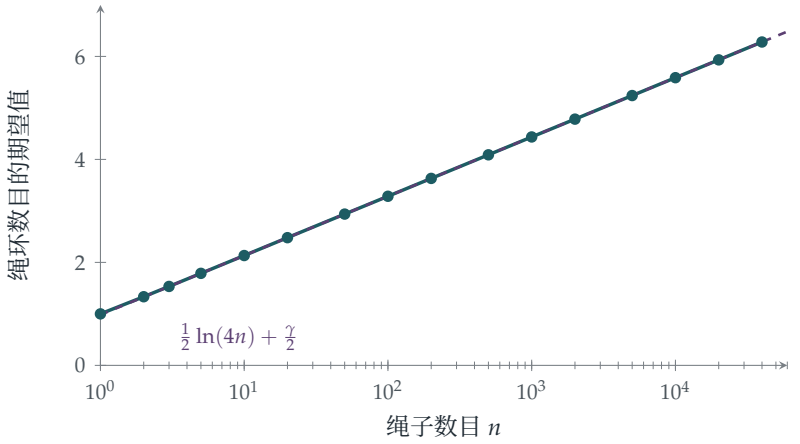


图 12.10.1 绳环数目的期望值与绳子数目的关系，横轴采用对数尺度。准确值（蓝绿色）落在直线 $\frac{1}{2}\ln(4n)+\frac{\gamma}{2}$（紫红色虚线）上，其中 $\gamma \approx 0.577$ 是欧拉常数。把绳子数目乘以 10，绳环只会多出约 1.15 个。

回顾

- **找出过程忘记了甚么。**每打一个结之后，不论过去的经过如何，箱中都是“k 条绳子”。
- **又是指示变量。**绳环的数目是指示变量之和，每个结对应一个。
- **有独立性便有方差。**由于每一步的概率都不取决于过去，各个指示变量互相独立。

背后的数学

对数增长

绳环数目的期望值以 $\frac{1}{2}\ln n$ 的速度增长，而方差与平均值相当接近（两者都是一些很小的概率之和）。根据中心极限定理（central limit theorem）（Section 12.24），当 n 很大时，绳环数目近似服从正态分布（normal distribution），平均值和方差都约为 $\frac{1}{2}\ln(4n)$。若有 40 000 条绳子，你通常会得到 2 至 11 个环。

随机置换的循环

同一个对数也出现在 n 件物件的随机置换（random permutation）的循环（cycle）数目之中：平均有 $H_n \approx \ln n$ 个（绳子问题是它的近亲，对应的是 $\frac{1}{2}\ln n$）。随机置换的循环很少：通常有一个巨大的循环包含了大部分元素，而最大循环所占元素的比例，平均约为 0.62，即 Golomb–Dickman 常数。随机地把绳子绑在一起，情况也是一样：大部分绳子最终都落在一两个长圈之中。这些事实属于随机组合结构（random combinatorial structures）

的研究范围，而这门学问又与数论（number theory）有关（一个随机整数的质因数，其表现就像一个随机置换的循环）。

随机纽结与聚合物

绑绳子问题是某些物理系统的简化模型。溶液中的长链聚合物分子会首尾相接，形成环和链，而成环的统计规律也遵从类似的法则。在纽结理论（knot theory）中，数学家问的是一个更难的问题：一个随机的闭合圈有多大机会打了结？对于很长的随机圈，答案趋向 1，这个结果由 De Witt Sumners 和 Stuart Whittington 于 1988 年证明。

思维技巧

找出无记忆的结构

如果每一步之后，系统所处的状态只取决于已经走了多少步，那么各步的表现便互相独立，期望值和方差都可以直接相加。

这正是优惠券收集问题（coupon collector）（Section 9.11）背后的想法，当中每个阶段都是一次全新的几何等待；它也在马可夫链（Markov chains）的理论中再次出现，在那里只有当前的状态才起作用。

延伸练习

1 **小的情况。**对于 $n = 2$，列出所有可能情况，并验证绳环数目的期望值是 $\frac{4}{3}$。

2 **一个大圈。**证明整个过程最终得到单一个包含所有绳子的圈的概率是 $\prod_{k=2}^{n}\left(1 - \frac{1}{2k-1}\right)$，并证明当 n 很大时，这个概率约为 $\sqrt{\frac{\pi}{4n}}$。（提示：找出这个乘积与 $\binom{2n}{n}$ 的关系。）

3 **置换的循环。**证明 n 件物件的随机置换的循环数目的期望值是 H_n，方法是逐个元素地构造这个置换（即“中国餐馆”构造法）。

4 **方差。**证明 $\mathrm{Var}(R) = \mathrm{E}(R) - \sum_{k=1}^{n} \frac{1}{(2k-1)^2}$，并证明被减去的那个和趋向 $\frac{\pi^2}{8}$。

延伸阅读

- **Peter Winkler, *Mathematical Puzzles: A Connoisseur's Collection***（A K Peters），书中收录了这道绑绳子谜题。

- **Richard Arratia, Andrew Barbour & Simon Tavaré, *Logarithmic Combinatorial Structures***（European Mathematical Society）。讲述循环、质数与随机结构，可留待日后阅读。
- **Colin Adams, *The Knot Book***（American Mathematical Society）。纽结理论入门，内容包括随机纽结。

12.11 邮票、杯子与 e

题目

STEP I 1996，第 13 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/96-s1-q13`

引子

有这样一种古怪的卖邮票方法：买家不断猜邮票藏在哪一个杯子下面，每一轮多加一个杯子；到他第一次猜错时，桌上有多少个杯子，他就要付多少英镑。平均而言，他要付 £e，约为 £2.72。

e 这个数在概率（probability）中出现之频繁，简直不可思议：没有人拿回自己衬衫的机会（Section 9.8）、秘书问题（secretary problem，Section 12.12）、泊松分布（Poisson distribution）。本题也许是它众多现身之中最简单的一次，而且原来它与另一道著名谜题有着完全相同的分布（distribution）：要把多少个介乎 0 与 1 之间的随机数相加，总和才会超过 1？平均而言，是 e 个。

动笔之前

想一想 若 $N = 4$，必须发生甚么事？概率是多少？

Jim 必须在有 2 个杯子时猜中（概率为 $\frac{1}{2}$），在有 3 个杯子时也猜中（$\frac{1}{3}$），然后在有 4 个杯子时猜错（$\frac{3}{4}$）。所以 $\mathrm{P}(N = 4) = \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{3}{4} = \frac{3}{4!} = \frac{1}{8}$。

逐步解题

阶段 1 **分布**

若 $N = k$（其中 $k \geq 2$），Jim 在有 $2, 3, \ldots, k-1$ 个杯子时都猜中，而在有 k 个杯子时猜错：

$$\mathrm{P}(N = k) = \frac{1}{2} \cdot \frac{1}{3} \cdots \frac{1}{k-1} \cdot \frac{k-1}{k} = \frac{k-1}{k!}.$$

（验算：$\frac{1}{2!} + \frac{2}{3!} + \frac{3}{4!} + \cdots = \sum_{k\geq 2} \left(\frac{1}{(k-1)!} - \frac{1}{k!} \right) = 1$，这是一个裂项和（telescoping sum）。）

阶段 2 期望值

想一想 化简 $k \cdot \frac{k-1}{k!}$。你会得到甚么级数（series）？

$$\mathrm{E}(N) = \sum_{k\geq 2} \frac{k(k-1)}{k!} = \sum_{k\geq 2} \frac{1}{(k-2)!} = \frac{1}{0!} + \frac{1}{1!} + \cdots = e. \qquad \blacksquare$$

阶段 3 方差

想一想 求 $\mathrm{E}(N^2)$：可写成 $k^2 = k(k-1) + k$，或直接化简 $\frac{k^2(k-1)}{k!}$。

$$\mathrm{E}(N^2) = \sum_{k\geq 2} \frac{k^2(k-1)}{k!} = \sum_{k\geq 2} \frac{k}{(k-2)!} = \sum_{j\geq 0} \frac{j+2}{j!} = \sum_{j\geq 1} \frac{1}{(j-1)!} + 2\sum_{j\geq 0} \frac{1}{j!} = e + 2e = 3e.$$

所以 $\mathrm{Var}(N) = 3e - e^2 = e(3-e) \approx 0.766$。 $\blacksquare$

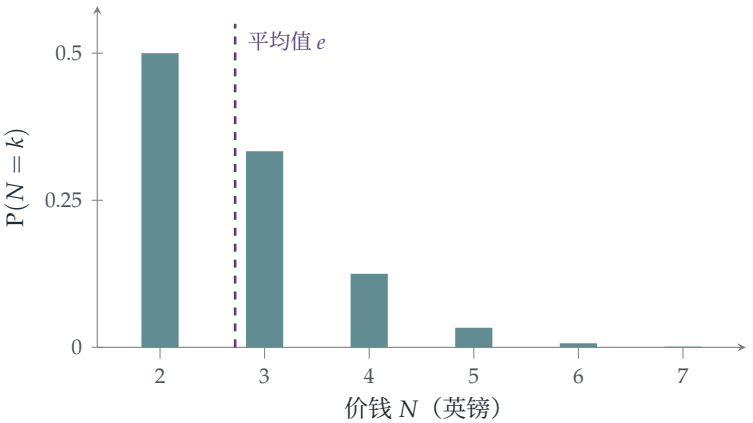


图 12.11.1 价钱的分布，$\mathrm{P}(N = k) = \frac{k-1}{k!}$。有一半机会 Jim 只需付 £2；那条又长又细的尾巴把平均值拉高至 £$e \approx$ £2.72。

回顾

▶ **概率的乘积。**一连串独立（independent）的成功之后接着一次失败，所得的乘积可化

简为 $\frac{k-1}{k!}$。

- **约去阶乘。**乘以 k 或 k^2 再约简，那些和式就变成了指数级数（exponential series）。
- **以平移求 $\mathrm{E}(N^2)$。**以 $j=k-2$ 重新编排下标，和式便分拆成两份 e。

背后的数学

换了装扮的同一个分布

从 $[0,1]$ 中均匀（uniformly）地随机选取数 $U_1, U_2, \ldots$，并设 M 为令它们的总和超过 1 所需的个数。首 k 个数之和小于 1 的概率，就是区域 $\{u_1+\cdots+u_k<1, u_i\geq 0\}$ 的体积，即 $\frac{1}{k!}$（面积为 $\frac{1}{2}$ 的三角形、体积为 $\frac{1}{6}$ 的四面体，如此类推）。所以

$$\mathrm{P}(M>k)=\frac{1}{k!}, \qquad \mathrm{P}(M=k)=\frac{1}{(k-1)!}-\frac{1}{k!}=\frac{k-1}{k!},$$

正正就是 N 的分布。在两个问题中都有 $\mathrm{P}(N>k)=\frac{1}{k!}$：就杯子而言，这是连续猜中 $k-1$ 次的概率。利用 $\mathrm{E}(N)=\sum_{k\geq 0}\mathrm{P}(N>k)$（Section 9.9 中的一条公式），一行便得出 $\mathrm{E}(N)=\sum\frac{1}{k!}=e$。

为何 e 不断出现

每当概率中含有 $\frac{1}{k!}$，不论它来自 k 件物件按某个特定次序排列、k 个概率依次为 $\frac{1}{2}, \frac{1}{3}, \ldots$ 的相继事件，还是单纯形（simplex）的体积，对 k 求和都会产生指数级数。错排（derangements，Section 9.8）给出 $\sum\frac{(-1)^k}{k!}=e^{-1}$；这里我们则有 $\sum\frac{1}{k!}=e$。至于形如 $\left(1-\frac{1}{n}\right)^n\to e^{-1}$ 的极限（limit），它们源自大量独立而不大可能发生的事件，正好解释了秘书问题和泊松分布中的 e。

蒙地卡罗 e

“总和超过 1”这个问题提供了一个以实验估计 e 的方法：产生随机数，数一数需要多少个才超过 1，重复多次，然后取平均值。它是蒲丰投针（Buffon's needle，Section 12.20）的近亲，后者用来估计 π。这类蒙地卡罗（Monte Carlo）方法以赌场命名，利用随机抽样来估计难以直接计算的量；如今它们广泛应用于科学的各个范畴，由粒子物理学到金融，主要用于计算高维积分。

思维技巧

认出级数

当期望值是一个含有阶乘的和式时，把各项化简，直至出现一个标准级数（指数级数、几何级数、二项级数）。

在大学，生成函数（generating function）把这种做法系统化：概率生成函数（probability generating function）$\mathrm{E}(t^N)$ 把整个分布包装起来，而它在 $t = 1$ 处的导数则给出平均值和方差。

延伸练习

1 **生成函数。**证明 $\mathrm{E}(t^N) = 1 + (t-1)e^t$，并用它求 $\mathrm{E}(N)$ 和 $\mathrm{E}(N(N-1))$。

2 **尾部公式。**对于取非负整数值的随机变量（random variable），证明 $\mathrm{E}(N) = \sum_{k\geq 0} \mathrm{P}(N > k)$，并用它一行求出 $\mathrm{E}(N)$。

3 **由一个杯子开始。**假如游戏由一个杯子开始（Jim 不可能猜错），分布会有甚么改变？价钱的期望值是多少？

4 **单纯形。**证明 $\{u_1 + u_2 + u_3 < 1, u_i \geq 0\}$ 的体积为 $\frac{1}{6}$，并用归纳法（induction）证明 k 维的情况为 $\frac{1}{k!}$。

延伸阅读

- **Eli Maor, *e: The Story of a Number***（Princeton University Press）。e 的历史，包括它在概率中的各次现身。
- **Geoffrey Grimmett & Dominic Welsh, *Probability: An Introduction***（Oxford University Press），关于生成函数。
- **Paul Nahin, *Digital Dice***（Princeton University Press）。以蒙地卡罗模拟解决的概率谜题。

12.12 何时停止寻找

题目

STEP I 2001，第 14 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/01-s1-q14`

引子

你正在逐一面试应征者，而且必须即场决定：取录眼前这一位，然后停止；还是永远拒绝这一位，再见下一位。你想要的是最好的那一位。如果太早取录，你很可能错过更好的人；如果等得太久，最好的那一位很可能已经走了。你应该怎样做？

答案是概率（probability）论中最著名的结果之一：先拒绝最初 37% 的应征者，然后取录第一位比这些人都好的应征者。这个做法大约有 37% 的机会成功，而 37% 正是 $\frac{1}{e}$。本题计算其中的基本概率，而与它配对的 STEP II 2006 题目（见练习）则分析这个策略（strategy）最简单的版本：只拒绝第一位应征者。

动笔之前

? **想一想** 在随机次序之下，全体之中最好的应征者落在最先面试的 n 位之中的概率是多少？

最好的应征者落在 N 个位置中任何一个的可能性均等，所以概率是 $\frac{n}{N}$。

逐步解题

阶段 1 最先 n 位中的最佳者就是全体最佳者

最先 n 位中的最佳者是全体最佳者，当且仅当全体最佳者在最先 n 位之中：概率为 $\frac{n}{N}$。■

阶段 2 最佳或次佳

想一想 最先 n 位中的最佳者是全体次佳者，当且仅当次佳者在最先 n 位之中而最佳者不在其中。这个概率是多少?

次佳者在最先 n 位之中的概率是 $\frac{n}{N}$；在此条件下，最佳者在其余 $N-1$ 个位置的其中一个，当中有 $N-n$ 个位置排在最先 n 位之后。所以

$$\mathrm{P}(\text{最佳或次佳}) = \frac{n}{N} + \frac{n}{N}\cdot\frac{N-n}{N-1}. \qquad \blacksquare$$

当 $N=4$，$n=2$ 时：$\frac{1}{2}+\frac{1}{2}\cdot\frac{2}{3}=\frac{5}{6}$。把最好的两位应征者在四个位置中的排法逐一列出（全部 12 种有序排法的可能性均等），除了两位顶尖应征者同时落在第 3 和第 4 位的 2 种排法之外，最先两位中的最佳者都是顶尖两位之一：$\frac{10}{12}=\frac{5}{6}$。

阶段 3 最优策略

想一想 策略：先拒绝最初 r 位应征者，然后取录第一位比他们全部都好的应征者。如果最佳应征者在位置 $k>r$，这个策略在甚么情况下会选中此人?

设最佳应征者在位置 k，这个策略选中此人，当且仅当位置 $r+1,\dots,k-1$ 上没有任何应征者胜过最初 r 位全部的人，即是说，最先 $k-1$ 位应征者中的最佳者在最初 r 位之中。根据第 1 阶段（把 N 换成 $k-1$），这件事的概率是 $\frac{r}{k-1}$。最佳应征者落在每个位置 k 的概率都是 $\frac{1}{N}$，所以

$$\mathrm{P}(\text{成功}) = \frac{1}{N}\sum_{k=r+1}^{N}\frac{r}{k-1} = \frac{r}{N}\Big(\frac{1}{r}+\frac{1}{r+1}+\cdots+\frac{1}{N-1}\Big).$$

想一想 当 N 很大时，记 $x=\frac{r}{N}$。这个和大约等于 $\ln\frac{N}{r}$。x 取甚么值时 $x\ln\frac{1}{x}$ 达到最大?

成功的概率大约是 $x\ln\frac{1}{x}$。求导可得，$\ln\frac{1}{x}-1=0$ 于 $x=\frac{1}{e}$ 时成立，而最大值是 $\frac{1}{e}\approx 0.368$（图 12.12.1）。

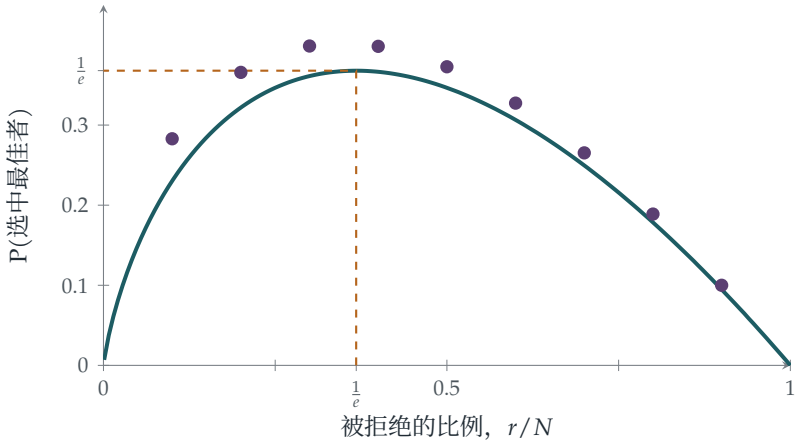


图 12.12.1 若先拒绝占比例 r/N 的最初一批应征者，选中最佳应征者的概率：极限曲线 $x \ln \frac{1}{x}$（蓝绿色），以及 $N = 10$ 时的精确值（紫红色）。最佳策略是拒绝大约 $\frac{1}{e} \approx 37\%$ 的应征者，成功的概率大约是 $\frac{1}{e}$，不论 N 有多大。

回顾

- **位置的对称性。**在随机次序之下，每位应征者落在任何位置的可能性均等，由此立即得出 $\frac{n}{N}$。
- **以最佳者的位置为条件。**计算策略的成功概率时，我们以最佳应征者所在的位置为条件分情况处理。
- **求和变成积分。**当 N 很大时，调和和（harmonic sum）变成了对数，再经一次求最大值便得出 $\frac{1}{e}$。

背后的数学

最优停止

秘书问题（secretary problem）是*最优停止*（optimal stopping）的奠基例子：根据随时间陆续到来的信息，决定何时采取行动，以使期望回报最大。可以证明，在*所有*策略之中（而不只是“先拒绝 r 位，然后取录下一位至今最佳者”这一类），这种阈值规则（threshold rule）是最优的，论证的方法是由最后一位应征者开始向前逆推。同样的数学也决定了何时卖出股票、何时行使金融期权，以及应该花多久去找泊车位。一个值得注意的特点是，在极限之下答案与 N 无关：不论应征者有一百位还是一百万位，你成功的概率都是 $\frac{1}{e}$。

各种变化

如果你只要选中顶尖两位应征者的其中一位便满意，最佳策略大约有 57% 的机会成功。如果你想做到的不是使选中最佳者的机会最大，而是使所选应征者的期望*名次*最优，最优策略便复杂得多，而期望名次趋向大约 3.87（Chow、Moriguti、Robbins 和 Samuels，1964）。如果你有一定的概率可以回头召回已被拒绝的应征者，又或者你知道应征者分数的分布（即“完全信息”问题），答案又会不同。练习中的雪糕题是最简单的情形：只拒绝一个，然后取下一个比它好的。

是谁解决的?

这个问题在 1950 年代流传于数学家之间，马丁·加德纳（Martin Gardner）于 1960 年以“googol 游戏”之名把它普及开来（在纸条上写下数字；你的对手逐张翻开，并且必须在最大的那个数出现时停下）。Thomas Ferguson 的文章“Who solved the secretary problem?”（1989）追溯了它错综复杂的历史；有一种说法指，天文学家开普勒（Johannes Kepler）在 1611 年挑选第二任妻子时，会见了十一位人选，最后向第五位求婚。

思维技巧

把策略参数化，然后求最优

当最佳策略未知时，考虑一族带有参数的自然策略，把成功概率写成该参数的函数，然后求最优。

在大学里，这种方法用于决策论、博弈论和统计学。要证明这一族之中的最佳策略就是全体之中的最佳策略，是更困难也更有趣的一步，通常要用*逆向归纳法*（backward induction）来完成。

延伸练习

1 **雪糕。**试做 STEP II 2006，第 13 题（`step.maths.org/questions/06-s2-q13`）：这是秘书问题的小型版本，规则很简单。拒绝第一个，然后取第一个胜过它的。

2 **较小的 N。**当 $N = 4$ 时，求 r 的最佳值和成功的概率。用列举法验算。

3 **最佳或次佳。**利用第 (ii) 部分，论证“最先 n 位中的最佳者是顶尖两位之一”的可能性远高于“最先 n 位中的最佳者就是最佳者”。这对于选中顶尖两位之一的策略有甚么启示?

4 **精确的最大值。**当 $N = 10$ 时，从图中验证 $r = 3$ 是最优的。当 $N = 100$ 时，证明 $r = 37$ 是最优的（准确至四位小数，$r = 36$、37、38 的成功概率分别是 0.3710、0.3710、0.3708）。

延伸阅读

- **Thomas Ferguson，“Who solved the secretary problem?”**，*Statistical Science* 4（1989）。历史与各种变化；可在网上免费阅读。
- **Brian Christian & Tom Griffiths，*Algorithms to Live By***（William Collins）。把最优停止及其他算法应用于日常决策。
- **Martin Gardner，*New Mathematical Diversions***（Mathematical Association of America）。googol 游戏。

12.13 三人决斗

题目

STEP II 1995，第 12 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/95-s2-q12`

引子

三个人用餐包决斗。Cuthbert 百发百中，Bertie 每五次掷中三次，而 Algernon 准头最差，每五次只掷中两次。Algernon 先掷。他应该瞄准谁？答案出人意表：谁也不瞄准。他应该故意把餐包掷偏。

原因在于策略。只要三人都仍在场，Bertie 和 Cuthbert 这两个危险的对手就会互相瞄准。假如 Algernon 淘汰了其中一人，余下的那个便会转而对付他。他故意掷失，就能让那两人先行分出胜负，然后由他向幸存者先掷一次。事实上，只要采用最佳方案，准头最差的人反而最有可能胜出。这场“三人决斗”（truel）是博弈论（game theory）中常被引用的例子，说明弱势有时反而是优势。

动笔之前

想一想 在一场二人决斗中，Percy 先掷。他要么一掷即中，要么两人都掷失，于是局面完全重演。试写出一条关于 Percy 胜出概率（probability）的方程。

设 P 为 Percy 胜出的概率。他一掷即中的概率是 p；否则（概率为 $1-p$）轮到 Rodney 掷，而 Rodney 掷失的概率是 $1-r$，此时局面回到起点。所以 $P = p + (1-p)(1-r)P$。

逐步解题

阶段 1 二人决斗

解方程，得 $P(1-(1-p)(1-r))=p$，而 $1-(1-p)(1-r)=p+r-pr$，所以

$$P=\frac{p}{p+r-pr}. \qquad \blacksquare$$

若 Algernon $(p=\frac{2}{5})$ 对 Bertie $(r=\frac{3}{5})$，由 Algernon 先掷，这个概率是 $\frac{2/5}{2/5+3/5-6/25}=\frac{10}{19}$；若由 Bertie 先掷，Algernon 胜出的概率是 $\left(1-\frac{3}{5}\right)\cdot\frac{10}{19}=\frac{4}{19}$。

阶段 2 Algernon 的三个选择

想一想 在每种情况下，还剩下谁？下一个轮到谁掷？

(i) *Algernon 掷中 Cuthbert*。接着由 Bertie 掷，目标是 Algernon。Algernon 要在 Bertie 先掷的二人决斗中胜出：概率为 $\frac{4}{19}$。

(ii) *Algernon 掷中 Bertie*。接着由 Cuthbert 掷，目标是 Algernon，而他从不失手。Algernon 的胜出机会是 0。

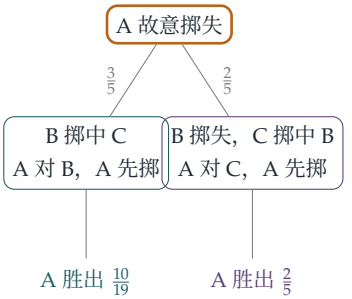
(iii) *Algernon 掷失*。Bertie 向 Cuthbert 掷。他有 $\frac{3}{5}$ 的概率掷中，剩下 Algernon 和 Bertie，并由 Algernon 先掷：Algernon 胜出的概率是 $\frac{10}{19}$。Bertie 有 $\frac{2}{5}$ 的概率掷失，然后 Cuthbert 掷中 Bertie，Algernon 便要面对 Cuthbert，而且只有一次出手机会：他必须掷中才能胜出，概率为 $\frac{2}{5}$。合计为

$$\frac{3}{5}\cdot\frac{10}{19}+\frac{2}{5}\cdot\frac{2}{5}=\frac{6}{19}+\frac{4}{25}=\frac{150+76}{475}=\frac{226}{475}\approx 0.476.$$

阶段 3 建议

想一想 如果 Algernon 瞄准 Cuthbert，他可能掷中（概率为 $\frac{2}{5}$），也可能掷失。他整体的胜出机会是多少？试与故意掷失比较。

瞄准 Cuthbert 的胜出概率是 $\frac{2}{5}\cdot\frac{4}{19}+\frac{3}{5}\cdot\frac{226}{475}=\frac{200+678}{2375}=\frac{878}{2375}\approx 0.370$。瞄准 Bertie 则更差。故意掷失的胜出概率是 $\frac{226}{475}\approx 0.476$，是三者中最高的。所以 Algernon 应该把餐包掷偏，这样他胜出的概率是 $\frac{226}{475}$。 $\blacksquare$



采用这个方案：
A 胜出 $\frac{226}{475} \approx 0.48$
B 胜出 $\frac{135}{475} \approx 0.28$
C 胜出 $\frac{114}{475} \approx 0.24$

图 12.13.1 Algernon 的最佳方案。他故意掷失，让两个较强的对手先行交锋，并确保自己能向幸存者先掷一次。结果准头最差的人最有可能胜出，而百发百中的人反而最不可能胜出。

回顾

- **重复出现的局面可以列出方程。**两人都掷失后，决斗“重新开始”，由此得出一条关于胜出概率的线性方程（linear equation）。
- **考虑对手的反应。**Algernon 的每个选择，都会改变 Bertie 和 Cuthbert 接下来的行动。
- **比较所有选择，包括甚么都不做。**最好的一着，正是看来最差的那一着。

背后的数学

策略与博弈论

三人决斗展示了博弈论的核心思想：最佳行动取决于其他参与者会如何回应。掷中最危险的对手 Cuthbert 看似理所当然，但这样做会改变局势，反而对 Algernon 不利。在本题中，Bertie 和 Cuthbert 的策略（strategy）是固定的（他们互相瞄准）；在完整的博弈论分析中，每个参与者都会自行选择策略，并且知道其他人也在这样做。这类分析始于冯·诺伊曼（John von Neumann）与摩根斯坦（Oskar Morgenstern）在 1944 年的工作，以及纳殊（John Nash）在 1950 年的工作，如今已成为经济学、演化生物学和政治学的基础。

强大反成弱点

三人决斗中谁会胜出？若 Algernon 采用最佳方案，Algernon 胜出的概率是 $\frac{226}{475}$，Bertie 是 $\frac{135}{475}$，Cuthbert 是 $\frac{114}{475}$。准头最好的人最不可能幸存，因为凡是可以选择目标的人都会瞄准他。同样的规律也见于有多名候选人的选举、企业之间的竞争，以及自然界：最强的竞争者可能成为首个被攻击的目标。Martin Shubik 在 1950 年代把三人决斗引入博弈论，

并借此探问：最适者是否必然生存？

思维技巧

多想一步

选择行动之前，先弄清楚它会怎样改变其他各人面对的局面，以及他们会如何回应。

到了大学，这会发展成博弈论中的逆向归纳法（backward induction）：先分析博弈的终局，再往回推，从而决定最佳的第一步。

延伸练习

1 **Bertie 与 Cuthbert。**验证：若 Algernon 故意掷失，Bertie 胜出的概率是 $\frac{135}{475}$，Cuthbert 胜出的概率是 $\frac{114}{475}$。

2 **准头较好的投手。**假如 Algernon 掷中的概率是 a 而不是 $\frac{2}{5}$，当 a 取哪些值时，故意掷失仍然是他最佳的第一步？

3 **Bertie 应该瞄准 Cuthbert 吗？**已知 Algernon 故意掷失，Bertie 瞄准 Cuthbert 而不瞄准 Algernon，是否对他自己有利？（假设 Cuthbert 的行为不变。）

4 **投掷次数有限的决斗。**在一场每人只可掷两次的二人决斗中，求 Percy 胜出的概率，并与不限次数的决斗比较。

延伸阅读

- **D. Marc Kilgour & Steven Brams，“The truel”**，*Mathematics Magazine* 70 (1997)。一篇浅白易读的三人决斗综述。
- **Avinash Dixit & Barry Nalebuff，*Thinking Strategically***（Norton）。写给一般读者的博弈论读物，当中也谈到三人决斗。
- **Ken Binmore，*Game Theory: A Very Short Introduction***（Oxford University Press）。

12.14 混合策略

题目

STEP I 1989，第 16 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/89-s1-q16`

引子

玩“包剪揾”时，任何固定的出法都会被对手利用：每次都出揾（石头），对手很快便学会出包（布）。唯一稳妥的策略就是随机出手。本题是一个更精细的版本：各种结果的得失并不相等，怎样随机化才对也并不明显。结果发现，只要以概率 $\frac{1}{4}$、$\frac{1}{2}$、$\frac{1}{4}$ 分别猜 1、2、3，不论对手怎样做，玩家都可以保证长远而言不会输钱。

这让我们初次窥见冯纽曼（John von Neumann）在 1928 年提出的**极小极大定理**（minimax theorem），它是博弈论（game theory）的奠基结果：在任何一方所得即另一方所失的二人博弈中，双方各有一个随机化策略（randomised strategy），能保证在对手发挥得完美无瑕时，仍取得可能范围内的最佳结果。

动笔之前

想一想 写出一个表，列出每一对猜测之下 B 赢得的金额（B 的猜测列于左侧，A 的猜测列于顶端）。

按规则逐一推算（例如，若 B 说 1 而 A 说 3，两数相差 2，说出较小数的 B 要付 £2）：

B \ A	1	2	3
1	0	+1	−2
2	−1	0	+1
3	+2	−1	0

这个表是**反对称**（antisymmetric）的（把两位玩家对调，正负号便会改变），所以这个博

弈是公平的，意思是双方都没有与生俱来的优势。

逐步解题

阶段 1 对付固定的猜测

若 A 每次都说 1，B 应该说 3，每局赢 £2。若 A 每次都说 2，B 说 1，赢 £1。若 A 每次都说 3，B 说 2，赢 £1。任何固定的猜测都会被对手利用。

阶段 2 对付随机的猜测

想一想 若 A 以均匀的概率随机猜测，当 B 说 j 时，B 的期望收益（expected gain）是多少？（求表中第 j 行的平均值。）

各行的平均值分别是：$j=1$ 时为 $\frac{1}{3}(0+1-2)=-\frac{1}{3}$，$j=2$ 时为 0，$j=3$ 时为 $\frac{1}{3}(2-1+0)=\frac{1}{3}$。所以 B 每局赢得金额的期望值是

$$-\tfrac{1}{3}b_1+0\cdot b_2+\tfrac{1}{3}b_3,$$

当 $b_3=1$ 时取得最大值：每次都说 3，平均每局赢 $\frac{1}{3}$ 镑。均匀的随机化并不能保护 A。

阶段 3 一般情况

想一想 现在设 A 以概率 a_j 说 j。把 B 赢得金额的期望值按 A 的猜测归项：a_1 的系数（即第 1 列以 b_j 为权的加权平均）是甚么？

沿着表中各列由上而下读，

$$\text{B 赢得金额的期望值}=(-b_2+2b_3)a_1+(b_1-b_3)a_2+(-2b_1+b_2)a_3,$$

所以 $X=-b_2+2b_3$，$Y=b_1-b_3$，$Z=-2b_1+b_2$。

想一想 你能令三者同时为零吗？（共有三条方程，但它们并非独立；为甚么？）

它们并非独立：$X+2Y+Z=0$ 恒成立（请自行验证）。所以由 $Y=0$ 得 $b_1=b_3$，由 $X=0$ 得 $b_2=2b_3$，两者合起来便迫使 $Z=0$ 亦成立。再配合 $b_1+b_2+b_3=1$：

$$b_1=\tfrac{1}{4},\qquad b_2=\tfrac{1}{2},\qquad b_3=\tfrac{1}{4},$$

于是 $X=Y=Z=0$。这样，不论 a_j 取甚么值，B 赢得金额的期望值都是 0，B 长远而言不会有任何损失。 ■

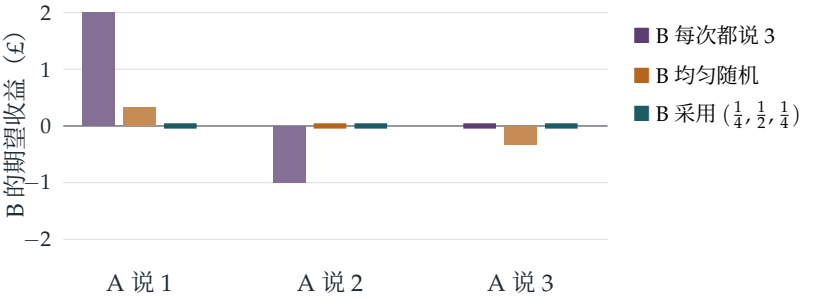


图 12.14.1 B 采用三种不同策略时，面对 A 每一种选择的期望收益。每次都说 3，遇上 A 说 1 会赢，遇上 A 说 2 却会输；均匀随机猜测，遇上 A 说 3 会输。策略 $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ 则不论 A 怎样做，收益都恰好是 0（图中以幼条表示），所以永远不会被对手利用。

回顾

- **写出得失表。**所有推理都围绕一个 3×3 矩阵的行和列。
- **纯策略（pure strategy）会被对手利用。**面对任何固定的猜测，甚至面对均匀随机的猜测，对手都有一个有利可图的应对方法。
- **拉平。**稳妥的策略令对手选甚么都没有分别：A 的每一种选择都带来相同的期望结果。

背后的数学

极小极大定理

在一个二人零和（zero-sum）博弈中，设得失矩阵（payoff matrix）为 M，B 选取概率 $\mathbf{b}$，A 选取概率 $\mathbf{a}$，则 B 赢得金额的期望值是 $\mathbf{b}^{\mathsf{T}} M \mathbf{a}$。冯纽曼的极小极大定理指出

$$\max_{\mathbf{b}} \min_{\mathbf{a}} \mathbf{b}^{\mathsf{T}} M \mathbf{a} = \min_{\mathbf{a}} \max_{\mathbf{b}} \mathbf{b}^{\mathsf{T}} M \mathbf{a}.$$

用文字说：B 预先公开自己的随机化策略而仍能保证取得的金额，等于 A 预先公开自己的策略而能把 B 压制到的金额。这个共同的数值称为博弈的值（value）。在本题中，由反对称性可知博弈的值是 0，而策略 $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ 对双方都是最优的。这个定理可用线性代数（linear algebra）和凸性（convexity）证明（它等价于线性规划（linear programming）的对偶定理），而它推广到多人博弈的版本，即纳殊均衡（Nash equilibrium），为纳殊（John Nash）赢得诺贝尔经济学奖。

为甚么要随机化?

这里的随机性无关运气，而是关乎信息。如果 B 的选择可以被预测，A 便能加以利用；B 以恰当的概率随机选择，既令自己的选择无从预测，又确保任何预测都帮不了对手。现实中的例子包括足球的十二码罚球——对职业球员的研究发现，主射球员和守门员在左右两边之间随机选择，比例与博弈论的最优值十分接近——以及网球的发球。

冯纽曼 1928 年的论文

冯纽曼在 1928 年证明了极小极大定理，论文题为“On the theory of games of strategy”（论策略博弈的理论），研究动机来自扑克，因为扑克中的“偷鸡”（bluffing，即虚张声势）正是一种随机化。1944 年，他与经济学家 Oskar Morgenstern 出版了 *Theory of Games and Economic Behavior*（博弈论与经济行为），博弈论自此成为一门学科。冯纽曼后来忆述，他一直视极小极大定理为关键结果，没有它，就没有值得发表的博弈理论。

思维技巧

以随机化消除对手的优势

当任何固定的选择都会被对手利用时，就随机选择，而所用的概率要令对手无论选哪一项都没有分别。

在大学里，博弈论、线性规划和算法理论都会用到这个想法。例如，随机算法（randomised algorithm）会作出任何对手都无法预测的随机选择，借此防范最坏情况的输入。

延伸练习

1 **A 的最佳应对。**若 B 采用 $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$，证明 A 的每一种策略所得的期望值都是 0。若 B 改用 $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$，求 A 的最佳应对。

2 **包剪揼。**写出“包剪揼”的得失矩阵，并求最优策略。为甚么这里的最优策略是均匀的，而本题的却不是?

3 **唯一性。**证明 $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ 是 B 能保证不输的唯一策略。（提示：它必须令 X、Y、Z 全部非负；利用 $X + 2Y + Z = 0$。）

4 **改动后的博弈。**若相差 2 时的罚款由 £2 改为 £3，求最优策略。

延伸阅读

- **Ken Binmore, *Playing for Real***（Oxford University Press）。一本详尽而平易近人的博弈论入门书。
- **Avinash Dixit & Barry Nalebuff, *The Art of Strategy***（Norton）。以体育和商业的例子讲解博弈论。
- **William Poundstone, *Prisoner's Dilemma***（Anchor）。冯纽曼、博弈论与冷战。

12.15 灭绝还是爆发

题目

STEP II 2003，第 12 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/03-s2-q12`

引子

1873 年，高尔顿（Francis Galton）在《*Educational Times*》上提出一个问题：为甚么贵族的姓氏那么容易失传？每个男人所生的儿子数目是随机的，而每个儿子都会把姓氏传下去。即使每个男人平均有多于一个儿子，这个姓氏就能永远流传下去吗？这个问题开创了分支过程（branching process）的理论，如今这套理论用来模拟核连锁反应、传染病的扩散、细胞群体的增长，以及突变能否存活下去。

本题建立的是最简单的模型：每个粒子分裂成 0、1 或 2 个粒子。粒子总数的期望值（expectation）按几何级数增长或缩减。但期望值并不是故事的全部。即使期望的粒子总数急剧膨胀，整个族系完全灭绝的概率仍然是正数，而一个漂亮的不动点（fixed point）论证就能把这个概率求出来。

动笔之前

? **想一想** 如果某一阶段有两个粒子，下一阶段粒子数目的分布是甚么？

两个粒子各自独立地产生后代，后代数目服从 $\mathrm{B}(2, p)$，所以总数服从 $\mathrm{B}(4, p)$：共有四次独立的“机会”，每次成功的概率为 p。

逐步解题

阶段 1　先是两个，然后又是两个

$\mathrm{P}(X_1=2)=p^2$，然后 $\mathrm{P}(X_2=2\mid X_1=2)=\binom{4}{2}p^2q^2=6p^2q^2$。所以

$$\mathrm{P}(X_1=2,X_2=2)=6p^4q^2. \qquad \blacksquare$$

阶段 2 逆向推理

> **想一想** $X_2=2$ 可以有多少种方式发生？（X_1 可以是甚么？）

要使 $X_2=2$，可以是 $X_1=1$（概率为 $2pq$），其后产生 2 个后代（p^2）；也可以是 $X_1=2$，其后由 $\mathrm{B}(4,p)$ 得出 2 个。（$X_1=0$ 则甚么也不剩。）所以 $\mathrm{P}(X_2=2)=2p^3q+6p^4q^2$，而

$$\mathrm{P}(X_1=2\mid X_2=2)=\frac{6p^4q^2}{2p^3q+6p^4q^2}=\frac{3pq}{1+3pq}.$$

令此式等于 $\frac{9}{25}$：$75pq=9+27pq$，所以 $pq=\frac{3}{16}$，即 $p(1-p)=\frac{3}{16}$，由此得 $p=\frac{1}{4}$ 或 $p=\frac{3}{4}$。 $\blacksquare$

阶段 3 粒子总数的期望值

> **想一想** 如果在第 $n-1$ 阶段有 k 个粒子，第 n 阶段粒子数目的期望值是多少？然后对 k 取平均。

每个粒子平均有 $\mathrm{E}(X)=2p$ 个后代，所以 k 个粒子平均有 $2pk$ 个后代：$\mathrm{E}(X_n\mid X_{n-1}=k)=2pk$。对 k 取平均，得 $\mathrm{E}(X_n)=2p\,\mathrm{E}(X_{n-1})$。由于 $\mathrm{E}(X_0)=1$，

$$\mathrm{E}(X_n)=(2p)^n.$$

当 $p=\frac{1}{4}$ 时，$(2p)^n=2^{-n}\to 0$；当 $p=\frac{3}{4}$ 时，$(2p)^n=\left(\frac{3}{2}\right)^n\to\infty$。 $\blacksquare$

回顾

- **独立二项变量之和。**k 个粒子各自给出 $\mathrm{B}(2,p)$，合起来就是 $\mathrm{B}(2k,p)$。
- **时间上的贝叶斯定理（Bayes' theorem）。**$\mathrm{P}(X_1=2\mid X_2=2)$ 是由较后的观察结果逆向推断较早的状态。
- **条件期望值（conditional expectation）。**以上一代的情况为条件，期望值就化成一个等比数列（geometric sequence）。

背后的数学

灭绝的概率

当 $p = \frac{3}{4}$ 时，粒子总数的期望值无止境地增长。这是否表示群体一定能存活下去？不一定：如果第一个粒子没有后代（概率为 $q^2 = \frac{1}{16}$），这个族系便立即终结。设 e 为最终灭绝的概率。第一个粒子有 k 个后代，而每个后代的族系各自独立地以概率 e 灭绝，所以

$$e = q^2 + 2pq\,e + p^2e^2 = (q + pe)^2 = \mathrm{G}(e),$$

其中 $\mathrm{G}(s) = (q + ps)^2$ 是后代数目的*概率生成函数*（probability generating function）。灭绝概率是 G 的一个不动点：事实上是最小的非负不动点。当 $p = \frac{3}{4}$ 时，由 $(1 + 3e)^2 = 16e$ 得 $9e^2 - 10e + 1 = 0$，所以 $e = \frac{1}{9}$ 或 1，而灭绝概率是 $\frac{1}{9}$。当 $p = \frac{1}{4}$ 时，在 $[0, 1]$ 内唯一的不动点是 $e = 1$：灭绝是必然的（图 12.15.1）。

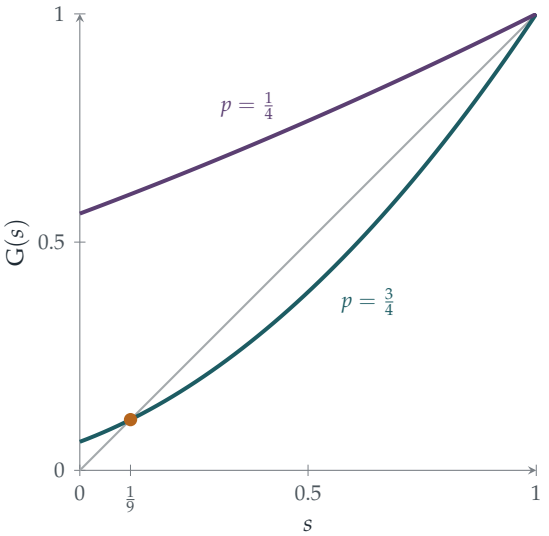


图 12.15.1 后代数目的生成函数 $\mathrm{G}(s) = (q + ps)^2$ 与直线 $y = s$。灭绝概率就是两者相交的最小一点。当 $p = \frac{1}{4}$（紫红色，平均值 $\frac{1}{2}$）时，该点是 $s = 1$：灭绝是必然的。当 $p = \frac{3}{4}$（蓝绿色，平均值 $\frac{3}{2}$）时，该点是 $s = \frac{1}{9}$（琥珀色）：群体以概率 $\frac{8}{9}$ 永远存活下去。

临界门槛

一般而言，如果后代数目的平均值小于 1，分支过程必然灭绝；平均值等于 1 时也是如此，唯一的例外是每个个体都恰好有一个后代的平凡情况（这时族系只是一个接一个地永远延续下去）；如果平均值大于 1，则分支过程以正概率存活下去。后代数目的平均值就是流行病学中的*繁殖数*（reproduction number）R：$R > 1$ 的疫症有可能大规模爆发，但即使如此，早期的爆发仍可能纯粹因为偶然而无疾而终，其概率正好由上面的不动点方程给出。在核物理学中，如果每次裂变平均恰好引发另一次裂变，连锁反应就处于*临界*

（critical）状态，而操作反应堆的学问，就在于让它一直保持在这个状态。

高尔顿、Watson 与 Bienaymé

高尔顿的问题在 1874 年由 Henry William Watson 牧师作答，但答案有部分是错的：他找到了不动点方程，却漏掉了第二个根，因而断定每个姓氏最终都会失传。其实正确的分析早在 1845 年已由法国统计学家 Irénée-Jules Bienaymé 提出，只是他的研究一直被人遗忘，直到 1970 年代才重见天日。这个模型现在称为*高尔顿—Watson 过程*（Galton–Watson process，有时称为 Bienaymé–Galton–Watson 过程），是每一门随机过程（random processes）课程中的核心例子。

思维技巧

以第一代为条件

一个会自我更新的过程（每个后代都开始一个独立的、与整个过程相同的复本），必定满足一条以第一步为条件而得出的方程。

方程 $e = \mathrm{G}(e)$ 是一个不动点问题，与 Section 2.2 和 Section 8.7 中遇到的属同一类。在大学里，生成函数和不动点是研究分支过程、随机树和随机网络的主要工具。

延伸练习

1 **一般情况下的灭绝。**设后代数目的分布为 $\mathrm{B}(2, p)$，证明当 $p \le \frac{1}{2}$ 时灭绝概率为 1，而当 $p > \frac{1}{2}$ 时灭绝概率为 $\left(\frac{q}{p}\right)^2$。

2 **方差（variance）。**证明 $\mathrm{Var}(X_n)$ 满足 $\mathrm{Var}(X_n) = (2p)^2\mathrm{Var}(X_{n-1}) + 2pq\,\mathrm{E}(X_{n-1})$，并求 $p = \frac{1}{2}$ 时的方差。

3 **在第 n 代或之前灭绝。**设 $e_n = \mathrm{P}(X_n = 0)$。证明 $e_n = \mathrm{G}(e_{n-1})$，并就 $p = \frac{3}{4}$ 计算 e_1、e_2、e_3。它们是否趋近 $\frac{1}{9}$？

4 **姓氏。**如果每个男人有 0、1、2 或 3 个儿子的概率各为 $\frac{1}{4}$，他的姓氏最终失传的概率是多少？

延伸阅读

- **Geoffrey Grimmett & David Stirzaker, *Probability and Random Processes***（Oxford University Press），论分支过程的一节。
- **Adam Kucharski, *The Rules of Contagion***（Profile）。讲述繁殖数与疫症爆发，适合一般读者。
- **David Kendall, "Branching processes since 1873"**，*Journal of the London Mathematical Society* 41 (1966)。讲述这段历史，包括 Bienaymé 的贡献。

12.16 顺向与逆向的马可夫链

题目

STEP III 1996，第 12 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/96-s3-q12`

引子

Ecks 教授的定理有一个规律：她下一个证明甚么，只取决于她上一个证明了甚么。一个过程的未来若只取决于现在而与过去无关，就称为**马可夫链**（*Markov chain*）。马可夫链无处不在：天气模型、蛇梯棋这种图版游戏、DNA 模型、Google 最初的 PageRank 算法，以至统计学家用来模拟复杂概率分布的方法，全都用得上它。

本题要求出各类定理在长远而言所占的比例，即**平稳分布**（*stationary distribution*）：长远来说，她的定理每六个之中有一个是新而正确的，而有一半是错的。然后题目把这条链在时间上倒过来运行，并问：已知今天的定理，昨天的定理是哪一类？

动笔之前

想一想 “每一行（row）之和为 1”这个条件是甚么意思？各列（column）之和是否也是 1？

第 i 行列出的是：已知现时的类型为 i，下一个定理属于各类型的概率；三者必有其一发生，所以每一行之和为 1。各列则不一定：第二列之和恰好是 1，但第一列之和是 0.6，第三列之和是 1.4。

逐步解题

阶段 1　一步

根据全概率定律（law of total probability），以现时定理的类型 i 为条件，

$$b_j = \sum_i \mathrm{P}(\text{现时}i)\,\mathrm{P}(\text{下一个}j \mid \text{现时}i) = a_1p_{1j} + a_2p_{2j} + a_3p_{3j}. \qquad \blacksquare$$

用矩阵的语言来说，行向量（row vector）$\mathbf{b} = \mathbf{a}P$。

阶段 2 平稳分布

想一想 我们要求 $\mathbf{a} = \mathbf{a}P$，且 $a_1 + a_2 + a_3 = 1$。把各方程写出来。哪些最简单？

各方程为

$$\begin{aligned} a_1 &= 0.3a_1 + 0.2a_2 + 0.1a_3, \\ a_2 &= 0.3a_1 + 0.4a_2 + 0.3a_3, \\ a_3 &= 0.4a_1 + 0.4a_2 + 0.6a_3. \end{aligned}$$

（它们并非独立：把三式相加，两边都得 $a_1 + a_2 + a_3$。）由第二式得 $0.6a_2 = 0.3(a_1 + a_3) = 0.3(1 - a_2)$，所以 $a_2 = \frac{1}{3}$。由第一式得 $7a_1 = 2a_2 + a_3 = \frac{2}{3} + a_3$，而 $a_1 + a_3 = \frac{2}{3}$，所以 $8a_1 = \frac{4}{3}$，$a_1 = \frac{1}{6}$，$a_3 = \frac{1}{2}$：

$$(a_1, a_2, a_3) = \left(\tfrac{1}{6}, \tfrac{1}{3}, \tfrac{1}{2}\right). \qquad \blacksquare$$

阶段 3 把链倒过来运行

想一想 $q_{ij} = \mathrm{P}(\text{上一个}i \mid \text{现时}j)$。运用贝叶斯定理（Bayes' theorem），上一个定理和现时的定理都采用平稳分布。

根据贝叶斯定理，

$$q_{ij} = \frac{\mathrm{P}(\text{上一个}i)\,\mathrm{P}(\text{现时}j \mid \text{上一个}i)}{\mathrm{P}(\text{现时}j)} = \frac{a_ip_{ij}}{a_j}.$$

例如 $q_{11} = \frac{\frac{1}{6}\times 0.3}{\frac{1}{6}} = 0.3$ 及 $q_{21} = \frac{\frac{1}{3}\times 0.2}{\frac{1}{6}} = 0.4$。全部结果如下（以 i 标示行，j 标示列）：

$$(q_{ij}) = \begin{pmatrix} 0.3 & 0.15 & \frac{2}{15} \\ 0.4 & 0.4 & \frac{4}{15} \\ 0.3 & 0.45 & 0.6 \end{pmatrix}. \qquad \blacksquare$$

现在是每一列之和为 1，因为第 j 列列出的是上一个定理各种可能的类型。

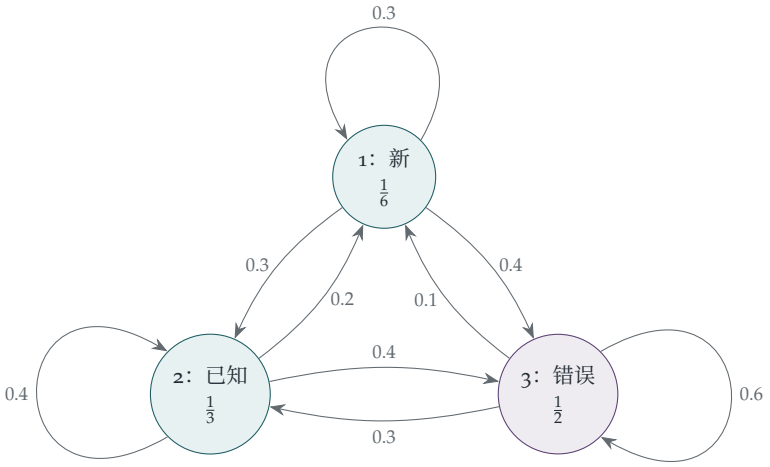


图 12.16.1 把 Ecks 教授的定理看成一条马可夫链。箭头显示下一个定理属于各类型的概率。圆圈内的分数是平稳分布：长远来说，每六个定理有一个是新而正确的，每三个有一个是已知的，而有一半是错的。

回顾

- **以矩阵表达全概率定律。**链每走一步，就是把一个行向量乘以矩阵 P。
- **平稳即特征向量（eigenvector）。**$\mathbf{a} = \mathbf{a}P$ 说明 $\mathbf{a}$ 是 P 的左特征向量，对应的特征值（eigenvalue）为 1。
- **用贝叶斯定理倒推。**逆向概率构成另一个随机矩阵（stochastic matrix），由正向的随机矩阵和平稳分布构造而成。

背后的数学

收敛至平衡

让 Ecks 教授从任何一个定理开始；再证明 n 个之后，类型的分布是 $\mathbf{a}_0 P^n$。对于像这样的一条链，即每种类型都可以紧接在任何一种类型之后，不论起点为何，P^n 都会收敛至一个每一行均等于平稳分布 $(\frac{1}{6}, \frac{1}{3}, \frac{1}{2})$ 的矩阵。收敛的速度由 P 的其他特征值（这里是 0.2 和 0.1）决定，它们以几何方式缩小：只需几个定理之后，她的过去就被遗忘了。这个关于马可夫链的*遍历定理*（*ergodic theorem*）是大学二年级概率课程的核心结果，马可夫链蒙地卡罗（Markov chain Monte Carlo）方法之所以行得通，靠的正是它：要从一个复杂的分布中抽样，就设计一条以该分布为平稳分布的马可夫链，然后让它运行一段长时间。

可逆性

逆向链 (q_{ij}) 与正向链 (p_{ij}) 并不相同：从她的定理的一份长纪录中，你可以分辨出时间的方向。若一条链的逆向版本与正向版本一致，即对所有 i、j 都有 $a_i p_{ij} = a_j p_{ji}$，就称为*可逆*（*reversible*）；网络上的随机游走（random walk）以及处于热平衡的物理系统的模型都属此类。可逆性是一个威力强大的性质，因为它令平稳分布很容易求出。

马可夫与普希金

马可夫（Andrey Markov）在 1906 年引入他的链，部分原因是要证明大数定律（law of large numbers）并不需要独立性；当时他正与数学家 Pavel Nekrasov 争论，后者声称独立性是必需的。1913 年，马可夫把他的想法应用于普希金（Pushkin）的诗体小说《叶甫盖尼·奥涅金》（*Eugene Onegin*）的首 20 000 个字母，把每个字母分为元音或辅音，并证明每个字母的类型取决于前一个字母的类型，而这种关系可用一条两个状态的链很好地描述。这是史上第一次对文本作统计分析，也是所有预测输入系统的始祖。

思维技巧

概率作向量，转移作矩阵

当一个系统以固定的概率在各状态之间转移时，把分布写成向量，把转移写成矩阵。这样，长远行为就变成一个关于特征向量的问题。

在大学里，这样做会把第 7 章的线性代数带进概率论。佩龙－弗罗宾尼斯定理（Perron–Frobenius theorem）保证平稳分布存在，而其他特征值则量度到达平稳分布有多快。

延伸练习

1 **兔子窝。**试做 STEP III 2003，第 13 题（`step.maths.org/questions/03-s3-q13`）：在一个正方形的四个顶点上的随机游走，每个顶点各有一个出口，以首步方程求解。

2 **两步。**若 Ecks 教授上一个定理是新而正确的，求再下一个定理属于三种类型的概率。

3 **特征值。**证明 P 的特征值为 1、0.2 和 0.1。

4 **可逆吗？**检查 Ecks 教授的链是否满足 $a_i p_{ij} = a_j p_{ji}$。答案告诉你甚么？

延伸阅读

- **James Norris, *Markov Chains***（Cambridge University Press）。标准的大学本科教科书。
- **Olle Häggström, *Finite Markov Chains and Algorithmic Applications***（Cambridge University Press）。篇幅短小，讲马可夫链蒙地卡罗尤其出色。
- **Brian Hayes, “First links in the Markov chain”**，*American Scientist* 101 (2013)。重述马可夫对普希金作品的分析。

12.17 巴拿赫的火柴盒

题目

STEP II 1992，第 16 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/92-s2-q16`

引子

相传数学家巴拿赫（Stefan Banach）烟瘾很大，两边口袋各放一盒火柴。每次要用火柴，他就随机选一个口袋。有一天，他发现其中一盒空了。另一盒还剩多少根火柴？这就是*巴拿赫火柴盒问题*（Banach's matchbox problem），由斯坦豪斯（Hugo Steinhaus）为向巴拿赫致意而设计；本题则是的士司机的版本，把火柴换成了糖果。

微妙之处在于“发现”一词。司机取走一包糖的最后一颗时，并不会察觉那包已经空了，要到下一次伸手去拿那包糖时才知道。所以这个事件是：她伸手去拿太妃糖，而全部 n 颗太妃糖早已吃完。计算这件事在第 N 次伸手时发生的方法数目，是二项分布（binomial distribution）的一道巧妙练习；而巴拿赫问题的答案有一个出人意表的特点：另一盒最可能剩下的火柴数目是 0（与 1 并列）。

动笔之前

想一想 若司机在第 N 次伸手时发现太妃糖已经吃完，第 N 次伸手必须是甚么？之前的 $N-1$ 次伸手又必须包含甚么？

第 N 次伸手必须是去拿太妃糖，而之前的 $N-1$ 次伸手必须刚好取走了 n 颗太妃糖（即全部）和 $N-1-n$ 颗薄荷糖。此外，她在此之前不可以已经发现薄荷糖空了。

逐步解题

阶段 1 发现太妃糖吃完了

想一想 头 $N-1$ 次伸手可以按甚么次序出现？$N-1-n$ 要符合甚么条件，才能确保她从未发现那包薄荷糖是空的？

那 n 次拿太妃糖的伸手，可以是头 $N-1$ 次伸手中的任何 n 次。只要她最多取了 m 颗薄荷糖，即 $N-1-n \le m$，她就从未在薄荷糖吃完之后再伸手去拿薄荷糖。因此，对于 $n+1 \le N \le n+m+1$，

$$\mathrm{P}(T_N) = \binom{N-1}{n} p^{n+1} q^{N-1-n},$$

其中 T_N 是事件“她在第 N 次伸手时发现太妃糖吃完了”（因子 p^{n+1} 已包括第 N 次伸手本身）。

阶段 2 发现薄荷糖吃完了

把两者的角色对调，用同样的论证可得，对于 $m+1 \le N \le n+m+1$，

$$\mathrm{P}(M_N) = \binom{N-1}{m} q^{m+1} p^{N-1-m}.$$

阶段 3 条件概率

想一想 已知在第 N 次伸手时有所发现。那次拿的是太妃糖的概率是多少？

事件 T_N 与 M_N 互斥（exclusive），所以

$$\mathrm{P}(\text{太妃糖} \mid \text{有所发现的伸手次数为} N) = \frac{\binom{N-1}{n} p^{n+1} q^{N-1-n}}{\binom{N-1}{n} p^{n+1} q^{N-1-n} + \binom{N-1}{m} q^{m+1} p^{N-1-m}},$$

其中二项式系数（binomial coefficient）在上述范围以外一律视为 0。（若 $N \le m$，则只有太妃糖可能已经吃完，概率为 1；同样，若 $N \le n$ 但 $N > m$，概率为 0。） ■

停一停，试一试 当 $n = m$ 且 $p = q = \frac{1}{2}$ 时，验算一下这条公式。根据对称性，对每一个可能的 N，答案都应该是 $\frac{1}{2}$。是否如此？

回顾

- **准确理解事件。**“发现”是指伸手去拿一包已空的糖，而不是取走最后一颗糖。

- **固定最后一步，再数其余。**第 N 次伸手已经固定；头 $N-1$ 次可以按任何次序出现，于是得出一个二项式系数。
- **检查边界。**限制条件 $N-1-n \le m$ 和 $N-1-m \le n$ 排除了不可能出现的过程。

背后的数学

巴拿赫的火柴盒

在巴拿赫的版本中，每盒起初有 n 根火柴，选中每个口袋的机会均等。当他首次发现其中一盒空了，设 K 为另一盒的火柴数目。若 $K=k$，他必定已从空的那盒取了 n 根，从另一盒取了 $n-k$ 根，次序不拘，然后再一次伸手去拿空的那盒；而两盒之中任何一盒都可以是空的那盒：

$$\mathrm{P}(K=k) = 2\binom{2n-k}{n}\left(\frac{1}{2}\right)^{2n-k+1} = \binom{2n-k}{n}\left(\frac{1}{2}\right)^{2n-k}.$$

这些概率随 k 增大而递减，唯一例外是 $\mathrm{P}(K=0)=\mathrm{P}(K=1)$（图 12.17.1）。剩余火柴数目的期望值（expected value）约为 $2\sqrt{n/\pi}-1$：若每盒有 50 根火柴，约为 7。这个 $\sqrt{n}$ 是经过 $2n$ 次随机选择之后，两盒之间差距的典型大小，与生日问题（Section 9.9）和随机漫步（random walk）中的平方根定律如出一辙。

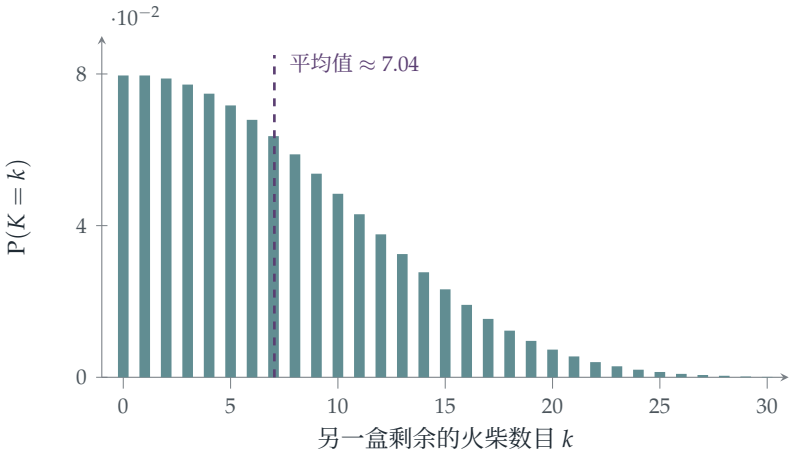


图 12.17.1 每盒有 50 根火柴的巴拿赫火柴盒问题：首次发现其中一盒空了的时候，另一盒剩余火柴数目的分布。最可能的数值是 0 和 1，两者概率相同，但平均值约为 7。

负二项分布

在一连串独立试验（independent trials）中，每次试验成功的概率为 p，则取得第 r 次成功所需的试验次数服从*负二项分布*（negative binomial distribution）：$\mathrm{P}(N=k)=$

$\binom{k-1}{r-1}p^r q^{k-r}$。的士司机的问题，是两个负二项过程（太妃糖和薄荷糖）之间的竞赛，每个过程所需的“成功”次数都比该包糖的数目多一次。负二项分布见于保险（索偿宗数）、生态学（动物聚集分布的数目）和遗传学，往往用来代替泊松分布（Poisson distribution），因为它更有弹性。

巴拿赫与苏格兰咖啡馆

巴拿赫是泛函分析（functional analysis）的奠基人之一，他的数学研究有不少是在 Lwów（今称 Lviv）的苏格兰咖啡馆（Scottish Café）里完成的。他和同事把问题写在一本由侍应保管的笔记簿上，这就是《苏格兰书》（Scottish Book），解题奖品从一杯啤酒到一只活鹅不等。斯坦豪斯是巴拿赫的恩师，他为了一篇向巴拿赫的吸烟习惯致意的演说，设计了这道火柴盒问题。最先在书刊中分析这个问题的，是费勒（William Feller）1950 年出版的著名概率论教科书。

思维技巧

固定最终事件，数出通往它的路径

要求一个过程在某一时刻以某种方式结束的概率，可先固定最后一步发生的事，再数出通往这一步的所有过程。

在大学里，这是负二项分布、随机漫步的首达时间（first-passage time），以及 Section 9.10 中反射原理（reflection principle）的基础。

延伸练习

1 **小规模情况。**设 $n = m = 1$ 及 $p = \frac{1}{2}$，列出所有可能的序列，并求 N 的分布。

2 **巴拿赫公式。**对 $n = 1$ 和 $n = 2$，验证 $\sum_{k=0}^{n} \binom{2n-k}{n} 2^{-(2n-k)} = 1$。

3 **并列最高。**证明在巴拿赫问题中，对每一个 n 都有 $\mathrm{P}(K = 0) = \mathrm{P}(K = 1)$。

4 **哪一包先吃完？**就的士司机的问题，当 $n = m = 2$ 及 $p = \frac{2}{3}$ 时，求太妃糖是首先被发现空了的那一包（不论何时发现）的概率。

延伸阅读

► **William Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1**（Wiley），

第 VI 章，正是这本书把巴拿赫问题介绍给全世界。

- **R. Daniel Mauldin (ed.), *The Scottish Book***（Birkhäuser）。收录苏格兰咖啡馆的问题，并附评注。
- **Geoffrey Grimmett & Dominic Welsh, *Probability: An Introduction***（Oxford University Press），内容涉及负二项分布。

12.18 无记忆性

题目

STEP III 2021，第 11 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2021 年 STEP 试卷（zip 文件）内的 STEP III 试卷：

`www.ocr.org.uk/Images/696588-2021-past-papers.zip`

引子

一颗放射性原子在实验室里待了一百万年，一直没有衰变。它是否“快要”衰变了？不是：它在下一秒内衰变的机会，与一颗刚刚产生的原子完全相同。放射性原子不会老化。直至衰变为止的等候时间服从指数分布（*exponential distribution*），而这个分布的决定性特征，就是它没有记忆。

这道题以一种令人印象深刻的形式展示了这种无记忆性（memorylessness）。把一个指数等候时间拆成整数个单位和小数部分。原来这两部分是独立（*independent*）的：知道你等了多少个完整的小时，对于分钟数毫无提示。这道题还会在练习中解释，为甚么你等巴士的时间，可以显得比两班巴士之间的间隔还要长。

动笔之前

想一想 对于 $x \geq 0$，$\mathrm{P}(X > x)$ 是甚么？$\mathrm{P}(X > s+t \mid X > s)$ 又是甚么？

$\mathrm{P}(X > x) = \int_x^\infty \lambda e^{-\lambda u}\,\mathrm{d}u = e^{-\lambda x}$。于是 $\mathrm{P}(X > s+t \mid X > s) = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = \mathrm{P}(X > t)$：已经等了 s 之后，余下的等候时间与从头开始等候的分布相同。这就是无记忆性质（*memoryless property*）。

逐步解题

阶段 1 整数部分

$Y = n$ 当且仅当 $n \le X < n+1$：

$$\mathrm{P}(Y = n) = e^{-\lambda n} - e^{-\lambda(n+1)} = (1 - e^{-\lambda})e^{-n\lambda}. \qquad \blacksquare$$

所以 Y 服从几何分布（geometric distribution）：每一个完整单位都以概率 $e^{-\lambda}$ 被“挨过”。

阶段 2 小数部分

想一想 $Z < z$ 表示 X 落在其中一个区间 $[n, n+z)$ 内。把这些区间的概率加起来。

$$\mathrm{P}(Z < z) = \sum_{n \ge 0} \left(e^{-\lambda n} - e^{-\lambda(n+z)}\right) = (1 - e^{-\lambda z}) \sum_{n \ge 0} e^{-\lambda n} = \frac{1 - e^{-\lambda z}}{1 - e^{-\lambda}}. \qquad \blacksquare$$

阶段 3 Z 的平均值

想一想 对于取值于 $[0,1]$ 的随机变量（random variable），$\mathrm{E}(Z) = \int_0^1 \mathrm{P}(Z > z)\,\mathrm{d}z$。运用这条公式。

设 $c = 1 - e^{-\lambda}$，则 $\mathrm{P}(Z > z) = 1 - \frac{1-e^{-\lambda z}}{c}$，所以

$$\mathrm{E}(Z) = 1 - \frac{1}{c}\left(1 - \frac{1 - e^{-\lambda}}{\lambda}\right) = 1 - \frac{1}{c} + \frac{1}{\lambda} = \frac{1}{\lambda} - \frac{e^{-\lambda}}{1 - e^{-\lambda}} = \frac{1}{\lambda} - \frac{1}{e^{\lambda} - 1}. \qquad \blacksquare$$

（当 λ 很小，即等候时间很长时，这个值趋向 $\frac{1}{2}$：小数部分几乎是均匀分布的。当 λ 很大时，它接近 $\frac{1}{\lambda}$，因为 X 几乎总是小于 1。）

阶段 4 独立性

$$\mathrm{P}(Y = n, z_1 < Z < z_2) = \mathrm{P}(n + z_1 < X < n + z_2) = e^{-\lambda n}\left(e^{-\lambda z_1} - e^{-\lambda z_2}\right).$$

想一想 利用第 (i) 及第 (ii) 部分，与 $\mathrm{P}(Y = n)\,\mathrm{P}(z_1 < Z < z_2)$ 作比较。

由 (i) 及 (ii)，$\mathrm{P}(Y = n)\mathrm{P}(z_1 < Z < z_2) = (1 - e^{-\lambda})e^{-n\lambda} \cdot \frac{e^{-\lambda z_1} - e^{-\lambda z_2}}{1 - e^{-\lambda}}$，两者相同。所以 Y 与 Z 独立。 $\blacksquare$

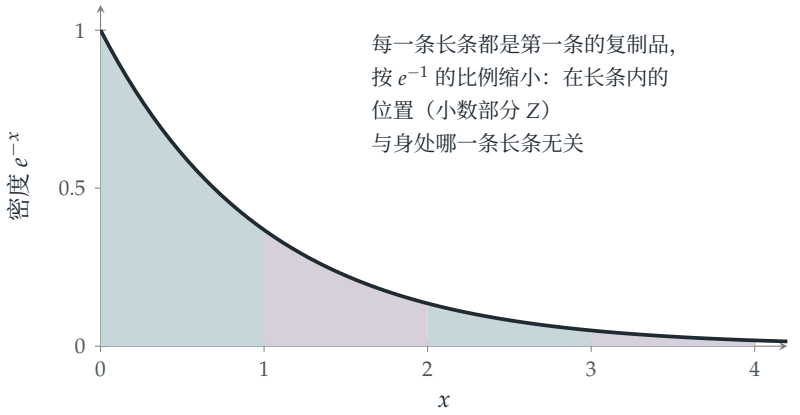


图 12.18.1 $\lambda = 1$ 的指数密度，分割成单位宽度的长条。由于 $e^{-(n+z)} = e^{-n}e^{-z}$，每一条长条的形状都相同。$X$ 落在哪一条长条（整数部分 Y），与它落在长条内的甚么位置（小数部分 Z），两者是独立的。

回顾

- **逐段求和。** Z 的分布和联合概率（joint probability），都是把各区间 $[n, n+1)$ 上的概率相加而得的。
- **尾部公式。** $\mathrm{E}(Z) = \int \mathrm{P}(Z > z)\,\mathrm{d}z$ 让我们不必求出密度。
- **能够分解即是独立。** 联合概率分拆成一个 n 的函数乘以一个 z_1、z_2 的函数。

背后的数学

指数分布为何特别

Y 与 Z 的独立性直接来自 $e^{-\lambda(n+z)} = e^{-\lambda n}e^{-\lambda z}$，而这正是无记忆性质的另一副面貌。事实上，指数分布是唯一没有记忆的连续分布：若对所有 $s, t \geq 0$ 都有 $\mathrm{P}(X > s+t) = \mathrm{P}(X > s)\mathrm{P}(X > t)$，则 $g(t) = \ln \mathrm{P}(X > t)$ 满足 $g(s+t) = g(s) + g(t)$，而（对于单调（monotonic）的 g）这就迫使 $g(t) = -\lambda t$。正因如此，指数分布可用来描述那些“随机”发生、没有任何累积过程的事件的等候时间：放射性衰变、顾客的到达、不会损耗的电子元件的故障。

最小值与等候时间悖论

若 X_1 与 X_2 是独立的指数随机变量，率分别为 λ_1 和 λ_2，则 $\mathrm{P}(\min > x) = e^{-\lambda_1 x}e^{-\lambda_2 x}$，所以最小值服从率为 $\lambda_1 + \lambda_2$ 的指数分布。等候两个独立随机事件中最先发生的一个，就

如同以合并后的率等候单一事件（STEP III 2002 Q13，见练习）。无记忆性也解释了*等候时间悖论*（*waiting-time paradox*）：若巴士随机到站，平均间隔为 10 分钟，则一位在随机时刻到达的乘客平均要等 10 分钟，而不是 5 分钟，因为余下的等候时间是无记忆的；而乘客到达时所处的那个间隔，平均长达 20 分钟，因为较长的间隔更容易被碰上。

卢瑟福（Rutherford）与泊松过程（Poisson process）

1910 年，欧内斯特·卢瑟福与汉斯·盖革（Hans Geiger）以 7.5 秒为一个时段，数算一个放射源所放出的阿尔法粒子（alpha particles），发现这些计数服从泊松分布（Poisson distribution）；数学家 Harry Bateman 证明了，只要各次放射以恒定的率独立发生，结果就理应如此。这样一来，相邻两次放射之间的时间便服从指数分布。*泊松过程*，即间隔服从指数分布的一连串事件，如今已成为描述时间上随机性的基本模型，应用于电话通话、保险索偿以及神经元的放电；下一节 Section 12.19 会把它推广到空间。

思维技巧

以因式分解检验独立性

若两个随机变量的联合概率可分解为其中一个变量的函数乘以另一个变量的函数，则这两个随机变量独立。

在大学里，联合密度（joint density）能否分解是检验独立性的标准方法，你亦会借此在多变量问题中辨认出独立的分量，例如 Section 12.22 中服从正态分布（normal distribution）的点的半径与角度。

延伸练习

1 **两条巴士路线。**试做 STEP III 2002，第 13 题（`step.maths.org/questions/02-s3-q13`）。这道题利用两个指数随机变量的最小值，求出等候下一班巴士的时间。

2 **离散的无记忆性。**证明几何分布是无记忆的：$\mathrm{P}(Y \geq m+n \mid Y \geq m) = \mathrm{P}(Y \geq n)$。

3 **等候时间悖论。**巴士的到站是一个率为 λ 的泊松过程。证明包含某固定时刻 t（t 很大）的那个间隔，其期望长度为 $\frac{2}{\lambda}$，即平均间隔的两倍。

4 Z **的密度。**求 Z 的概率密度（probability density），并就 $\lambda = 1$ 和 $\lambda = 5$ 绘出其草图。

延伸阅读

- **Geoffrey Grimmett & Dominic Welsh, *Probability: An Introduction***（Oxford University Press），论及指数分布与泊松过程。
- **Sheldon Ross, *Introduction to Probability Models***（Academic Press）。指数分布、泊松过程与排队。
- **David Spiegelhalter, *The Art of Uncertainty***（Pelican）。日常生活中的随机性与等候。

12.19 平面上的随机点

题目

STEP II 2012，第 13 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/12-s2-q13`

引子

假如超级市场、树木、恒星或流动电话发射站“随机”散布在某个区域内，你要走多远才到达最近的一个？答案只取决于密度 k（每单位面积的平均数目），而且有一个独特的形状，就是瑞利分布（*Rayleigh distribution*）。平均而言，最近的一个位于距离 $\frac{1}{2\sqrt{k}}$ 处：若每平方公里有一间超级市场，这距离就是 500 米。

这个模型称为空间泊松过程（*spatial Poisson process*），即时间上随机到达（Section 12.18）的二维版本。计算过程简短而优美，但它真正的价值在于充当基准：生态学家把实际的最近邻距离与这条公式比较，以判断树木或鸟巢究竟是聚集成群、随机散布，还是刻意彼此隔开。

动笔之前

想一想 一个平均值为 μ 的泊松随机变量（Poisson random variable），取值 0 的概率是多少？

$\mathrm{P}(0) = e^{-\mu}$。半径为 y 的圆面积是 πy^2，所以圆内超级市场数目的平均值为 $k\pi y^2$。

逐步解题

阶段 1 附近没有超级市场

在距离 y 以内没有超级市场的概率是 $e^{-\pi k y^2}$。

阶段 2 最近距离的分布

? **想一想** $Y \geq y$ 表示在距离 y 以内没有超级市场。那么 $\mathrm{P}(Y < y)$ 是甚么?

$\mathrm{P}(Y < y) = 1 - e^{-\pi k y^2}$，求导后便得到密度函数（density）

$$\mathrm{f}(y) = 2\pi k y e^{-\pi k y^2}, \qquad y \geq 0. \qquad \blacksquare$$

阶段 3 平均值

? **想一想** 可以用分部积分法（integration by parts）计算 $y\,\mathrm{f}(y)$ 的积分，也可以利用 $\mathrm{E}(Y) = \int_0^\infty \mathrm{P}(Y > y)\,\mathrm{d}y$。然后作代换，以便使用题目给出的高斯积分（Gaussian integral）。

$$\mathrm{E}(Y) = \int_0^\infty e^{-\pi k y^2}\,\mathrm{d}y.$$

代入 $x = y\sqrt{2\pi k}$，使得 $\pi k y^2 = \frac{1}{2}x^2$，便得 $\mathrm{E}(Y) = \frac{1}{\sqrt{2\pi k}}\sqrt{\frac{\pi}{2}} = \frac{1}{2\sqrt{k}}$。 $\blacksquare$

阶段 4 方差

$$\mathrm{E}(Y^2) = \int_0^\infty y^2 \cdot 2\pi k y e^{-\pi k y^2}\,\mathrm{d}y.$$

代入 $u = \pi k y^2$（因此 $\mathrm{d}u = 2\pi k y\,\mathrm{d}y$），上式变成 $\int_0^\infty \frac{u}{\pi k} e^{-u}\,\mathrm{d}u = \frac{1}{\pi k}$。因此

$$\mathrm{Var}(Y) = \frac{1}{\pi k} - \frac{1}{4k} = \frac{4-\pi}{4\pi k}. \qquad \blacksquare$$

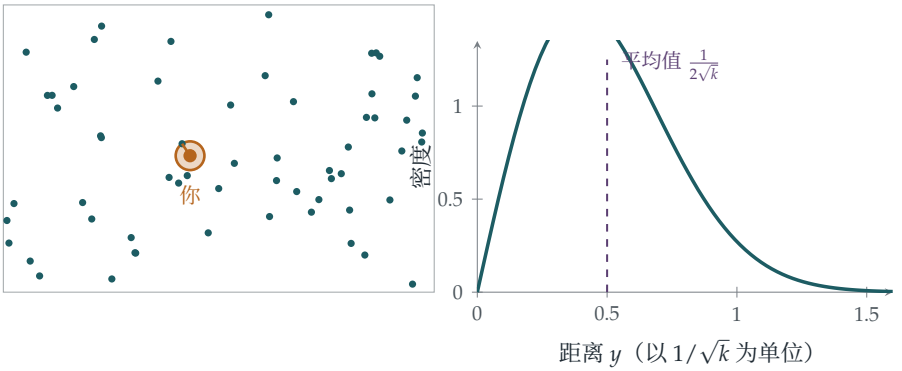


图 12.19.1 左：六十个点随机散布在面积为 24 的区域内（所以 $k = 2.5$），以及由选定的一点到其中最近一点的距离（琥珀色）。右：该距离的瑞利密度函数 $2\pi k y e^{-\pi k y^2}$，距离以 $1/\sqrt{k}$ 为单位。其平均值为 $\frac{1}{2\sqrt{k}}$。

回顾

- **把距离问题化为计数问题。**“最近的一个比 y 更远”就是“在 y 以内一个也没有”。
- **先求累积分布，再求密度。** $\mathrm{P}(Y < y)$ 可以直接写出；求导后便得到密度函数。
- **以代换化为已知积分。**每个矩（moment）都化成了高斯积分或指数积分。

背后的数学

平面上的泊松过程

强度（intensity）为 k 的空间泊松过程，是一种随机散布的点集，满足：互不相交的区域内的点数彼此独立，而面积为 A 的区域内的点数服从平均值为 kA 的泊松分布。把极大量的点独立而均匀地投放到一片极大的区域上，并保持适当的平均密度，得到的就是这种过程：这正是“稀有事件定律”（law of rare events），它同样解释了错排（derangement）的现象（Section 9.8）。它是完全空间随机性（complete spatial randomness）的自然模型，实际的分布形态都以它为比较标准。*Clark–Evans 检验*（*Clark–Evans test*）由两位生态学家于 1954 年提出，把观察所得的平均最近邻距离与 $\frac{1}{2\sqrt{k}}$ 比较：小得多表示有聚集现象（种子落在母树附近），大得多则表示间隔规则（有领域行为的鸟类，或植物之间争夺水分）。

三维的情况

在三维空间中作同样的计算，可得到最近的恒星、分子或星系的距离满足 $\mathrm{P}(Y > y) = e^{-\frac{4}{3}\pi k y^3}$。钱德拉塞卡（Subrahmanyan Chandrasekhar）在 1943 年用它来研究一颗恒星受随机分布的邻近恒星所施的引力，物理学家则用它来计算气体分子的平均自由程（mean free path）。

瑞利

密度函数 $\frac{y}{\sigma^2} e^{-y^2/(2\sigma^2)}$ 以瑞利勋爵（Lord Rayleigh）命名，他在 1880 年研究声学理论时遇上它：它描述了许多相位随机的波叠加后的振幅。若一点的两个坐标是独立的正态随机变量（normal random variable），它也给出该点与原点距离的分布，这是理解它为何在这里出现的另一个角度（练习 3）。今天，它用来模拟流动电话信号的衰落，因为信号是沿着许多相位随机的路径到达接收器的。

思维技巧

先求分布函数

要求最小值、最大值或最近距离的分布时，$\mathrm{P}(Y > y)$ 往往很容易直接写出。求导便得到密度函数。

这个方法会在 Section 12.25 再次出现，用来处理许多随机寿命中最长的一个。在大学里，这是处理任何“顺序统计量”（order statistic）的标准第一步。

延伸练习

1 **中位数。**求到最近超级市场的距离的中位数（median），并与平均值比较。

2 **第二近的。**求到第二近的超级市场的距离的分布。（提示：$\mathrm{P}(Y_2 > y)$ 是在 y 以内最多只有一间超级市场的概率。）

3 **正态坐标。**若 U 和 V 是独立的 $\mathrm{N}(0, \sigma^2)$ 随机变量，证明 $\sqrt{U^2 + V^2}$ 的密度函数为 $\frac{r}{\sigma^2} e^{-r^2/(2\sigma^2)}$。$\sigma$ 取何值时会得到超级市场的分布？

4 **Clark–Evans。**在一项调查中，一片面积 1 km^2 的树林内有 400 棵树，而一棵树到其最近邻的平均距离为 35 m。这与完全空间随机性相符吗？还是显示有聚集或间隔的现象？

延伸阅读

- **Peter Diggle, *Statistical Analysis of Spatial Point Patterns***（Arnold）。这方面的标准参考书，载有 Clark–Evans 检验及其后续发展的方法。
- **John Kingman, *Poisson Processes***（Oxford University Press）。一本优美的短篇著作，讨论任意维数的泊松过程。
- **Geoffrey Grimmett & David Stirzaker, *Probability and Random Processes***（Oxford University Press），论泊松过程的一章。

12.20 蒲丰投针

题目

STEP I 1993，第 16 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/93-s1-q16`

引子

把一根针抛到铺有平行地板条的地板上。针横跨板缝的机会竟然与 π 有关，尽管眼前根本没有任何圆形。抛很多次并数一数横跨板缝的次数，就可以用实验来估算 π。这就是*布丰投针问题*（*Buffon's needle*），由布丰伯爵乔治-路易·勒克莱尔（Georges-Louis Leclerc, Comte de Buffon）于 1733 年提出，并于 1777 年由他本人解决：它是*几何概率*（*geometric probability*）的第一个问题，也是所有蒙地卡罗方法（Monte Carlo method）的始祖。

这个 π 来自角度：针的方向在半圈之内均匀分布（uniform），把针横跨平行线方向的“影子”长度对所有方向取平均，便得出因子 $\frac{2}{\pi}$。本题则以 $\cos^{-1}$ 的积分来完成这个计算。

动笔之前

想一想 针的中心与最近一条直线的距离为 x，而针与这些直线的垂直线所成的角为 y。针朝着该直线能伸展多远？

针的每一半长度为 a，它在垂直于直线方向上的分量是 $a\cos y$。因此，针与最近的直线相交，当且仅当 $a\cos y \geq x$。（由于 $a < b$，针不可能触及其他直线。）

逐步解题

阶段 1 积分

设 $t=\cos y$，当 t 由 0 变到 1 时，$y=\cos^{-1}t$ 由 $\frac{\pi}{2}$ 变到 0，而 $\mathrm{d}t=-\sin y\,\mathrm{d}y$。所以

$$\int_0^1 \cos^{-1}t\,\mathrm{d}t=\int_0^{\pi/2} y\sin y\,\mathrm{d}y=[-y\cos y]_0^{\pi/2}+\int_0^{\pi/2}\cos y\,\mathrm{d}y=0+1=1. \qquad \blacksquare$$

（或者：这个积分是 $\cos^{-1}$ 的图像与两条坐标轴之间的面积，它等于 $y=\cos x$ 由 0 到 $\frac{\pi}{2}$ 的曲线下面积，两者只是沿 $y=x$ 反射而已。）

阶段 2 给定位置

想一想 对于固定的 x，哪些角度 y 会使针与直线相交？这些角度的概率是多少？

若 $x\geq a$，针无法触及直线。若 $x<a$，相交的条件是 $\cos y\geq\frac{x}{a}$，即 $y\leq\cos^{-1}\frac{x}{a}$。由于 Y 在 $[0,\frac{\pi}{2}]$ 上均匀分布，

$$\mathrm{P}(\text{相交}\mid X=x)=\frac{2}{\pi}\cos^{-1}\frac{x}{a}\qquad(0\leq x<a). \qquad \blacksquare$$

阶段 3 对位置取平均

X 在 $[0,b]$ 上均匀分布，其密度（density）为 $\frac{1}{b}$，所以

$$\mathrm{P}(\text{相交})=\frac{1}{b}\int_0^a\frac{2}{\pi}\cos^{-1}\frac{x}{a}\,\mathrm{d}x=\frac{2a}{\pi b}\int_0^1\cos^{-1}t\,\mathrm{d}t=\frac{2a}{\pi b}, \qquad \blacksquare$$

其中用了代换 $x=at$。

停一停，试一试 若针的长度恰好等于半间距（$a=\frac{b}{2}$，即针的长度 $2a$ 等于 b），相交的概率是 $\frac{1}{\pi}$。要把 π 估算至小数点后两位，你需要抛多少次？



图 12.20.1 三十根长度为 $2a=1.4$ 的针，抛在相距 $2b=2$ 的直线上。与直线相交的针以紫红色表示。这里有 17 根相交，而期望值是 13.4 根：小样本只能给出粗略的估算，而布丰投针正是出名缓慢的 π 计算方法。

回顾

- **两个独立的坐标。**针的位置和角度互相独立（independent）且均匀分布，所以所求概率是一个二重积分（double integral），可以逐个变量计算。
- **先取条件，再求平均。**先固定 x，对角度取平均；然后再对 x 取平均。
- **代换显露出常数。**以 $x = at$ 作缩放，答案便化约为单一个通用的积分。

背后的数学

布丰的面条

以下的方法几乎不用计算就能看出答案。把针屈曲成任意形状，成为一根“面条”，并设 C 为它与直线相交的次数。根据期望值的线性（linearity of expectation），一条曲线相交次数的期望值，等于它各小段的期望值之和，所以与长度成正比：$\mathrm{E}(C) = \kappa L$，其中 κ 是某个常数。要求出 κ，可取面条为一个直径等于直线间距 $2b$ 的圆：它总是恰好相交两次，而其长度为 $2\pi b$。所以 $\kappa \cdot 2\pi b = 2$，得出 $\kappa = \frac{1}{\pi b}$。对于长度为 $2a < 2b$ 的直针，它最多相交一次，故 $\mathrm{P}(\text{相交}) = \mathrm{E}(C) = \frac{2a}{\pi b}$。这个论证出自 Joseph-Émile Barbier（1860 年），是概率推理中的瑰宝：那个 π 来自用来校准常数的圆。

估算 π，以及错误示范

若抛出 n 根针，其中 c 根相交，则 $\frac{c}{n} \approx \frac{2a}{\pi b}$，而 $\pi \approx \frac{2an}{bc}$。误差只会按 $\frac{1}{\sqrt{n}}$ 的速度缩小，所以要把 π 算至小数点后三位，大约需要抛一百万次。1901 年，意大利数学家 Mario Lazzarini 声称抛了 3408 次便得出 $\pi \approx \frac{355}{113} = 3.1415929\ldots$（准确至小数点后六位）。统计学家指出，这个结果好得令人难以置信：抛 3408 次，再配合一根精心挑选的针，恰好 1808 次相交便会得出 $\frac{355}{113}$，而只要在适当的时刻停止实验，就能制造出这个结果。这是关于选择性报告的经典警世故事。

布丰

布丰是一位博物学家，曾任巴黎皇家植物园园长，并著有一套 36 卷的自然史。他于 1733 年向法国科学院（Académie des Sciences）提出投针问题（以及一个使用方形地砖的版本，即“franc-carreau”游戏），并于 1777 年发表解答。他的目的是要说明，概率中不单可以运用算术，也可以运用几何。积分几何（integral geometry）——研究随机直线及其与曲线和曲面的交点的理论——正是由此发展而来，如今应用于体视学（stereology），即从二维切片估算三维结构（例如组织中的细胞）的大小和形状的科学。

思维技巧

把概率看成区域上的积分

当结果取决于数个独立的连续随机变量时，一个事件的概率就是联合密度（*joint density*）在该事件发生的区域上的积分。

在大学里，这成为连续概率的基础：联合密度、边际分布（marginal distribution）和变量代换（change of variable），这些全都会在 Section 12.21 和 Section 12.22 再次出现。

延伸练习

1 **长针。**若 $a > b$，针可能与多于一条直线相交。证明相交次数的期望值仍然是 $\frac{2a}{\pi b}$。

2 **面条。**把一根长度为 $2a$ 的针屈成直角（两臂各长 a）。相交次数的期望值是多少？

3 **方形地砖。**把一枚半径为 r 的硬币抛到铺有边长为 $s > 2r$ 的方形地砖的地板上。求硬币完全落在一块地砖之内的概率。

4 **Lazzarini 的针。**设 $a = \frac{5}{6}b$（使得 $\frac{2a}{\pi b} = \frac{5}{3\pi}$），证明抛 3408 次而有 1808 次相交，会给出估算值 $\pi \approx \frac{355}{113}$。

延伸阅读

- **Daniel Klain & Gian-Carlo Rota, *Introduction to Geometric Probability***（Cambridge University Press），此书由布丰投针问题讲起。
- **Lee Badger, "Lazzarini's lucky approximation of π"**，*Mathematics Magazine* 67 (1994)。一个可疑实验背后的统计学。
- **Paul Nahin, *Digital Dice***（Princeton University Press）。以蒙地卡罗方法处理概率问题。

12.21 折断一根棒

题目

STEP III 1990，第 16 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/90-s3-q16`

引子

在一根木棒上随机选两点，把它折断。三段木棒能否组成三角形？著名的答案是：概率（probability）为 $\frac{1}{4}$。如果能组成三角形，它是锐角（acute）三角形还是钝角（obtuse）三角形？通常是钝角三角形：所组成的三角形中，只有大约三分之一是锐角三角形。这类关于随机图形的问题，自十九世纪起便令概率学家乐此不疲，也是练习处理连续概率（continuous probability）的绝佳场地：区域内的一个随机点、一个界定子区域的条件，以及一个计算其面积的积分。

本题采用两种不同的折棒方法，并在每种情况下问出现钝角的可能性有多大。答案 $3-2\sqrt{2}$ 和 $3-4\ln 2$ 都是出人意表的数，却源自简单的几何。

动笔之前

想一想 三个长度能组成三角形，当且仅当每一个都小于其余两个之和。如果三个长度之和为 L，这对最长的一个有何含意？边 c 所对的角何时是钝角？

$c < a+b$ 连同 $a+b+c=L$，即 $c<\frac{L}{2}$；所以三段能组成三角形，正正是当每一段都短于总长度的一半。根据余弦公式（cosine rule），c 所对的角是钝角，正正是当 $c^2 > a^2+b^2$。

逐步解题

阶段 1 先折一次，再对半分

三段的长度是 X、$\frac{1-X}{2}$、$\frac{1-X}{2}$。两段相等的木棒必定短于 $\frac{1}{2}$，所以能组成三角形当且仅当 $X<\frac{1}{2}$。这个三角形是等腰（isosceles）三角形；两个相等的角都是锐角，所以只有当 X 所对的角是钝角时，它才是钝角三角形：

$$X^2>2\Big(\frac{1-X}{2}\Big)^2=\frac{(1-X)^2}{2},\qquad \text{即}\qquad \sqrt{2}X>1-X,\quad X>\frac{1}{1+\sqrt{2}}=\sqrt{2}-1.$$

已知 $X<\frac{1}{2}$ 时，X 在 $[0,\frac{1}{2}]$ 上均匀分布（uniform），所以

$$\mathrm{P}(\text{钝角}\mid\text{三角形})=\frac{\frac{1}{2}-(\sqrt{2}-1)}{\frac{1}{2}}=3-2\sqrt{2}\approx 0.172.\qquad\blacksquare$$

阶段 2 第二种构造方法

想一想 设 (x,y) 位于以 $(1,0)$、$(1,1)$、$(0,1)$ 为顶点的三角形内，对边长 x、y 和 $2-x-y$ 检查三条三角不等式（triangle inequality）。

该区域是 $\{x\le 1, y\le 1, x+y\ge 1\}$。三条不等式分别是：$x+y\ge 2-x-y$，即 $x+y\ge 1$；$x+(2-x-y)\ge y$，即 $y\le 1$；以及 $y+(2-x-y)\ge x$，即 $x\le 1$。它们在区域内全部成立（在区域的边上等号成立，此时三角形是退化（degenerate）的）。 ■

想一想 B 处的角所对的边是 $CA=y$。把条件 $y^2>x^2+(2-x-y)^2$ 写成关于 y 的不等式。

展开得 $y^2>x^2+4+x^2+y^2-4x-4y+2xy$，所以 $4y-2xy>2x^2-4x+4$，即 $2y(2-x)>2(x^2-2x+2)$。由于 $2-x>0$，

$$\angle ABC\text{ 为钝角}\iff y>\frac{x^2-2x+2}{2-x}.\qquad\blacksquare$$

阶段 3 面积

想一想 以 $2-x$ 除 x^2-2x+2。然后求区域中位于曲线上方那部分的面积，再除以区域的面积 $\frac{1}{2}$。

由于 $x^2-2x+2=-x(2-x)+2$，曲线是 $y=-x+\frac{2}{2-x}$。它由 $(0,1)$ 伸延至 $(1,1)$，而当 $0<x<1$ 时，它位于直线 $y=1-x$ 之上、$y=1$ 之下（图 12.21.1）。钝角区域位于曲线与 $y=1$ 之间，面积为

$$\int_0^1\Big(1+x-\frac{2}{2-x}\Big)\mathrm{d}x=1+\frac{1}{2}+\big[2\ln(2-x)\big]_0^1=\frac{3}{2}-2\ln 2.$$

除以 $\frac{1}{2}$：$\mathrm{P}(\angle ABC\text{ 为钝角})=3-4\ln 2\approx 0.227$。 ■

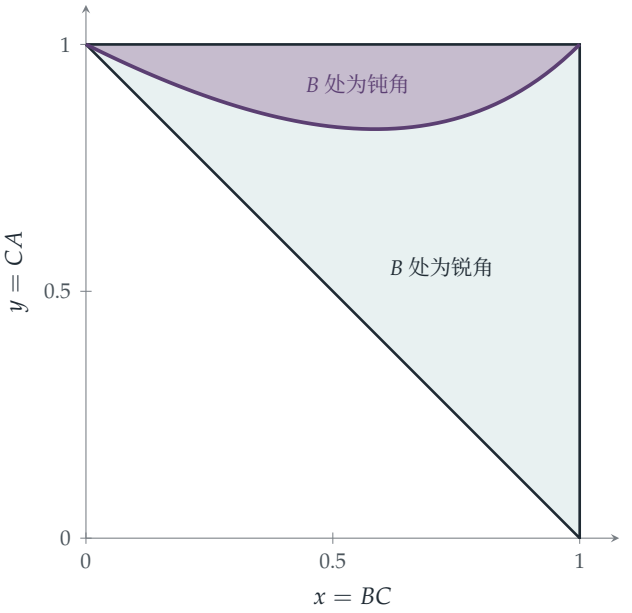


图 12.21.1 第 (ii) 部分中 (x, y) 所有可能取值构成的三角形。在曲线 $y = \frac{x^2-2x+2}{2-x}$ 上方（紫色区域），B 处的角是钝角；这个区域占三角形的比例为 $3 - 4\ln 2 \approx 0.23$。

回顾

- **把几何转化为不等式。**“能组成三角形”和“是钝角三角形”都化成了各长度之间的不等式。
- **连续变量的条件概率（conditional probability）。**已知 $X < \frac{1}{2}$ 时，X 在较小的区间上均匀分布。
- **概率即面积之比。**对于区域内均匀分布的点，概率等于面积除以总面积。

背后的数学

经典的折棒问题

在两个独立而均匀分布的点 U 和 V 处把木棒折断。数对 (U, V) 在单位正方形上均匀分布，而三段能组成三角形，正正是当每一段都短于 $\frac{1}{2}$：这个条件切去四个小三角形，总面积为 $\frac{3}{4}$，余下的概率是 $\frac{1}{4}$。已知能组成三角形时，它是锐角三角形的概率是 $\frac{1}{4}$ 当中的 $3\ln 2 - 2 \approx 0.079$，即大约 32%。这三段木棒是均匀随机分割（uniform random partition）的一个例子，它们的联合分布（joint distribution）是狄利克雷分布（Dirichlet distribution），

这种分布出现在贝叶斯统计学（Bayesian statistics）中，也出现在描述资源如何随机分配的模型中。

怎样折断木棒，大有关系

请留意，题目中的两种方法，对听起来相似的问题给出了不同的答案 $3-2\sqrt{2}$ 和 $3-4\ln 2$，而两者又都与经典方法的答案不同。正如贝特朗（Bertrand）的弦（Section 12.3）一样，在指明具体程序之前，“随机”一词毫无意义。

随机三角形

关于随机三角形的问题，是维多利亚时代流行的趣题。路易斯·卡罗（Lewis Carroll）的 *Pillow Problems*（1893 年）问及“随机”选取的三角形是钝角三角形的概率，而他的答案，正如本节的答案一样，完全取决于他所选择的随机性模型。关于随机三角形的问题一直带来惊喜：例如，在正方形内均匀随机选取三点，它们组成钝角三角形的概率约为 0.725，这个数值到二十世纪才有人准确算出。

思维技巧

样本空间（sample space）的几何

把随机结果表示为区域内的一点，把事件表示为子区域，并以面积之比（或密度（*density*）的积分）计算概率。

在大学里，这幅图像是多元概率（multivariate probability）的基础，也是变量代换公式（change-of-variables formula）的基础；这条公式说明坐标改变时，密度如何随之变换。

延伸练习

1 **经典问题。**证明：若在两个独立而均匀分布的点处把木棒折断，则各段能组成三角形的概率是 $\frac{1}{4}$。

2 **折断较长的一段。**在一个均匀分布的点处把木棒折断，然后在较长的一段上取一个均匀分布的点，再把它折断。证明三段能组成三角形的概率是 $2\ln 2-1\approx 0.386$。

3 **锐角还是钝角。**在第 (ii) 部分中，求三角形 ABC 在 C 处（对边为 $AB=2-x-y$）有钝角的概率。由此求它有任何一个角是钝角的概率。

4 **最长一段的期望长度。**对于经典的折棒问题，证明最长一段的期望（expected）长度是 $\frac{11}{18}$。

延伸阅读

- **Martin Gardner, *The Colossal Book of Mathematics***（Norton），书中收录了折棒问题。
- **Frederick Mosteller, *Fifty Challenging Problems in Probability***（Dover）。
- **Lewis Carroll, *Pillow Problems***（Dover，与 *A Tangled Tale* 合订）。维多利亚时代的趣题，包括随机三角形。

12.22 没有平均值的分布

题目

STEP III 1995，第 12 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/95-s3-q12`

引子

把许多次量度的结果取平均，随机误差便会互相抵消：平均值会稳定下来，趋向真实的数值。这条*大数定律*（*law of large numbers*）是整个统计学的基石。但它暗藏一个假设：这些量度必须有有限的平均值（mean）。本题构造了一个再自然不过的随机变量（random variable），即两个独立（independent）标准正态（standard normal）变量之比，而它根本没有平均值。它的平均永远不会稳定下来；事实上，一百万个这样的变量的平均，其分布与单独一个的分布完全相同。

这就是*柯西分布*（*Cauchy distribution*）。它的密度（density）看来像一个人畜无害的钟形，但它的尾部实在太厚，偶然出现的极大数值足以左右任何平均。本题用一个漂亮的几何论证把它找出来：两个坐标之比就是一条直线的斜率，而径向对称令这条直线的角度呈均匀分布。

动笔之前

想一想 若 (U, V) 是一个随机点，其密度只取决于它与原点的距离，那么由原点到该点的直线与坐标轴所成的角 Θ，你能说些甚么？

把平面旋转并不会改变密度，所以每个方向的可能性都相同：角 Θ 在 $[0, 2\pi)$ 上呈均匀（uniform）分布。而 $\frac{U}{V}$ 只取决于方向，与距离无关。

逐步解题

阶段 1 联合密度（joint density）

由于 X 与 Y 独立，它们的联合密度就是各自密度的乘积：

$$\mathrm{f}(x,y)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\cdot\frac{1}{\sqrt{2\pi}}e^{-y^2/2}=\frac{1}{2\pi}e^{-\frac{1}{2}(x^2+y^2)}.$$

它只通过 x^2+y^2（即与原点距离的平方）取决于 x 和 y：在以原点为圆心的圆上，密度是常数。这就是“径向对称”（radially symmetrical）的意思。

阶段 2 运用对称性

> **想一想** $\frac{U}{V}<k$ 可以在两种情况下发生，视乎 V 的正负号而定。这两种情况与旋转 180° 的变换 $(u,v)\mapsto(-u,-v)$ 有甚么关系？

$\frac{U}{V}<k$ 的意思是：要么 $V>0$ 且 $U<kV$，要么 $V<0$ 且 $U>kV$。旋转 $(u,v)\mapsto(-u,-v)$ 把第一个区域映射到第二个区域上，并且保持密度不变。所以两者的概率相同，于是

$$\mathrm{P}\Big(\frac{U}{V}<k\Big)=2\mathrm{P}(U<kV,V>0).\qquad\blacksquare$$

阶段 3 一个角

> **想一想** 在以 u 为横轴、v 为纵轴的平面上，画出区域 $\{V>0,U<kV\}$。它是一个顶点在原点的楔形。它的角度是多少？

区域 $V>0$ 是上半平面。通过原点的直线 $u=kv$ 与正 v 轴所成的角为 $\arctan k$（当 $k>0$ 时偏向正 u 的一方）。上半平面中位于这条直线左方的部分，是一个角度为 $\frac{\pi}{2}+\arctan k$ 的楔形。根据径向对称，楔形的概率等于它的角度除以 2π，所以

$$\mathrm{P}\Big(\frac{U}{V}<k\Big)=2\cdot\frac{\frac{\pi}{2}+\arctan k}{2\pi}=\frac{1}{2}+\frac{1}{\pi}\arctan k.$$

对 k 求导，便得到密度

$$\mathrm{g}(k)=\frac{1}{\pi(1+k^2)}.\qquad\blacksquare$$

回顾

- **独立的正态变量具有径向对称性。**两个正态密度的乘积只取决于距离。
- **比值即方向。**$\frac{U}{V}$ 只取决于该点的角度，而这个角度呈均匀分布。
- **楔形的概率。**在径向对称之下，楔形的概率与它的角度成正比。

背后的数学

没有平均值

柯西分布的平均值本应是 $\int_{-\infty}^{\infty} \frac{k}{\pi(1+k^2)}\,\mathrm{d}k$，但 $\int_0^K \frac{k}{1+k^2}\,\mathrm{d}k = \frac{1}{2}\ln(1+K^2) \to \infty$：这个积分并不收敛（converge），所以平均值不存在。（对称性或会令人以为答案是 0，但正负两部分都是无穷大，而 $\infty - \infty$ 并没有值。）方差（variance）同样是无穷大。

更糟的是，n 个独立柯西变量的平均仍然是柯西变量，而且分布完全相同。（其中一个看出这点的方法，是利用特征函数（characteristic function）$\mathrm{E}(e^{itX}) = e^{-|t|}$。）所以取平均完全无补于事：一百万次量度的平均，并不比单独一次可靠。图 12.22.1 显示了实际情况：正态样本的累计平均会稳定下来，而柯西样本的累计平均却永远东歪西倒，不时被偶然出现的极大数值撞得偏离原位。

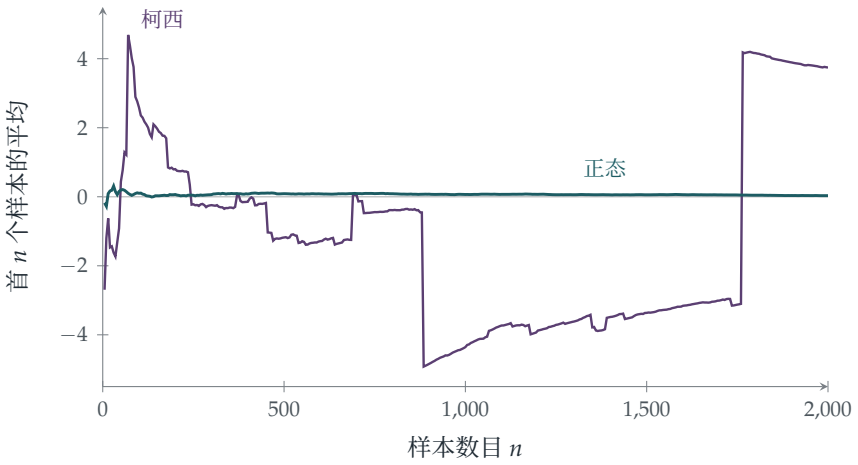


图 12.22.1 2000 个独立样本的累计平均。对于标准正态样本（蓝绿色），平均值一如大数定律所预测，稳定下来并趋向 0。对于柯西样本（紫红色），它永远不会稳定：每隔一段时间，单单一个极大的数值就会把它抛到另一个位置。

柯西变量从何而来

拿一支电筒，朝一个均匀随机的方向照向一面墙：光点在墙上的位置便服从柯西分布。在物理学中，同一个密度称为*劳仑兹分布*（*Lorentzian*），用来描述谱线和共振峰的形状（比较 Section 11.1）。在金融学中，像柯西分布这类厚尾（heavy-tailed）分布，会用来为偶尔出现的股灾建立模型，而正态模型往往严重低估这类事件。这对统计学是一个严肃的教训：建基于平均值和方差的方法，遇上厚尾数据便会失效，这时就需要稳健的方法，例如使用中位数（median）（柯西分布的中位数完全没有问题）。

泊松、柯西与一场论战

这个分布在 1824 年由泊松（Siméon Denis Poisson）研究过，他注意到它违反大数定律。柯西（Augustin-Louis Cauchy）则在 1853 年与 Irénée-Jules Bienaymé 就最小二乘法（method of least squares）公开论战时遇上它：柯西用它来说明，当误差属于厚尾时，最小二乘法可以是一个很差的方法。这个分布最终冠上了柯西的名字，正是 Stephen Stigler 所称的“史蒂格勒命名定律”（Stigler's law of eponymy）的一例：没有任何科学发现是以其最初发现者命名的。

思维技巧

检查期望值是否存在

使用平均值或方差之前，先确定定义它的积分是收敛的。对称性并不能令一个发散的积分变成零。

在大学里，可积性条件是概率论中每一条定理的一部分：大数定律需要有限的平均值，中心极限定理（central limit theorem）（Section 12.24）需要有限的方差。当这些条件不成立时，柯西分布就是标准的反例。

延伸练习

1 **电筒。**一支电筒与一面长而直的墙相距 1，所指的方向与垂直线成角 Θ，其中 Θ 在 $(-\frac{\pi}{2}, \frac{\pi}{2})$ 上呈均匀分布。证明光点的位置 $\tan\Theta$ 服从柯西分布。

2 **倒数。**证明若 C 服从柯西分布，则 $\frac{1}{C}$ 亦然。（提示：$\frac{U}{V}$ 与 $\frac{V}{U}$。）

3 **中位数行得通。**证明柯西分布的中位数是 0，而它的四分位数（quartiles）是 ± 1。

4 **半径与角度。**设 X、Y 为独立的标准正态变量，通过转换成极坐标（polar coordinates），证明 $R^2 = X^2 + Y^2$ 服从平均值为 2 的指数分布（exponential distribution），并且与角度独立。（这是用来产生正态随机数的 Box–Muller 方法的基础。）

延伸阅读

- **Geoffrey Grimmett & Dominic Welsh, *Probability: An Introduction***（Oxford University Press），论随机变量的变换。

- **Nassim Nicholas Taleb, *The Black Swan***（Penguin）。讲述厚尾现象及其重要性，适合一般读者。
- **Stephen Stigler, *The History of Statistics***（Harvard University Press）。书中记述了柯西与 Bienaymé 的论战。

12.23 由概率得出斯特灵公式

题目

STEP II 2016，第 13 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/16-s2-q13`

引子

100! 有多大? 要逐项乘出来得花点工夫：它有 158 位数字。斯特灵公式（*Stirling's formula*）$n! \approx \sqrt{2\pi n}\,(n/e)^n$ 只用一行，就能给出误差在 0.1% 以内的答案。它是数学中最有用的近似式之一，在概率（probability）、统计物理和算法分析中经常用到。

它有很多种证明（proof），可以用积分（integral）（Section 4.4），也可以用欧拉–麦克劳林公式（Euler–Maclaurin formula）。本题给出的是最出人意表的一种：用概率来证明。平均值（mean）为 n 的泊松分布（Poisson distribution）近似于正态分布（normal distribution），在峰顶比较两者，$n!$ 的值就被确定下来。斯特灵公式中的 $\sqrt{2\pi}$，正正就是正态密度（normal density）中的 $\sqrt{2\pi}$。

动笔之前

想一想 设离散随机变量（discrete random variable）X 的平均值为 μ、方差（variance）为 σ^2，并且近似于正态分布，那么 $\mathrm{P}(X = \mu)$ 大约是多少？（试想一个宽度为 1、以 μ 为中心的区间的概率。）

按正态近似（normal approximation），$\mathrm{P}(X = \mu) \approx \mathrm{P}(\mu - \frac{1}{2} < X < \mu + \frac{1}{2})$，这大约等于宽度 1 乘以峰顶处的密度：$\frac{1}{\sigma\sqrt{2\pi}}$。

逐步解题

阶段 1 十六枚硬币

$X \sim \mathrm{B}(16, \tfrac{1}{2})$ 的平均值为 8，方差为 $16 \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} = 4$，所以 $\sigma = 2$。由正态近似，

$$\mathrm{P}(X = 8) \approx \frac{1}{2\sqrt{2\pi}} \approx 0.1995.$$

准确值是 $\binom{16}{8}/2^{16} = \tfrac{12870}{65536} \approx 0.1964$，两者相差不到 2%。

阶段 2 中央二项式系数

> **想一想** $\mathrm{B}(2n, \tfrac{1}{2})$ 的平均值和标准差（standard deviation）是多少？分别写出 $\mathrm{P}(X = n)$ 的准确值和近似值。

平均值是 n，方差是 $\tfrac{n}{2}$，所以 $\sigma = \sqrt{n/2}$。准确值是 $\mathrm{P}(X = n) = \binom{2n}{n}\tfrac{1}{2^{2n}}$；近似值是 $\tfrac{1}{\sqrt{n/2}\sqrt{2\pi}} = \tfrac{1}{\sqrt{\pi n}}$。令两者相等，

$$\frac{(2n)!}{(n!)^2\, 2^{2n}} \approx \frac{1}{\sqrt{\pi n}}, \qquad (2n)! \approx \frac{2^{2n}(n!)^2}{\sqrt{n\pi}}. \qquad \blacksquare$$

阶段 3 斯特灵公式

> **想一想** 参数为 n 的泊松分布，其平均值和方差都等于 n。分别写出 $\mathrm{P}(X = n)$ 的准确值和近似值。

准确值是 $\mathrm{P}(X = n) = e^{-n}\tfrac{n^n}{n!}$。当 n 很大时，泊松分布近似于正态分布（它是 n 个独立的 Poisson(1) 变量之和），其中 $\sigma = \sqrt{n}$，所以 $\mathrm{P}(X = n) \approx \tfrac{1}{\sqrt{2\pi n}}$。令两者相等，

$$n! \approx \sqrt{2\pi n}\, e^{-n} n^n. \qquad \blacksquare$$

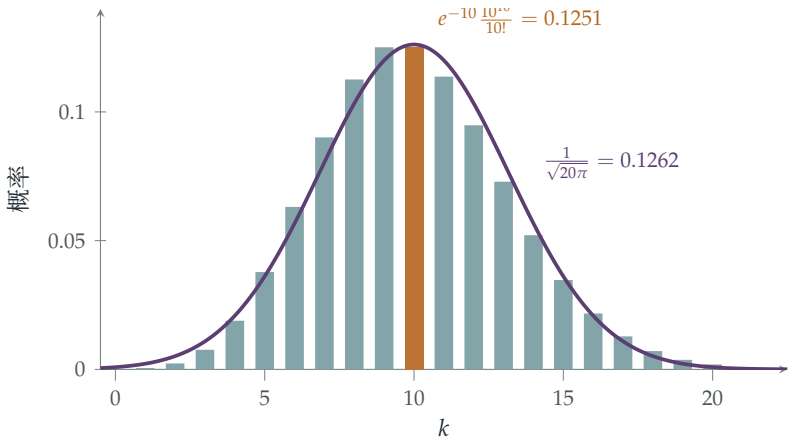


图 12.23.1 平均值为 10 的泊松分布（长条），以及平均值和方差与之相同的正态密度（紫色曲线）。在峰顶，泊松概率 $e^{-10}10^{10}/10!$（琥珀色）与正态密度 $\frac{1}{\sqrt{2\pi\cdot 10}}$ 十分接近。把这个近似等式移项整理，就是斯特灵公式。

回顾

- **一点上的概率 ≈ 密度乘以宽度。**对于格点分布（lattice distribution），某一点上的正态密度近似于该点的概率。
- **准确公式遇上近似式。**二项分布（binomial distribution）和泊松分布的概率是准确的；令它们等于各自的正态近似，便得到阶乘（factorial）的估计。
- **按公式的形状选择分布。**$\binom{2n}{n}$ 令人想到 $\mathrm{B}(2n, \frac{1}{2})$；$\frac{n^n}{n!}$ 令人想到 $\mathrm{Poisson}(n)$。

背后的数学

这个近似有多准确？

这个论证只是启发式（heuristic）的：它在*单独一点上*使用正态近似，而一般的中心极限定理（central limit theorem）并不足以支持这样做。它有一个严格的版本，称为*局部中心极限定理*（*local central limit theorem*），所做的正是这件事，并使这个论证成为真正的证明。斯特灵公式其实出奇地准确：比值 $\frac{n!}{\sqrt{2\pi n}(n/e)^n}$ 大约是 $1 + \frac{1}{12n}$，所以即使 $n = 1$，它给出的也是 0.92 而不是 1；而当 $n = 10$ 时，它只偏小了 0.8%。完整的级数（series）$1 + \frac{1}{12n} + \frac{1}{288n^2} - \cdots$ 是*渐近级数*（*asymptotic series*）的一个例子：对于大的 n，它给出的近似值越来越好；但对于固定的 n，它最终会发散（diverge）。

阶乘与正态分布为何相连

斯特灵公式和正态密度中的 $\sqrt{2\pi}$ 是同一个常数，因为 $\int_{-\infty}^{\infty} e^{-x^2/2}\,\mathrm{d}x = \sqrt{2\pi}$。棣莫弗（Abraham de Moivre）在 1733 年正是循这条路找到二项分布的正态近似的：他需要在 n 很大时近似 $\binom{n}{k}$；他找到了 $n! \approx Cn^{n+1/2}e^{-n}$ 这个形式，却没有找到那个常数。斯特灵（James Stirling）证明了 $C = \sqrt{2\pi}$。正态分布是从阶乘的渐近性质中诞生的。

棣莫弗与斯特灵

棣莫弗是一位逃亡到伦敦的法国新教徒，靠当私人教师和为赌徒出谋献策维生。他在 1718 年出版了 *The Doctrine of Chances*（《机会论》），并在 1733 年的一本小册子中找到二项分布的正态近似，这是正态曲线首次面世。苏格兰数学家斯特灵在他的 *Methodus Differentialis*（1730）中发表了该常数的值，棣莫弗亦大方地把功劳归于他。阶乘的近似至今仍是活跃的课题：组合数学（combinatorics）和密码学（cryptography）都需要对极大的 n 计算 $n!$。

思维技巧

用概率去证明与概率无关的事情

把一个准确的概率与一个近似值作比较，可以得知公式中那些数的信息。

这种“概率方法（probabilistic method）”是双向的。这里概率给出了阶乘的估计；艾狄胥（Erdős）则用它来证明某些图（graph）的存在（Section 9.2）。到了大学，你会见到概率被用来证明分析、数论和组合数学中的定理。

延伸练习

1 **检验公式。**用斯特灵公式估计 10! 和 20!,并与真确值 3 628 800 和 2 432 902 008 176 640 000 比较。

2 **由 (iii) 到 (ii)。**证明第 (iii) 部分的斯特灵公式可推出第 (ii) 部分的近似式。

3 **100! 的位数。**利用 $\log_{10} n! \approx n\log_{10}\frac{n}{e} + \frac{1}{2}\log_{10}(2\pi n)$ 证明 100! 有 158 位数字。

4 **掷硬币打成平手。**利用第 (ii) 部分，证明掷一枚公平硬币 1000 次而恰好得到 500 次正面的概率约为 0.025。

延伸阅读

- **William Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1**（Wiley），第 II 章，当中证明了斯特灵公式。
- **Ronald Graham, Donald Knuth & Oren Patashnik, *Concrete Mathematics***（Addison-Wesley），第 9 章，讲述渐近分析。
- **Stephen Stigler, *The History of Statistics***（Harvard University Press），讲述棣莫弗与正态曲线的诞生。

12.24 中心极限定理

题目

STEP III 2008，第 12 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/08-s3-q12`

引子

把许多微小而独立的随机效应加起来，不论每一个效应本身服从甚么分布，总和都会服从钟形的正态曲线。这就是*中心极限定理*（central limit theorem），它也许是概率论中最重要的定理。它解释了为何身高、量度误差和考试分数都大致服从正态分布（normal distribution），也解释了为何统计学家即使对背后的分布一无所知，仍能说出一个平均值有多可靠，只要该分布的方差（variance）有限便可（试与 Section 12.22 比较）。

本题在一个特殊情况下证明这个定理。一只青蛙随机跳跃，每次跳跃服从双边指数分布（two-sided exponential distribution），其密度函数（density）有一个尖峰，完全不像钟形。跳了许多次之后，牠的总位移却服从正态分布。所用的工具是*矩母函数*（moment generating function），它把独立变量之和化为乘积，再取对数，又把乘积化为和。

关于课程大纲的说明

矩母函数已于 2019 年从 STEP 考试大纲中删除。随机变量（random variable）X 的矩母函数是 $\mathrm{M}_X(\theta) = \mathrm{E}(e^{\theta X})$。它在 0 处的各阶导数给出各阶矩（moments）：$\mathrm{M}'(0) = \mathrm{E}(X)$，$\mathrm{M}''(0) = \mathrm{E}(X^2)$。若 X 与 Y 独立（independent），则 $\mathrm{M}_{X+Y} = \mathrm{M}_X\mathrm{M}_Y$，而且 $\mathrm{M}_{cX}(\theta) = \mathrm{M}_X(c\theta)$。若矩母函数收敛，则（在适当的意义下）分布也收敛。

动笔之前

想一想 密度函数 $\frac{1}{2}e^{-|x|}$ 是对称的。它的图像是甚么样子？它的平均值（mean）是多少？

它是一条帐篷形的曲线，在 0 处有一个高度为 $\frac{1}{2}$ 的尖峰，并向两边按指数衰减（图 12.24.1）。由对称性可知平均值为 0。

逐步解题

阶段 1 矩母函数

当 $|\theta| < 1$ 时，把积分在 0 处分开，

$$\mathrm{M}_X(\theta) = \frac{1}{2}\int_{-\infty}^{0} e^{(1+\theta)x}\,\mathrm{d}x + \frac{1}{2}\int_{0}^{\infty} e^{-(1-\theta)x}\,\mathrm{d}x = \frac{1}{2}\left(\frac{1}{1+\theta} + \frac{1}{1-\theta}\right) = \frac{1}{1-\theta^2}. \qquad \blacksquare$$

展开得 $\mathrm{M}_X(\theta) = 1 + \theta^2 + \theta^4 + \cdots$，所以 $\mathrm{E}(X) = 0$，$\mathrm{E}(X^2) = 2$（即 $\frac{\theta^2}{2!}$ 的系数）。方差为 2。

阶段 2 青蛙的位移

想一想 $Y = X_1 + \cdots + X_n$ 的矩母函数是甚么？$Y/\sqrt{2n}$ 的又是甚么？

由独立性得 $\mathrm{M}_Y(\theta) = (1-\theta^2)^{-n}$，再作缩放，

$$\mathrm{M}_{Y/\sqrt{2n}}(\theta) = \mathrm{M}_Y\left(\frac{\theta}{\sqrt{2n}}\right) = \left(1 - \frac{\theta^2}{2n}\right)^{-n}.$$

（除以 Y 的标准差（standard deviation）$\sqrt{2n}$ 作缩放，令方差变为 1。）

阶段 3 极限

想一想 取对数，并利用当 u 很小时 $\ln(1-u) \approx -u$。

$\ln \mathrm{M}_{Y/\sqrt{2n}}(\theta) = -n\ln\left(1 - \frac{\theta^2}{2n}\right) = -n\left(-\frac{\theta^2}{2n} - \frac{\theta^4}{8n^2} - \cdots\right) \to \frac{\theta^2}{2}$，所以

$$\mathrm{M}_{Y/\sqrt{2n}}(\theta) \to e^{\theta^2/2}, \qquad \blacksquare$$

这正是标准正态分布的矩母函数。

阶段 4 青蛙的机会

想一想 当 n 很大时，Y 近似服从 $\mathrm{N}(0, 2n)$。何时有 $\mathrm{P}(|Y| \geq 25) = 0.05$?

对于标准正态变量 Z，$\mathrm{P}(|Z| \geq 1.96) = 0.05$。因此我们需要 $\frac{25}{\sqrt{2n}} = 1.96$，由此得 $2n = \left(\frac{25}{1.96}\right)^2 \approx 162.7$，即 $n \approx 81.3$。由于落在远处的机会随 n 增大而增加，要有至少 5% 的机会，最少的跳跃次数约为 82。 ■

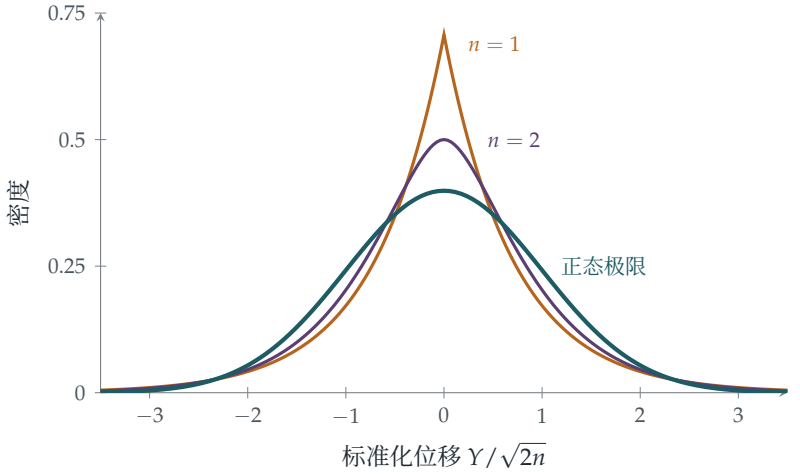


图 12.24.1 青蛙跳了 $n = 1$ 次（琥珀色：一个拉普拉斯（Laplace）尖峰）和 $n = 2$ 次（紫红色）之后，其标准化位移的密度函数，与标准正态密度函数（蓝绿色）的比较。只跳了两次，尖峰已经变得圆滑；中心极限定理指出，当 $n \to \infty$ 时，这些曲线收敛至正态曲线。

回顾

- **变换把和化为乘积。**独立变量之和的矩母函数，等于各变量矩母函数的乘积。
- **对数把乘积化为和。**$\ln\left(1 - \frac{\theta^2}{2n}\right)^{-n}$ 变成了一个和，其极限很容易求得。
- **标准化（standardise）。**除以标准差 $\sqrt{2n}$ 令方差保持为 1，所以极限既不是 0，也不是无穷大。

背后的数学

一般的定理

同样的论证适用于任何平均值为 0、方差 σ^2 有限的分布（还须有矩母函数，不过这个条件可以避开）。这时 $\mathrm{M}(\theta) = 1 + \frac{1}{2}\sigma^2\theta^2 + o(\theta^2)$，而且

$$\left(\mathrm{M}\left(\frac{\theta}{\sigma\sqrt{n}}\right)\right)^n = \left(1 + \frac{\theta^2}{2n} + o\left(\frac{1}{n}\right)\right)^n \to e^{\theta^2/2}.$$

取极限之后只有方差留下来：原来分布的形状已被遗忘。这就是正态分布具有普遍性的原因。在大学里，这个定理是用特征函数（characteristic function）$\mathrm{E}(e^{i\theta X})$ 来证明的，特征函数必定存在，而分布的收敛则由莱维连续性定理（Lévy's continuity theorem）赋予严格的意义。

有多快?

中心极限定理说明极限情况下会发生甚么，却没有说明收敛得有多快。*Berry–Esseen* 定理（Berry–Esseen theorem）给出误差的上界：一个常数乘以 $\frac{\mathrm{E}|X|^3}{\sigma^3\sqrt{n}}$。对于像青蛙跳跃那样的对称分布，跳了几十次之后近似已相当准确。对于偏态（skewed）分布，或者在极端的尾部，即使 n 很大，近似也可能很差：正态近似会低估极罕见事件发生的机会，这个事实对保险和金融都有影响。

从棣美弗到 Lindeberg

棣美弗（De Moivre）在 1733 年发现了二项分布（binomial distribution）的正态近似（Section 12.23），拉普拉斯在 1810 年把它推广至更一般的变量之和，并首先提出量度误差是许多微小效应之和这个想法。严格的证明要到很久之后才出现：李亚普诺夫（Aleksandr Lyapunov）在 1901 年给出证明，Jarl Lindeberg 则在 1922 年给出证明，并找到了基本上是令定理成立的最弱条件。“中心极限定理”这个名称是波利亚（George Pólya）在 1920 年起的，意思是这个定理在概率论中处于中心地位。

思维技巧

变换、化简、再变换回来

一个困难的运算（把独立随机变量相加，即对它们的密度函数作卷积（*convolution*）），经过变换之后变成一个容易的运算（乘法）。在那里完成工作，然后诠释所得的结果。

同样的策略也是生成函数（generating functions）（Section 9.4）、分析学和工程学中的傅里叶（Fourier）变换和拉普拉斯变换，以及把乘法化为加法的对数背后

的原理。

延伸练习

1 **两次跳跃的精确结果。**证明 $Y = X_1 + X_2$ 的密度函数为 $\frac{1}{4}(1 + |y|)e^{-|y|}$，并验证其方差为 4。

2 **从二项到正态。**设 $X \sim \mathrm{B}(1, \frac{1}{2})$，并把它平移使平均值为 0（即取值 $\pm\frac{1}{2}$），求出 $\mathrm{M}_X(\theta) = \cosh\frac{\theta}{2}$，并证明 n 个这样的变量的标准化和的矩母函数趋向 $e^{\theta^2/2}$。

3 **青蛙的精确结果。**利用电脑（或以 $n = 2$ 时的精确密度函数作检验），看看当 $n = 82$ 时正态近似有多准确。

4 **为何柯西不行？**柯西分布（Cauchy distribution）（Section 12.22）没有矩母函数。论证中哪一步失效？n 个柯西变量的平均值又会怎样？

延伸阅读

- **Geoffrey Grimmett & Dominic Welsh, *Probability: An Introduction***（Oxford University Press），关于生成函数和极限定理的一章。
- **Hans Fischer, *A History of the Central Limit Theorem***（Springer）。从拉普拉斯到 Lindeberg。
- **David Spiegelhalter, *The Art of Statistics***（Pelican），讲述为何正态分布无处不在。

12.25 众多之中的最大值

题目

STEP III 2025，第 11 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2025 年 STEP 试卷（zip 文件）内的 STEP III 试卷：

`www.ocr.org.uk/Images/750486-2025-past-papers.zip`

引子

下一个世纪最高的洪水会有多高？一千个灯泡之中，寿命最长的那个能用多久？在 n 名参赛者之中，跑得最快的马拉松选手会有多快？这类关于极值（extremes）的问题，即许多随机数值之中最大或最小的一个，在工程、保险和气候科学中都极为重要，而极值的表现与平均值截然不同。

本题要找出 n 个指数分布（exponential distribution）寿命中最大者的分布。它的典型大小按 $\ln n$ 增长，虽然缓慢，却没有上限。然而它的离散程度却完全不会增长：一千个寿命中的最大值，其可预测的程度与十个寿命中的最大值大致相同。在极限情况下，它的分布形状就是耿贝尔分布（Gumbel distribution），也就是极值世界中与正态曲线相对应的角色。

动笔之前

想一想 $X_1, \ldots, X_n$ 中的最大者不超过 y，恰好等同于发生了甚么事？

$Y \le y$ 恰好在每个 $X_i \le y$ 时成立。利用独立性（independence），这个概率便化为一个乘积。

逐步解题

阶段 1 最大值的分布

每个 X_i 在 $y \ge 0$ 时都有 $\mathrm{F}(y) = 1 - e^{-\lambda y}$，由独立性可得

$$\mathrm{G}(y) = \mathrm{P}(X_1 \le y, \dots, X_n \le y) = (1 - e^{-\lambda y})^n. \qquad \blacksquare$$

阶段 2 分位数

想一想 由 $\mathrm{G}(L) = \alpha$ 解出 L。然后解 $\mathrm{G}(U) = 1 - \alpha$。

由 $\left(1 - e^{-\lambda L}\right)^n = \alpha$ 得 $e^{-\lambda L} = 1 - \alpha^{1/n}$，所以

$$L(\alpha) = -\frac{1}{\lambda}\ln\!\left(1 - \alpha^{1/n}\right), \qquad U(\alpha) = -\frac{1}{\lambda}\ln\!\left(1 - (1-\alpha)^{1/n}\right). \qquad \blacksquare$$

阶段 3 n 很大的情况

想一想 把 $\alpha^{1/n} = e^{(\ln \alpha)/n}$ 写出来。当 n 很大时，指数很小。那么 $1 - \alpha^{1/n}$ 近似于甚么？

$\alpha^{1/n} = e^{(\ln\alpha)/n} \approx 1 + \frac{\ln\alpha}{n}$，所以 $1 - \alpha^{1/n} \approx \frac{-\ln\alpha}{n} = \frac{\ln(1/\alpha)}{n}$。因此

$$\lambda L(\alpha) \approx -\ln\frac{\ln(1/\alpha)}{n} = \ln n - \ln\!\left(\ln\frac{1}{\alpha}\right). \qquad \blacksquare$$

同理，$\lambda U(\alpha) \approx \ln n - \ln\!\left(\ln\frac{1}{1-\alpha}\right)$。

阶段 4 增长与离散程度

中位数（median）是 $L(\frac{1}{2})$，而 $\lambda L(\frac{1}{2}) \approx \ln n - \ln\ln 2 \to \infty$。但是

$$\lambda\big(U(\alpha) - L(\alpha)\big) \approx \ln\!\left(\ln\frac{1}{\alpha}\right) - \ln\!\left(\ln\frac{1}{1-\alpha}\right),$$

当中的 $\ln n$ 项已互相抵消：区间的阔度与 n 无关。 $\blacksquare$

阶段 5 一个 90% 区间

想一想 取 $\alpha = 0.05$。$\ln\frac{1}{\alpha} = \ln 20$ 和 $\ln\frac{1}{0.95}$ 大约是多少？

当 $\alpha = 0.05$ 时，$\ln 20 \approx 3$，所以 $\ln\ln\frac{1}{\alpha} \approx \ln 3$。另外 $\ln\frac{1}{0.95} = -\ln(1-0.05) \approx 0.05 = \frac{1}{20}$，所以 $\ln\ln\frac{1}{1-\alpha} \approx \ln\frac{1}{20} \approx -3$。因此阔度约为 $\frac{1}{\lambda}(\ln 3 + 3) \approx \frac{4.1}{\lambda}$，而 Y 以概率 0.9 落在 $[L(0.05), U(0.05)]$ 之内，这个区间的阔度约为 $4\lambda^{-1}$。 $\blacksquare$

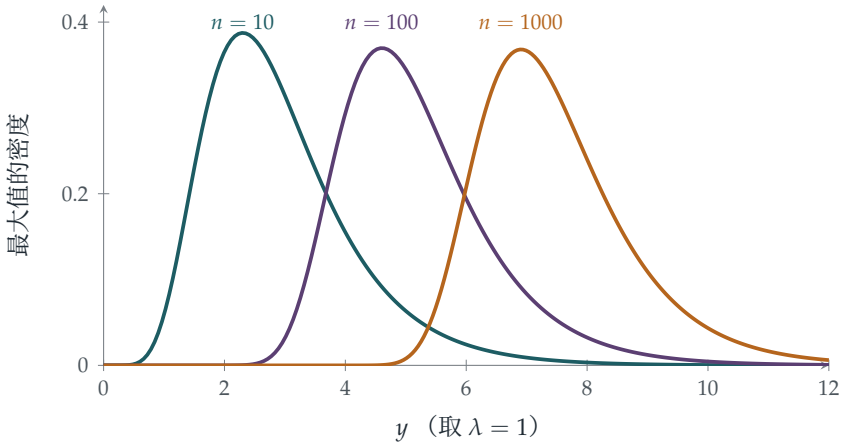


图 12.25.1 n 个指数分布寿命（平均值为 1）中最大者的密度，分别对应 $n = 10$、100 和 1000。n 每增至十倍，曲线便向右平移 $\ln 10 \approx 2.3$，但其形状和阔度基本上保持不变：这就是平移了 $\ln n$ 的耿贝尔分布。

回顾

- **所有值都不超过 y 时，最大值亦然。**独立性使最大值的分布函数成为一个幂。
- **以反函数求分位数。**解 $G(y) = \alpha$ 便得出区间两端的精确公式。
- **先近似，再比较。**L 和 U 的近似式都含有 $\ln n$，在计算阔度时互相抵消。

背后的数学

耿贝尔分布

由于 $P(\lambda Y - \ln n \le x) = \left(1 - \frac{e^{-x}}{n}\right)^n \to \exp(-e^{-x})$，平移后的最大值 $\lambda Y - \ln n$ 在极限下服从耿贝尔分布，其分布函数（distribution function）为 $\exp(-e^{-x})$。它的平均值是欧拉常数（Euler's constant）$\gamma \approx 0.577$。你其实早已遇过它：集齐全部 n 款优惠券所需的时间（Section 9.11）就是 n 个（接近独立的）等待时间中的最大值，这正是它约为 $n \ln n + \gamma n$ 的原因，而其波动属耿贝尔型，数量级为 n。

极值理论

费雪（Fisher）、Tippett 和格涅坚科（Gnedenko）有一个非凡的定理（1928–1943）：许多独立同分布（independent identically distributed）的变量，其最大值经适当平移和缩放后，极限形状只可能有三种：耿贝尔型（适用于轻尾分布，例如指数分布或正态分布）、弗雷歇（Fréchet）型（适用于重尾分布，例如柯西（Cauchy）分布）和韦布尔（Weibull）型

（适用于有硬性上限的变量）。这是中心极限定理（central limit theorem）（Section 12.24）在极值方面的对应结果，所属的学问称为极值理论（extreme value theory）。工程师用它来设计防洪设施和海上平台，以抵御“百年一遇的巨浪”；保险公司用它为巨灾保障定价；气候科学家则用它来估计热浪正在怎样变化。

耿贝尔与洪水

Emil Julius Gumbel（耿贝尔）是德国统计学家，直言不讳地反对纳粹，1932 年被迫离开大学教席，最终逃往美国。他在 1930 和 1940 年代把极值理论应用于河流泛滥的研究。他的著作 *Statistics of Extremes*（1958）使这门学问成为实用的工具。荷兰的三角洲工程（Delta Works）是在 1953 年北海洪灾之后为保护荷兰而兴建的，设计时就运用了极值统计学，以抵御预计约 10 000 年一遇的风暴潮。

思维技巧

最大值与最小值的分布

对于独立变量的最大值，$\mathrm{P}(\max \le y) = \prod \mathrm{P}(X_i \le y)$；对于最小值，$\mathrm{P}(\min > y) = \prod \mathrm{P}(X_i > y)$。

这一个想法就能处理最近的超级市场（Section 12.19）、第一班巴士（Section 12.18）和寿命最长的灯泡。在大学里，它是次序统计量（order statistics）和极值理论的起点。

延伸练习

1 **最大值的期望值。**证明 $\mathrm{E}(Y) = \frac{1}{\lambda}\left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right)$。（提示：根据无记忆性（memorylessness），n 个寿命中的第一个结束之后，余下的 $n-1$ 个会重新开始。）

2 **最小值。**证明 $X_1, \ldots, X_n$ 中的最小者服从参数为 $n\lambda$ 的指数分布。

3 **中位数的精确值。**求 Y 的精确中位数，并在 $n = 10$ 时与近似值 $\frac{1}{\lambda}(\ln n - \ln \ln 2)$ 比较。

4 **均匀分布的最大值。**若 $X_1, \ldots, X_n$ 在 $[0,1]$ 上均匀分布（uniform），求它们的最大值 M 的分布，并证明当 n 很大时，$n(1-M)$ 近似服从指数分布。这属于三种极值形状中的哪一种?

延伸阅读

- **Stuart Coles, *An Introduction to Statistical Modeling of Extreme Values***（Springer）。标准的入门书，附有洪水和海平面的真实数据。
- **Emil Julius Gumbel, *Statistics of Extremes***（Dover）。经典之作。
- **David Spiegelhalter, *The Art of Uncertainty***（Pelican），谈罕见事件以及应该怎样思考它们。

我们学到了甚么

本章开首我们说过：不确定性是可以精确地推理的，但单凭直觉却不行。我们提出了三个问题：题目问的究竟是甚么？巧妙的切入点在哪里？长远而言会发生甚么事？以下是本章二十五道题目所揭示的答案。

1. 题目问的究竟是甚么？

首四节谈的都是如何把问题弄清楚。“已知取出的是一块有编号的圆片”与“已知取出的是编号为 3 的圆片”，两者给出不同的答案（Section 12.1）。一种罕见疾病的检测即使准确度达 90%，检测呈阳性的病人真正患病的可能性也只有 9%，因为基本比率（base rate）举足轻重（Section 12.2）。“随机的弦”在我们说明它是如何选取之前，根本没有期望长度可言（Section 12.3），而折断的木棒（Section 12.21）说明的也是同一道理。此外，“独立”（independent）原来有好几种意思：事件可以两两独立而并非全体独立（Section 12.4），而不相关（uncorrelated）的变量也不一定独立。在每一个例子中，解决之道都是确切地写下有哪些可能的结果，以及它们是如何产生的。

2. 巧妙的切入点在哪里？

许多题目都既有冗长的计算，也有简短的洞见。

- **对称性（symmetry）。**过程本身无法分辨的两个座位（Section 12.5）；“正面较多”与“反面较多”是互补的镜像（Section 12.6）；每一对选手在任何一场比赛中相遇的可能性均等（Section 12.7）；对任何硬币而言，HT 与 TH 出现的可能性相同（Section 12.9）；径向对称的点，其方向服从均匀分布（Section 12.22）。
- **指示变量（indicator）与线性（linearity）。**把配对（Section 12.4）、绳结（Section 12.10）和比赛场数（Section 12.7）写成指示变量之和来计数，毋须求出完整的分布，便可得到平均值和方差。
- **状态与首步分析（first-step analysis）。**Penney 游戏（Section 12.8）、兔子洞穴与 Ecks 教授（Section 12.16），以及姓氏的消亡（Section 12.15），全都是以下一步为条件来求解的。
- **策略。**三人决斗（truel）中枪法最差的人应该故意射失（Section 12.13）；随机化的策略无法被对手利用（Section 12.14）；先拒绝最初 $\frac{1}{e}$ 的应征者是最优的做法（Section 12.12）。

3. 长远而言会发生甚么事?

本章最后的部分追踪随机性如何随时间演变,以及在大量试验中的表现。分支过程(branching process）要么消亡，要么爆发式增长，而即使是爆发式增长的过程，也可能因偶然而消亡（Section 12.15）。马可夫链（Markov chain）会趋于一个平稳分布（stationary distribution），并忘却自己的过去（Section 12.16）。服从指数分布的等候时间则完全没有记忆（Section 12.18）。平均值通常会稳定下来，但柯西（Cauchy）分布却不然，因为它没有平均值（Section 12.22）。根据中心极限定理（central limit theorem），大量独立效应之和会趋于正态分布（Section 12.24），而这种正态性之中更隐藏着 $n!$ 的斯特灵（Stirling）公式（Section 12.23）。最大值的表现又截然不同：它们按 $\ln n$ 的速度增长，离散程度却完全不会增长（Section 12.25）。而常数 e、π 和 γ 则不断出现，在杯子与邮票之中，也在针与绳圈之中。

4. 概率的工具箱

方法	你在哪里用过
写下样本空间（sample space）	条件概率（12.1）；贝特朗（Bertrand）（12.3）
贝叶斯定理（Bayes' theorem）与赔率（odds）	基本比率（12.2）；马可夫逆推（12.16）
对称性	12.5、12.6、12.7、12.22
指示变量与线性	12.4、12.7、12.10
首步分析	Penney（12.8）；兔子洞穴（12.16）；分支过程（12.15）
补事件（complement）与界	公平的决定（12.9）
以级数求期望值	邮票与 e（12.11）
参数化并求最优	秘书问题（12.12）
博弈论（game theory）式推理	三人决斗（12.13）；混合策略（12.14）
计数通往终点的路径	火柴盒（12.17）
先求分布函数（distribution function）	12.18、12.19、12.25
把概率看成面积或积分	蒲丰（Buffon）（12.20）；折断的木棒（12.21）
正态近似	斯特灵（12.23）；中心极限定理（12.24）

下一步

- 第 13 章由概率转向统计：从数据估计未知的量，由 Section 13.1 开始。
- Section 9.8、Section 9.9 和 Section 9.11 的计数问题，是本章内容在组合数学方面的一面。
- 马可夫链用到第 7 章的线性代数，而分支过程的不动点则用到 Section 8.7 的动力系统理论。

如果你想在入读大学之前再进一步

- **Joseph Blitzstein & Jessica Hwang, *Introduction to Probability***：一流的入门课程，可在网上免费阅读，并附有讲课影片。
- **Geoffrey Grimmett & Dominic Welsh, *Probability: An Introduction***：简洁而严谨，属大学一年级程度。
- **Frederick Mosteller, *Fifty Challenging Problems in Probability***：经典谜题，附完整解答。
- **David Spiegelhalter, *The Art of Statistics* 及 *The Art of Uncertainty***：现实世界中的概率与风险。

拉普拉斯（Laplace）说过，概率就是化为计算的常识。本章说明了，之所以往往需要计算，正正是因为常识会失灵；而最好的计算不止于得出数字，它们还能解释个中道理。一种对称性、一个指示变量、一个选得恰当的状态，或一条极限定理，都能把一道谜题化为一个洞见，而大学的概率课程正是建基于这些洞见之上。

延伸练习答案

以下只是简短的答案和提示，并非完整解答：它们给出结果和关键步骤，方便你核对自己的做法。请先认真做过每道练习，然后才看答案。

12.1 精确地取条件

1. 在 (iv) 和 (v) 中，其余 $k-1$ 块圆片都是从余下的 $b+n-1$ 块中随机抽出的，其中 $n-1$ 块有编号，所以两者都等于 $\frac{(n-1)\binom{b}{k-2}}{\binom{b+n-1}{k-1}}$。(vi) 的答案是 $\frac{\binom{n}{2}\binom{b}{k-2}}{\binom{b+n}{k}-\binom{b}{k}}$。设 $N=b+n$，两者只在 $(2N-nk)\binom{N-1}{k-1}=2k\binom{b}{k}$ 时才相等。这只会在个别特殊情况下发生，例如 $(b,n,k)=(2,4,3)$、$(3,3,4)$ 或 $(6,11,3)$，又或者是 $n=2$ 而所有圆片都被抽出这种平凡情况。对于题目所给的数值，等式并不成立。

2. 两个孩子的（性别，星期几）组合共有 $14^2=196$ 种，各种组合可能性相等，其中 $196-13^2=27$ 种包含至少一个星期二出生的男孩。在这些组合之中，$7^2-6^2=13$ 种是两个男孩。所以概率是 $\frac{13}{27}$。

3. 你选了 1 号门。若汽车在 1 号门后（概率 $\frac{1}{3}$），换门会输；若汽车在 2 号或 3 号门后，主持人必须打开另一扇后面是山羊的门，换门便会赢：概率为 $\frac{2}{3}$。若主持人随机打开一扇门，而门后碰巧是山羊，那么根据你所见，“汽车在你选的门后”和“汽车在另一扇关着的门后”这两个结果的可能性相等：概率为 $\frac{1}{2}$。

4. $\mathrm{P}(\geq 2 \mid \geq 1)=\frac{1-\mathrm{P}(0)-\mathrm{P}(1)}{1-\mathrm{P}(0)}\approx 0.370$，但 $\mathrm{P}(\geq 2 \mid \text{葵扇 A})=1-\binom{48}{12}/\binom{51}{12}\approx 0.561$。知道是哪一张 A，会令手上有第二张 A 的可能性增加。

12.2 基本比率谬误

1. 若尸体在那里，两者都说“有尸体”的概率是 pq；若不在，则是 $(1-p)(1-q)$。先验胜算（prior odds）为 $1:1$，所以后验概率（posterior probability）是 $\frac{pq}{pq+(1-p)(1-q)}=\frac{pq}{1-p-q+2pq}$。至于第二部分，在三种可能性相等的情况中，有两种情况下“A 在那里”为真，所以 $\mathrm{P}(\text{两者都在}\mid\text{两者都说}A)=\frac{pq}{2pq+(1-p)(1-q)}$。

2. 我们需要 $0.9\pi>\frac{1}{11}(1-\pi)$，即 $\pi>\frac{10}{109}\approx 9.2\%$。

3. 每一次 AL 测试呈阳性，都会把胜算乘以 $\frac{0.9}{1/11}=9.9$。由先验胜算 $\frac{1}{99}$ 开始：一次阳性之后，胜算为 0.1（概率 $\frac{1}{11}$）；两次之后，胜算为 0.99（概率约 0.497）；三次之后，胜算为 9.8（概率约 0.907）。

4. 在 100 名患者之中，99 人测试呈阳性；在 999 900 名没有患病的人之中，约有 9999 人呈阳性。所以约有 10 100 人被告知可能患病，而当中只有 99 人（约 1%）真正患病。

12.3 “随机”是甚么意思？

1. 弦（chord）比内接等边三角形的边长 $\sqrt{3}a$ 长的条件分别是：它的远端落在与近端相对的那三分之一圆周上（$\frac{1}{3}$）；它与 O 的距离小于 $\frac{a}{2}$（$\frac{1}{2}$）；它的中点落在半径为 $\frac{a}{2}$ 的圆内（$\frac{1}{4}$）。

2. 以 C 为中点的弦垂直于 OC，所以它就是通过 C 而角度为 $t = \frac{\pi}{2}$ 的弦，而 $L(\frac{\pi}{2}) = \frac{4a}{3}\left(0 + \frac{1}{2}\sec^2\frac{\pi}{4}\right) = \frac{4a}{3}$。

3. 弦长为 $2\sqrt{a^2 - d^2}$，平均值为 $\frac{2}{a}\int_0^a \sqrt{a^2 - d^2}\,\mathrm{d}d = \frac{2}{a}\cdot\frac{\pi a^2}{4} = \frac{\pi a}{2}$。

4. 随机端点：$\frac{2}{\pi\sqrt{4a^2 - \ell^2}}$。随机距离：$\frac{\ell}{2a\sqrt{4a^2 - \ell^2}}$。随机中点：$\frac{\ell}{2a^2}$。

12.4 两两独立，却非相互独立

1. 展开后得 $\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = (a - pq)(x_1 - x_2)(y_1 - y_2)$，所以 $a = pq$；其余三格随之确定，于是两者独立。例子：在 $(0,0)$ 放概率 $\frac{1}{3}$，在 $(\pm1, \pm1)$ 各点放概率 $\frac{1}{6}$。两个边际分布（marginal distribution）都是均匀的，而且 $\mathrm{E}(AB) = 0$，但 $\mathrm{P}(A = 0, B = 1) = 0 \neq \frac{1}{9}$。

2. 共有 21 对，每对相同的概率是 $\frac{1}{6}$，而各指示变量（indicator）两两独立（pairwise independent）。所以平均值是 3.5，方差（variance）是 $21 \cdot \frac{1}{6} \cdot \frac{5}{6} \approx 2.92$，标准差约为 1.71。得分 7 大约比平均值高出两个标准差。

3. 平均值是 $\binom{k}{2}\frac{1}{365}$。各指示变量两两独立，与骰子的情况完全一样，所以方差是 $\binom{k}{2}\frac{1}{365}\cdot\frac{364}{365}$。（它们并非相互独立（mutually independent）：若 1, 2 相同，2, 3 也相同，则 1, 3 亦相同。）

4. 取八个可能性相等的结果，$A = B = \{1,2,3,4\}$，$C = \{1,5,6,7\}$。那么 $\mathrm{P}(A \cap B \cap C) = \frac{1}{8}$，但 $\mathrm{P}(A \cap B) = \frac{1}{2} \neq \frac{1}{4}$。

12.5 最后一位乘客的座位

1. 对所有 $n \geq 3$，答案都是 $\frac{1}{3}$。最后一个座位现在是 S_1、S_2 和 S_n 其中之一，而每一次随机选择对这三个座位都是对称的。

2. 若 T_1 坐了 S_1，没有人要让位。若 T_1 坐了 S_3，则 T_1 和 T_3 都坐错了位。若 T_1 坐了 S_2，便由 T_2 选择：选 S_1 有 2 人坐错位，选 S_3 则有 3 人。期望值（expectation）是 $\frac{1}{3}\left(0 + 2 + \frac{5}{2}\right) = \frac{3}{2}$，与公式给出的 $\frac{2}{3} + \frac{1}{3} + \frac{1}{2}$ 一致。

3. 当 $r \geq 2$ 时 $P_r = \frac{1}{2}$，于是 $\frac{1}{n}\left(1 + \frac{n-2}{2}\right) = \frac{1}{2}$，所以递推关系（recurrence）与 $P_n = \frac{1}{2}$ 相容。

4. 最后一个空位必定是 S_1 或 S_n（任何其他 S_k 最迟在 T_k 登机时已被占去）。所以若它不是 S_n，就是 S_1。

12.6 多一枚硬币

1. 设 H' 为 Bob 首 $n+1$ 次掷币的结果，X 为他最后一次掷币的结果。若 $H' > H_A$，或者 $H' = H_A$ 而 X 是正面，Bob 便胜出。所以概率是 $\mathrm{P}(H' > H_A) + \frac{1}{2}\mathrm{P}(H' = H_A)$，而根据“多一枚硬币”的结果，$\mathrm{P}(H' > H_A) = \frac{1}{2}$。

2. 不成立。当 $n = 0$ 时，Bob 胜出的概率是 p。那个巧妙的证明把正面和反面对调，而这只有对公平硬币才是一种对称。

3. $H_A = H_B$ 恰好在 $H_A + (\text{Bob 的反面数目}) = n$ 时成立，即一个 $\mathrm{B}(2n, \frac{1}{2})$ 变量等于 n。数值：$\frac{1}{2}$、$\frac{3}{8}$、$\frac{5}{16}$。

4. 不成立：掷出六点就像一枚 $p = \frac{1}{6}$ 的硬币。当 $n = 0$ 时概率是 $\frac{1}{6}$，而一般而言概率低于 $\frac{1}{2}$。

12.7 淘汰赛中谁会遇上谁

1. 第 k 轮有 2^{n-k} 场比赛，而根据对称性，每场比赛由 $\binom{2^n}{2}$ 对球员中任何一对对赛的可能性都相等。把整个淘汰赛的 $2^n - 1$ 场比赛加起来，得 $\frac{2^n-1}{\binom{2^n}{2}} = \frac{2}{2^n}$。

2. 不会：在固定的对阵表中随机编排位置，对称性依然保持，同样的计数方法仍然适用。

3. X 在第 k 轮的对手来自一组 2^{k-1} 名球员，每人从该组胜出的可能性相等。所以 Y 是该对手的概率为 $\frac{2^{k-1}}{2^n-1} \cdot \frac{1}{2^{k-1}} = \frac{1}{2^n-1}$，而 n 轮的总和是 $\frac{n}{2^n-1}$。

4. 共有 $\binom{2n}{2} = n(2n-1)$ 场比赛；每一对都会相遇，所以概率是 1。

12.8 Penney 游戏

1. HHT 只有在头两次掷出 HH 时才会胜出；否则在凑成 HH 之前已出现一个 T，THH 必定先出现。所以 HHT 胜出的概率是 $\frac{1}{4}$。四人一起玩时：HH 之后，HHT 必胜；TT 之后，TTH 必胜；由 HT 和 TH 开始，HTT 和 THH 的胜出机会分别是 $\frac{2}{3}$ 和 $\frac{1}{3}$（或倒过来），于是每名玩家的胜出概率是 $\frac{1}{4} \cdot 1$ 或 $\frac{1}{4}\left(\frac{2}{3} + \frac{1}{3}\right)$，即 $\frac{1}{4}$。TTH 击败 THH 的概率是 $\frac{2}{3}$（第 (i) 部分）。

2. 对于 HH：$e_0 = 1 + \frac{1}{2}e_{\mathrm{H}} + \frac{1}{2}e_0$ 及 $e_{\mathrm{H}} = 1 + \frac{1}{2}e_0$，得 $e_0 = 6$。对于 HT：一旦有了一个 H，便等待一个 T（平均掷两次），而第一个 H 平均要掷两次：共 4 次。

3. HHH 和 TTT 分别被 THH 和 HTT 击败，概率为 $\frac{7}{8}$；HHT 和 TTH 分别被 THH 和 HTT 击败，概率为 $\frac{3}{4}$；其余四个（HTH、HTT、THH、THT）分别被 HHT、HHT、TTH、TTH 击败，概率为 $\frac{2}{3}$。HHH 和 TTT 最容易被击败。

4. 设 $A = \mathrm{HTT}$，$B = \mathrm{HHT}$，各相关数（correlation）为 $AA = 4$、$AB = 0$、$BB = 4$、$BA = 2$。B 胜出的胜算是 $(AA - AB) : (BB - BA) = 4 : 2$，即概率为 $\frac{2}{3}$。

12.9 用不公平的硬币作公平的决定

1. 所需轮数服从几何分布（geometric distribution）：第 (i) 部分的平均值是 $\frac{1}{p^2+q^2}$，而冯·诺伊曼(von Neumann)的方法(HT 或 TH)的平均值是 $\frac{1}{2pq}$。由于 $p^2 + q^2 - 2pq = (p-q)^2$，对所有 $p \neq \frac{1}{2}$，冯·诺伊曼的方法都较慢。不过，公平的却是它。

2. 用冯·诺伊曼的技巧制造公平的位元（bit）。每次取两个：00、01 和 10 分别代表三个选项，11 则表示重新开始。

3. 把掷币结果逐次与二进制展开式 $\frac{1}{3} = 0.010101\ldots$ 比较（把 H 读作 1，T 读作 0）。在第一次出现不同的那一掷，若掷币结果小于该位数字（T 对着 1），便宣布“正面”，否则宣布“反面”。这样得出正面的概率恰好是 $\frac{1}{3}$，平均用 2 次掷币。

4. 在未能决定的一轮之后，其中一面有两个，另一面有一个，而整个过程是这两个状态上的马可夫链（Markov chain）。解之得 $\mathrm{P}(\text{酒店}) = \frac{p^3}{p^3+q^3}$，当 $p = \frac{1}{2}$ 时等于 $\frac{1}{2}$：这时是公平的，其他情况则不然。

12.10 把绳头系起来

1. 第一次打结形成一个环的概率是 $\frac{1}{3}$（其余三个绳端之中，有一个属于同一条绳），而最后一次打结则必定形成一个环。环的数目：2 个的概率是 $\frac{1}{3}$，1 个的概率是 $\frac{2}{3}$；平均值是 $\frac{4}{3}$。

2. 余下 k 条绳时，一次打结不形成环的概率是 $1 - \frac{1}{2k-1}$。乘积 $\prod_{k=2}^{n} \frac{2k-2}{2k-1}$ 等于 $\frac{4^n}{2n\binom{2n}{n}}$，约为 $\frac{\sqrt{\pi n}}{2n} = \sqrt{\frac{\pi}{4n}}$。

3. 依次插入元素 $1, 2, \ldots, n$：元素 k 要么开始一个新的循环（cycle）（概率 $\frac{1}{k}$），要么放在之前 $k-1$ 个元素其中一个之后。把各指示变量加起来便得 H_n。

4. 各次打结的成环指示变量互相独立，概率为 $\frac{1}{2k-1}$，所以方差是 $\sum \frac{1}{2k-1}\left(1 - \frac{1}{2k-1}\right)$。另外 $\sum_{k\geq 1} \frac{1}{(2k-1)^2} = \frac{\pi^2}{6} - \frac{1}{4} \cdot \frac{\pi^2}{6} = \frac{\pi^2}{8}$。

12.11 邮票、杯子与 e

1. $\sum_{k\geq 2} \frac{k-1}{k!} t^k = te^t - (e^t - 1) = 1 + (t-1)e^t$。求导后得 $\mathrm{E}(N) = G'(1) = \mathrm{e}$ 及 $\mathrm{E}(N(N-1)) = G''(1) = 2\mathrm{e}$，所以 $\mathrm{Var}(N) = 3\mathrm{e} - \mathrm{e}^2$。

2. $\mathrm{P}(N=j)$ 在 $k=0,\ldots,j-1$ 的 $\mathrm{P}(N>k)$ 中出现，即共 j 次。这里 $\mathrm{P}(N>k)=\frac{1}{k!}$，所以 $\mathrm{E}(N)=\sum\frac{1}{k!}=\mathrm{e}$。

3. 完全没有改变：只有一个杯时，Jim 必定“猜”中。期望价格仍然是 e。

4. $\int_0^1\frac{(1-u)^2}{2}\,\mathrm{d}u=\frac{1}{6}$。一般而言，在 $u_k=u$ 处的截面是一个按 $1-u$ 缩放的 $(k-1)$ 维单纯形（simplex），所以 $V_k=V_{k-1}\int_0^1(1-u)^{k-1}\,\mathrm{d}u=\frac{V_{k-1}}{k}$。

12.12 何时停止寻找

1. 若最大的在第 $k\geq 2$ 位，你恰好在第一个雪糕是首 $k-1$ 个之中最大时才会得到它，概率为 $\frac{1}{k-1}$。对 k 取平均得 $\frac{1}{n}\sum_{k=2}^{n}\frac{1}{k-1}$，当 $n=4$ 时是 $\frac{11}{24}$。列出全部 24 种次序，可知得到第二、第三和第四大的概率分别是 $\frac{7}{24}$、$\frac{1}{6}$ 和 $\frac{1}{12}$。

2. 拒绝 $r=0,1,2,3$ 人分别给出 $\frac{1}{4}$、$\frac{11}{24}$、$\frac{5}{12}$、$\frac{1}{4}$。所以应拒绝一名候选人，成功概率为 $\frac{11}{24}$。

3. 首 n 人之中最好的一个位列头两名的概率是 $\frac{n}{N}+\frac{n}{N}\cdot\frac{N-n}{N-1}$，当 n 很小时接近 $\frac{n}{N}$ 的两倍。若头两名任何一个都可以接受，你便可以少看一些候选人才作决定。（以此为目标的最佳策略，成功概率约为 0.57。）

4. 设 $\mathrm{P}(r)=\frac{r}{N}\sum_{k=r}^{N-1}\frac{1}{k}$：当 $N=10$ 时，$r=2,3,4$ 给出 0.366、0.399、0.398；当 $N=100$ 时，$r=36,37,38$ 给出 0.3710、0.3710、0.3708，其中 $r=37$ 仅仅领先。

12.13 三人决斗

1. Algernon 掷失后，Bertie 掷向 Cuthbert。若他掷中 $(\frac{3}{5})$，便是 Algernon 对 Bertie，由 Algernon 先掷，Bertie 胜出的概率是 $1-\frac{2/5}{2/5+3/5-6/25}=\frac{9}{19}$。若他掷失，Cuthbert 会掷中 Bertie，然后除非 Algernon 先掷中 $(\frac{3}{5})$，否则 Cuthbert 胜出。所以 Bertie 胜出的概率是 $\frac{3}{5}\cdot\frac{9}{19}=\frac{135}{475}$，Cuthbert 则是 $\frac{2}{5}\cdot\frac{3}{5}=\frac{114}{475}$。

2. 对所有 $a<1$ 都成立。掷失给出 $\frac{3}{5}\cdot\frac{a}{a+\frac{3}{5}-\frac{3}{5}a}+\frac{2}{5}a$；掷中 Cuthbert 给出 $1-\frac{3/5}{3/5+a-\frac{3}{5}a}$（这时 Bertie 先向他掷），这个值总是较小；掷中 Bertie 则给出 0。

3. 不应该。若 Bertie 掷向 Algernon 而掷中，Cuthbert 随即掷中 Bertie；若他掷失，Cuthbert 照样掷中 Bertie。无论如何 Bertie 都一无所得，而掷向 Cuthbert 则有 $\frac{27}{95}$。

4. $p+(1-p)(1-r)p$，它小于 $\frac{p}{p+r-pr}$，即整个等比级数（geometric series）之和。

12.14 混合策略

1. 取 $b=(\frac{1}{4},\frac{1}{2},\frac{1}{4})$，则 $X=Y=Z=0$，所以无论对方怎样做，B 都赢得 0。对着 $(\frac{1}{3},\frac{1}{3},\frac{1}{3})$，A 猜 1、2 和 3 的期望赢款分别是 $-\frac{1}{3}$、0 和 $\frac{1}{3}$，所以 A 应该每次都猜 3。

2. 收益矩阵（payoff matrix）的对角线上是 0，其余位置是 ± 1，把三个策略循环置换后矩阵不变，所以最优策略是 $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$。在原题中，相差 1 和相差 2 的罚则不同，所以各策略不能互换。

3. 若 X、Y、$Z \geq 0$ 且 $X + 2Y + Z = 0$，则三者都是 0。由此必有 $b_1 = b_3$ 及 $b_2 = 2b_3$。

4. 现在 $X = -b_2 + 3b_3$、$Y = b_1 - b_3$ 及 $Z = -3b_1 + b_2$，而 $X + 3Y + Z = 0$。令三者都等于零，得 $(\frac{1}{5}, \frac{3}{5}, \frac{1}{5})$。

12.15 灭绝还是爆发

1. 灭绝概率（extinction probability）是 $s = (q + ps)^2$ 在 $[0, 1]$ 内的最小根。两个根是 1 和 $\left(\frac{q}{p}\right)^2$，而第二个根恰好在 $p > \frac{1}{2}$ 时小于 1。

2. 给定 X_{n-1}, X_n 的平均值是 $2pX_{n-1}$，方差是 $2pqX_{n-1}$，所以 $\mathrm{Var}(X_n) = \mathrm{E}(2pqX_{n-1}) + (2p)^2\mathrm{Var}(X_{n-1})$。当 $p = \frac{1}{2}$ 时：$\mathrm{Var}(X_n) = \mathrm{Var}(X_{n-1}) + \frac{1}{2}$，所以 $\mathrm{Var}(X_n) = \frac{n}{2}$。

3. 在第 n 代或之前灭绝，表示第一代每一颗粒子的后代都在 $n - 1$ 代之内消亡。取 $p = \frac{3}{4}$：$e_1 = 0.0625$，$e_2 \approx 0.0881$，$e_3 \approx 0.0999$，递增并趋向 $\frac{1}{9} \approx 0.111$。

4. $s = \frac{1}{4}(1 + s + s^2 + s^3)$ 可因式分解为 $(s - 1)(s^2 + 2s - 1) = 0$，所以灭绝概率是 $\sqrt{2} - 1 \approx 0.414$。

12.16 顺向与逆向的马可夫链

1. 根据对称性 $p_B = p_D$；另外 $p_A = \frac{1}{3} + \frac{2}{3}p_B$、$p_B = \frac{1}{3}p_A + \frac{1}{3}p_C$ 及 $p_C = \frac{2}{3}p_B$。所以 $p_A = \frac{7}{15}$。由 A 出发，不经过 C 而在 A 离开：$x_A = \frac{1}{3} + \frac{2}{3}x_B$ 及 $x_B = \frac{1}{3}x_A$，所以 $x_A = \frac{3}{7}$。条件概率是 $\frac{3/7}{7/15} = \frac{45}{49}$。

2. 把第一行乘以 P：$(0.19, 0.33, 0.48)$。

3. 迹（trace）是 1.3，行列式（determinant）是 0.02，而 1 是一个特征值（eigenvalue）（各行之和为 1）。另外两个特征值之和为 0.3，积为 0.02：它们是 0.2 和 0.1。

4. $a_1 p_{12} = \frac{1}{20}$ 但 $a_2 p_{21} = \frac{1}{15}$，所以这条链不是可逆的（reversible）：倒过来运行时，它的转移概率并不相同（第 (iii) 部分的 q_{ij} 与 p_{ij} 不同）。

12.17 巴拿赫的火柴盒

1. 若第二次伸手要拿的与第一次是同一种，她当下便发现那包已空（$N = 2$）；否则她拿走另一颗糖果，并在第三次伸手时发现有一包已空。所以 $N = 2$ 或 3，概率各为 $\frac{1}{2}$。

2. $n = 1$：$\frac{1}{2} + \frac{1}{2} = 1$。$n = 2$：$\frac{6}{16} + \frac{6}{16} + \frac{4}{16} = 1$。

3. $\binom{2n-1}{n}2^{-(2n-1)} = \frac{1}{2}\binom{2n}{n}\cdot 2\cdot 4^{-n} = \binom{2n}{n}4^{-n}$。

4. 就余下的拖肥糖和薄荷糖数目作首步递推（first-step recursion），得 $\frac{64}{81} \approx 0.79$。

12.18 无记忆性

1. $\mathrm{P}(\min > t) = \mathrm{e}^{-\lambda_1 t}\mathrm{e}^{-\lambda_2 t}$。等候下一班巴士的时间服从指数分布（exponential distribution），率为每分钟 $\frac{1}{15}+\frac{1}{30}=\frac{1}{10}$，所以其中位数（median）是 $10\ln 2 \approx 6.9$ 分钟。

2. $\mathrm{P}(Y \geq k) = (1-p)^k$，所以比值是 $(1-p)^n$。

3. 包含 t 的那段间隔，由回溯至上一班巴士的时间和前望至下一班巴士的时间组成，两者都服从平均值为 $\frac{1}{\lambda}$ 的指数分布（当 t 很大时），所以间隔的期望长度是 $\frac{2}{\lambda}$。较长的间隔较有可能包含某个指定时刻。

4. 在 $[0,1]$ 上为 $\frac{\lambda \mathrm{e}^{-\lambda z}}{1-\mathrm{e}^{-\lambda}}$。当 $\lambda = 1$ 时，它由约 1.58 缓缓下降至 0.58；当 $\lambda = 5$ 时，它集中在 0 附近，由约 5.03 下降至 0.03。

12.19 平面上的随机点

1. 由 $1-\mathrm{e}^{-\pi k y^2} = \frac{1}{2}$ 得中位数为 $\sqrt{\frac{\ln 2}{\pi k}} \approx \frac{0.470}{\sqrt{k}}$，比平均值 $\frac{1}{2\sqrt{k}}$ 略小。

2. $\mathrm{P}(Y_2 > y) = \mathrm{e}^{-\pi k y^2}(1+\pi k y^2)$，所以概率密度（density）是 $2\pi^2 k^2 y^3 \mathrm{e}^{-\pi k y^2}$。

3. 在极坐标下，R 的密度是 $\frac{1}{2\pi\sigma^2}\mathrm{e}^{-r^2/(2\sigma^2)}\cdot 2\pi r$。当 $\sigma^2 = \frac{1}{2\pi k}$ 时，它与超级市场的分布吻合。

4. 取 $k = 4\times 10^{-4}$ 每 m^2，完全随机的假设预测平均值为 $\frac{1}{2\sqrt{k}} = 25\,\mathrm{m}$。观察所得的 35 m 较大，这显示树木是互相疏开的（竞争），而不是聚集成丛。

12.20 蒲丰投针

1. 把针切成很多小段。相交次数的期望值具有可加性，而每一小段的贡献与其长度成正比，所以无论 a 是多少，总和都是 $\frac{2a}{\pi b}$。

2. 根据同样的可加性（蒲丰投面条，Buffon's noodle），仍然是 $\frac{2a}{\pi b}$。

3. 圆心与每条边的距离都必须至少为 r：$\frac{(s-2r)^2}{s^2}$。

4. 由 $\frac{1808}{3408} \approx \frac{5}{3\pi}$ 得 $\pi \approx \frac{5\times 3408}{3\times 1808} = \frac{17\,040}{5424} = \frac{355}{113}$：完美得令人起疑。

12.21 折断一根棒

1. 要组成三角形，每一段都必须短于 $\frac{1}{2}$。在由折断点组成的单位正方形中，这个条件划出一个面积为 $\frac{1}{4}$ 的区域。

2. 假设第一个折断点在 $x < \frac{1}{2}$（另一情况是对称的），而较长的一段 $1-x$ 再被均匀地折断。新的两段都必须短于 $\frac{1}{2}$，其概率为 $\frac{x}{1-x}$。所以答案是 $2\int_0^{1/2} \frac{x}{1-x}\,\mathrm{d}x = 2\ln 2 - 1$。

3. 当 $(2-x-y)^2 > x^2+y^2$，即 $y < \frac{2-2x}{2-x}$ 时，C 处的角是钝角。这条曲线与 $x+y=1$ 之间的面积是 $\frac{3}{2} - 2\ln 2$，所以概率是 $3-4\ln 2$，与 B 处的相同。这并非巧合：三条边各自最多为 1，总和为 2，并且是均匀分布的，所以它们可以互换。因此 $\mathrm{P}(\text{钝角三角形}) = 9 - 12\ln 2 \approx 0.68$。

4. 根据容斥原理（inclusion–exclusion），最长的一段 M 满足 $\mathrm{P}(M > m) = 3(1-m)^2 - 3(1-2m)_+^2 + (1-3m)_+^2$，其中 $x_+ = \max(x, 0)$。在 $[0,1]$ 上积分得 $\mathrm{E}(M) = 1 - \frac{1}{2} + \frac{1}{9} = \frac{11}{18}$。

12.22 没有平均值的分布

1. $\mathrm{P}(\tan\Theta \le k) = \frac{1}{\pi}\bigl(\tan^{-1}k + \frac{\pi}{2}\bigr)$，求导后得 $\frac{1}{\pi(1+k^2)}$。

2. 把 U 和 V 对调，径向对称的密度保持不变，所以 $\frac{V}{U}$ 与 $\frac{U}{V}$ 有相同的分布。

3. 密度是对称的，所以中位数是 0；而 $\mathrm{P}(C \le 1) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}1 = \frac{3}{4}$。

4. $\frac{1}{2\pi}\mathrm{e}^{-r^2/2}\,r\,\mathrm{d}r\,\mathrm{d}\theta$，而令 $s = r^2$ 后这就是 $\frac{1}{2\pi}\,\mathrm{d}\theta \cdot \frac{1}{2}\mathrm{e}^{-s/2}\,\mathrm{d}s$：角度是均匀分布的，$R^2$ 服从平均值为 2 的指数分布，而且两者独立。

12.23 由概率得出斯特灵公式

1. $\sqrt{20\pi}(10/\mathrm{e})^{10} \approx 3\,598\,696$（约低 0.8%）及 $\sqrt{40\pi}(20/\mathrm{e})^{20} \approx 2.4228 \times 10^{18}$（约低 0.4%）。

2. $\frac{(2n)!}{(n!)^2} \approx \frac{\sqrt{4\pi n}(2n/\mathrm{e})^{2n}}{2\pi n(n/\mathrm{e})^{2n}} = \frac{4^n}{\sqrt{\pi n}}$。

3. $100\log_{10}\frac{100}{\mathrm{e}} + \frac{1}{2}\log_{10}(200\pi) \approx 157.97$，所以 100! 有 158 位数字。

4. $\binom{1000}{500}2^{-1000} \approx \frac{1}{\sqrt{500\pi}} \approx 0.0252$。

12.24 中心极限定理

1. 作卷积（convolution），得 $\frac{1}{4}\int \mathrm{e}^{-|x|-|y-x|}\,\mathrm{d}x = \frac{1}{4}(1+|y|)\mathrm{e}^{-|y|}$。方差是 $2+2=4$。

2. $\mathrm{E}(\mathrm{e}^{\theta X}) = \frac{1}{2}(\mathrm{e}^{\theta/2} + \mathrm{e}^{-\theta/2}) = \cosh\frac{\theta}{2}$。标准化后的和 $\frac{2}{\sqrt{n}}\sum X_i$ 的母函数（generating function）是 $\cosh^n\frac{\theta}{\sqrt{n}}$，其对数为 $n\ln\bigl(1 + \frac{\theta^2}{2n} + \cdots\bigr) \to \frac{\theta^2}{2}$。

3. 把 Y 写成两个独立的 $\Gamma(82,1)$ 变量之差，再作数值积分，得 $\mathrm{P}(|Y| \geq 25) \approx 0.051$：正态近似（normal approximation）非常准确。

4. 它没有矩母函数（moment generating function），甚至连平均值也没有。事实上，n 个独立柯西（Cauchy）变量的平均值本身也服从柯西分布：取平均完全不会令它变得集中。

12.25 众多之中的最大值

1. n 个寿命之中，最先完结的一个平均需时 $\frac{1}{n\lambda}$；之后余下 $n-1$ 个，一切重新开始，如此类推。所以 $\mathrm{E}(Y) = \frac{1}{\lambda}\left(\frac{1}{n} + \frac{1}{n-1} + \cdots + 1\right)$。

2. $\mathrm{P}(\min > t) = (\mathrm{e}^{-\lambda t})^n = \mathrm{e}^{-n\lambda t}$。

3. 由 $(1-\mathrm{e}^{-\lambda y})^n = \frac{1}{2}$ 得 $\lambda y = -\ln(1-2^{-1/n})$。当 $n=10$ 时这是 2.704，而近似值为 $\ln 10 - \ln\ln 2 \approx 2.669$。

4. $\mathrm{P}(M \leq m) = m^n$，所以 $\mathrm{P}\big(n(1-M) > s\big) = \left(1-\frac{s}{n}\right)^n \to \mathrm{e}^{-s}$。这属于韦布尔（Weibull）型：各变量有上界，而最大值会挤向这个上界。

章

13

统计与现实世界中的随机性

概率由模型推向数据，统计则由数据返回现实世界。

本章内容

13.1	解读新闻标题	STEP I 1992
13.2	计数田鼠	STEP II 2000
13.3	德国坦克问题	STEP III 1998
13.4	平均数还是中位数?	STEP II 2000
13.5	问尴尬的问题	STEP II 2023
13.6	累积起来的误差	STEP II 1993
13.7	制造相关性	STEP III 2010
A	练习答案	第 1029 页

从数据中学习

所有模型都是错的，但有些是有用的。

— 统计学家 George Box

在概率（probability）中，我们知道骰子怎样被动了手脚，然后问会发生甚么事。在统计（statistics）中，我们看到已发生的事，然后问骰子必定是怎样被动了手脚。概率由因推果，统计则由果溯因。对总体的每一个估计、每一次民意调查、每一项临床试验、每一次科学测量，都是一个统计推断（statistical inference）：根据有限而带有杂讯的数据，对现实世界作出结论。

与数学的大部分分支相比，统计是一门年轻的学科。它的现代面貌是在二十世纪初由皮尔逊（Karl Pearson）、费雪（Ronald Fisher）、奈曼（Jerzy Neyman）等人建立的，背后的推动力往往是实际问题：酿造啤酒（t 检验就是在健力士（Guinness）啤酒厂发明的）、种植农作物、测试药物。时至今日，随着数据量爆炸式增长，统计及其近亲机器学习（machine learning）已成为数学最重要的应用之一。

统计为何重要

有关药物、公共卫生、经济政策、气候和司法的决定，全都建基于统计证据。误读这些证据会带来实实在在的后果：病人被假阳性（false positive）结果吓坏（Section 12.2），政策建基于一次偶然的波动（Section 13.1），电脑系统中的微小误差不断累积（Section 13.6）。明白数据能告诉你甚么、不能告诉你甚么，以及任何结论有多不确定，是作为一个知情公民的一部分，对科学家、医生、工程师和经济学家而言更是专业上的必需。

你在大学会遇到甚么

统计的入门课程涵盖估计（estimation，即为未知量找出好的估计值，并说明其精确程度）、假设检验（hypothesis testing，即判断数据是否与某个假设相符）、置信区间（confidence interval）、回归（regression，即把一个变量与其他变量联系起来），以及实验和调查的设计。核心概念是*抽样分布*（sampling distribution，即估计值本身是一个随机变量（random variable））、*偏差*（bias）与*方差*（variance），以及*似然*（likelihood）。其后的课程会加入贝叶斯推断（Bayesian inference）、多元分析（multivariate analysis）、时间序列（time series）和机器学习的方法，而机器学习其实就是规模极其庞大的统计。

贯穿本章的问题

三大问题

1. **是真有其事，还是只是杂讯？** 单凭偶然，我们应预期有多大的变动？一项变化要到甚么程度才值得认真看待？
2. **我们应该怎样估计？** 有了数据，一个未知量的最佳估计是甚么？我们又怎样比较不同的方法？
3. **随机性怎样进入数据？** 数据的收集、保护或处理方式，会对我们的结论造成甚么影响？

本章先从以批判的眼光阅读报章上的统计数字开始（Section 13.1），然后转向估计：用捉放法（capture–recapture）计数田鼠的数目、由序号估计总数，以及比较平均值（mean）与中位数（median）。最后是三道关于随机性怎样进入数据的题目：保护隐私的随机化调查、会累积或互相抵消的舍入误差（rounding error），以及构造具有指定相关性的变量。

本章的概念

§	课题	概念
13.1	解读统计数字	比率、自然变异、趋势
13.2	捉放法	最大似然（maximum likelihood）
13.3	序号	偏差、方差与充分性（sufficiency）
13.4	估计中心位置	效率（efficiency）与稳健性（robustness）
13.5	随机化回答（randomised response）	混合分布（mixture）与总方差（total variance）
13.6	舍入误差	系统误差（systematic error）与随机误差（random error）
13.7	相关（correlation）	楚列斯基（Cholesky）构造法

Section 13.7 会用到协方差（covariance）和相关，这两个课题已于 2019 年从 STEP 考试范围中删去；该节会解释所需的知识。

怎样阅读本章

请记住，每一个估计值都是随机变量：假如数据有所不同，估计值也会随之不同。对于本章的每一种方法，都要问一问：重复很多次之后，它的分布会是怎样的？平均而言它是否正确？它的变动又有多大？这一个问题，正是统计思维的核心。

13.1 解读新闻标题

题目

STEP I 1992，第 14 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/92-s1-q14`

引子

统计学（statistics）是从数据中学习的科学，而它的第一课就是：数字本身不会说话。“安全六倍”、“升幅惊人”、“死伤日益惨重”：新闻标题每天都把数字变成论断，而这些论断往往是错的；错不在数字本身有假，而在于比较的对象不对，或是把寻常的随机波动（random variation）误当成趋势。

这道题完全没有公式，在 STEP 中甚为罕见。它要求你像统计学家那样思考：要证明每个标题成立，我们需要知道甚么？单凭偶然，我们又应预期有多大的波动？答案要用到泊松分布（Poisson distribution）（Section 12.18），而当中的推理正是本章其余部分的基础。

动笔之前

? **想一想** 假如死亡事件以稳定的比率独立而随机地发生，每年的总数应有多大的变动？若平均每年为 1080，每年与平均值的典型偏差是多少？

独立罕见事件的计数近似服从泊松分布，而平均值为 μ 的泊松计数，其方差（variance）为 μ，标准差（standard deviation）为 $\sqrt{\mu}$。取 $\mu = 1080$，标准差约为 33：每年总数落在约 1015 至 1145 之间，都不足为奇。

逐步解题

阶段 1 “坐飞机比步行安全六倍”

> **想一想**　每年有 1080 名行人死亡，180 人死于空难。六倍，究竟是甚么的六倍？

这个标题比较的是死亡总数，但安全与否关乎的是*风险*：每单位风险承担量（exposure）的死亡人数。几乎人人每天都步行，坐飞机的人却相对较少，而且只是偶一为之。公平的比较应采用每公里行程、每小时行程或每次行程的死亡人数，而这几种算法可以得出截然不同的答案（按每公里计，航空旅行极为安全；按每次行程计则没有那么出色，因为大多数空难死亡都发生在起飞和降落之时）。题目所给的数字根本不能“证明”两者孰更安全。此外，“在道路意外中丧生的行人”是个狭窄的类别；整体的道路死亡人数要大得多。

阶段 2　“我们的天空更安全：1990 年仅 125 人丧生”

> **想一想**　空难死亡像独立的罕见事件吗？你预期每年总数的变动有多大？

空难死亡是成群出现的：一次坠机就可能夺去数百人的性命。因此每年的总数主要取决于坠机的次数，而这个次数很小，于是总数每年的变动极大。某一年有 125 人死亡，相对于平均的 180 人，与情况毫无改变完全吻合：只要多一次坠机，这个数字就会彻底改观。一年的数据不能显示趋势。

阶段 3　“道路死伤日益惨重：星期二 7 人丧生”

> **想一想**　平均每天有多少名行人丧生？若每天的死亡人数服从泊松分布，一天之内出现 7 人或以上的情况有多常见？

平均每天有 $\frac{1080}{365} \approx 2.96$ 名行人死亡。若每天的计数服从这个平均值的泊松分布，则 $P(\geq 7) \approx 0.031$：纯粹出于偶然，每年就约有 11 天会有 7 名或以上的行人死亡。（而且“道路死伤”想必还包括司机和乘客，那么一天之内有 7 人死于道路意外就更不值得大惊小怪了。）单单一天的数字，对趋势毫无说明作用。这个标题从一个变动极大的数列中挑出一个极端值，是典型的谬误。

阶段 4　“升幅惊人：1990 年 1350 名行人丧生”

> **想一想**　把 1350 与泊松标准差比较。这纯属偶然吗？若不是，还有甚么可以解释它？这又足以称为“升幅惊人”吗？

这是全题最有分量的部分。在稳定的泊松模型之下，1350 比平均值 1080 高出 270，约为 8 个标准差：这极不可能是偶然所致。所以很可能确实有些事情改变了。但这个标题仍然言过其实，原因有以下几方面：

- *被平均值掩盖的趋势。*1080 这个数字是 16 年的平均值。假如死亡人数由 1974 年的约 900 稳步上升至 1989 年的约 1260，平均值仍然是 1080，而 1990 年的 1350 只不过是

延续既有的趋势，并非突如其来的“惊人”升幅（图 13.1.1）。我们需要逐年的数字。

- 风险承担量。人口增长、步行的人增多、交通更繁忙或人口老化，都会令死亡人数上升，而道路按每次行程计却没有变得更危险。按人均计算，或按每公里步行路程计算的比率，才是恰当的量度。
- 定义。死亡个案的记录或分类方式一旦改变（例如意外发生后多久之内死亡才计算在内），就可以造成数字的跳升。
- 过度离散（*over-dispersion*）。现实中每年计数的变动比泊松模型所显示的要大，因为各种条件（天气、经济活动）都在变。对“8 个标准差”这个说法应有所保留。

升幅很可能是真实的，但它是否惊人、成因为何，则需要更多数据才能判断。

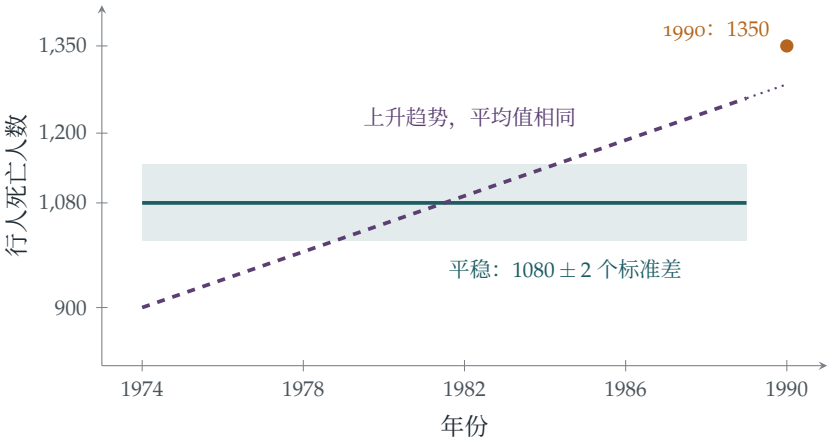


图 13.1.1 两段截然不同的历史，16 年的平均值同样是 1080。假如死亡人数一直平稳（蓝绿色，连同泊松分布典型波动的范围带），1990 年的 1350 就是急剧的跳升。假如死亡人数一直稳步上升（紫红色），1350 就只是略高于趋势。单凭平均值，无法分辨这两种情况。

回顾

- **同类才可相比。**风险需要有分母：每人、每公里或每次行程的死亡人数。
- **认识自然波动。**泊松分布的变异性，以及成群出现的事件那大得多的变异性，让我们知道甚么时候的变化才值得留意。
- **平均值会掩盖趋势。**单凭一个概括的数字，无法描述一个量如何随时间变化。

背后的数学

信号与杂讯

每一个统计问题，归根究底都是要把*信号*（signal，即真实的效应或趋势）与*杂讯*（noise，即随机波动）分开。泊松模型为独立事件计数中的杂讯提供了一把尺：标准差为 $\sqrt{\mu}$。罕见事件的*相对*波动很大：计数为 9 时，波动约为 ±3，即其大小的三分之一；而计数为 10 000 时，波动只有约 1%。正因如此，小数目（某一天的死亡人数、某一年的坠机次数）才会屡屡被人过度解读，而小镇上的癌症“群组”最终也多半只是偶然。

均值回归

还有另一个陷阱。道路安全摄影机往往安装在刚刚经历了异常恶劣一年的地点。翌年，意外减少，功劳就归于摄影机。但异常恶劣的一年过后，本来就很可能迎来较为平常的一年：这就是*均值回归*（regression to the mean），最早由高尔顿（Francis Galton）在 1880 年代提出。要把摄影机的效果与均值回归分开，就必须与没有安装摄影机的类似地点作比较，这正是对照实验（controlled experiment）背后的理念。

南丁格尔的统计学

南丁格尔（Florence Nightingale）以克里米亚战争中的护士身分闻名于世，同时也是统计学的先驱。她在 1858 年绘制的“玫瑰图”显示，死于可预防疾病的士兵远多于死于创伤的士兵，并说服了政府改革军队的卫生状况。她坚持使用比率（每年每千名士兵的死亡人数）而非原始总数，并坚持同类才可相比，这些正是本题所考核的要点。她是首位获选为皇家统计学会（Royal Statistical Society）会士的女性。

思维技巧

问一句：“跟甚么相比?”

在接受一个统计论断之前，先问清楚：比较的对象是甚么，分母是甚么，以及单凭偶然会预期有多大的波动。

到了大学，统计学课程会把这些问题化为精确的方法：假设检验（hypothesis test）探究一项观察结果是否与偶然相符；置信区间（confidence interval）说明一个量测量得有多精确；而实验设计（design of experiments）则确保比较公平。

延伸练习

1 **泊松分布的尾部。**设 X 服从平均值为 2.96 的泊松分布，验证 $\mathrm{P}(X \geq 7) \approx 0.031$。你预期每年有多少天会有 9 名或以上的行人死亡?

2 **成群出现的死亡。**假设坠机事故按泊松过程（Poisson process）发生，比率为每年 1.2 宗，而每宗夺去 150 人的性命。求每年空难死亡人数的平均值和标准差。一年之内有 125 人或以下死亡，是否不寻常？有没有可能某一年恰好有 125 人死亡?

3 **显著的升幅?**假如 1974–1989 年的数字一直平稳地维持在 1080 ± 33，试（利用正态近似（normal approximation））求单一年份出现 1350 或以上的概率。

4 **拟一个更好的标题。**为四个标题逐一改写，各拟一个数据真正支持的标题。

延伸阅读

- **David Spiegelhalter, *The Art of Statistics***（Pelican）。讲述如何从数据中学习，并附有大量取自新闻的例子。
- **Tim Harford, *How to Make the World Add Up***（Bridge Street）。思考传媒中的数字的十条法则。
- **Darrell Huff, *How to Lie with Statistics***（Penguin）。1954 年出版的短篇经典，至今仍不过时。

13.2 计数田鼠

题目

STEP II 2000，第 13 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/00-s2-q13`

引子

一个岛上住了多少只田鼠？逐只去数是不可能的：牠们会躲藏，会走动，而且长得一模一样。不过，你可以先捉一些，做上记号，放走，然后再捉一批。假如有 200 只做了记号，而第二次捉到的 200 只当中有 11 只带有记号，那么整个种群中大约有 $\frac{11}{200}$ 带有记号，所以种群的大小约为 $200 \times \frac{200}{11} \approx 3636$。

这种*标记重捕法*（capture–recapture）用于估计鱼类、鸟类和鲸鱼的数量，也用于估计吸毒者、无家者和战争伤亡者的人数，即任何无法完全计数的总体。这道题用*最大似然*（maximum likelihood）原理为这个直观的估计打下稳固的基础。最大似然是统计估计中最重要的思想：选取使观察到的数据出现概率最大的那个未知量的值。

动笔之前

想一想 种群共有 N 只田鼠，其中 200 只带有记号。现随机抽取 200 只作为样本。恰好有 11 只带有记号的概率是多少？

从 200 只有记号的田鼠中选出 11 只，再从 $N-200$ 只没有记号的田鼠中选出 189 只，而等可能的样本共有 $\binom{N}{200}$ 个：

$$p_N = \frac{\binom{200}{11}\binom{N-200}{189}}{\binom{N}{200}}.$$

这就是*超几何分布*（hypergeometric distribution）。

逐步解题

阶段 1　似然

把含有 N 的二项式系数（binomial coefficient）展开写出，

$$\binom{N-200}{189} = \frac{(N-200)!}{189!\,(N-389)!}, \qquad \binom{N}{200} = \frac{N!}{200!\,(N-200)!},$$

所以

$$p_N = \binom{200}{11}\frac{200!}{189!}\cdot\frac{((N-200)!)^2}{N!\,(N-389)!} = k\frac{((N-200)!)^2}{N!\,(N-389)!}. \qquad \blacksquare$$

把数据固定，视之为未知量 N 的函数时，p_N 称为*似然*（likelihood）。

阶段 2　对整数求最大值

> **想一想**　N 是整数，所以用微积分并不方便。不如比较 p_{N+1} 与 p_N。两者之比何时大于 1？

大部分阶乘都会互相约去：

$$\frac{p_{N+1}}{p_N} = \frac{(N-199)^2}{(N+1)(N-388)}.$$

当 $(N-199)^2 > (N+1)(N-388)$ 时，这个比大于 1，即 $N^2 - 398N + 39601 > N^2 - 387N - 388$，亦即 $11N < 39989$，$N < 3635.4$。所以 p_N 一直递增至 $N = 3636$，其后递减：最大似然估计值为

$$\hat{N} = 3636. \qquad \blacksquare$$

这正是 $\frac{200\times 200}{11} = 3636.4$ 的整数部分，也就是那个直观的估计。

阶段 3　第三个样本

抽取第二个样本之后，其中 189 只没有记号的田鼠也被做上记号，所以现在共有 389 只田鼠带有记号。取 $N = 3636$，则有 $3636 - 389 = 3247$ 只没有记号。第三个含 200 只田鼠的样本中恰好有 j 只带有记号的概率为

$$\frac{\binom{389}{j}\binom{3247}{200-j}}{\binom{3636}{200}}.$$

这些概率（对 $j = 0, 1, \ldots, 200$）之和必定等于 1，由此即得该恒等式。 $\blacksquare$

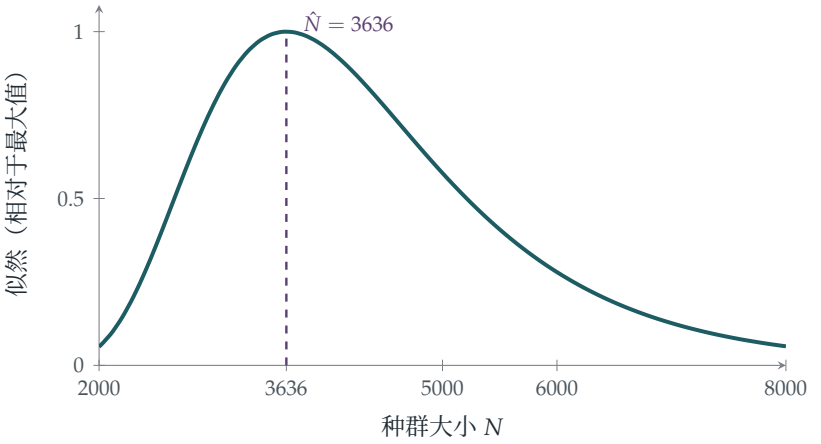


图 13.2.1 在第二个样本中发现 11 只有记号田鼠的似然 p_N，表示为种群大小 N 的函数，并按比例缩放使其最大值为 1。它在 $N = 3636$ 处达到峰值，但曲线宽阔而且偏斜：由大约 2500 至 6500 的种群大小，都与数据相当吻合。

回顾

- **把数据的概率看成未知量的函数。**似然把惯常的观点倒转过来：数据固定不变，参数（parameter）才是变量。
- **用比值求离散最大值。**比较相邻两项，便找出了一个整数变量函数的最大值所在。
- **概率之和为 1。**把一个分布对所有结果求和，便证明了一个组合恒等式（范德蒙恒等式，Vandermonde's identity）。

背后的数学

最大似然

最大似然原理由费雪（Ronald Fisher）在 1910 至 1920 年代发展出来，内容是：以使观察到的数据出现概率最大的那个值，作为未知参数的估计。它是至今应用最广的估计方法，由以曲线拟合数据，到训练机器学习中的神经网络，无一不用；神经网络的“损失函数”（loss function）通常就是负对数似然。费雪证明了，对于大样本，最大似然估计值近似服从正态分布（normal distribution），而且其精确程度不逊于任何合理的估计量（estimator），前提是模型具备某些良好性质，而这些性质对大多数日常模型都成立。不过也有重要的例外。在下一节（Section 13.3）中，未知量是可能出现的最大值，即分布的一个端点，在那里最大似然估计值的表现大不相同：它的误差并不近似服从正态分布，而且误差是以 $\frac{1}{n}$ 而非 $\frac{1}{\sqrt{n}}$ 的速度缩小。

这个估计有多精确？

图 13.2.1 显示似然曲线很宽阔：11 次重捕所提供的信息并不多。粗略计算一下：第二个样本中有记号田鼠的数目近似服从泊松分布（Poisson distribution），平均值为 $\frac{200\times 200}{N}$，所以其标准差（standard deviation）约为 $\sqrt{11}\approx 3.3$，相对误差约为 30%，而这个误差会传递到 N 的估计上。若要把 N 的估计误差控制在 10% 以内，生物学家需要大约 100 次重捕。统计学家还会修正这个简单估计的轻微偏差（bias），采用*查普曼估计量*（Chapman estimator）$\frac{(M+1)(C+1)}{R+1}-1$，其中 M 和 C 是两次捕捉的数量，R 是重捕到的数目。

拉普拉斯计数法国人口

拉普拉斯（Laplace）在 1802 年用了同样的想法，在没有人口普查的情况下估计法国的人口。他知道全国一年内登记的出生人数，并且就一批抽样的市镇而言，同时知道其人口和出生人数。他假设人口与出生人数之比在各地都相同，从而估计法国人口约为 28 百万，甚至给出了不确定程度的度量。后来这个方法在动物种群研究中被重新发现：C. G. J. Petersen 在 1890 年代于丹麦的峡湾为鲽鱼做记号，Frederick Lincoln 则在 1930 年代使用鸟环；这个方法至今仍常被称为*林肯—彼得森*（Lincoln–Petersen）估计量。

思维技巧

选取使数据最有可能出现的参数

要估计一个未知量，先把你所观察到的结果的概率写成该未知量的函数，然后求其最大值。

在大学里，最大似然是统计推断（statistical inference）的骨干，并有一套理论说明估计值有多精确（费雪信息量，Fisher information），以及如何比较互相竞争的模型（似然比检验，likelihood ratio test）。

延伸练习

1 **一般公式。**若有 M 只动物做了记号，而第二个含 C 只动物的样本中有 R 只带有记号，证明 N 的最大似然估计值是 $\frac{MC}{R}$ 的整数部分。

2 **范德蒙恒等式。**考虑从 a 名女士和 b 名男士中选出 c 人组成委员会的方法数目，以此证明 $\sum_j \binom{a}{j}\binom{b}{c-j}=\binom{a+b}{c}$。

3 **掷硬币。**一枚硬币掷出正面的概率 p 未知，掷 n 次后观察到 h 次正面。证明 p 的最大似然估计值是 $\frac{h}{n}$。

4 **没有重捕。**如果第二个样本中没有任何带记号的田鼠，最大似然估计值会变成怎样？这说明了这个方法的甚么特点？

延伸阅读

- **David Spiegelhalter, *The Art of Statistics***（Pelican），谈估计与不确定性。
- **George Seber, *The Estimation of Animal Abundance***（Blackburn）。标记重捕法的标准参考书。
- **John Rice, *Mathematical Statistics and Data Analysis***（Duxbury）。数理统计的入门课本，内容包括最大似然。

13.3 德国坦克问题

题目

STEP III 1998，第 14 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/98-s3-q14`

引子

第二次世界大战期间，盟军情报部门想知道德国每月生产多少辆坦克。间谍的报告显示每月约 1400 辆。统计学家则另辟蹊径：缴获的坦克在变速箱和其他零件上都刻有序号，而这些序号实际上就是从 $1, 2, \ldots, N$ 中抽出的一个随机样本（random sample）。单凭这些数字，他们估计产量约为每月 250 辆。按照一般引用的数字，战后查阅的德国纪录显示为 245 辆。统计学家的估计远比间谍准确。

本题比较两种利用序号估计 N 的方法。两者平均而言都是正确的，但其中一种运用数据的效率远高于另一种；弄清楚是哪一种，就会引出估计理论（estimation theory）的核心思想：在平均而言正确的估计量（estimator）当中，应选取波动最小的一个。

动笔之前

想一想 设 M 是 $[0, N]$ 上 n 个独立均匀分布（uniform distribution）取值中的最大者，那么 $\mathrm{P}(M \le m)$ 是多少？

$M \le m$ 当且仅当每个 $X_i \le m$，所以当 $0 \le m \le N$ 时，$\mathrm{P}(M \le m) = \left(\frac{m}{N}\right)^n$（可与 Section 12.25 比较）。求导后可知，M 的概率密度（density）为 $\frac{nm^{n-1}}{N^n}$。

逐步解题

阶段 1 两者皆无偏

对于 M：

$$\mathrm{E}(M) = \int_0^N m \cdot \frac{n m^{n-1}}{N^n}\,\mathrm{d}m = \frac{n}{n+1}N, \qquad \text{所以} \qquad \mathrm{E}(Z_1) = \frac{n+1}{n} \cdot \frac{n}{n+1}N = N.$$

对于平均值：每个 X_i 的平均值为 $\frac{N}{2}$，所以 $\mathrm{E}(Z_2) = \frac{2}{n} \cdot n \cdot \frac{N}{2} = N$。 ■

两个估计量都是无偏（unbiased）的：平均而言，它们会给出正确的答案。因子 $\frac{n+1}{n}$ 用来修正这个事实：观察到的最大序号总是比 N 略小一点。

阶段 2 两者的方差

想一想 求 $\mathrm{E}(M^2)$，从而求 $\mathrm{Var}(M)$。至于 Z_2，单一个 $[0, N]$ 上的均匀变量的方差（variance）是多少？

$\mathrm{E}(M^2) = \int_0^N m^2 \cdot \frac{n m^{n-1}}{N^n}\,\mathrm{d}m = \frac{n}{n+2}N^2$，所以

$$\mathrm{Var}(M) = \frac{n}{n+2}N^2 - \frac{n^2}{(n+1)^2}N^2 = \frac{nN^2}{(n+1)^2(n+2)},$$
$$\mathrm{Var}(Z_1) = \frac{(n+1)^2}{n^2}\mathrm{Var}(M) = \frac{N^2}{n(n+2)}.$$

$[0, N]$ 上的均匀变量的方差为 $\frac{N^2}{12}$，而各个 X_i 互相独立，所以

$$\mathrm{Var}(Z_2) = \frac{4}{n^2} \cdot n \cdot \frac{N^2}{12} = \frac{N^2}{3n}.$$ ■

阶段 3 哪一个较好？

想一想 比较 $\frac{1}{n(n+2)}$ 与 $\frac{1}{3n}$。当 n 增大时，两者如何变化？

两者之比为 $\frac{\mathrm{Var}(Z_1)}{\mathrm{Var}(Z_2)} = \frac{3}{n+2}$，当 $n \geq 2$ 时小于 1（当 $n = 1$ 时则相等，此时两个估计量完全相同）。所以 Z_1 较好，而且 n 愈大就好得愈多：它的标准差（standard deviation）以 $\frac{1}{n}$ 的速度缩小，而 Z_2 的标准差只以 $\frac{1}{\sqrt{n}}$ 的速度缩小（图 13.3.1）。若有 $n = 10$ 个序号，Z_1 的标准差约为 N 的 9%，而 Z_2 的约为 18%。

选用 Z_1 还有另一个理由：Z_2 有时会得出荒谬的结果。若序号为 10、20 和 300，则 $Z_2 = \frac{2}{3} \cdot 330 = 220$，比我们实际见过的一个序号还要小。而 $Z_1 = \frac{4}{3} \cdot 300 = 400$ 永远不会小于最大的观察值。 ■

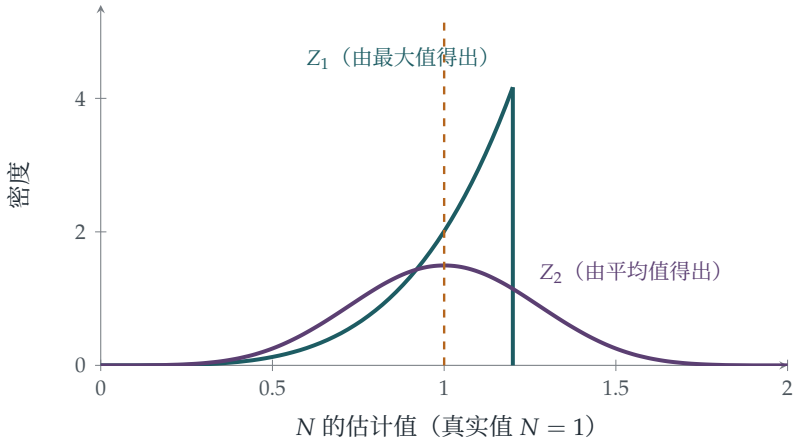


图 13.3.1 两个估计量的抽样分布（sampling distribution），其中序号数目 $n = 5$，真实值 $N = 1$。两者的平均值都恰好是 1，但 Z_1（蓝绿色）比 Z_2（紫红色）更紧密地集中在真实值附近，而且它永远不会低于观察到的最大序号。

回顾

- **无偏性。**每个估计量都经过缩放，使其期望值（expected value）恰好等于 N。
- **比较方差。**在无偏估计量当中，方差较小的一个较为可靠。
- **选用信息量最大的统计量。**最大值包含了样本中关于 N 的全部信息；平均值则浪费了其中一大部分。

背后的数学

充分性

为甚么最大值好这么多？一旦知道最大值 M，其余各值不论 N 是多少，都只是在 $[0, M]$ 上均匀分布：它们再没有带来关于 N 的任何信息。最大值是一个*充分统计量*（sufficient statistic）。C. R. Rao 与 David Blackwell 的一条定理（1945–47）指出，任何无偏估计量，只要换成它在给定充分统计量下的期望值，都可以得到改进，或至少不会变差；在这里，这个做法正好把 Z_2 变成 Z_1。*Lehmann–Scheffé 定理*进一步证明，Z_1 是方差最小的无偏估计量。还要留意，Z_1 的方差以 $\frac{1}{n^2}$ 的速度下降，比大多数估计量的 $\frac{1}{n}$ 更快：这是因为 N 是分布的端点，数据会挤向这个端点。

离散版本

真实的序号是整数 $1, 2, \dots, N$，并且是以不放回（without replacement）方式抽样的。这时最佳的无偏估计是 $\hat{N} = m + \frac{m}{k} - 1$，其中 m 是 k 个观察到的序号中的最大者：即最大值加上各观察值之间的平均间距。盟军统计学家运用这类估计，主要根据变速箱和车轮的序号，估计德国的产量约为每月 246 辆坦克；相比之下，其后在德国纪录中找到的数字是 245 辆，而情报报告所显示的是 1400 辆。

历史上的序号

此后这个方法曾多次派上用场，例如分析员曾根据最先售出的几件产品的序号，估计某款新消费产品的生产数量。制造商汲取了教训：现在不少制造商都会把序号编码，使人无法把它们当作简单的顺序编号来解读。Richard Ruggles 与 Henry Brodie 于 1947 年在 *Journal of the American Statistical Association* 上记述了这项战时工作。

思维技巧

以抽样分布评价估计量

估计量是一个随机变量（*random variable*）。要问它平均而言是否正确（偏差），以及它的波动有多大（方差），然后选取误差较小的一个。

在大学，这会引出有效估计（efficient estimation）的理论：Cramér–Rao 下界（Cramér–Rao bound）指出无偏估计量的方差最小可以小到甚么程度，而充分性则说明如何达到这个下界。

延伸练习

1. **均方误差（mean squared error）。**有偏估计量 M 满足 $\mathrm{E}(M-N)^2 = \mathrm{Var}(M) + \big(\mathrm{E}(M) - N\big)^2$。计算此值，并与 $\mathrm{Var}(Z_1)$ 比较。求使 $\mathrm{E}(cM - N)^2$ 取最小值的常数 c。

2. **最小值也可以。**证明 $(n+1)\min\{X_i\}$ 也是 N 的无偏估计量，并求它的方差。这个估计量好用吗？

3. **离散公式。**对于序号 $1, \dots, N$，以不放回方式抽取 k 个，其中最大者为 m。证明 $\mathrm{P}(\max \le m) = \binom{m}{k} / \binom{N}{k}$，并验证当 $k = 1$ 时 $m + \frac{m}{k} - 1$ 是无偏的。

4. **坦克。**若 5 辆缴获的坦克的序号为 19、40、42、60、98，试估计坦克的生产数量。

延伸阅读

- **Richard Ruggles & Henry Brodie, "An empirical approach to economic intelligence in World War II"**，*Journal of the American Statistical Association* 42 (1947)。最初的记述。
- **John Rice, *Mathematical Statistics and Data Analysis***（Duxbury），论及偏差、方差与充分性。
- **David Spiegelhalter, *The Art of Statistics***（Pelican）。

13.4 平均数还是中位数?

题目

STEP II 2000，第 14 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/00-s2-q14`

引子

要估计一组量度数据的中心，你可以取平均值（mean），也可以取中位数（median）。哪一个较好？答案视乎数据而定。假如量度结果偶尔会出现离谱的误差，中位数就可靠得多：一个荒谬的数值足以把平均值拉到任何地方，却几乎动摇不了中位数。至于规规矩矩的数据，通常是平均值胜出。本题要量度的，正是在数据服从均匀分布（uniform distribution）时，平均值究竟胜出多少：中位数的方差（variance）是平均值的三倍，所以要令中位数与平均值同样精确，就需要三倍数量的观测值。

这个比较引入了*效率*（efficiency）的概念，它是选择统计方法时的主要考虑之一；同时也用到 Beta 积分（beta integral），这个积分在 Section 4.4 也出现过。

动笔之前

想一想 为甚么中位数的密度（density）与 $y^n(1-y)^n$ 成正比？（要 Y 接近 y，$2n+1$ 个值之中须有多少个在其下，又有多少个大于 y？）

中位数要落在 y，就必须有一个值在 y，n 个值在其下（各自的概率为 y），另有 n 个值在其上（各自的概率为 $1-y$）。这就得出因子 $y^n(1-y)^n$；常数则是数算哪些值扮演哪个角色的选法数目。

逐步解题

阶段 1　常数

密度的积分必须等于 1：

$$k\int_0^1 y^n(1-y)^n\,\mathrm{d}y = k\cdot\frac{n!\,n!}{(2n+1)!} = 1, \qquad k = \frac{(2n+1)!}{(n!)^2}. \quad \blacksquare$$

阶段 2　平均值与方差

密度关于 $\frac{1}{2}$ 对称（把 y 换成 $1-y$ 即可看出），所以 $\mathrm{E}(Y) = \frac{1}{2}$。

> **想一想**　利用 Beta 积分计算 $\mathrm{E}(Y^2) = k\int_0^1 y^{n+2}(1-y)^n\,\mathrm{d}y$，并化简当中的阶乘。

$$\mathrm{E}(Y^2) = \frac{(2n+1)!}{(n!)^2}\cdot\frac{(n+2)!\,n!}{(2n+3)!} = \frac{(n+2)(n+1)}{(2n+3)(2n+2)} = \frac{n+2}{2(2n+3)}.$$

因此

$$\mathrm{Var}(Y) = \frac{n+2}{2(2n+3)} - \frac{1}{4} = \frac{2(n+2)-(2n+3)}{4(2n+3)} = \frac{1}{4(2n+3)}. \quad \blacksquare$$

阶段 3　与平均值比较

> **想一想**　$\bar{X}$ 是 $2n+1$ 个独立均匀变量的平均，它的平均值和方差是多少？

每个 X_i 的平均值为 $\frac{1}{2}$，方差为 $\frac{1}{12}$，所以 $\mathrm{E}(\bar{X}) = \frac{1}{2}$，而 $\mathrm{Var}(\bar{X}) = \frac{1}{12(2n+1)}$。当 n 很大时，

$$\mathrm{Var}(Y) \approx \frac{1}{8n}, \qquad \mathrm{Var}(\bar{X}) \approx \frac{1}{24n}.$$

> **想一想**　假设 Y 和 $\bar{X}$ 都服从正态分布（normal distribution），平均值为 $\frac{1}{2}$，标准差（standard deviation）分别为 σ_Y 和 $\sigma_{\bar{X}}$，试用标准正态分布函数 Φ 表示这两个概率。

$\mathrm{P}\big(|Y-\frac{1}{2}| < \frac{d}{\sqrt{n}}\big) = 2\Phi\big(\frac{d}{\sqrt{n}\,\sigma_Y}\big) - 1$，而 $\sqrt{n}\,\sigma_Y \to \frac{1}{\sqrt{8}}$；$\bar{X}$ 的情况相似，其中 $\sqrt{n}\,\sigma_{\bar{X}} \to \frac{1}{\sqrt{24}}$。所以两个概率分别趋向 $2\Phi(d\sqrt{8}) - 1$ 和 $2\Phi(d\sqrt{24}) - 1$。由于 Φ 是递增函数，而且 $\sqrt{24} > \sqrt{8}$，后者较大，因此对所有足够大的 n，不等式都成立。■平均值较有可能接近真正的中心。

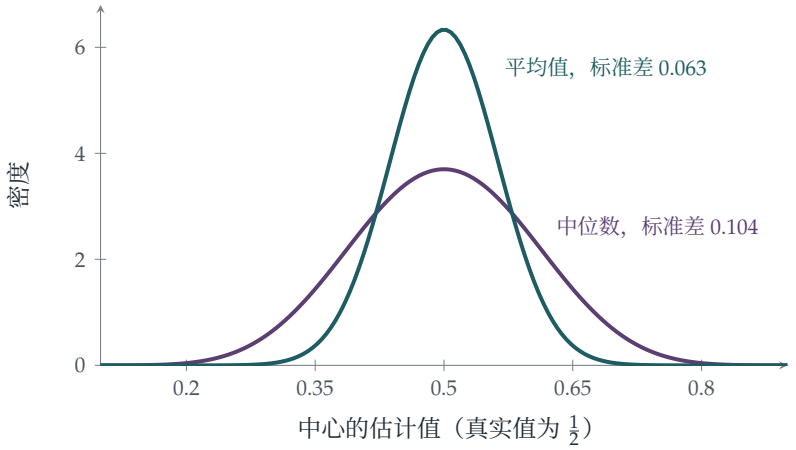


图 13.4.1 21 个独立均匀观测值的中位数（紫红色）和平均值（蓝绿色）的抽样分布（sampling distribution）。两者都以真实值 $\frac{1}{2}$ 为中心，但平均值集中得多：它的方差大约是中位数方差的三分之一。

回顾

- **Beta 积分。**$\int_0^1 y^r(1-y)^s\,\mathrm{d}y$ 这一条公式,同时给出了归一化常数(normalising constant)和各阶矩（moments)。
- **对称性给出平均值。**求 $\mathrm{E}(Y)$ 根本不用积分。
- **在合适的尺度上比较离散程度。**把标准差乘以 $\sqrt{n}$，比较的结果便与 n 无关。

背后的数学

效率

两个无偏估计量（unbiased estimator）的*相对效率*（relative efficiency）是它们方差之比；它说明其中一个需要多少倍的观测值，才能与另一个看齐。对于均匀分布的数据，中位数相对于平均值的效率是 $\frac{1}{3}$。对于正态分布的数据，这个数字是 $\frac{2}{\pi} \approx 0.64$：中位数大约需要多 57% 的观测值。但遇上厚尾（heavy-tailed）数据，形势就逆转了。对于柯西（Cauchy）分布（Section 12.22），平均值毫无用处，因为 n 个观测值的平均并不比单独一个观测值好，而中位数却表现良好。对于拉普拉斯（Laplace）分布（Section 12.24），中位数的效率是平均值的两倍。

稳健统计学

中位数是*稳健*（robust）的：即使把接近一半的数据改成任意数值，它也不会移动太远。平均值却没有这种保障：一个错误的数值就足以把它毁掉。在 1960 及 1970 年代，John Tukey、Peter Huber 等人发展出*稳健统计学*（robust statistics），即一些估计量，对干净的数据而言效率几乎与平均值一样高，却又能抵御离群值（outlier），例如*截尾平均值*（trimmed mean）（舍去最大和最小的 10%，再取余下数据的平均），跳水、体操等运动项目的评分就采用这种方法。

拉普拉斯与中位数

拉普拉斯在 1774 年证明，假如误差服从双指数（拉普拉斯）分布，则真实值的最佳估计是中位数；高斯（Gauss）在 1809 年证明，假如误差服从正态分布，则最佳估计是平均值，他并由此得出最小二乘法（method of least squares）。谁是谁非，取决于误差的表现如何；每遇上一种新的数据，统计学家至今仍会为这个问题争论一番。

思维技巧

以变异程度比较不同方法

两种方法可能平均而言都正确；要比较的是它们的答案变动有多大，以及数据增加时这种变动收缩得有多快。

在大学里，渐近效率（asymptotic efficiency），即比较大样本下的方差，是在不同统计方法之间作取舍的标准工具；与之并用的还有稳健性，以应付模型可能出错的情况。

延伸练习

1 **中位数的密度。**用数算的方法推导密度 $\frac{(2n+1)!}{(n!)^2}y^n(1-y)^n$：一个值落在 $[y, y+\delta y]$ 内，n 个在其下，n 个在其上。

2 **三个观测值。**当 $n=1$（三个均匀观测值）时，求 $\mathrm{Var}(Y)$ 和 $\mathrm{Var}(\bar{X})$ 的准确值。哪一个较小？

3 **中列数。**中心的另一个估计量是中列数（midrange）$\frac{1}{2}(\max+\min)$。证明对于均匀分布的数据，若有 $m=2n+1$ 个观测值，它的方差是 $\frac{1}{2(m+1)(m+2)}$。它与平均值相比如何？（试与 Section 13.3 比较。）

4 **一个离群值。**十一个量度结果为 9.8、9.9、10.0、10.0、10.1、10.1、10.2、10.2、10.3、10.4 和 104（打字错误）。比较平均值和中位数。

延伸阅读

- **John Rice, *Mathematical Statistics and Data Analysis***（Duxbury），论及顺序统计量（order statistics）和效率。
- **Rand Wilcox, *Fundamentals of Modern Statistical Methods***（Springer）。为非专家讲解稳健方法为何重要。
- **Stephen Stigler, *The Seven Pillars of Statistical Wisdom***（Harvard University Press）。从历史角度看数据的汇总、平均值和中位数。

13.5 问尴尬的问题

题目

STEP II 2023，第 11 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2023 年 STEP 试卷（zip 文件）内的 STEP II 试卷：

`www.ocr.org.uk/Images/707102-2023-past-papers.zip`

引子

怎样才能知道有多少比例的学生曾在考试中作弊，或者有多少比例的雇员曾偷取公司财物？直接问的话，很多人都会说谎。1965 年，统计学家 Stanley Warner 想出了一个巧妙的解决办法：随机化回答（randomised response）。每个人私下掷一枚硬币或一颗骰子，由结果决定他是如实回答那条敏感问题，还是回答与之相反的问题。访问员永远不知道是哪一种情况，所以任何一个人的答案都不足以成为对他不利的证据；然而，由于随机化的概率（probability）是已知的，整体的比例仍然可以估计出来。

本题探讨的正是这背后的数学。让每个人独立地（independently）随机化（第 (ii) 部分），所得的总数性质良好，足以确定该比例。用同一次掷硬币为所有人随机化（第 (iii) 部分），所得的总数虽然平均值（mean）相同，分布（distribution）却截然不同，呈双峰形状，几乎不能告诉我们任何事。随机性放在哪里，至关重要。

动笔之前

想一想 Y 以概率 p 等于 X_1，以概率 q 等于 X_2。$\mathrm{E}(Y)$ 是甚么？$\mathrm{E}(Y^2)$ 又是甚么？

以选中哪一个变量为条件，可得 $\mathrm{E}(Y) = p\mu_1 + q\mu_2$，以及 $\mathrm{E}(Y^2) = p\,\mathrm{E}(X_1^2) + q\,\mathrm{E}(X_2^2) = p(\sigma_1^2 + \mu_1^2) + q(\sigma_2^2 + \mu_2^2)$。

逐步解题

阶段 1　混合分布的方差

$$\mathrm{Var}(Y) = p(\sigma_1^2 + \mu_1^2) + q(\sigma_2^2 + \mu_2^2) - (p\mu_1 + q\mu_2)^2.$$

含 μ 的各项为 $p\mu_1^2 + q\mu_2^2 - p^2\mu_1^2 - 2pq\mu_1\mu_2 - q^2\mu_2^2 = pq\mu_1^2 + pq\mu_2^2 - 2pq\mu_1\mu_2$，当中用到 $p - p^2 = pq$ 和 $q - q^2 = pq$。因此

$$\mathrm{Var}(Y) = p\sigma_1^2 + q\sigma_2^2 + pq(\mu_1 - \mu_2)^2. \qquad \blacksquare$$

混合分布（mixture）的方差（variance），等于各成分内部方差的平均，再加上各成分平均值之间的方差。

阶段 2　独立随机化

> **想一想**　$\mathrm{P}(B = 1)$ 是多少？B 的 n 个值是否独立？

$\mathrm{P}(B = 1) = \frac{1}{2}\cdot\frac{1}{6} + \frac{1}{2}\cdot\frac{5}{6} = \frac{1}{2}$。这 n 个值互相独立，每个的成功概率都是 $\frac{1}{2}$，所以 $Z_1 \sim \mathrm{B}(n, \frac{1}{2})$，其平均值为 $\frac{n}{2}$，标准差（standard deviation）为 $\frac{\sqrt{n}}{2}$。

与平均值相差不超过 10%，即是 $|Z_1 - \frac{n}{2}| < \frac{n}{20}$。以标准单位表示，这就是 $|Z| < \frac{n/20}{\sqrt{n}/2} = \frac{\sqrt{n}}{10}$，而此值趋向无穷大，所以该概率趋向 1。 $\blacksquare$

阶段 3　所有人共用一枚硬币

> **想一想**　现在只掷一次硬币，便决定全部 n 颗骰子的情况。已知掷得正面，Z_2 的分布是甚么？已知掷得反面呢？试用第 (i) 部分的结果。

已知掷得正面，$Z_2 \sim \mathrm{B}(n, \frac{1}{6})$：平均值为 $\frac{n}{6}$，方差为 $\frac{5n}{36}$。已知掷得反面，$Z_2 \sim \mathrm{B}(n, \frac{5}{6})$：平均值为 $\frac{5n}{6}$，方差为 $\frac{5n}{36}$。在第 (i) 部分取 $p = q = \frac{1}{2}$，得

$$\mathrm{E}(Z_2) = \frac{n}{2}, \qquad \mathrm{Var}(Z_2) = \frac{5n}{36} + \frac{1}{4}\left(\frac{4n}{6}\right)^2 = \frac{5n}{36} + \frac{n^2}{9}.$$

> **想一想**　大部分的概率集中在哪里？在平均值 $\frac{n}{2}$ 附近的概率多吗？

Z_2 有两个峰，分别在 $\frac{n}{6}$ 和 $\frac{5n}{6}$ 附近，每个峰的阔度约为 $\sqrt{n}$（图 13.5.1），而两峰之间几乎甚么都没有。具有相同平均值和方差的正态分布（normal distribution）则只有一个峰，中心位于 $\frac{n}{2}$，正正是 Z_2 几乎从不出现的位置。要 Z_2 与 $\frac{n}{2}$ 相差不超过 10%，一个服从 $\mathrm{B}(n, \frac{1}{6})$ 或 $\mathrm{B}(n, \frac{5}{6})$ 的变量必须落在 $[0.45n, 0.55n]$ 之内。对 $\mathrm{B}(n, \frac{1}{6})$ 而言，这比其平均值高

出至少 $0.28n$，约为 $\frac{0.28n}{0.37\sqrt{n}} \approx 0.75\sqrt{n}$ 个标准差，所以该概率趋向 0；$\mathrm{B}(n, \frac{5}{6})$ 的情况亦类似。 ■

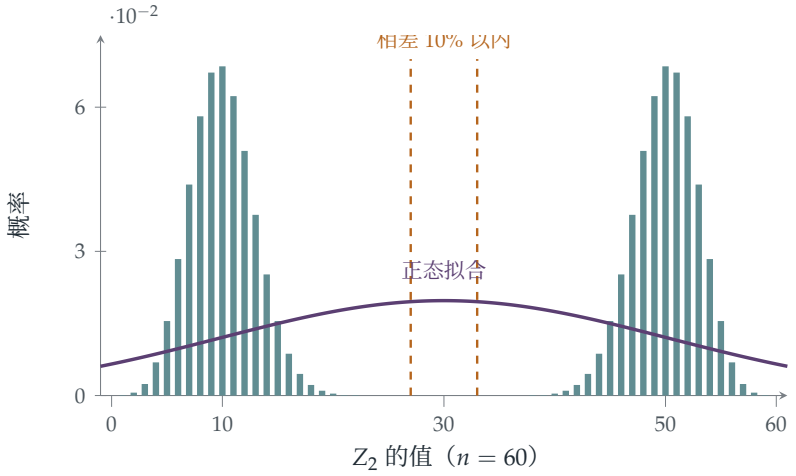


图 13.5.1 $n = 60$ 颗骰子时 Z_2 的分布（长条）：两个峰分别位于 10 和 50，来自那一次掷硬币的两种结果。具有相同平均值（30）和方差的正态分布（紫红色）拟合得很差，而且几乎没有概率落在与平均值相差 10% 以内的范围（两条琥珀色线之间）。

回顾

- **以隐藏的选择为条件。**混合分布的平均值和方差，都是以选中哪一个成分为条件而求得的。
- **内部方差加上组间方差。**额外的一项 $pq(\mu_1 - \mu_2)^2$ 量度了两个成分之间的离散程度。
- **平均值和方差并非全部。**Z_2 与 Z_1 的平均值相同，但形状完全不同。

背后的数学

随机化回答

在 Warner 的方案中，每位受访者私下作出选择，以概率 p 回答“你是否属于 A 组?”，以概率 $1 - p$ 回答“你是否不属于 A 组?”。若属于 A 的人所占比例为 π，则答“是”的概率为

$$\lambda = p\pi + (1 - p)(1 - \pi),$$

因此由观察所得答“是”的比例 $\hat{\lambda}$，我们可以估计 $\hat{\pi} = \frac{\hat{\lambda} - (1-p)}{2p-1}$（前提是 $p \neq \frac{1}{2}$）。本题第 (ii) 部分正是这种情况，由每个人自己的骰子来决定。隐私的代价是精确度：除以 $2p - 1$

会令方差增大，所以随机化调查所需的受访者比直接调查更多。不过每个人都是独立地随机化的，所以大数定律（law of large numbers）能够发挥作用。第 (iii) 部分显示了共用随机化时会出甚么问题：那一枚硬币所带来的不确定性，永远不会被平均抵销。

差分隐私

随机化回答是*差分隐私*（differential privacy）的前身。差分隐私是 Cynthia Dwork 及其同事于 2006 年提出的一个关于隐私的数学定义：如果加入或移除任何一个人的数据，只会令任何输出的概率改变至多一个很小的倍数，该程序便算是保障隐私的。科技公司用它来收集使用统计数据，美国人口普查局亦在 2020 年人口普查中采用了它，在公布的数字上加入经仔细校准的随机杂讯，令任何个人的答案都无法被还原。

全方差定律

第 (i) 部分的恒等式是*全方差定律*（law of total variance）的一个特例：$\operatorname{Var}(Y) = \operatorname{E}\big(\operatorname{Var}(Y \mid C)\big) + \operatorname{Var}\big(\operatorname{E}(Y \mid C)\big)$，其中 C 是隐藏的选择。它是*方差分析*（analysis of variance，ANOVA）的基础。方差分析由费雪（Ronald Fisher）于 1920 年代在 Rothamsted 为农业实验而引入，它把农作物产量的变异分拆为两部分：可由肥料或品种解释的部分，以及未能解释的部分。

思维技巧

要看整个分布，而不只是它的概括数值

平均值和方差能很好地描述单峰分布，但对混合分布却可能造成严重误导。要问一问：概率实际上落在哪里。

在大学里，混合模型（mixture model）用来描述由几个群组组成的总体，而全方差定律则是统计学的标准工具，从调查设计到阶层模型（hierarchical model）都会用到。

延伸练习

1 **双胞胎。**试做 STEP II 2005，第 12 题（`step.maths.org/questions/05-s2-q12`）。这其实是一道乔装了的随机化回答谜题：在甚么情况下，一个答案完全不能透露硬币的结果？

2 **估计 π。**在一项 $p = \frac{3}{4}$ 的 Warner 调查中，有 400 人作答，其中 160 人答“是”。试估计 A 组所占的比例 π，并求出你的估计值的近似标准误差（standard error）。

3 **全方差。**设 C 为取多个值的离散隐藏变量，证明全方差定律。

4 **共用硬币。**在第 (iii) 部分中，$Z_2 > \frac{n}{2}$ 的概率是多少？试解释为何当 n 很大时，这个概率接近 $\frac{1}{2}$，尽管 Z_2 几乎从不接近 $\frac{n}{2}$。

延伸阅读

- **Stanley Warner, “Randomized response: a survey technique for eliminating evasive answer bias”**，*Journal of the American Statistical Association* 60 (1965)。
- **Cynthia Dwork & Aaron Roth, *The Algorithmic Foundations of Differential Privacy***（可于网上免费阅读）。现代理论；开首几章不难读懂。
- **David Spiegelhalter, *The Art of Statistics***（Pelican），谈调查及其陷阱。

13.6 累积起来的误差

题目

STEP II 1993，第 15 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/93-s2-q15`

引子

1982 年，温哥华证券交易所推出了一个新指数，起始值为 1000。每当有股价变动，指数便会重新计算，并截断（truncate）至小数点后三位。每次截断平均损失 0.0005，看来微不足道。可是每天大约要重新计算 3000 次，二十二个月后，尽管大市上升，指数却只剩下约 520。经正确重算后，指数一夜之间跃升至约 1100。

这道题目解释了个中原因。截断误差（truncation error）全部朝同一方向，因此会稳步累积，与运算次数成正比。舍入误差（rounding error）则有正有负，大部分互相抵销：它们的增长速度只相当于运算次数的平方根，这正是中心极限定理（central limit theorem）（Section 12.24）所预测的随机漫步（random walk）行为。

动笔之前

想一想 若某数的小数部分 U 在 $[0,1)$ 上均匀分布（uniform），把该数截断时误差是多少？舍入时又是多少？两者的平均值（mean）和方差（variance）各是多少？

截断是以 $\lfloor a \rfloor$ 取代 a，误差（真确值减近似值）为 U，在 $[0,1)$ 上均匀分布：平均值为 $\frac{1}{2}$，方差为 $\frac{1}{12}$。舍入所得的误差在 $(-\frac{1}{2},\frac{1}{2}]$ 上均匀分布：平均值为 0，方差为 $\frac{1}{12}$。

逐步解题

阶段 1　总误差

加了 1500 个数之后，总误差是 1500 个独立（independent）误差之和，根据中心极限定理，它近似于正态分布（normal distribution）：

- LEP（舍入）：平均值为 0，方差为 $\frac{1500}{12} = 125$，标准差（standard deviation）为 $\sqrt{125} \approx 11.18$。
- VOZ（截断）：平均值为 $\frac{1500}{2} = 750$，标准差同样是 11.18。

阶段 2　出现大误差的概率

> **想一想**　对 LEP 而言，若 $E \sim \mathrm{N}(0, 125)$，$\mathrm{P}(|E| > 15)$ 是多少？对 VOZ 而言，误差几乎必然落在哪里？

对 LEP 而言，$\frac{15}{11.18} \approx 1.342$，所以

$$\mathrm{P}(|E| > 15) \approx 2\big(1 - \Phi(1.342)\big) \approx 2 \times 0.090 = 0.18.$$

对 VOZ 而言，误差约为 750 ± 11；其绝对值小于 15 的概率基本上是零，所以 $\mathrm{P}(|E| > 15) \approx 1$。 ■

阶段 3　可以做多少次加法？

> **想一想**　做了 n 次加法后，LEP 的误差的标准差为 $\sqrt{n/12}$。何时有 $\mathrm{P}(|E| < 10) \geq 0.9$？

我们需要 $\frac{10}{\sqrt{n/12}} \geq 1.645$（因为 $\mathrm{P}(|Z| < 1.645) = 0.9$），即 $\frac{n}{12} \leq \left(\frac{10}{1.645}\right)^2 \approx 36.95$，所以 $n \leq 443$。LEP 可以做大约 443 次加法。 ■

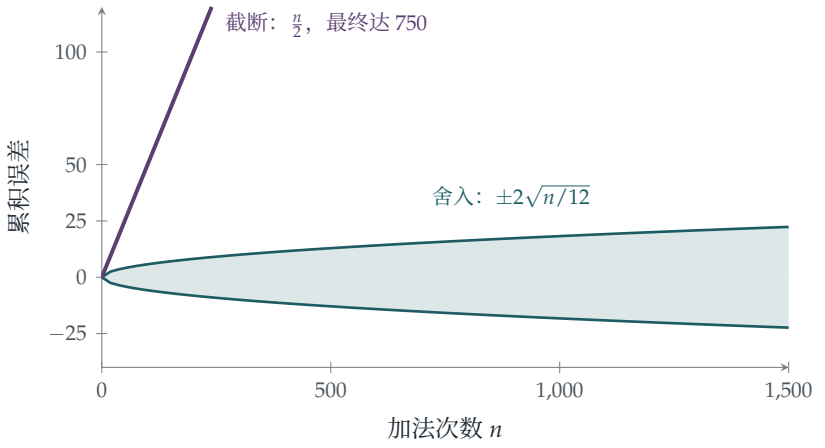


图 13.6.1 误差如何累积。截断误差（紫红色）全部同号，与 n 成正比增长：做了 1500 次加法后，总误差约为 750。舍入误差（蓝绿色带状区域，即零的两侧各两个标准差）平均而言互相抵销，增长速度只相当于 $\sqrt{n}$：做了 1500 次加法后，误差通常在约 ± 22 之内。

回顾

- **为每个误差建立模型。**小数部分均匀分布，令每个误差都有简单的分布。
- **把平均值和方差相加。**平均值可以相加（得出偏差 $\frac{n}{2}$），而对于独立的误差，方差同样可以相加。
- **正态近似。**中心极限定理把 1500 个均匀误差之和化为正态分布。

背后的数学

偏差与方差

截断是*有偏*（biased）的：其误差的平均值不是零，而偏差（bias）会线性累积。舍入是*无偏*（unbiased）的，其误差的表现有如随机漫步，增长速度相当于 $\sqrt{n}$。经过一百万次运算后，两者的分别就是数量级为 500 000 的误差与数量级为 300 的误差之别。这与实验中系统误差（systematic error，例如仪器校准失准）和随机测量误差（random error）之间的区别如出一辙：把多次测量结果取平均，可以减少随机误差，对系统误差却毫无作用。

浮点数

所有电脑都只能以有限的精确度储存实数，每次算术运算都要把结果舍入。现代电脑遵从 IEEE 754 标准，把结果舍入至最接近的可表示数；若刚好位于两数正中，则取末位为

偶数的一个，目的正是令舍入平均而言无偏。即使如此，把大量数字相加时若不小心处理，仍会损失准确度；William Kahan 的*补偿求和*（compensated summation）算法（1965 年）会记录失去的低位数字，再把它们加回去。1991 年波斯湾战争期间，一个爱国者导弹连未能拦截一枚来袭导弹，原因是它的时钟以十分之一秒为单位计时，而这个数无法以二进制准确储存；连续运作 100 小时后，截断误差已增至约三分之一秒，足以令目标位置的估算偏差超过半公里。

温哥华指数

温哥华证券交易所指数于 1982 年 1 月设立，起始值为 1000.000。到 1983 年 11 月，尽管大市普遍上升，指数却已跌至约 524。顾问发现，指数每天重新计算数千次，而每次都截断（而非舍入）至小数点后三位。经过一个周末的正确重算，指数重开时约为 1098。这件事已成为数值计算课程中的经典警示。

思维技巧

分清系统误差与随机误差

正负号一致的误差会线性累积；平均值为零的独立误差则按平方根的速度累积。先认清你面对的是哪一种。

在大学，数值分析（numerical analysis）研究误差如何在计算过程中传播，统计学则区分偏差与方差；“偏差–方差权衡”（bias–variance trade-off）是现代机器学习的核心概念。

延伸练习

1. **修正 VOZ。**若 VOZ 在最后把 $\frac{n}{2}$ 加到答案上，其误差的分布是甚么？它现在是否与 LEP 一样好？

2. **舍入至小数点后一位。**若把各数舍入至小数点后一位，而被舍去的数字均匀分布，求把 10 000 个数相加时总误差的标准差。

3. **温哥华指数。**若指数每天重新计算 3000 次，每次截断平均损失 0.0005，那么在一年 250 个交易日内，指数会损失多少？

4. **精确分布。**求两个舍入误差之和的精确密度函数（density，呈三角形）。若是三个误差之和，试绘出其草图。它多快便开始貌似正态分布？

延伸阅读

- **Nicholas Higham, *Accuracy and Stability of Numerical Algorithms***（SIAM）。关于舍入误差的权威著作；第一章浅白易读。
- **David Goldberg, "What every computer scientist should know about floating-point arithmetic"**, *ACM Computing Surveys* 23 (1991)。经典之作，可在网上免费阅读。
- **Tim Harford, *How to Make the World Add Up***（Bridge Street），探讨数据中的系统误差。

13.7 制造相关性

题目

STEP III 2010，第 13 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/10-s3-q13`

引子

一间银行想知道，手上的股票投资组合（portfolio）有多大机会亏蚀。股价往往一起变动：一只股票下跌，其他股票也倾向跟着下跌。要模拟数以千计的可能未来，银行必须产生一批随机数，而这些随机数之间的相关程度要恰到好处。如果手上只有互相独立（independent）的随机变量（random variable），怎样才能制造出具有指定相关系数（correlation）的随机变量呢？

这道题就三个变量的情形，一步一步给出了配方。每个新变量都是先前各个独立材料的组合，再加上一份全新的独立材料，用来补足方差（variance）。这个构造其实是线性代数（linear algebra）中一个著名结果的化身，即*科列斯基分解*（Cholesky decomposition），它同时揭示了一个隐藏的限制：并非每一组相关系数都有可能出现。

关于课程范围的说明

协方差（covariance）和相关系数已于 2019 年从 STEP 的考试范围中删去。协方差是 $\operatorname{Cov}(U,V) = \mathrm{E}(UV) - \mathrm{E}(U)\mathrm{E}(V)$；独立变量的协方差为零，而且协方差对每个变元都是线性的：$\operatorname{Cov}(aU+bV,W) = a\operatorname{Cov}(U,W) + b\operatorname{Cov}(V,W)$。此外 $\operatorname{Var}(U) = \operatorname{Cov}(U,U)$。相关系数必定介乎 -1 与 1 之间。

动笔之前

想一想　Z_1 和 Z_2 互相独立。它们的协方差是多少？相关系数又是多少？

独立变量满足 $\mathrm{E}(Z_1Z_2) = \mathrm{E}(Z_1)\,\mathrm{E}(Z_2)$。所以协方差是 0，而 $\mathrm{Corr}(Z_1, Z_2) = 0$。

逐步解题

阶段 1　两个变量

$\mathrm{E}(Y_2) = 0$。由于 Z_1 和 Z_2 互相独立，方差均为 1，

$$\mathrm{Var}(Y_2) = \rho_{12}^2 + (1-\rho_{12}^2) = 1, \qquad \mathrm{Cov}(Y_1, Y_2) = \mathrm{Cov}(Z_1, \rho_{12}Z_1 + \sqrt{1-\rho_{12}^2}\,Z_2) = \rho_{12}.$$

两个方差都是 1，所以 $\mathrm{Corr}(Y_1, Y_2) = \rho_{12}$。 ■

想一想　把 Z_1 和 Z_2 想像成互相垂直的单位向量。Y_2 在哪里？Y_1 与 Y_2 之间夹角的余弦是多少？

$Y_2 = \rho_{12}Z_1 + \sqrt{1-\rho_{12}^2}\,Z_2$ 是一个单位向量，与 $Z_1 = Y_1$ 的夹角为 θ，其中 $\cos\theta = \rho_{12}$。把变量看成向量，相关系数就是它们之间夹角的余弦（图 13.7.1）。

阶段 2　第三个变量

想一想　试以 a、b、c 写出 $\mathrm{Var}(Y_3) = 1$、$\mathrm{Cov}(Y_1, Y_3) = \rho_{13}$、$\mathrm{Cov}(Y_2, Y_3) = \rho_{23}$ 这三个条件。

利用线性和独立性：

$$a^2 + b^2 + c^2 = 1, \qquad a = \rho_{13}, \qquad \rho_{12}a + \sqrt{1-\rho_{12}^2}\,b = \rho_{23}.$$

所以

$$a = \rho_{13}, \qquad b = \frac{\rho_{23} - \rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}}, \qquad c = \sqrt{1 - \rho_{13}^2 - \frac{(\rho_{23} - \rho_{12}\rho_{13})^2}{1-\rho_{12}^2}}.$$ ■

（由于 $c \geq 0$，这里取非负平方根。）

阶段 3　任意的均值和方差

缩放和平移都不会改变相关系数，所以

$$X_i = \mu_i + \sigma_i Y_i \qquad (i = 1, 2, 3)$$

的均值为 μ_i，方差为 σ_i^2，相关系数为 ρ_{ij}。 ■

> **停一停，试一试** c 的公式要求平方根内的量非负。试取 $\rho_{12} = \rho_{13} = \rho_{23} = -0.6$。哪里出了问题？三个变量不可能两两之间都有很强的负相关，为甚么这是合理的？

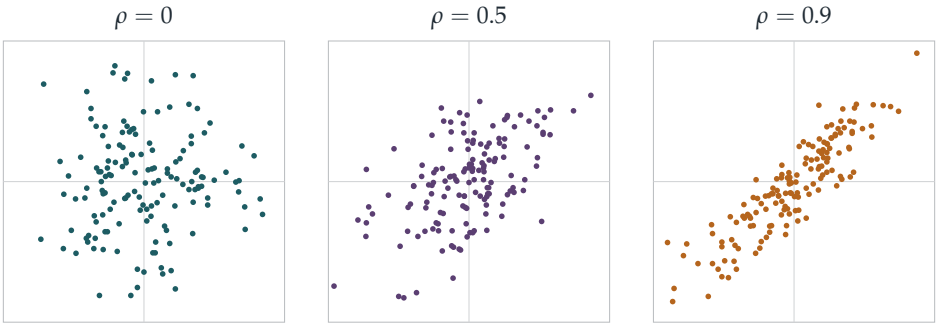


图 13.7.1 (Y_1, Y_2) 的样本，由独立的标准正态（standard normal）变量 Z_1, Z_2 按配方 $Y_1 = Z_1$, $Y_2 = \rho Z_1 + \sqrt{1-\rho^2}\, Z_2$ 制成，分别对应 $\rho = 0$、0.5 和 0.9。同样的独立材料，可以造出我们任意选定的相关系数。

回顾

- **每次只构造一个变量。**每个新变量都从先前的材料中取用适当的份量，再加入一份全新的独立材料。
- **协方差是线性的。**所有条件都化成了关于系数的简单方程。
- **相关系数是余弦。**把标准化后的变量看成单位向量，相关系数就变成了夹角。

背后的数学

科列斯基分解

用矩阵形式来写，这个构造就是 $\mathbf{Y} = L\mathbf{Z}$，其中的矩阵是下三角（lower triangular）矩阵

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \rho_{12} & \sqrt{1-\rho_{12}^2} & 0 \\ a & b & c \end{pmatrix},$$

而 $\mathbf{Y}$ 的相关矩阵（correlation matrix）是 $R = LL^{\mathsf{T}}$。把一个对称矩阵写成 LL^{T} 的形式，就是科列斯基分解，由法国军事工程师科列斯基（André-Louis Cholesky）约于 1910 年

发现，用来求解大地测量中的方程组（他在第一次世界大战中阵亡，这项成果在他身后才发表）。只要 R 是正定（positive definite）的，这个分解就行得通，而计算过程正是我们刚才用手做的：逐行求解。在金融、工程和气候模型中，这是模拟相关变量的标准方法。

哪些相关系数是可能的？

c 的平方根内的数必须非负。展开后，条件是

$$1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12}\rho_{13}\rho_{23} \geq 0,$$

这正是相关矩阵的行列式（determinant）非负的条件。从几何上看：三个单位向量两两之间的夹角为 θ_{ij}（其中 $\cos\theta_{ij} = \rho_{ij}$），它们必须真的能在空间中拼合起来。如果 $\rho_{12} = \rho_{13} = \rho_{23} = -0.6$，各个夹角都约为 127°，而三个两两相隔 127° 的向量，夹角之和就要超过 360°：这是不可能的。一般而言，一个矩阵要成为相关矩阵，必须是半正定（positive semi-definite）的，即所有特征值（eigenvalue）均非负（Section 7.2）。如果各个相关系数是根据不完整的数据分别估计出来的，这个条件就可能不成立，统计学家于是要找出最接近的有效相关矩阵，这个问题由 Nicholas Higham 在 2002 年解决。

高尔顿、皮尔逊与相关系数

高尔顿（Francis Galton）在 1888 年研究父母与子女身高的关系时引入了相关的概念，而皮尔逊（Karl Pearson）在 1896 年给出了本题所用的积矩（product moment）公式。高尔顿还注意到均值回归（regression to the mean）的现象（Section 13.1）。皮尔逊于 1911 年在伦敦大学学院（University College London）创立了全世界第一个大学统计学系。“相关不等于因果”这句警告，几乎与相关系数本身一样古老。

思维技巧

逐步构造，每次加入全新的独立材料

要构造具有指定关系的对象，可以逐个加入，每一个都是先前各个对象的组合，再加上一些全新而独立的东西。

这是线性代数中的格拉姆–施密特（Gram–Schmidt）正交化过程倒过来做，也就是科列斯基分解。在大学里，它会出现在最小二乘法（least squares）、多元正态分布（multivariate normal distribution）的理论，以及随机过程的模拟之中。

延伸练习

1 **不可能的相关系数。**证明：除非 $\rho \geq -\frac{1}{2}$，否则三个随机变量不可能两两之间的相关系数都是 ρ。对于 n 个变量，相应的结论是甚么？

2 **检查行列式。**证明 $1 - \rho_{13}^2 - b^2 \geq 0$ 等价于 $\det R \geq 0$，其中 R 是 3×3 相关矩阵。

3 **投资组合的风险。**两只股票回报的标准差（standard deviation）分别为 $\sigma_1 = 0.2$ 和 $\sigma_2 = 0.3$，相关系数为 ρ。若投资组合中两只股票各占一半，求 $\rho = -1$、0 和 1 时该组合的标准差。分散投资达到了甚么效果?

4 **还原材料。**给定题目中的 Y_1、Y_2，试以 Y_1 和 Y_2 表示 Z_2。（这就是格拉姆–施密特过程中的一步。）

延伸阅读

- **Gilbert Strang,** ***Introduction to Linear Algebra***（Wellesley-Cambridge），讲述正定矩阵和科列斯基分解。
- **Paul Glasserman,** ***Monte Carlo Methods in Financial Engineering***（Springer）。讲述实际应用中如何模拟相关变量；内容较深，但开头几章不难读。
- **Theodore Porter,** ***Karl Pearson: The Scientific Life in a Statistical Age***（Princeton University Press）。

我们学到了甚么

本章开首说过，概率（probability）是由模型走向数据，而统计（statistics）则是由数据回到现实世界。我们提出了三个问题：这是真有其事，还是纯属杂讯？应该怎样估计？随机性是怎样进入数据的？以下是本章七道题目所揭示的答案。

1. 是真有其事，还是纯属杂讯？

在 Section 13.1 中，四则新闻标题把数字变成了论断。比较总数而不理会分母、单凭一年或一日的数据就看出趋势、把趋势隐藏在平均数之中，这些全都是错误；而泊松模型（Poisson model）则提供了一把尺，用来量度计数数据的自然变异。上升了 8 个标准差（standard deviation），几乎肯定是真有其事；某个星期二有七人死亡，则几乎肯定不是。再加上均值回归（regression to the mean）和公平比较的必要，整幅图画便完整了。

2. 应该怎样估计？

有三道题目发展了估计的理论。

- **最大似然（maximum likelihood）**（Section 13.2）：选取令观察所得数据出现的概率最大的未知数值。在捉放法（capture–recapture）中，它给出了自然的估计值 $\frac{200\times 200}{11} \approx 3636$，而宽阔的似然函数则显示出这个估计有多不确定。
- **偏差（bias）与方差（variance）**（Section 13.3）：潜艇数目有两个无偏估计量（unbiased estimator），一个以最大值为基础，另一个以平均值为基础。最大值是充分统计量（sufficient statistic），精确得多，其方差按 $\frac{1}{n^2}$ 而非 $\frac{1}{n}$ 的速度缩小。同样的推理，在战时估计德军坦克数量一事上，胜过了情报部门。
- **效率（efficiency）与稳健性（robustness）**（Section 13.4）：对于均匀分布的数据，平均值的效率是中位数（median）的三倍；但对于重尾（heavy-tailed）数据，则是中位数胜出，而且它不受极端误差影响，十分稳健。

3. 随机性是怎样进入数据的？

随机性可以是刻意加入的：随机化回答（randomised response）（Section 13.5）既保障每位受访者的隐私，又仍然可以对总体作出估计，条件是每个人各自独立地进行随机化；假如所有人共用同一枚硬币，便会得出一个双峰的混合分布，几乎不能告诉我们任何东

西。随机性也会在处理数据的过程中进入：四舍五入的误差会互相抵销，只按 $\sqrt{n}$ 的速度增长，而截尾误差（truncation error）则会线性地累积，温哥华证券交易所（Vancouver Stock Exchange）便亲身领教过（Section 13.6）。随机性还可以经人手设计出来：把独立的成分以 Cholesky 构造法（Cholesky construction）组合起来，便可产生具有任何一组有效相关系数（correlation）的变量（Section 13.7），但并非每一组相关系数都是有效的。

4. 统计学的工具箱

方法	在哪里用过
泊松变异	新闻标题（13.1）
似然及其最大值	田鼠（13.2）
无偏性与方差	序号（13.3）；中位数（13.4）
充分统计量	序号（13.3）
正态近似（normal approximation）	13.4、13.5、13.6
全方差定律（law of total variance）	随机化回答（13.5）
偏差与随机误差	四舍五入（13.6）
协方差（covariance）的线性性质	相关系数（13.7）

下一步往哪里去

- 贝叶斯定理（Bayes' theorem）（Section 12.2）是贝叶斯统计学的基础，而贝叶斯统计学正是本章各种方法以外最主要的另一套方法。
- 中心极限定理（central limit theorem）（Section 12.24）为本章处处用到的正态近似提供了理据；柯西分布（Cauchy distribution）（Section 12.22）则显示这些近似在甚么时候失效。
- 相关矩阵（correlation matrix）是半正定矩阵（positive semi-definite matrix），可以用 Section 7.2 的特征向量（eigenvector）来研究。

如果你想在入读大学之前再走远一点

- **David Spiegelhalter, *The Art of Statistics***：为一般读者而写的统计思维入门书之中最好的一本。
- **Tim Harford, *How to Make the World Add Up***：教你怎样解读新闻中的数字。
- **John Rice, *Mathematical Statistics and Data Analysis***：大学数理统计的第一门课程。
- **Stephen Stigler, *The Seven Pillars of Statistical Wisdom***：统计学的重大思想及

其来源。

George Box 说过，所有模型都是错的，但有些是有用的。本章尝试说明怎样令模型变得有用：知道应该预期多大的变异，以方法在多次重复之下的表现来评价方法，并且思考数据是怎样得来的。正是这些习惯，把数字变成知识。

最后的话

全书到此结束。在这十三章里，我们以 STEP 题目为门径，走进大学一年级课程的数学：数论与分析、微积分与复数、代数与线性代数、微分方程与动力系统、组合数学、力学、概率与统计。每一道题目都是一份邀请，请你追问它背后藏着甚么，而答案几乎每一次都是一整门学科。如果你曾经手执铅笔，认真做过其中哪怕只是几节，思考页边提出的问题，并尝试每节末的习题，你就已经开始像数学家那样思考了。余下的旅程，就由你自己去走。

延伸练习答案

以下只是简短的答案和提示，并非完整解答：这里只给出结果和关键步骤，方便你核对自己的做法。每道练习都应先认真做过，然后才看答案。

13.1 解读新闻标题

1. $1 - \sum_{k=0}^{6} \mathrm{e}^{-2.96} \frac{2.96^k}{k!} \approx 0.0315$，所以一年之中大约有 11 天会出现七宗或以上的死亡个案。$\mathrm{P}(X \geq 9) \approx 0.0035$：大约一年一天。

2. 死亡人数为 $150K$，其中 K 服从平均值（mean）为 1.2 的泊松分布（Poisson distribution）：平均值为 180，方差（variance）为 $150^2 \times 1.2 = 27\,000$，标准差（standard deviation）约为 164。死亡人数为 125 或以下的年份，就是没有发生空难的年份，其概率为 $e^{-1.2} \approx 0.30$：完全是寻常事。但在这个模型中，死亡人数不可能恰好是 125，因为死亡人数只能是 150 的倍数。平均值和标准差都显示不出这一点：个案扎堆出现，令分布变得一团一团，全部概率都集中在 0、150、300 等数值上。

3. $z = \frac{1349.5-1080}{33} \approx 8.2$：概率的数量级为 10^{-16}。稳定不变的发生率无法解释 1350 这个数字；必定是有些事情真的改变了（又或者是统计死亡人数的方法改变了）。

4. 例如：(i)“单凭总数，无法判断乘飞机与步行哪样较安全”；(ii)“1990 年空难死亡人数属正常范围”；(iii)“星期二七名行人死亡：不幸的一天，但这样的日子一年约有十一次”；(iv)“1990 年行人死亡人数急升：必须追查原因”。

13.2 计数田鼠

1. $\frac{p_N}{p_{N-1}} = \frac{(N-M)(N-C)}{N(N-M-C+R)}$，而这个比值不小于 1，当且仅当 $NR \leq MC$。所以 p_N 一直递增至 $N = \frac{MC}{R}$，然后递减，而最大值就在该数的整数部分取得。

2. 从 a 名女性和 b 名男性中选出 c 人组成委员会，按其中女性的人数 j 来数：对每个 j，共有 $\binom{a}{j}\binom{b}{c-j}$ 种。

3. $\frac{\mathrm{d}}{\mathrm{d}p}\big(h \ln p + (n-h)\ln(1-p)\big) = \frac{h}{p} - \frac{n-h}{1-p} = 0$ 给出 $p = \frac{h}{n}$。

4. 当 $R = 0$ 时，p_N 永远递增，估计值是无限大。这个方法只有在第二个样本中有足够多已标记的动物时才行得通；重复捕获的数目少的话，估计值会非常不稳定。

13.3 德国坦克问题

1. 设 M 为最大值，则 $\mathrm{E}(M) = \frac{nN}{n+1}$ 及 $\mathrm{Var}(M) = \frac{nN^2}{(n+1)^2(n+2)}$，所以 $\mathrm{E}(M-N)^2 = \frac{2N^2}{(n+1)(n+2)}$。这是 $\mathrm{Var}(Z_1) = \frac{N^2}{n(n+2)}$ 的 $\frac{2n}{n+1}$ 倍：较差。最佳的倍数是 $c = \frac{\mathrm{NE}(M)}{\mathrm{E}(M^2)} = \frac{n+2}{n+1}$，其均方误差（mean squared error）为 $\frac{N^2}{(n+1)^2}$，比 Z_1 稍佳。

2. $\mathrm{E}(\min) = \frac{N}{n+1}$，而 min 的方差与 M 的相同，所以 $\mathrm{Var}\big((n+1)\min\big) = \frac{nN^2}{n+2}$。这个方差不会随 n 增大而缩小：是个差劣的估计量（estimator）。

3. 最大值不超过 m，当且仅当全部 k 个数都来自 $\{1,\dots,m\}$。当 $k=1$ 时，估计量是 $2X-1$，其平均值为 $2\cdot\frac{N+1}{2}-1 = N$。

4. $98 + \frac{98}{5} - 1 = 116.6$：大约 117 辆坦克。

13.4 平均数还是中位数?

1. 先选定哪一个值是中位数（median）（有 $2n+1$ 种选法），再选定其余各值中哪 n 个在它之下（有 $\binom{2n}{n}$ 种选法）。所求概率为 $(2n+1)\binom{2n}{n}y^n(1-y)^n\,\delta y = \frac{(2n+1)!}{(n!)^2}y^n(1-y)^n\,\delta y$。

2. $\mathrm{Var}(Y) = \frac{1}{4(2n+3)} = \frac{1}{20}$ 及 $\mathrm{Var}(\bar{X}) = \frac{1}{12(2n+1)} = \frac{1}{36}$。平均值较佳。

3. $\mathrm{Var}(\max) = \mathrm{Var}(\min) = \frac{m}{(m+1)^2(m+2)}$ 及 $\mathrm{Cov}(\max,\min) = \frac{1}{(m+1)^2(m+2)}$，由此得出 $\frac{1}{2(m+1)(m+2)}$。它缩小的速度与 $\frac{1}{m^2}$ 同阶，远快于平均值的 $\frac{1}{12m}$：对于均匀分布（uniform distribution）的数据，信息都在极端值之中，与坦克问题的情况一样。

4. 平均值约为 18.6，被一个打字错误完全破坏；中位数是 10.1，不受影响。

13.5 问尴尬的问题

1. 设 $t = ap + bq$，即答案属实的概率。回答“是”的概率为 $\frac{1}{2}t + \frac{1}{2}(1-t) = \frac{1}{2}$，所以 $\mathrm{P}(\text{正面} \mid \text{是}) = t$，而它等于 $\frac{1}{2}$，当且仅当 $2t = 1$。若由同一名孪生儿回答两次，则不论硬币掷出的是正面还是反面，先答“是”后答“否”的概率都是 $pa(1-a)+qb(1-b)$，所以后验概率（posterior probability）是 $\frac{1}{2}$。两次都答“是”的概率分别为 pa^2+qb^2（正面）和 $p(1-a)^2+q(1-b)^2$（反面），两者相等当且仅当 $2(ap+bq)=1$。

2. $\mathrm{P}(\text{是}) = \frac{3}{4}\pi + \frac{1}{4}(1-\pi)$，所以 $\hat{\pi} = \frac{0.4-0.25}{0.5} = 0.3$。标准误（standard error）为 $\frac{\sqrt{0.4\times0.6/400}}{0.5} \approx 0.049$。

3. $\mathrm{E}(Y^2) = \sum_c \mathrm{P}(C=c)\big(\mathrm{Var}(Y \mid c) + \mathrm{E}(Y \mid c)^2\big)$；减去 $\mathrm{E}(Y)^2 = \big(\sum_c \mathrm{P}(C=c)\mathrm{E}(Y \mid c)\big)^2$ 便得 $\mathrm{E}(\mathrm{Var}(Y \mid C)) + \mathrm{Var}(\mathrm{E}(Y \mid C))$。

4. Z_2 接近 $\frac{n}{6}$（正面）或 $\frac{5n}{6}$（反面），两种情况的概率各为 $\frac{1}{2}$。所以 $\mathrm{P}(Z_2 > \frac{n}{2}) \to \frac{1}{2}$，尽管 Z_2 几乎从不接近 $\frac{n}{2}$。

13.6 累积起来的误差

1. 每个截尾误差(truncation error)都在 $(-1,0]$ 上均匀分布;各加上 $\frac{1}{2}$ 后,便在 $(-\frac{1}{2},\frac{1}{2}]$ 上均匀分布,与四舍五入(rounding)的情况完全一样。所以修正后的 VOZ 与 LEP 一样好。

2. 每个误差都在 $(-0.05,0.05)$ 上均匀分布,方差为 $\frac{0.1^2}{12}$,所以总和的标准差为 $\sqrt{10\,000\times 0.01/12}\approx 2.9$。

3. $3000\times 250\times 0.0005=375$ 点。

4. 两个误差时,是 $[-1,1]$ 上的 $1-|x|$。三个误差时:当 $|x|\le\frac{1}{2}$ 时为 $\frac{3}{4}-x^2$,当 $\frac{1}{2}\le|x|\le\frac{3}{2}$ 时为 $\frac{1}{2}\left(\frac{3}{2}-|x|\right)^2$,已经很接近正态曲线(normal curve)。

13.7 制造相关性

1. 对于标准化变量(standardised variables),$\operatorname{Var}(X_1+X_2+X_3)=3+6\rho\ge 0$,所以 $\rho\ge-\frac{1}{2}$。对于 n 个变量,$n+n(n-1)\rho\ge 0$ 给出 $\rho\ge-\frac{1}{n-1}$。

2. 取 $b=\frac{\rho_{23}-\rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}}$,则 $(1-\rho_{12}^2)(1-\rho_{13}^2-b^2)=1-\rho_{12}^2-\rho_{13}^2-\rho_{23}^2+2\rho_{12}\rho_{13}\rho_{23}=\det R$。由于 $1-\rho_{12}^2>0$,这两个条件等价。

3. $\operatorname{Var}=\frac{1}{4}(0.04+0.09+0.12\rho)$:当 $\rho=-1$、0、1 时,标准差分别为 0.05、0.18 和 0.25。除非两只股票完全相关(perfectly correlated),否则这个组合的风险低于两者风险的平均值(0.25)。

4. $Z_2=\dfrac{Y_2-\rho_{12}Y_1}{\sqrt{1-\rho_{12}^2}}$。

Beyond STEP

STEP 是大多数中学生一生中应考过最艰深的数学考试，也是最出色的考试之一：题目简短、陌生而深刻，当中最精彩的，其实是乔装了的大学数学小品。

本书精选最有趣的 STEP 试题，由 1987 年首次开考之前的样本试卷（specimen papers），到 2025 年的试卷，并以每一道题目作为一扇门。每道题目都会仔细解答，沿途提出优秀学生会问的问题，然后带你穿过这扇门：这个想法从何而来，到了大学会演变成甚么，以及接下来该读些甚么。

本书为正在准备 STEP、有天分又有志投身数学的学生而写。他们想要的不止是模拟试卷：他们想开始像数学系本科生那样思考，并一探前方有甚么正等着自己。

- 共 13 章，涵盖纯数学（pure mathematics）、力学（mechanics）、概率（probability）与统计（statistics）
- 136 道 STEP 历届试题，每题均附引导式解答
- 每道题目背后的大学数学
- 附答案的延伸练习，以及延伸阅读

可免费阅读、下载及分享，网址为

beyondstep.keithmcnulty.org

文字内容已贡献予公有领域（CC0 1.0）。STEP 试题 © Cambridge University Press & Assessment。

