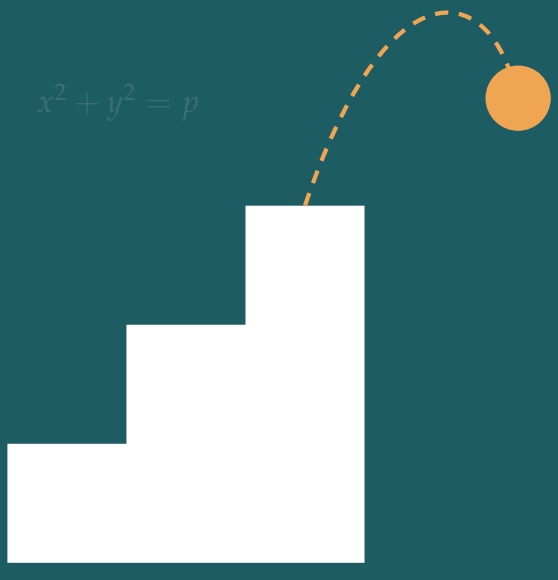
STEP 备试 · 大学数学

Beyond STEP

STEP I, II 及 III

2000 年详细解答



KEITH McNULTY

Beyond STEP

以历届 STEP 试题为门
走进大学数学

STEP I, II 及 III

2000 年详细解答

Keith McNulty

Beyond STEP：以历届 STEP 试题为门，走进大学数学

2000 年详细解答

2000 年 STEP 试卷（STEP I, II 及 III）全部 42 题的详细解答。自样本试卷以来每一道 STEP 题目的详细解答、*Beyond STEP* 全书以及各章的单行本，均可在 `beyondstep.keithmcnulty.org` 免费阅读及下载。

阅读解答之前，请先认真尝试解题。每个答案均已用数值方法核对，但解答数量如此之多，难免会有疏漏：如果发现有可疑之处，请相信自己的演算，并仔细检查。

由机器从英文翻译而成；尚未经审阅。

2026 年 11 月首次于网上发表。

本 PDF 内凡属作者本人创作的内容，包括正文、解答、插图、练习及答案，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、改编及分享，毋须征求许可。详见 `creativecommons.org/publicdomain/zero/1.0`。

本书讨论的 STEP 试题，版权属 Cambridge University Press & Assessment 所有。本书并无转载这些试题：每一题均附连结，通往官方来源的题目，即 STEP Support Programme 的试题资料库或 OCR 的历届试卷。

如有错漏，欢迎到 github.com/keithmcnulty/step_book_mcnulty 提交 issue 指正。

目录

STEP I

第 1 题 对数与首位数字 纯数学 . . . 3

第 2 题 二项展开式的系数 纯数学 . . . 6

第 3 题 阶梯函数的积分 纯数学 . . . 10

第 4 题 积分的界限 纯数学 . . . 13

第 5 题 三维向量 纯数学 . . . 16

第 6 题 直线所围成的区域 纯数学 . . . 19

第 7 题 导数的界限 纯数学 . . . 23

第 8 题 绝对值的积分 纯数学 . . . 26

第 9 题 加速参考系中的抛体 力学 . . . 30

第 10 题 碰撞与恢复系数 力学 . . . 34

第 11 题 梯子的摩擦力平衡 力学 . . . 37

第 12 题 几何分布 概率与统计 . . . 41

第 13 题 条件概率 概率与统计 . . . 44

第 14 题 均匀分布的矩 概率与统计 . . . 47

STEP II

第 1 题 埃及分数 纯数学 . . . 50

第 2 题 多项式的重根 纯数学 . . . 54

第 3 题 误差传递 纯数学 . . . 57

第 4 题 福角与反正切恒等式 纯数学 . . . 61

第 5 题 最小二乘近似 纯数学 . . . 65

第 6 题 半角代换 纯数学 . . . 69

第 7 题 直线与圆锥 纯数学 . . . 73

第 8 题 可分离变量微分方程 纯数学 . . . 76

第 9 题 有空气阻力的运动 力学 79
第 10 题 连接滑轮系统 力学 83
第 11 题 弹性绳与简谐运动 力学 87
第 12 题 随机变量之和 概率与统计 91
第 13 题 捕捉再捕捉估计 概率与统计 95
第 14 题 顺位数的分布 概率与统计 98

STEP III

第 1 题 双曲线的切线 纯数学 102
第 2 题 代换积分法 纯数学 107
第 3 题 复数与几何 纯数学 110
第 4 题 有理函数的描绘 纯数学 114
第 5 题 线性变换 纯数学 119
第 6 题 四次多项式的因式分解 纯数学 123
第 7 题 不等式与极限 纯数学 127
第 8 题 递推关系 纯数学 130
第 9 题 弹性斜碰撞 力学 134
第 10 题 极限摩擦力平衡 力学 138
第 11 题 绕粗糙支点的转动 力学 142
第 12 题 泊松近似 概率与统计 146
第 13 题 概率生成函数 概率与统计 150
第 14 题 方差与独立性 概率与统计 154

详细解答

STEP I 2000，第 1 题

纯数学 · 对数与首位数字

> **STEP I 2000，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/00-s1-q1

关键思路

一个大数 N 的首位数字（leading digit）完全取决于 $\log_{10} N$ 的小数部分（fractional part）。若 $\log_{10} N = k + f$，其中 k 为整数而 $0 \le f < 1$，则 $N = 10^f \times 10^k$，而 N 的首位数字是 d，当且仅当 $\log_{10} d \le f < \log_{10}(d+1)$。因此我们只需要数字 1 至 9 的对数（logarithm），它们都可由 $\log_{10} 2$ 和 $\log_{10} 3$ 求得；此外还要留心我们能够容许多大的舍入误差（rounding error）。

解答

以下 log 一律表示 $\log_{10}$。区间 $[0,1)$ 分成九段，每段对应一个可能的首位数字：



小数部分 f 落在标示为 d 的一段内，首位数字就是 d。

第 (i) 部分

由于 $5 = 10/2$ 及 $6 = 2 \times 3$，

$$\log 5 = 1 - \log 2 = 1 - 0.301029996 = 0.698970004 \approx 0.699,$$
$$\log 6 = \log 2 + \log 3 = 0.301029996 + 0.477121255 = 0.778151251 \approx 0.778,$$

两者均准确至小数点后三位。

现在

$$\log 3^{100} = 100 \log 3 = 47.7121255.$$

题目所给的 $\log 3$ 的值，误差不超过 5×10^{-10}，所以上式的误差不超过 5×10^{-8}；必定有 $47.712 < \log 3^{100} < 47.713$。与

$$\log\left(5 \times 10^{47}\right) = 47 + \log 5 \approx 47.699 \quad \text{及} \quad \log\left(6 \times 10^{47}\right) = 47 + \log 6 \approx 47.778,$$

比较，可得

$$\log\left(5 \times 10^{47}\right) < \log 3^{100} < \log\left(6 \times 10^{47}\right).$$

两边的余裕（约 0.013 和 0.066）远远大于任何舍入误差，而 log 是严格递增（strictly increasing）的，所以

$$5 \times 10^{47} < 3^{100} < 6 \times 10^{47},$$

正如所求。因此 3^{100} 是 $5\ldots. \times 10^{47}$：它的首位数字是 $\boxed{5}$（而且它共有 48 位数字）。

第 (ii) 部分

我们沿用同一个想法，取 $\log 2 = 0.301029996$，其误差至多为 5×10^{-10}。乘以 10^n 会把这个误差也乘以 10^n，所以我们必须检查小数部分是否仍然稳妥地留在数线上的同一段之内。

2^{1000}. $\log 2^{1000} = 301.029996$, 误差至多为 5×10^{-7}。小数部分约为 0.030, 介乎 $\log 1 = 0$ 与 $\log 2 \approx 0.301$ 之间，所以 $2^{1000} = 1\ldots. \times 10^{301}$，其首位数字是 $\boxed{1}$。

$2^{10\,000}$. $\log 2^{10\,000} = 3010.29996$，误差至多为 5×10^{-6}。小数部分介乎 0.299955 与 0.299965 之间。

> **想一想** 小数部分 0.29996 非常接近 $\log 2 = 0.30103$。我们能肯定首位数字不是 2 吗?

能够: 两者相差 $0.30103 - 0.29996 \approx 0.0011$, 是最大可能误差 5×10^{-6} 的二百多倍。所以小数部分确实小于 $\log 2$，而 $10^{3010} < 2^{10\,000} < 2 \times 10^{3010}$。（等价地，$2^{10\,000} < 2 \times 10^{3010}$ 即是说 $2^{9999} < 10^{3010}$，亦即 $\log 2 < 3010/9999 = 0.3010301\ldots$，而这是成立的。）首位数字是 $\boxed{1}$。

$2^{100\,000}$. $\log 2^{100\,000} = 30102.9996$，误差至多为 5×10^{-5}，所以小数部分介乎 0.99955 与 0.99965 之间。我们需要知道它落在数线上的哪一段，而最顶的一段始于

$$\log 9 = 2 \log 3 = 0.954242510.$$

由于 $0.954 < 0.99955$ 及 $0.99965 < 1$，我们得到 $9 \times 10^{30102} < 2^{100\,000} < 10^{30103}$，所以首位数字是 $\boxed{9}$。

想一想 误差会不会把小数部分推过 1，令这个数变成 $1.\ldots \times 10^{30103}$?

除非 $100\,000 \log 2$ 的真确值至少是 30103 才会如此，而这需要它有 0.0004 的误差，是最大误差 5×10^{-5} 的八倍。所以首位数字是 9，而 $2^{100\,000}$ 共有 30 103 位数字。

验算

用电脑作精确的整数运算，可证实以上所有结果：3^{100} 有 48 位数字，开首为 $515\ldots$；2^{1000} 开首为 $107\ldots$（共 302 位数字）；$2^{10\,000}$ 开首为 $199\ldots$（共 3011 位数字）；而 $2^{100\,000}$ 开首为 $999\ldots$（共 30 103 位数字）。再用一个简单例子检验这个方法是否合理：$\log 2^{10} = 3.0103$ 的小数部分为 $0.0103 < \log 2$，与 $2^{10} = 1024$ 吻合。

常见错误

- 从对数的整数部分（integer part）读出首位数字。整数部分只能告诉我们位数（位数比它多一）；首位数字藏在小数部分之中。
- 忘记乘以 100 000 时，$\log 2$ 的舍入误差也会同样被放大。在这里九位小数仅仅足够，而且值得说明原因。
- 看见 $f = 0.9996 \approx 1$ 便断定首位数字是 1。小数部分略小于 1，表示这个数略小于十的某个幂，所以它以 9 开首。

STEP 之外

2^n 的首位数字由 $n \log 2$ 的小数部分决定。由于 $\log 2$ 是无理数（irrational），这些小数部分在 $[0,1)$ 内是等分布（equidistributed）的（外尔（Weyl）等分布定理），所以在 2 的幂之中，以数字 d 开首的所占比例趋向 $\log(1 + 1/d)$：这就是本福特定律（Benford's law），而大约 30% 的 2 的幂以 1 开首。至于无理性与逼近这方面的故事，达文波特（Davenport）的 *The Higher Arithmetic* 是一本十分值得一读的好书。

详细解答

STEP I 2000，第 2 题

纯数学 · 二项展开式的系数

> **STEP I 2000，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/00-s1-q2

关键思路

把两个带分式项的二项式(binomial)直接展开,再逐一寻找相配的幂,过程相当繁琐。较简洁的做法是先把 x 的负数幂抽出来，令每个括号都变成只含 x 的单一种幂（x^6 或 x^2）的多项式(polynomial)。这样一来,凭着因式分解(factorisation)$x^6-1=(x^2-1)(x^4+x^2+1)$，便会发现第二个乘积其实就是第一个乘积的化身，所以题目中“由此”（hence）的部分几乎不费工夫。

解答

第一个展开式

从每个括号中抽出 x 的最低次幂，

$$\left(x^4-\frac{1}{x^2}\right)^5=\left(\frac{x^6-1}{x^2}\right)^5=x^{-10}\,(x^6-1)^5,$$

$$\left(x-\frac{1}{x}\right)^6=\left(\frac{x^2-1}{x}\right)^6=x^{-6}\,(x^2-1)^6.$$

因此整个算式等于 $x^{-16}P(x)$，其中

$$P(x)=(x^6-1)^5\,(x^2-1)^6.$$

所以原展开式中 x^n 的系数（coefficient），就是多项式 $P(x)$ 中 x^{n+16} 的系数。

根据二项式定理（binomial theorem），

$$(x^6-1)^5=\sum_{j=0}^{5}\binom{5}{j}(-1)^{5-j}x^{6j},\qquad (x^2-1)^6=\sum_{k=0}^{6}\binom{6}{k}(-1)^{6-k}x^{2k},$$

故此 $P(x)$ 中 x^m 的系数，就是

$$\binom{5}{j}(-1)^{5-j}\binom{6}{k}(-1)^{6-k}$$

对所有符合 $6j+2k=m$、$0\le j\le 5$、$0\le k\le 6$ 的数对求和的结果。

x^{-12} 的系数。 我们需要 $m=-12+16=4$，即 $6j+2k=4$。由于 $6j\le 4$，必有 $j=0$，所以唯一的数对是 $(j,k)=(0,2)$，得出

$$\binom{5}{0}(-1)^5\cdot\binom{6}{2}(-1)^4=(-1)(15)=-15,$$

正如题目所求。

x^2 的系数。 这次 $m=2+16=18$，所以 $6j+2k=18$，即 $3j+k=9$，而 $k\le 6$。这容许 $j=1,2,3$：

(j,k)	来自 $(x^6-1)^5$	来自 $(x^2-1)^6$	乘积
$(3,0)$	$\binom{5}{3}(-1)^2=10$	$\binom{6}{0}(-1)^6=1$	10
$(2,3)$	$\binom{5}{2}(-1)^3=-10$	$\binom{6}{3}(-1)^3=-20$	200
$(1,6)$	$\binom{5}{1}(-1)^4=5$	$\binom{6}{6}(-1)^0=1$	5

所以 x^2 的系数是 $10+200+5=215$。

> ? **想一想** 为甚么 $j=0$ 或 $j=4$ 不可能对 x^{18} 有贡献?

若 $j=0$，便需要 $k=9$；若 $j=4$，便需要 $k=-3$；两者都不在 $0\le k\le 6$ 的范围内。在这一步很容易漏掉某个数对（或者凭空多出一个），所以有系统地列出所有可能情况是值得的。

第二个展开式

> ? **想一想** x^4+x^2+1 与第一部分有甚么关系?

把 x^6-1 看成以 x^2 表示的立方差（difference of cubes）时，它就是另一个因式：

$$x^6-1=(x^2)^3-1=(x^2-1)(x^4+x^2+1).$$

把 $(x^2-1)^{11}$ 拆成 $(x^2-1)^5(x^2-1)^6$，

$$(x^2-1)^{11}(x^4+x^2+1)^5=\left[(x^2-1)(x^4+x^2+1)\right]^5(x^2-1)^6=(x^6-1)^5(x^2-1)^6=P(x).$$

所以第二个展开式正正就是 $P(x)$，即第一个展开式乘以 x^{16}。

x^4 的系数。 这就是 $P(x)$ 中 x^4 的系数，我们在上面已经求得（它对应第一个展开式中的 x^{-12}）。答案是 -15。

x^{38} 的系数。 这对应 $6j+2k=38$，即 $3j+k=19$，而 $j\le 5$、$k\le 6$。由于 $k\le 6$，我们需要 $3j\ge 13$，所以 $j=5$、$k=4$，得出

$$\binom{5}{5}(-1)^0\cdot\binom{6}{4}(-1)^2=15.$$

所以 x^{38} 的系数是 15。

其实有一个巧妙的方法，不用计数也能看出这个结果。P 的次数（degree）是 $30+12=42$，而把 x 换成 $1/x$ 可得

$$x^{42}P(1/x)=(1-x^6)^5(1-x^2)^6=(-1)^{11}P(x)=-P(x).$$

比较系数可知，P 中 x^{42-m} 的系数等于 x^m 的系数的负数。取 $m=4$，x^{38} 的系数就是 $-(-15)=15$，与上面的结果吻合。

验算

- 把 $x=1$ 代入 $P(x)$：其值为零，所以 P 的所有系数之和是 0。如果你把它完全展开，这是一个快捷的合理性检查。
- 我们用电脑把两个乘积精确地乘开（以系数列表的形式）。在第一个乘积中，x^{-12} 的系数是 -15，x^2 的系数是 215；在第二个乘积中，x^4 的系数是 -15，x^{38} 的系数是 15。我们亦确认了第二个乘积的次数是 42，并且等于第一个乘积乘以 x^{16}。
- 把随机选取的有理数值代入原式中的 x，所得结果与计算出来的展开式完全吻合。

常见错误

- 正负号出错：若选取的幂是 x^{6j} 和 x^{2k}，该项的正负号是 $(-1)^{5-j}(-1)^{6-k}=-(-1)^{j+k}$，而不是 $(-1)^{j+k}$。用错了的话，每个答案的正负号都会颠倒。
- 计算 x^2 的系数时漏掉 $(1,6)$ 或 $(3,0)$ 这两项，结果得出 200 或 205。
- 没有留意到因式分解，直接展开 $(x^2-1)^{11}(x^4+x^2+1)^5$：这样做是可行的，但冗长而且容易出错。

STEP 之外

把多个和式的乘积写成单一个和式，再从中抽出 x 的某一个幂，正是生成函数（generating function）背后的想法。乘积中的某个系数，数算的是由各部分凑出某个总数的方法数目，与我们在上面计数数对 (j,k) 的做法如出一辙。对称关系 $x^{42}P(1/x) = -P(x)$ 说明 P 是一个（反）回文多项式（(anti)palindromic polynomial）；这类多项式会在分圆多项式（cyclotomic polynomial）和单位根（roots of unity）的研究中出现，而 $x^6 - 1$ 正是其中最初步的例子。想做更多同类风格的题目，Siklos 的 *Advanced Problems in Mathematics* 是很好的下一步；而 Allenby 的 *Rings, Fields and Groups* 则对多项式的因式分解有完整而严谨的论述。

详细解答

STEP I 2000，第 3 题

纯数学 · 阶梯函数的积分

STEP I 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-s1-q3

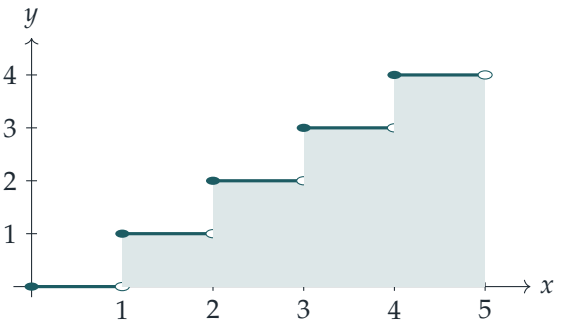
关键思路

任何量的整数部分（integer part）都是一个阶梯函数（step function）：它在各个区间上保持不变，然后向上跳升一。因此这个积分不过是若干个长方形面积之和，真正要做的工夫只是找出每一级由哪里开始、到哪里结束。对于 $[\mathrm{e}^x]$，跳跃发生在 e^x 越过整数之处，即 $x = \ln k$，而由此得出的对数之和恰好可以逐项相消（telescope）。

解答

$y = [x]$ 的图像

在每个区间 $k \le x < k+1$（其中 k 为整数）上，我们有 $[x] = k$。所以图像是由水平线段组成的阶梯，每段长度为 1，左端是实心点（函数取到该值），右端是空心点（函数在该处向上跳升 1）。



积分就是阴影部分的面积，由阔度为 1、高度分别为 $0, 1, 2, 3, 4$ 的长方形组成（函数在

$x = 1, 2, 3, 4$ 这几个单独的点上的取值对面积毫无影响)。因此

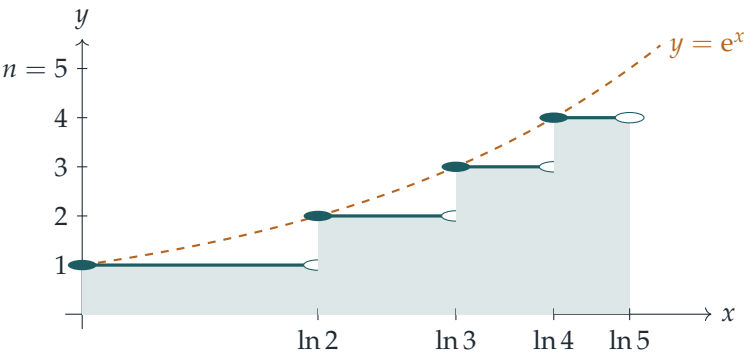
$$\int_0^5 [x]\,\mathrm{d}x = 0 + 1 + 2 + 3 + 4 = 10.$$

$y = [\mathrm{e}^x]$ 的图像

当 $0 \leq x < \ln n$ 时，我们有 $1 \leq \mathrm{e}^x < n$。由于 e^x 是递增的，

$$[\mathrm{e}^x] = k \iff k \leq \mathrm{e}^x < k+1 \iff \ln k \leq x < \ln(k+1),$$

其中 $k = 1, 2, \ldots, n-1$。所以图像仍然是每级高度为 1 的阶梯，但现在各级随着 x 增大而愈来愈短：高度为 k 的那一级阔度为 $\ln(k+1) - \ln k = \ln\left(1 + \frac{1}{k}\right)$。下面的草图所示为 $n = 5$ 的情况，曲线 $y = \mathrm{e}^x$ 以虚线表示；阶梯在每一级的左端与曲线相接。



积分

把各长方形的面积相加，即高度 k 乘以阔度 $\ln(k+1) - \ln k$，

$$\int_0^{\ln n} [\mathrm{e}^x]\,\mathrm{d}x = \sum_{k=1}^{n-1} k\big(\ln(k+1) - \ln k\big) = \sum_{k=1}^{n-1} k\ln(k+1) - \sum_{k=1}^{n-1} k\ln k.$$

想一想 怎样才能令这两个和互相抵消?

每个 $\ln j$ 在两个和中都会出现，只是系数不同。在第一个和中代入 $j = k+1$，便可把两者对齐：

$$\begin{aligned}\sum_{k=1}^{n-1} k\ln(k+1) - \sum_{k=1}^{n-1} k\ln k &= \sum_{j=2}^{n}(j-1)\ln j - \sum_{j=1}^{n-1} j\ln j \\ &= (n-1)\ln n + \sum_{j=2}^{n-1}\big((j-1) - j\big)\ln j - 1\cdot\ln 1 \\ &= (n-1)\ln n - \sum_{j=2}^{n-1}\ln j.\end{aligned}$$

这里我们从第一个和中抽出 $j=n$ 的一项，并从第二个和中抽出 $j=1$ 的一项（该项为零，因为 $\ln 1=0$）。现在 $\sum_{j=2}^{n-1}\ln j=\ln\big((n-1)!\big)$，而 $(n-1)\ln n=n\ln n-\ln n$，所以

$$\int_0^{\ln n}[\mathrm{e}^x]\,\mathrm{d}x=n\ln n-\big(\ln n+\ln((n-1)!)\big)=n\ln n-\ln(n!),$$

正是所求。（当 $n=1$ 时，积分区间的长度为零，等式两边都是 0。）

图像证明（选读）

阔度为 $\ln n$、高度为 n 的长方形，面积是 $n\ln n$。它位于阶梯上方的部分，由高度为 1 的水平长条组成：在高度 $m-1$ 与 m 之间，其中 $m=1,\ldots,n$，长条由 $x=0$ 伸延至 $x=\ln m$，因为 $[\mathrm{e}^x]<m$ 恰好在 $x<\ln m$ 时成立。这长条的面积是 $\ln m$，所以阶梯上方区域的面积为 $\ln 1+\ln 2+\cdots+\ln n=\ln(n!)$。从 $n\ln n$ 中减去它，便再次得到所求结果。这就是“改为对 y 而不是对 x 积分”的离散版本。

验算

当 $n=2$ 时，阶梯只是一个高度为 1、阔度为 $\ln 2$ 的长方形，而公式给出 $2\ln 2-\ln 2=\ln 2$。当 $n=4$ 时，公式给出 $4\ln 4-\ln 24\approx 2.3671$；以中点法则（midpoint rule）计算 $\int_0^{\ln 4}[\mathrm{e}^x]\,\mathrm{d}x$，结果与之吻合至小数点后五位，而对于直至 8 的每一个 n 亦然。答案还必须小于 $\int_0^{\ln n}\mathrm{e}^x\,\mathrm{d}x=n-1$，因为阶梯位于曲线之下：当 $n=4$ 时，$2.37<3$。

常见错误

- 把高度为 k 的那一级的阔度当作 1（照搬第一部分），而不是 $\ln(k+1)-\ln k$。
- 让阶梯由高度 0 开始：在 $0\le x<\ln n$ 上我们有 $\mathrm{e}^x\ge 1$，所以最低的一级位于高度 1。
- 移动求和指标时出错，漏掉了 $(n-1)\ln n$ 这一项；或者没有留意到它必须与 $\ln((n-1)!)$ 合并，才能得出 $\ln(n!)$。

STEP 之外

由于阶梯位于 $y=\mathrm{e}^x$ 之下，积分小于 $n-1$，整理后即得 $\ln(n!)>n\ln n-n+1$。以同样的方法从上方给阶梯定界，便会通往斯特灵公式（Stirling's formula）$\ln(n!)\approx n\ln n-n$，这是分析学、概率论和组合数学中的一个关键估计。由长方形之和转换为长条之和，也是“分部求和法”（summation by parts）的一个例子，它是分部积分法（integration by parts）在离散情况下的近亲。若想为阶梯函数的积分建立严格的基础，Abbott 的 *Understanding Analysis*（分析学入门读本）是下一步的好选择。

详细解答

STEP I 2000，第 4 题

纯数学 · 积分的界限

> **STEP I 2000，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/00-s1-q4

关键思路

用初等方法难以直接对 $1/(x^2+1)^4$ 积分，所以题目把它拆成两部分：一部分恰好是某个函数的导数（exact derivative），积分时不费吹灰之力；另一部分是很小的余项。第 (ii) 部分负责建立这个分拆，第 (i) 部分则提供余项的上界。要估计一个积分的界（bound），通常的做法是先估计被积函数（integrand）的界，再乘以区间的长度。

解答

第 (i) 部分

设 $f(x)=\dfrac{x^6}{(x^2+1)^4}$。由于 f 只取决于 x^2，令 $t=x^2$，则 t 与 x 一样在 $[0,1]$ 上取值，并且 $f=g(t)$，其中

$$g(t)=\frac{t^3}{(1+t)^4}.$$

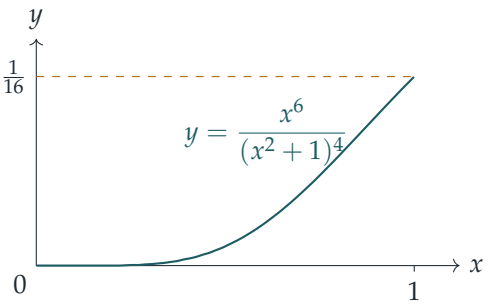
由商法则（quotient rule），

$$g'(t)=\frac{3t^2(1+t)^4-4t^3(1+t)^3}{(1+t)^8}=\frac{t^2\big(3(1+t)-4t\big)}{(1+t)^5}=\frac{t^2(3-t)}{(1+t)^5}.$$

当 $0<t\le 1$ 时，每个因子都是正数（特别地，$3-t\ge 2$），所以 $g'(t)>0$，而 g 在 $[0,1]$ 上严格递增（strictly increasing）。因此它的最大值在 $t=1$ 处取得，即 $x=1$：

$$f(1)=\frac{1}{2^4}=\frac{1}{16},$$

这正是所要证明的。此外，处处都有 $f(x)\ge 0$，而最小值是 $f(0)=0$。



第 (ii) 部分

用商法则对括号内的式子求导。设 $N = Ax^5 + Bx^3 + Cx$，分母为 $(x^2+1)^3$，其导数为 $6x(x^2+1)^2$，于是可以约去一个因子 $(x^2+1)^2$：

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{N}{(x^2+1)^3}\right) = \frac{(5Ax^4+3Bx^2+C)(x^2+1) - 6x(Ax^5+Bx^3+Cx)}{(x^2+1)^4}.$$

展开分子，并按 x 的幂次整理：

$$\begin{aligned} &5Ax^6 + 5Ax^4 + 3Bx^4 + 3Bx^2 + Cx^2 + C - 6Ax^6 - 6Bx^4 - 6Cx^2 \\ &\quad = -Ax^6 + (5A-3B)x^4 + (3B-5C)x^2 + C. \end{aligned}$$

现在各项的分母都是 $(x^2+1)^4$，所以恒等式（identity）对所有 x 成立，当且仅当两边的分子作为多项式相等：

$$(D-A)x^6 + (5A-3B)x^4 + (3B-5C)x^2 + C = 1.$$

由低次项起逐项比较系数（comparing coefficients）：$C = 1$；然后 $3B = 5C = 5$；然后 $5A = 3B = 5$；最后 $D = A$。所以

$$A = 1, \qquad B = \tfrac{5}{3}, \qquad C = 1, \qquad D = 1.$$

第 (iii) 部分

把第 (ii) 部分的恒等式在 $[0,1]$ 上积分。右边第一项恰好是一个导数，所以

$$\begin{aligned} \int_0^1 \frac{\mathrm{d}x}{(x^2+1)^4} &= \left[\frac{x^5 + \frac{5}{3}x^3 + x}{(x^2+1)^3}\right]_0^1 + \int_0^1 \frac{x^6}{(x^2+1)^4}\,\mathrm{d}x \\ &= \frac{1+\frac{5}{3}+1}{8} - 0 + J = \frac{11}{24} + J, \end{aligned}$$

其中 $J = \displaystyle\int_0^1 \frac{x^6}{(x^2+1)^4}\,\mathrm{d}x$。

想一想 怎样才能不用算出 J 而估计它的界?

由第 (i) 部分，J 的被积函数满足 $0 \le f(x) \le \frac{1}{16}$，对 $[0,1]$ 内的每个 x 都成立。把这些不等式在长度为 1 的区间上积分，得

$$0 \le J \le \frac{1}{16} \times 1 = \frac{1}{16}.$$

各项同时加上 $\frac{11}{24}$，

$$\frac{11}{24} \le \int_0^1 \frac{\mathrm{d}x}{(x^2+1)^4} \le \frac{11}{24} + \frac{1}{16},$$

这就是所求的结果。(其实两个不等式都是严格的，因为 f 连续，而且并非在整个区间上都等于任何一个界，但题目只要求证明 $\le$。)

验算

- 把 $x = 0$ 代入第 (ii) 部分的恒等式：导数项等于 $C = 1$，而含 D 的项为零，得 $1 = 1$。当 $x = 1$ 时，导数的分子是 $-1+0+0+1 = 0$，所以恒等式变成 $\frac{1}{16} = \frac{D}{16}$，与 $D = 1$ 吻合。
- 在数值上，用森逊法则（Simpson's rule）可得 $\int_0^1 \frac{\mathrm{d}x}{(x^2+1)^4} \approx 0.47460$，稳稳落在 $\frac{11}{24} \approx 0.45833$ 与 $\frac{11}{24} + \frac{1}{16} \approx 0.52083$ 之间。余项 $J \approx 0.01627$ 远小于 $\frac{1}{16}$，因为 f 除了在 $x = 1$ 附近之外都非常小。

常见错误

- 使用商法则时忘记约去公因子 $(x^2+1)^2$，结果分母变成 $(x^2+1)^6$，令比较系数的工作变得一团糟。
- 在第 (i) 部分，只检查 $x = 1$ 是驻点（stationary point）或端点（endpoint）：你必须证明 f 是递增的（或比较所有可能的点），才能断定端点的值是最大值。

STEP 之外

代换 (substitution) $x = \tan\theta$ 可把这个积分化为 $\int_0^{\pi/4} \cos^6\theta \, \mathrm{d}\theta$，再用归约公式 (reduction formula) 便可求出准确值：其值为 $\frac{11}{48} + \frac{5\pi}{64}$，与我们的数值估计吻合。把被积函数拆成“一个导数加上一个很小的量”，正是归约公式背后的想法；而通过估计被积函数的界来估计积分的界，则是实分析（real analysis）中最基本的估计方法。Bartle 与 Sherbert 合著的 *Introduction to Real Analysis*（实分析入门）对这些估计有严谨细致的阐述。

详细解答

STEP I 2000，第 5 题

纯数学 · 三维向量

STEP I 2000，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/00-s1-q5

关键思路

一切都发生在斜面上，所以先在这个平面内设置两个互相垂直的单位向量（unit vector）：一个沿最大斜率线（line of greatest slope）指向正上方，另一个是水平的。Arthur 和 Bertha 的位置就是这两个向量的简单组合，而和差化积公式（sum-to-product formulae）可以立刻把两者之差整理好。至于与铅垂线所成的角，我们只需要知道平面内这两个方向各自与铅垂方向的关系，θ 正是在这里出现的。

解答

建立设置

设 $\mathbf{e}$ 为沿着通过 O 的最陡直线指向上方的单位向量，并设 $\mathbf{f}$ 为平面内与 $\mathbf{e}$ 垂直、指向两人所在一侧的单位向量。由于 $\mathbf{f}$ 与最大斜率线垂直，所以它是水平的。设 $\mathbf{k}$ 为铅垂向上的单位向量。那么

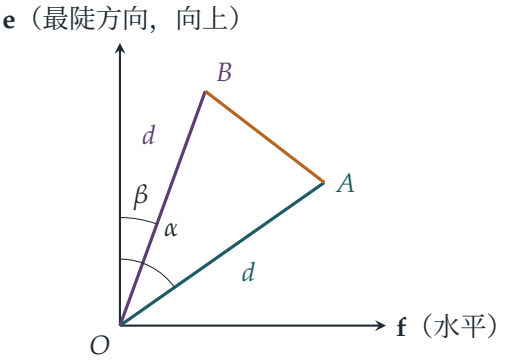
$$\mathbf{e}\cdot\mathbf{k}=\sin\theta,\qquad \mathbf{f}\cdot\mathbf{k}=0,\qquad \mathbf{e}\cdot\mathbf{f}=0.$$

（第一式成立，是因为 $\mathbf{e}$ 与水平面成角 θ，因而与 $\mathbf{k}$ 成角 $\frac{\pi}{2}-\theta$。）

两人步行的速率相同，所以当 Arthur 走了距离 d 时，Bertha 也走了同样的距离。他们相对于 O 的位置是

$$\mathbf{a}=d(\cos\alpha\,\mathbf{e}+\sin\alpha\,\mathbf{f}),\qquad \mathbf{b}=d(\cos\beta\,\mathbf{e}+\sin\beta\,\mathbf{f}).$$

两人都在最陡直线的同一侧向上走，即 $0\le\alpha,\beta<\frac{\pi}{2}$。



此图画的是斜面本身，即正对着斜面看过去的样子。

两人之间的距离

利用和差化积公式，

$$\begin{aligned}\mathbf{a}-\mathbf{b} &= d\big[(\cos\alpha-\cos\beta)\,\mathbf{e}+(\sin\alpha-\sin\beta)\,\mathbf{f}\big]\\ &= d\big[-2\sin\tfrac{1}{2}(\alpha+\beta)\sin\tfrac{1}{2}(\alpha-\beta)\,\mathbf{e}+2\cos\tfrac{1}{2}(\alpha+\beta)\sin\tfrac{1}{2}(\alpha-\beta)\,\mathbf{f}\big]\\ &= 2d\sin\tfrac{1}{2}(\alpha-\beta)\,\big[-\sin\tfrac{1}{2}(\alpha+\beta)\,\mathbf{e}+\cos\tfrac{1}{2}(\alpha+\beta)\,\mathbf{f}\big].\end{aligned}$$

记 $s=\frac{1}{2}(\alpha+\beta)$。向量 $\mathbf{u}=-\sin s\,\mathbf{e}+\cos s\,\mathbf{f}$ 是单位向量，因为 $\mathbf{e}$ 和 $\mathbf{f}$ 是互相垂直的单位向量，而且 $\sin^2 s+\cos^2 s=1$。因此

$$AB=|\mathbf{a}-\mathbf{b}|=2d\left|\sin\tfrac{1}{2}(\alpha-\beta)\right|,$$

正是所求。（从几何上看：OAB 是等腰三角形（isosceles triangle），两条等边长 d，顶角为 $|\alpha-\beta|$；由 O 作垂线，可得相同的答案。）

与铅垂线所成的角

当 $\alpha\neq\beta$ 时，由于 $0<|\alpha-\beta|<\frac{\pi}{2}$，我们有 $\sin\frac{1}{2}(\alpha-\beta)\neq 0$，所以 $A\neq B$，直线 AB 有明确定义。上面的计算显示，这条直线平行于单位向量 $\mathbf{u}$，或者等价地说，平行于 $\mathbf{v}=-\mathbf{u}=\sin s\,\mathbf{e}-\cos s\,\mathbf{f}$。

? **想一想** 一条直线有两个方向。我们应该拿哪一个来与铅垂方向比较?

直线与铅垂线所成的角是指锐角（或直角），所以我们取直线上指向上方的那个方向，或者至少是不指向下方的那个方向。现在

$$\mathbf{v}\cdot\mathbf{k}=\sin s\,(\mathbf{e}\cdot\mathbf{k})-\cos s\,(\mathbf{f}\cdot\mathbf{k})=\sin\theta\sin s.$$

由于 $0 \le s < \frac{\pi}{2}$ 及 $0 < \theta < \frac{\pi}{2}$，这个值非负，所以 $\mathbf{v}$ 就是向上的方向。因为 $\mathbf{v}$ 和 $\mathbf{k}$ 都是单位向量，两者之间的夹角 ϕ 满足

$$\cos\phi = \mathbf{v}\cdot\mathbf{k} = \sin\theta \sin\tfrac{1}{2}(\alpha+\beta),$$

这就是所求的结果。留意结果与 d 无关：两人一路走，直线 AB 一路平行于自身移动。

验算

- 若 Arthur 沿水平方向走（$\alpha = \frac{\pi}{2}$，即边界情况），而 Bertha 笔直向上走（$\beta = 0$），则 $AB = 2d\sin\frac{\pi}{4} = \sqrt{2}\,d$，这正是把毕氏定理（Pythagoras' theorem）用于两条互相垂直、长度同为 d 的路径所得的结果。取 $\theta = 30^\circ$ 及 $d = 1$，用坐标可得 $A = (1,0,0)$，$B = (0, \frac{\sqrt{3}}{2}, \frac{1}{2})$，所以由 B 到 A 的铅垂落差是 $\frac{1}{2}$，而两点相距 $\sqrt{2}$：$\cos\phi = \frac{1}{2\sqrt{2}} \approx 0.354$，与 $\sin 30^\circ \sin 45^\circ$ 吻合，所以 $\phi \approx 69.3^\circ$。
- 若平面是水平的（$\theta = 0$），则 $\cos\phi = 0$：直线是水平的，理应如此。
- 用电脑在真正的三维坐标中随机选取 10^5 组 θ、α、β、d 作检验，结果与两条公式的误差都在 10^{-12} 以内。

常见错误

- 量度了平面内 AB 与最陡直线之间的角，而不是空间中与铅垂线所成的角；答案中必须出现 θ。
- 忘记加绝对值（modulus）：当 $\beta > \alpha$ 时，$\sin\frac{1}{2}(\alpha-\beta)$ 是负数，但距离不会是负数。
- 取了 AB 上向下的方向，得出 $\cos\phi = -\sin\theta\sin s$；直线与铅垂线所成的角永远不会是钝角。

STEP 之外

这里的方法，是选取一组切合问题的标准正交基（orthonormal basis）（先取平面内的两个向量，再找出它们与周围空间中铅垂方向的关系），然后由点积（dot product）读出长度和角度。在大学一年级的线性代数课程中，这会发展成基变换（change of basis）和正交投影（orthogonal projection）的概念：$\mathbf{v}\cdot\mathbf{k}$ 就是 $\mathbf{v}$ 在铅垂方向上的投影的长度。Strang 的 *Introduction to Linear Algebra*（线性代数入门）细致地阐述了这些概念，并配有大量几何内容。

详细解答

STEP I 2000，第 6 题

纯数学 · 直线所围成的区域

> **STEP I 2000，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/00-s1-q6

关键思路

一个关于 x 和 y 的二次式，若能因式分解（factorise）为两个一次因式，所描述的就是一对直线。两个因式的积为正，当且仅当两个因式同号；两者异号时积为负。因此每个不等式（inequality）所选出的，都是两条相交直线之间一对相对的“楔形”区域。只要把两个区域画在同一组坐标轴上，要找一个公共点，一看便知。

解答

第一个因式分解与区域

展开右边，

$$\begin{aligned}(x-y+2)(x+y-1) &= x^2+xy-x-xy-y^2+y+2x+2y-2\\ &= x^2-y^2+x+3y-2,\end{aligned}$$

正如所求。所以把第一个不等式两边同时减去 2 后，得

$$(x-y+2)(x+y-1)>0.$$

边界由两条直线组成：

$$L_1:\ y=x+2, \qquad L_2:\ y=1-x,$$

它们的斜率（gradient）分别为 1 和 -1，并在 $x+2=1-x$ 处相交，即交于 $\left(-\frac{1}{2},\frac{3}{2}\right)$。

两个因式同号时，积为正：

- 两者皆正：$y<x+2$ 且 $y>1-x$，即在 L_1 之下、L_2 之上。这是交点*右方*的楔形（例如它包含 $(5,0)$）；

- 两者皆负：$y > x+2$ 且 $y < 1-x$，即在 L_1 之上、L_2 之下。这是交点*左方*的楔形（它包含 $(-5,0)$）。

所以这个区域就是直线 $y = x+2$ 与 $y = 1-x$ 之间左右相对的一对楔形，而直线本身并不包括在内（在直线上积为 0，而非 > 0）。简单测试一下：在原点，该式的值为 $-2 < 0$，而原点位于下方的楔形内，这个楔形确实应被排除。

第二个区域

> **想一想** 我们怎样知道这一个也能因式分解？它的因式又应该是甚么样子？

由二次项 $x^2 - 4y^2 = (x+2y)(x-2y)$ 可以想到尝试

$$x^2 - 4y^2 + 3x - 2y + 2 = (x+2y+a)(x-2y+b).$$

右边展开后是 $x^2 - 4y^2 + (a+b)x + (2b-2a)y + ab$。比较系数（coefficient）得 $a+b=3$ 和 $b-a=-1$，所以 $a=2$，$b=1$，而常数项 $ab=2$ 亦同样吻合。因此，把第二个不等式两边同时加上 2，得

$$x^2 - 4y^2 + 3x - 2y < -2 \iff (x+2y+2)(x-2y+1) < 0.$$

边界直线为

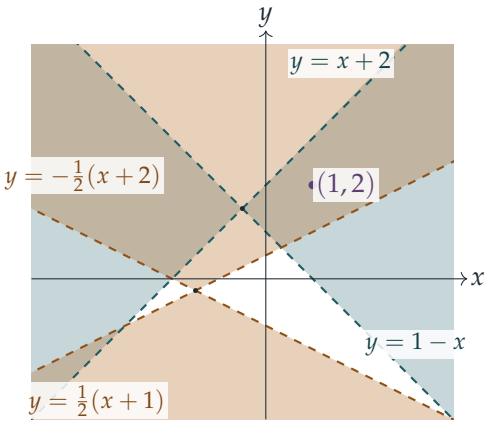
$$L_3:\ y = -\tfrac{1}{2}(x+2), \qquad L_4:\ y = \tfrac{1}{2}(x+1),$$

斜率分别为 $-\frac{1}{2}$ 和 $\frac{1}{2}$，并在 $-x-2 = x+1$ 处相交，即交于 $\left(-\frac{3}{2}, -\frac{1}{4}\right)$。

这次我们需要两个因式*异号*：

- $x+2y+2 > 0$ 且 $x-2y+1 < 0$：$y > -\frac{1}{2}(x+2)$ 且 $y > \frac{1}{2}(x+1)$，即在两条直线之上。这是*上方*的楔形（它包含 $(0,5)$）；
- $x+2y+2 < 0$ 且 $x-2y+1 > 0$：在两条直线之下。这是*下方*的楔形（它包含 $(0,-5)$）。

所以第二个区域就是直线 L_3 与 L_4 之间上下相对的一对楔形，边界同样不包括在内。



在草图中，第一个区域以蓝绿色涂上阴影（左右两个楔形），第二个区域则用琥珀色（上下两个楔形），所以颜色较深的重叠部分就是两个不等式同时成立之处；虚线并不包括在内。

同时位于两个区域的一点

这两对楔形必定有重叠：右边的蓝绿色楔形在斜率为 ± 1 的两条直线之间向右张开，而上方的琥珀色楔形则包含了位于斜率为 $\pm\frac{1}{2}$ 的两条较平缓直线之上的所有点。当 $x > -\frac{1}{2}$ 时，蓝绿色楔形的上边 $y = x + 2$ 比直线 $y = \frac{1}{2}(x+1)$ 上升得更快，所以两者之间留有空间。

例如取 $(x, y) = (1, 2)$。则

$$x^2 - y^2 + x + 3y = 1 - 4 + 1 + 6 = 4 > 2$$

且

$$x^2 - 4y^2 + 3x - 2y = 1 - 16 + 3 - 4 = -16 < -2,$$

所以 $(1, 2)$ 同时满足两个不等式。（用因式核对：$(x - y + 2)(x + y - 1) = 1 \cdot 2 > 0$ 且 $(x + 2y + 2)(x - 2y + 1) = 7 \cdot (-2) < 0$。）

验算

- 每个因式分解都可以随意取一点来核对：在 $(3, 1)$，$x^2 - y^2 + x + 3y - 2 = 12$ 而 $(x - y + 2)(x + y - 1) = 4 \cdot 3 = 12$；在同一点，$x^2 - 4y^2 + 3x - 2y + 2 = 14$ 而 $(x + 2y + 2)(x - 2y + 1) = 7 \cdot 2 = 14$。
- 在每个楔形内各测试一点，例如 $(0, -5)$ 给出 $x^2 - 4y^2 + 3x - 2y = -90 < -2$，确认下方的琥珀色楔形属于第二个区域。

- 在 $[-10,10]^2$ 内随机抽取 200 000 个点作数值检验，结果显示不等式与上述楔形的描述完全一致，没有任何出入。

常见错误

- 把错误的一对楔形涂上阴影。务必测试一点（原点最方便），而不要单靠“之上”和“之下”来判断。
- 把常数移项时忘记了 -2 和 $+2$：第一个不等式是 $(x-y+2)(x+y-1)>0$，第二个是 $(x+2y+2)(x-2y+1)<0$。
- 把边界直线画成实线：这些不等式是严格不等式，所以直线并不包括在内。

STEP 之外

一个二元二次式 $ax^2+bxy+cy^2+dx+ey+f$ 定义了一条圆锥曲线（conic）；当它能分解为两个一次因式时，这条圆锥曲线是退化（degenerate）的，即一对直线，而这种情况恰好在由各系数构成的某个 3×3 行列式（determinant）等于零时出现。二次型（quadratic form）的正负号分布（这里的 x^2-y^2 和 x^2-4y^2 都是不定的（indefinite）），会在大学一年级的线性代数（linear algebra）中，通过对称矩阵（symmetric matrix）及其特征值（eigenvalue）来研究。Strang 的 *Introduction to Linear Algebra*（《线性代数导论》）是了解这个课题的好去处。

详细解答

STEP I 2000，第 7 题

纯数学 · 导数的界限

STEP I 2000，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/00-s1-q7

关键思路

求导之后，条件 $\mathrm{f}'(x) \geq 0$ 就变成 “a 不小于某个关于 x 的函数”。因此一切取决于该函数的最大值。这个函数只与 x^2 有关，经代换（substitution）$u = 1 + x^2$ 后，它成为 $1/u$ 的二次式。再用配方法（completing the square）便可求得最大值，并清楚看出最大值在何处取得。

解答

导数

由商法则（quotient rule），

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{x^3}{1+x^2}\right) = \frac{3x^2(1+x^2) - x^3 \cdot 2x}{(1+x^2)^2} = \frac{x^4 + 3x^2}{(1+x^2)^2},$$

所以

$$\mathrm{f}'(x) = a - g(x), \qquad \text{其中} g(x) = \frac{x^2(x^2+3)}{(1+x^2)^2}.$$

因此，$\mathrm{f}'(x) \geq 0$ 对所有 x 成立，当且仅当 $a \geq g(x)$ 对所有 x 成立。我们需要求出 g 的最大值。

g 的最大值

设 $u = 1 + x^2$，则 $u \geq 1$ 且 $x^2 = u - 1$。于是

$$g = \frac{(u-1)(u+2)}{u^2} = \frac{u^2 + u - 2}{u^2} = 1 + \frac{1}{u} - \frac{2}{u^2}.$$

想一想 为甚么用 $1/u$ 而不是 u 来表示这个式子会较好?

以 $s = 1/u$ 表示，$g = 1 + s - 2s^2$，这只是一个二次式，而二次式的最大值可由配方法求得，无须用到微积分:

$$1 + s - 2s^2 = \frac{9}{8} - 2\left(s - \frac{1}{4}\right)^2.$$

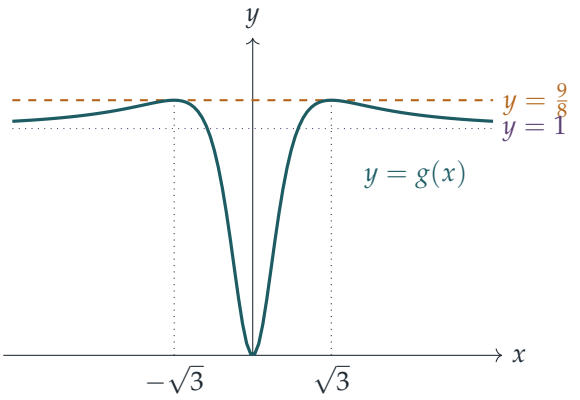
由于平方永不为负，对每个 x 都有 $g \le \frac{9}{8}$，等号仅在 $s = \frac{1}{4}$ 时成立，即 $u = 4$，$x = \pm\sqrt{3}$。把结果写回 x 的形式（利用 $s - \frac{1}{4} = \frac{3-x^2}{4(1+x^2)}$），便得到导数的一个简洁恒等式（identity）:

$$\mathrm{f}'(x) = \left(a - \frac{9}{8}\right) + \frac{(3-x^2)^2}{8(1+x^2)^2}.$$

结论

若 $a \ge \frac{9}{8}$, 第一个括号非负, 而第二项是一个平方除以正数, 所以同样非负。因此 $\mathrm{f}'(x) \ge 0$ 对所有 x 成立，正如所求。

常数 $\frac{9}{8}$ 也是最佳的：若 $a < \frac{9}{8}$，则 $\mathrm{f}'(\sqrt{3}) = a - \frac{9}{8} < 0$。当 $a = \frac{9}{8}$ 时，f' 只在 $x = \pm\sqrt{3}$ 处为零，所以 f 仍然是严格递增（strictly increasing）的，并有两个驻拐点（stationary points of inflection）。



g 的图像由 0 开始上升，越过其极限（limit）1，在 $x = \pm\sqrt{3}$ 时达到顶点 $\frac{9}{8}$，然后回落并趋向 1。导数 $\mathrm{f}'(x)$ 就是直线 $y = a$ 高出这条曲线的高度。

验算

▶ 在 $x = \sqrt{3}$ 处：$g = \dfrac{3 \cdot 6}{4^2} = \dfrac{18}{16} = \dfrac{9}{8}$。✓

- 当 $x \to \infty$ 时，$g \to 1 < \frac{9}{8}$。这是合理的，因为当 x 很大时 $\frac{x^3}{1+x^2}$ 的表现与 x 相若，所以其斜率趋向 1。
- 数值上，取 $a = \frac{9}{8}$，在 $[-100, 100]$ 上以步长 0.001 的网格计算，$\mathrm{f}'(x)$ 的最小值约为 2×10^{-10}（基本上是 0，在 $x \approx \pm\sqrt{3}$ 处）。取 $a = 1.1$，则 $\mathrm{f}'(\sqrt{3}) = -0.025 < 0$。

常见错误

- 把 g 在无穷远处的极限 1 当作它的最大值。曲线会越过这个极限，所以答案是 $\frac{9}{8}$ 而不是 1。
- 直接对 g 求导，结果迷失在繁复的代数运算之中。代换 $s = 1/(1+x^2)$ 可把问题化为一行就能完成的配方。

STEP 之外

在这里，一个“对所有 x”的命题被化为求上确界（supremum）的问题：$\mathrm{f}' \geq 0$ 处处成立，当且仅当 $a \geq \sup g$。g 在无穷远处有极限，却在有限的点上取得最大值，这正好提醒我们：极限不一定是界（bound）。在分析学（analysis）中，你会证明导数非负的函数是递增的（借助中值定理（Mean Value Theorem）），这就是为甚么这道题其实是在问 f 是否单调（monotonic）。Abbott 的 *Understanding Analysis* 对这些概念有细致的阐述。

详细解答

STEP I 2000，第 8 题

纯数学 · 绝对值的积分

STEP I 2000，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/00-s1-q8

关键思路

处理积分（integral）内的绝对值（modulus）符号，方法是找出符号内的式子在何处变号。在相邻两个变号点之间的每个区间上，$|f|$ 不是 f 就是 $-f$，所以我们在这些点把积分范围拆开，然后对普通的函数逐段积分。真正要花工夫的，是找出零点的位置：xe^x 可凭观察得出，三次式要靠因式分解（factorising），而 $\sin x + \cos x$ 则要写成单一个正弦波。

解答

第一个积分

由于对所有 x 都有 $e^x > 0$，xe^x 的正负号与 x 的正负号相同。因此在 $[-1,0]$ 上有 $xe^x \le 0$ 及 $|xe^x| = -xe^x$，而在 $[0,1]$ 上则有 $xe^x \ge 0$ 及 $|xe^x| = xe^x$。在 0 处把积分范围拆开，

$$\int_{-1}^{1} |xe^x|\,\mathrm{d}x = \int_{-1}^{0} (-xe^x)\,\mathrm{d}x + \int_{0}^{1} xe^x\,\mathrm{d}x = -\int_{-1}^{0} xe^x\,\mathrm{d}x + \int_{0}^{1} xe^x\,\mathrm{d}x,$$

这正是所需的恒等式。

要计算其值，可用分部积分法（integration by parts），取 $u = x$ 及 $\mathrm{d}v/\mathrm{d}x = e^x$：

$$\int xe^x\,\mathrm{d}x = xe^x - \int e^x\,\mathrm{d}x = (x-1)e^x + c.$$

因此

$$\int_{-1}^{0} xe^x\,\mathrm{d}x = \left[(x-1)e^x\right]_{-1}^{0} = -1 - (-2e^{-1}) = \frac{2}{e} - 1,$$

$$\int_{0}^{1} xe^x\,\mathrm{d}x = \left[(x-1)e^x\right]_{0}^{1} = 0 - (-1) = 1.$$

其中第一个积分是负数（这是必然的，因为被积函数（integrand）在该区间上是负的），于是

$$\int_{-1}^{1} |xe^x|\,dx = \left(1-\frac{2}{e}\right)+1 = 2-\frac{2}{e}.$$

第 (i) 部分

设 $p(x)=x^3-2x^2-x+2$。把各项分组，

$$p(x)=x^2(x-2)-(x-2)=(x-2)(x^2-1)=(x+1)(x-1)(x-2).$$

在 $[0,4]$ 内的零点只有 $x=1$ 和 $x=2$。首项系数为正、且有单根（simple root）$-1,1,2$ 的三次式，在 -1 与 1 之间为正，在 1 与 2 之间为负，在 2 之后为正。所以

$$|p(x)| = \begin{cases} p(x), & 0\le x\le 1,\\ -p(x), & 1\le x\le 2,\\ p(x), & 2\le x\le 4. \end{cases}$$

取原函数（antiderivative）

$$F(x)=\tfrac{1}{4}x^4-\tfrac{2}{3}x^3-\tfrac{1}{2}x^2+2x,$$

则有

$$F(0)=0, \qquad F(1)=\tfrac{1}{4}-\tfrac{2}{3}-\tfrac{1}{2}+2=\tfrac{13}{12},$$
$$F(2)=4-\tfrac{16}{3}-2+4=\tfrac{2}{3}, \qquad F(4)=64-\tfrac{128}{3}-8+8=\tfrac{64}{3}.$$

? **想一想** 为甚么不能直接计算 $F(4)-F(0)$，然后取其绝对值?

因为绝对值必须作用在被积函数上，而不是作用在答案上。在 $[1,2]$ 上这个三次式是负的，$F(4)-F(0)$ 会把那块面积减去，而不是加上。所以我们把三段分开处理：

$$\begin{aligned}\int_0^4 |p(x)|\,dx &= \big(F(1)-F(0)\big)-\big(F(2)-F(1)\big)+\big(F(4)-F(2)\big)\\ &= \tfrac{13}{12}-\big(\tfrac{2}{3}-\tfrac{13}{12}\big)+\big(\tfrac{64}{3}-\tfrac{2}{3}\big)\\ &= \tfrac{13}{12}+\tfrac{5}{12}+\tfrac{62}{3}=\tfrac{3}{2}+\tfrac{62}{3}=\frac{133}{6}.\end{aligned}$$

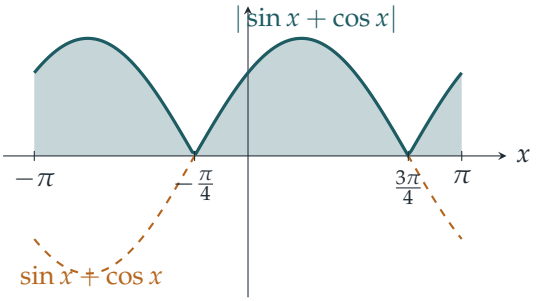
每一项都是正数，这是应有的结果：$\frac{13}{12}$、$\frac{5}{12}$ 和 $\frac{62}{3}$ 就是三块各自的面积。（作为比较，$F(4)-F(0)=\frac{64}{3}$，比正确答案少了 $2\times\frac{5}{12}$。）

第 (ii) 部分

利用 $\sin(x+\frac{\pi}{4})=\frac{1}{\sqrt{2}}(\sin x+\cos x)$，

$$\sin x+\cos x=\sqrt{2}\,\sin\big(x+\tfrac{\pi}{4}\big).$$

在 $[-\pi,\pi]$ 内，此式在 $x=-\frac{\pi}{4}$ 及 $x=\frac{3\pi}{4}$ 处为零。它在 $[-\pi,-\frac{\pi}{4}]$ 和 $[\frac{3\pi}{4},\pi]$ 上为负，在两者之间为正。



我们可以像之前那样拆成三段，但其实有一条更简洁的路。作代换（substitution）$u=x+\frac{\pi}{4}$：

$$\int_{-\pi}^{\pi}|\sin x+\cos x|\,\mathrm{d}x=\sqrt{2}\int_{-3\pi/4}^{5\pi/4}|\sin u|\,\mathrm{d}u.$$

函数 $|\sin u|$ 的周期（period）为 π，而新的积分范围长度为 2π，所以刚好涵盖两个周期。周期函数在整数个周期上的积分，与起点在哪里无关，所以

$$\sqrt{2}\int_{-3\pi/4}^{5\pi/4}|\sin u|\,\mathrm{d}u=\sqrt{2}\times 2\int_{0}^{\pi}\sin u\,\mathrm{d}u=\sqrt{2}\times 2\times 2=4\sqrt{2}.$$

如果你较喜欢直接拆开来做，$G(x)=\sin x-\cos x$ 是 $\sin x+\cos x$ 的一个原函数，其中 $G(-\pi)=1$、$G(-\frac{\pi}{4})=-\sqrt{2}$、$G(\frac{3\pi}{4})=\sqrt{2}$ 及 $G(\pi)=1$。三块面积分别是 $1+\sqrt{2}$、$2\sqrt{2}$ 和 $\sqrt{2}-1$，加起来是 $4\sqrt{2}$。

验算

- 在每一段上（被积函数在该段上是光滑的）使用辛普森法则（Simpson's rule），第一个积分得 1.264241…，与 $2-2/\mathrm{e}\approx 1.264241$ 吻合。
- 对于 (i)，数值积分（numerical integration）得 22.16667，与 $\frac{133}{6}$ 吻合。它也比 $F(4)-F(0)=\frac{64}{3}\approx 21.33$ 略大，这是必然的，因为绝对值把一小块负面积变成了正面积。
- 对于 (ii)，数值积分得 5.656854，与 $4\sqrt{2}$ 吻合。再作一个合理性检查：$|\sin x+\cos x|$ 的形状与 $\sqrt{2}|\sin x|$ 相同，后者的平均值为 $\frac{2\sqrt{2}}{\pi}$，而 $2\pi\times\frac{2\sqrt{2}}{\pi}=4\sqrt{2}$。

常见错误

- 直接跨过变号点积分，到最后才取绝对值。这样会把每一个负的区域都少算了。

- 在 (i) 中遗漏了 $x = 1$ 这个根。务必彻底因式分解，或至少用因式定理（factor theorem）检查细小的整数根。
- 计算 $\left[(x-1)\mathrm{e}^x\right]_{-1}^{0}$ 时弄错正负号：在 $x = -1$ 处的值是 $-2\mathrm{e}^{-1}$，所以减去它即是加上 $2/\mathrm{e}$。

STEP 之外

积分 $\int_a^b |f(x)|\,\mathrm{d}x$ 是曲线与坐标轴之间的总面积。到了大学，它会以 f 的 L^1 范数（norm）的身分再次出现，这是量度一个函数的大小、以及两个函数之间距离的一种方法。证明只要 f 可积（integrable）则 $|f|$ 亦可积，以及证明 $\left|\int f\right| \le \int |f|$，是实分析（real analysis）入门阶段的标准习题。Abbott 的 *Understanding Analysis*（实分析入门教材）对黎曼积分（Riemann integral）的建构相当严谨，足以证明这两个结果。

详细解答

STEP I 2000，第 9 题

力学 · 加速参考系中的抛体

STEP I 2000，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/00-s1-q9

关键思路

射程是在车厢内量度的，所以应在车厢的参考系（frame of reference）中考虑问题。由于车厢以加速度 a 向北加速，炮弹相对于地板的加速度是向下的 g 加上向南的 a，而向南正是大炮发射的方向。这仍然是一个恒加速度的抛体运动（projectile motion），只是多了一个水平项。只要把射程用 2θ 表示出来，求其最大值便只需一行微分。

解答

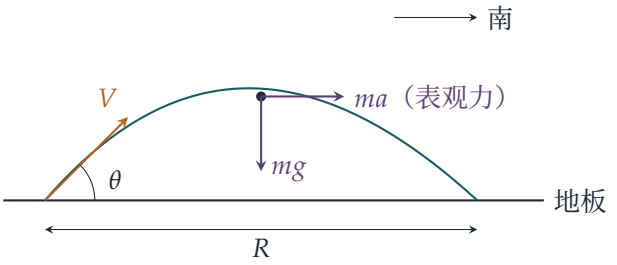
车厢内的射程

把炮弹当作质点（particle），并忽略空气阻力。假设炮弹从地板高度发射并落在地板上，而且车厢够长，炮弹会落在车厢之内。取 x 为水平方向并指向南（沿发射方向），y 为铅垂向上，两者均相对于车厢量度。

在地面参考系中，炮弹的加速度为向下的 g，并没有水平加速度。车厢的加速度为向北的 a。因此炮弹相对于车厢的加速度为

$$\ddot{x} = 0 - (-a) = a, \qquad \ddot{y} = -g.$$

换一个说法，在车厢的参考系中，质量为 m 的炮弹除了受本身的重量作用外，还受到一个向南的表观力（apparent force）ma。



相对于车厢的初速度为向南的 $V\cos\theta$ 和向上的 $V\sin\theta$，所以

$$x = Vt\cos\theta + \tfrac{1}{2}at^2, \qquad y = Vt\sin\theta - \tfrac{1}{2}gt^2.$$

铅垂方向的运动不受 a 影响：炮弹在 $y = 0$ 且 $t > 0$ 时着地，即在 $T = \dfrac{2V\sin\theta}{g}$ 时着地。因此它在车厢内的射程为

$$\begin{aligned} R = x(T) &= \frac{2V^2\sin\theta\cos\theta}{g} + \frac{a}{2}\cdot\frac{4V^2\sin^2\theta}{g^2} \\ &= \frac{V^2}{g}\left(\sin 2\theta + \frac{a}{g}(1-\cos 2\theta)\right), \end{aligned}$$

当中用到 $2\sin\theta\cos\theta = \sin 2\theta$ 和 $2\sin^2\theta = 1 - \cos 2\theta$。

求射程的最大值

对 θ 求导，得

$$\frac{\mathrm{d}R}{\mathrm{d}\theta} = \frac{2V^2}{g}\left(\cos 2\theta + \frac{a}{g}\sin 2\theta\right),$$

它在以下情况等于零：

$$\boxed{\tan 2\theta = -\frac{g}{a}.}$$

想一想 正切是负数？这是否表示哪里出了错？

并没有出错。当 $0 < \theta < \frac{\pi}{2}$ 时，我们有 $0 < 2\theta < \pi$，而 $\tan 2\theta < 0$ 只不过表示 2θ 位于 $(\frac{\pi}{2}, \pi)$ 之内，所以 $\theta > \frac{\pi}{4}$。这是合理的：只要炮弹仍在空中，额外的推动 a 便一直起作用，所以发射角比 45° 稍高一点、换取更长的飞行时间是划算的。若 $a = 0$，便有 $\cos 2\theta = 0$，得出大家熟悉的 $\theta = \frac{\pi}{4}$。

要确认这是最大值，设 $\tan\varphi = \dfrac{a}{g}$，其中 $0 < \varphi < \frac{\pi}{2}$。于是

$$\sin 2\theta - \frac{a}{g}\cos 2\theta = \frac{\sin(2\theta - \varphi)}{\cos\varphi}, \qquad \text{所以} \qquad R = \frac{V^2}{g}\left(\frac{\sin(2\theta-\varphi)}{\cos\varphi} + \frac{a}{g}\right).$$

当 $2\theta-\varphi=\frac{\pi}{2}$ 时，此式取最大值，即 $2\theta=\frac{\pi}{2}+\varphi$，而这正是 $\tan 2\theta=-g/a$ 在 $(\frac{\pi}{2},\pi)$ 内的根。所以最佳发射角的准确值为

$$\theta=\frac{\pi}{4}+\frac{1}{2}\arctan\frac{a}{g},$$

又由于 $\cos\varphi=g/\sqrt{g^2+a^2}$，最大射程的准确值为

$$R_{\max}=\frac{V^2}{g}\left(\frac{\sqrt{g^2+a^2}}{g}+\frac{a}{g}\right)=\frac{V^2}{g^2}\left(a+\sqrt{g^2+a^2}\right).$$

a 很小的情况

当 u 很小时，$\arctan u=u+O(u^3)$。取 $u=a/g$，得

$$\theta=\frac{\pi}{4}+\frac{1}{2}\left(\frac{a}{g}+O\left(\frac{a^3}{g^3}\right)\right)\approx\frac{\pi}{4}+\frac{a}{2g},$$

这正是所需的结果。至于射程，$\sqrt{g^2+a^2}=g\sqrt{1+a^2/g^2}=g+O(a^2/g)$，所以

$$R_{\max}=\frac{V^2}{g^2}\left(a+g+O(a^2/g)\right)\approx\frac{V^2}{g}+\frac{V^2a}{g^2},$$

这就是第二个所需的结果。两个近似式都准确至 a/g 的一阶（first order）；略去的各项为 $(a/g)^2$ 阶或更小。

如果你想继续使用 $\tan 2\theta$，可设 $2\theta=\frac{\pi}{2}+\varepsilon$。于是 $\tan 2\theta=-\cot\varepsilon\approx-1/\varepsilon$，所以 $\varepsilon\approx a/g$。把 $\sin 2\theta=\cos\varepsilon\approx 1$ 和 $\cos 2\theta=-\sin\varepsilon\approx-\varepsilon$ 代入 R，可得准确至一阶的 $R\approx\frac{V^2}{g}\left(1+\frac{a}{g}\right)$。

验算

- $a=0$：公式化简为 $\theta=\frac{\pi}{4}$ 和 $R_{\max}=V^2/g$，即水平地面上的标准结果。
- 量纲（dimensions）：a/g 是无量纲的，而 V^2/g 是长度，正如所料。
- 数值上，取 $V=7$、$g=9.8$ 和 $a=1.3$，在地面参考系中逐步计算运动，再减去车厢的位置，所得结果与 R 的公式吻合（例如在 $\theta=0.7$ 时 $R\approx 5.478$）。对若干个随机选取的 a，以穷举法搜索 θ，发现最大值出现在 $\tan 2\theta=-g/a$ 之处，而 $R_{\max}$ 与 $\frac{V^2}{g^2}(a+\sqrt{g^2+a^2})$ 吻合至 10 位有效数字。当 $a=0.05$ 时，最佳发射角的准确值 $0.787949\ldots$ 与 $\frac{\pi}{4}+\frac{a}{2g}$ 吻合至小数点后七位。

常见错误

- 弄错 a 的正负号。车厢加速的方向与炮弹的发射方向相反，所以相对于地板，炮弹向南的速率会增加：$\frac{1}{2}aT^2$ 这一项是加上去的，而不是减去的。

- 量度了地面上的射程，而不是车厢内的射程。题目问的是沿车厢地板的距离。
- 用计算机求得 $2\theta = \arctan(-g/a)$，这样会得出一个负的角。正确的根位于 $(\frac{\pi}{2}, \pi)$ 之内。

STEP 之外

在加速参考系中考虑问题，并加上一个表观力 $-m\mathbf{a}$，是假想力（fictitious force）最简单的例子。旋转参考系也以同样的方式引入离心力（centrifugal force）和科里奥利力（Coriolis force）。在本题中，表观力只不过令“等效重力”倾斜，使其大小变为 $\sqrt{g^2+a^2}$，这是理解答案中为何出现 $\sqrt{g^2+a^2}$ 的一种方法。Kleppner 与 Kolenkow 合著的 *An Introduction to Mechanics*（力学入门）对非惯性参考系（non-inertial frame）有清晰的讲解。

详细解答

STEP I 2000，第 10 题

力学 · 碰撞与恢复系数

STEP I 2000，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/00-s1-q10

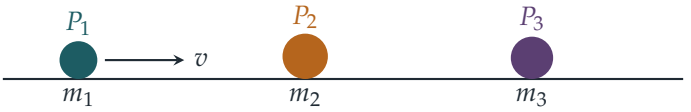
关键思路

每次碰撞都给我们两条方程：动量守恒（conservation of momentum）和牛顿恢复定律（Newton's law of restitution）。这里碰撞的结果已部分知道（P_1 停下；P_3 以速率 v 离开），所以我们不是求各个速度，而是求未知的系数。可行条件则来自任何恢复系数（coefficient of restitution）都必须符合的唯一物理限制 $0 \le e \le 1$，而最终能量成为唯一可自由变动的质量 m_2 的函数，我们在这些条件所容许的区间上求它的最大值和最小值。

解答

模型

我们把三个物体视为在光滑桌面上沿同一直线运动的质点（particle），因此没有水平外力作用，每次碰撞中动量都守恒。取 P_1 初始运动的方向为正方向。



第一次碰撞

设 P_2 以速度 u 离开。由于 P_1 变为静止，

$$\text{动量：}\quad m_1 v = m_2 u \quad\Longrightarrow\quad u = \frac{m_1 v}{m_2},$$

$$\text{恢复定律：}\quad u - 0 = e(v - 0) \quad\Longrightarrow\quad e = \frac{u}{v} = \frac{m_1}{m_2}.$$

第二次碰撞

P_2 现在以速率 u 撞向静止的 P_3。设 P_2 碰撞后的速度为 w；题目告诉我们 P_3 的速度为 v。由动量守恒得

$$m_2u = m_2w + m_3v \quad\Longrightarrow\quad m_1v = m_2w + m_3v \quad\Longrightarrow\quad w = \frac{(m_1 - m_3)v}{m_2},$$

其中用到第一次碰撞所得的 $m_2u = m_1v$。由恢复定律（分离速率 $= e' \times$ 接近速率）得

$$v - w = e'u.$$

代入 w 和 u，并将全式乘以 m_2/v（它不等于零）：

$$m_2 - (m_1 - m_3) = e'm_1 \quad\Longrightarrow\quad e' = \frac{m_2 + m_3 - m_1}{m_1},$$

这就是所要求的结果。

这样的碰撞何时可能发生?

任何恢复系数都介乎 0 与 1 之间。应用 $0 \le e' \le 1$，并乘以 $m_1 > 0$：

$$0 \le m_2 + m_3 - m_1 \le m_1 \quad\Longleftrightarrow\quad m_1 \le m_2 + m_3 \le 2m_1,$$

这正是题目要求的不等式。所以它是这些碰撞可能发生的必要条件。

> **?** **想一想** 我们是否已用上每一个限制条件？e 又如何?

我们还需要 $0 \le e \le 1$。由于 $e = m_1/m_2$，这个条件成立当且仅当 $m_2 \ge m_1$。这个条件并不属于题目要求我们证明的结果（事实上，有了它，左边的不等式 $m_2 + m_3 \ge m_1$ 便自动成立），但它在最后一部分十分重要。综合以上各点，这些碰撞可能发生当且仅当

$$m_1 \le m_2 \le 2m_1 - m_3.$$

要这个区间非空，我们需要 $m_3 \le m_1$。这时 $w = (m_1 - m_3)v/m_2 \ge 0$，所以第二次碰撞后 P_2 继续向前运动（或停下），永远不会回头撞上静止的 P_1。同时 $v - w = e'u \ge 0$，所以 P_2 不会再追上 P_3。因此总共只有这两次碰撞。

最终能量

最后 P_1 静止，P_2 的速度为 w，P_3 的速度为 v，所以最终动能（kinetic energy）为

$$E = \tfrac{1}{2}m_2w^2 + \tfrac{1}{2}m_3v^2 = \frac{(m_1 - m_3)^2v^2}{2m_2} + \tfrac{1}{2}m_3v^2.$$

当 m_1、m_3 和 v 固定时，只有 m_2 可以变动，范围是 $m_1 \le m_2 \le 2m_1 - m_3$。第一项是一个非负常数除以 m_2，所以 E 是 m_2 的递减函数（若 $m_1 = m_3$ 则为常数，不过这时区间本来就只有 $m_2 = m_1$ 这一点）。

最大能量在 $m_2 = m_1$ 时出现（第一次碰撞为完全弹性（perfectly elastic），$e = 1$）：

$$E_{\max} = \frac{(m_1 - m_3)^2 v^2}{2m_1} + \frac{m_3 v^2}{2} = \frac{(m_1^2 - m_1 m_3 + m_3^2)\, v^2}{2m_1}.$$

最小能量在 $m_2 = 2m_1 - m_3$ 时出现（第二次碰撞为完全弹性，$e' = 1$）：

$$E_{\min} = \frac{(m_1 - m_3)^2 v^2}{2(2m_1 - m_3)} + \frac{m_3 v^2}{2} = \frac{(m_1^2 - 2m_1 m_3 + m_3^2 + 2m_1 m_3 - m_3^2) v^2}{2(2m_1 - m_3)} = \frac{m_1^2 v^2}{2(2m_1 - m_3)}.$$

验算

- 碰撞中能量永远不会增加，所以我们需要 $E_{\max} \le \frac{1}{2} m_1 v^2$。这即是 $m_1^2 - m_1 m_3 + m_3^2 \le m_1^2$，亦即 $m_3(m_3 - m_1) \le 0$，由于 $m_3 \le m_1$，此式成立。
- 可以验证 $E_{\max} - E_{\min} = \dfrac{(m_1 - m_3)^3 v^2}{2m_1(2m_1 - m_3)} \ge 0$，而且当 $m_1 = m_3$ 时两个答案都等于 $\frac{1}{2} m_1 v^2$（这时 $m_2 = m_1$，两次碰撞都是弹性碰撞，没有能量损失）。
- 数值验算：取 $m_1 = 2$、$m_3 = 1$、$v = 1$，初始能量为 1，$E_{\max} = \frac{3}{4}$，$E_{\min} = \frac{2}{3}$。以精确算术让 m_2 逐步走遍 $[2, 3]$，可重现这两个极值，并确认 $e, e' \in [0, 1]$ 在整个区间内成立，在区间外则不成立。

常见错误

- 在 $m_1 - m_3 \le m_2 \le 2m_1 - m_3$ 上求最大值和最小值（只考虑 e'）。这样会得出“最大值” $\frac{1}{2} m_1 v^2$，即全部初始能量，但对应的 m_2 值使 $e = m_1 / m_2 > 1$，这是不可能的。
- 把恢复定律弄反了：是分离速率等于 e 乘以接近速率。

STEP 之外

这里的模式是：每个限制条件为参数划出一个容许的区间，然后我们在该区间上求一个单调函数的最大值和最小值。这是约束优化（constrained optimisation）最简单的情形，由此可发展出线性规划（linear programming）和拉格朗日乘数（Lagrange multipliers）。在物理方面，碰撞在质心参考系（centre-of-mass frame）中简洁得多：在那里，恢复系数只是把每个速度乘以 $-e$，而损失的能量为 $\frac{1}{2}\mu(1 - e^2) \times$（相对速率）^{2}，其中 μ 为约化质量（reduced mass）。Kleppner 与 Kolenkow 的 *An Introduction to Mechanics*（力学入门）正是以这个观点阐述碰撞。

详细解答

STEP I 2000，第 11 题

力学 · 梯子的摩擦力平衡

STEP I 2000，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/00-s1-q11

关键思路

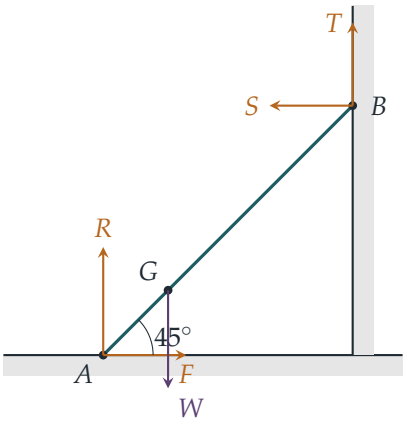
这是一道两端都有摩擦力的梯子问题。“在 A 和 B 两处同时达到极限”这个条件，令每个摩擦力（frictional force）都可以用对应的法向反作用力（normal reaction）表示，于是两条分解力的方程便决定了全部四个接触力。这样一来，力矩（moment）方程就成了多出来的一个条件：它要么碰巧成立，要么不成立。在第一部分，我们证明它不成立。在第二部分，我们把额外质量的位置当作未知数，令它成立。

解答

模型与方程

我们把杆看成一根刚性杆（rigid rod），其重量 $W = 5g\,\mathrm{N}$ 作用于 G。A（地面）和 B（墙壁）两处的接触面都是粗糙的。各个力如下：

- 在 A：法向反作用力 R 向上，摩擦力 F 沿水平方向；
- 在 B：法向反作用力 S 沿水平方向离开墙壁，摩擦力 T 沿铅垂方向（取向上为正，所以 $T < 0$ 即表示向下）；
- 重量 W 在 G 处向下。



水平方向的力只有 F 和 S，而 $S \geq 0$ 把杆推离墙壁。所以 F 必须指向墙壁，并且

$$F = S. \tag{1}$$

沿铅垂方向分解，

$$R + T = W. \tag{2}$$

对 A 取力矩时，设 $AB = 0.81$，$c = \cos 45^\circ = \sin 45^\circ$。那么 B 与 A 的水平距离是 $0.81c$，高度也是 $0.81c$，而 G 与 A 的水平距离是 $0.21c$。比较顺时针和逆时针方向的力矩，

$$S\,(0.81c) + T\,(0.81c) = W\,(0.21c),$$

约去 c 后得

$$0.81\,(S + T) = 0.21\,W. \tag{3}$$

45° 的好处就在这里：因子 c 可以从每一项中约去。

摩擦力不可能在两端同时达到极限

假设摩擦力在两端都达到极限。那么 $F = 0.2R$，而 $|T| = 1.0\,S = S$。

> ? **想一想** 当 B 处的摩擦力达到极限时，它指向哪个方向?

最自然的情况是杆即将滑动，A 远离墙壁而 B 向下滑，这时 $T = +S$。为求周全，我们把两种符号都检查一遍。

情况 $T = S$。由 (1)，$S = F = 0.2R$，而 (2) 给出 $R + 0.2R = W$。所以 $R = \frac{5}{6}W$，$S = T = \frac{1}{6}W$。于是 (3) 的左方是

$$0.81\left(\tfrac{1}{6}W + \tfrac{1}{6}W\right) = 0.27\,W \neq 0.21\,W.$$

所以力矩方程不成立。

情况 $T = -S$。(3) 的左方是 0，但右方是 $0.21W > 0$，这是不可能的。

在两种情况下，杆都不可能处于平衡（equilibrium）。所以摩擦力不可能在 A 和 B 两处同时达到极限，这正是所要证明的。

（在非极限的情况下，平衡当然是可能的：例如 $T = S = \frac{7}{54}W$ 满足 (3)，并给出 $R = \frac{47}{54}W$，而且 $F = S \leq 0.2R$。）

在 P 加上质量

设 $AP = x\,\mathrm{m}$。额外的重量同样是 $W = 5g\,\mathrm{N}$，作用于 P，所以总重量是 $2W$。方程 (1) 和 (2) 变成

$$F = S, \qquad R + T = 2W,$$

而对 A 的力矩方程多了一项 $W\,(xc)$。约去 c 后，它变成

$$0.81\,(S + T) = 0.21\,W + x\,W. \tag{4}$$

现在摩擦力在两端都达到极限。$T = -S$ 的情况同样不可能，因为这时 (4) 的左方是 0，右方却是正数。所以 $T = S$ 且 $F = 0.2R$，因此

$$S = 0.2R, \qquad R + 0.2R = 2W \quad \Longrightarrow \quad R = \tfrac{5}{3}W, \quad S = T = \tfrac{1}{3}W.$$

代入 (4) 得

$$0.81 \times \tfrac{2}{3}\,W = 0.54\,W = 0.21\,W + x\,W,$$

所以 $x = 0.33$。因此

$$AP = 0.33\ \mathrm{m}.$$

这个值介乎 0 与 0.81 之间，所以 P 在杆上。它在 G 之外（也在中点之外），所以额外的质量把合重心（centre of gravity）沿杆向上推，而这往往会令梯子滑倒。

验算

- 质量放在 P 时，以 W 为单位，各力为 $R = \frac{5}{3}$ 和 $F = S = T = \frac{1}{3}$。于是 $F/R = 0.2$，$T/S = 1$，所以两处都达到极限。铅垂方向有 $\frac{5}{3} + \frac{1}{3} = 2$，而力矩给出 $0.81 \times \frac{2}{3} = 0.54 = 0.21 + 0.33$。
- 合重心与 A 的距离是 $\frac{1}{2}(0.21 + 0.33) = 0.27\,\mathrm{m}$。这正是第一部分中，要令摩擦力在两端同时达到极限所需的位置（$0.27W$ 对比 $0.21W$）。这是合理的，因为摩擦系数不变，只是总重量增加了一倍。
- 质量和 g 在整个计算中都会约去，所以答案与它们无关。以 $g = 9.8$ 和真正的 45° 力矩作数值代入，所得残差的数量级为 10^{-15}。

常见错误

- 把 A 处摩擦力的方向取错。由于 S 是另外唯一的水平力，F 必须指向墙壁。
- 忘记 B 处的摩擦力是沿铅垂方向的，所以它会出现在铅垂方向的分解式中，令 R 小于总重量。
- 用了杆的中点而不是 G。这根杆并不均匀。

STEP 之外

平面上的刚体（rigid body）给出三条平衡方程，所以一道有四个未知接触力的问题是静不定（statically indeterminate）的。单靠刚体静力学，无法得知摩擦力怎样在两端之间分配。加上“在两端同时达到极限”这个条件，就多了两条方程，令方程组变成超定（over-determined），而这正是第一部分无解的原因。在工程学上，静不定的问题是通过为接触点的弹性建立模型来解决的。Kleppner 与 Kolenkow 的 *An Introduction to Mechanics*（力学入门）以大学一年级的程度，仔细讨论了刚体静力学和摩擦力。

详细解答

STEP I 2000，第 12 题

概率与统计 · 几何分布

STEP I 2000，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/00-s1-q12

关键思路

每一种情况都先问：已知之前每一次尝试都失败，下一次成功的机会是多少？然后把首 $n-1$ 次尝试失败的概率（probability）与第 n 次成功的概率相乘。三种规则得出的答案形态截然不同：一种是条件概率（conditional probability）不断增大，一种是由一开始就保持不变，还有一种是第一次之后才保持不变。

全篇以 N 表示所需的尝试次数，因此我们要求的是 $\mathrm{P}(N=n)$，其中 $n=1,2,3,\ldots$。我们把能开门的那条钥匙称为正确钥匙。

解答

第 (i) 部分

这里我绝不重复使用任何钥匙，所以我是按一个均匀随机的次序逐一试钥匙。根据对称性（symmetry），正确钥匙落在这个次序中 k 个位置的任何一个，可能性都相同。因此

$$\mathrm{P}(N=n)=\frac{1}{k}\quad \text{其中} n=1,2,\ldots,k,\qquad \mathrm{P}(N=n)=0 \text{ 当} n>k.$$

如果觉得这个对称性论证来得太取巧，可以改为把条件概率相乘。失败 j 次之后，还有 $k-j$ 条未试过的钥匙，其中一条是正确的，所以对于 $1\le n\le k$

$$\mathrm{P}(N=n)=\frac{k-1}{k}\cdot\frac{k-2}{k-1}\cdots\frac{k-n+1}{k-n+2}\cdot\frac{1}{k-n+1}=\frac{1}{k},$$

因为这个乘积会逐项相消（telescope）。

第 (ii) 部分

现在每一次尝试都是从全部 k 条钥匙中重新、独立（independent）地选取，所以不论之前发生过甚么，每一次尝试成功的概率都是 $\frac{1}{k}$。要刚好尝试 n 次，首 $n-1$ 次必须失败，而第 n 次必须成功：

$$\mathrm{P}(N=n)=\left(\frac{k-1}{k}\right)^{n-1}\frac{1}{k},\qquad n=1,2,3,\ldots$$

这是参数为 $\frac{1}{k}$ 的几何分布（geometric distribution），而 n 并没有上限。

第 (iii) 部分

第一次尝试成功的概率是 $\frac{1}{k}$，所以 $\mathrm{P}(N=1)=\frac{1}{k}$。

现在假设第 m 次尝试（对某个 $m\geq 1$）失败了。那么第 m 次所试的是一条错的钥匙。在第 $m+1$ 次尝试，我从其余 $k-1$ 条钥匙中均匀地选取。

> **想一想** 第 m 次尝试之前的经过，会否影响第 $m+1$ 次尝试成功的机会？

不会。唯一有关系的是哪一条钥匙被排除，而被排除的永远是一条错的钥匙。所以可供选择的 $k-1$ 条钥匙永远包括正确的那一条，而不论之前发生过甚么，第 $m+1$ 次尝试成功的概率都是 $\frac{1}{k-1}$，失败的概率都是 $\frac{k-2}{k-1}$。

所以对于 $n\geq 2$：第一次尝试失败（概率为 $\frac{k-1}{k}$），第 $2,\ldots,n-1$ 次尝试失败（每次概率为 $\frac{k-2}{k-1}$），而第 n 次尝试成功（概率为 $\frac{1}{k-1}$）。因此

$$\mathrm{P}(N=n)=\frac{k-1}{k}\left(\frac{k-2}{k-1}\right)^{n-2}\frac{1}{k-1}=\frac{1}{k}\left(\frac{k-2}{k-1}\right)^{n-2},\qquad n\geq 2,$$

另外还有 $\mathrm{P}(N=1)=\frac{1}{k}$。

有趣的是，同样有 $\mathrm{P}(N=2)=\frac{1}{k}$，与第 (i) 部分的结果一样。（当 $k=2$ 时，这个规则迫使我轮流使用两条钥匙，所以 N 是 1 或 2，概率各为 $\frac{1}{2}$；只要把 0^0 理解为 1，公式便给出这个结果。）

验算

- 每个分布的总和都是 1。在 (iii) 中，$\frac{1}{k}+\frac{1}{k}\sum_{j\geq 0}\left(\frac{k-2}{k-1}\right)^j=\frac{1}{k}+\frac{1}{k}(k-1)=1$。
- 各个平均值（mean）都合乎情理：(i) 是 $\frac{k+1}{2}$，(ii) 是 k，(iii) 是 $1+\frac{(k-1)^2}{k}$。与 (ii) 相比，不重复使用上一条钥匙略有帮助，但远远不及完全不重复使用任何钥匙。
- 我们对 $k=2,\ldots,5$ 和 $n\leq 7$，以分数列举每一个选择序列，精确地核对了全部三条公

式。就 $k = 5$ 的 (iii) 进行 2×10^5 次模拟 (simulation) 试验，得出 $\mathrm{P}(N = 3) \approx 0.1496$，而精确值是 $\frac{1}{5} \cdot \frac{3}{4} = 0.15$。

常见错误

- 在 (iii) 中，第一次尝试也用了 $\frac{k-2}{k-1}$。第一次是从全部 k 条钥匙中选取，所以第一个因子是 $\frac{k-1}{k}$。
- 在 (i) 中，忘记说明当 $n > k$ 时 $\mathrm{P}(N = n) = 0$。
- 在 (iii) 中，以为被排除的钥匙可能是正确的那一条。其实被排除的永远只是刚刚试过而失败的钥匙。

STEP 之外

第 (iii) 部分是一条小型的马可夫链（*Markov chain*）。接下来发生甚么只取决于当前的状态（state）（这里是指上一次试的是哪一条钥匙），而与整段历史无关。找出一个令过去变得无关重要的状态，正是马可夫链理论的核心技巧，它把繁复的条件概率化为简单的乘积。Grimmett 与 Welsh 的 *Probability: An Introduction*（概率入门教材）正是以这个程度介绍马可夫链。

详细解答

STEP I 2000，第 13 题

概率与统计 • 条件概率

> **STEP I 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-s1-q13

关键思路

整个问题归根究柢就是追踪父母各自传下来的一个基因。在 (i) 中，两个 AB 型父母所生的每个孩子，是 AA 或 BB 的概率（probability）同样那么小，而四个孩子互相独立（independent），所以我们只需数出排列方式的数目，再把概率相乘。第 (ii) 部分是“逆向”的条件概率（conditional probability）：我们知道结果（孩子），要找的是原因（父亲）。这正是贝叶斯定理（Bayes' theorem），而按父亲三种可能的基因型（genotype）画出树形图（tree diagram），就能把计算整理得井井有条。

解答

第 (i) 部分

父母各自都是 AB，所以各自以概率 $\frac{1}{2}$ 传下 A 或 B，而且父母二人互相独立。对单独一个孩子而言，

$$\mathrm{P}(AA) = \tfrac{1}{2}\cdot\tfrac{1}{2} = \tfrac{1}{4}, \qquad \mathrm{P}(BB) = \tfrac{1}{4}, \qquad \mathrm{P}(AB) = \tfrac{1}{2},$$

其中最后一个已把两种次序都计算在内（A 来自母亲而 B 来自父亲，或者反过来）。

不同孩子的遗传互相独立。四个孩子中两个是 AA、两个是 BB 的某一个特定次序，例如 AA, AA, BB, BB，其概率为 $\left(\frac{1}{4}\right)^4$。从四个孩子中选出哪两个是 AA 的方法数目是 $\binom{4}{2} = 6$。所以

$$\mathrm{P}(\text{两个}AA\text{ 和两个}BB) = 6\times\left(\tfrac{1}{4}\right)^4 = \frac{6}{256} = \frac{3}{128},$$

正如题目所求。

第 (ii) 部分

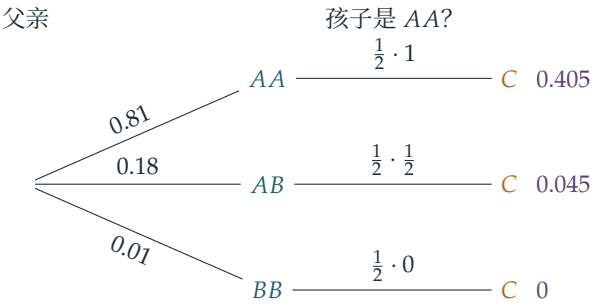
设 F_{AA}、F_{AB}、F_{BB} 分别为我父亲是 AA、AB、BB 的事件，并设 C 为我是 AA 的事件。我们把父亲视为从人口中随机抽出的一员，而且这个选取与母亲的基因型无关，所以

$$\mathrm{P}(F_{AA}) = 0.81, \qquad \mathrm{P}(F_{AB}) = 0.18, \qquad \mathrm{P}(F_{BB}) = 0.01.$$

> **想一想** 为甚么 0.18 还不是答案?

因为我们有额外的资料：我是 AA。这完全排除了父亲是 BB 的可能，同时令父亲是 AA 的可能性增加，因为 AA 型的父亲必定传下 A。所以我们必须以 C 为条件。

我要成为 AA，就需要从母亲得到一个 A，也从父亲得到一个 A。我母亲是 AB，所以不论父亲是哪一种基因型，她传下 A 的概率都是 $\frac{1}{2}$。父亲传下 A 的概率则视乎他的基因型，分别为 1、$\frac{1}{2}$ 或 0。树形图如下：



根据全概率公式（law of total probability），

$$\begin{aligned}\mathrm{P}(C) &= 0.81 \cdot \tfrac{1}{2} \cdot 1 + 0.18 \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} + 0.01 \cdot \tfrac{1}{2} \cdot 0 \\ &= 0.405 + 0.045 + 0 = 0.45.\end{aligned}$$

再根据贝叶斯定理，

$$\mathrm{P}(F_{AB} \mid C) = \frac{\mathrm{P}(F_{AB})\,\mathrm{P}(C \mid F_{AB})}{\mathrm{P}(C)} = \frac{0.045}{0.45} = \frac{1}{10}.$$

所以我父亲是 AB 的概率是 $\frac{1}{10}$。

验算

- 在 (i) 中，四个孩子各种基因型数目组合的概率，总和应为 1；$\binom{4}{2}$ 这个因子，正是 $\left(\frac{1}{4} + \frac{1}{2} + \frac{1}{4}\right)^4$ 的多项式展开（multinomial expansion）中出现的那一个，而 $3/128 \approx 0.0234$ 就是其中一项。把八个基因的全部 2^8 种遗传方式逐一列举，所得的结果正是 $3/128$。

- 在 (ii) 中，答案由先验概率（prior）0.18 降至 0.1，这是合理的：父亲是 BB 的可能已被排除，而父亲是 AA 的可能性则有所增加。模拟 10^6 个家庭，所得结果为 0.0998。

常见错误

- 在 (i) 中漏掉因子 $\binom{4}{2} = 6$，得出 $1/256$。
- 在 (ii) 中答 0.18，或者计算了 $\mathrm{P}(C \mid F_{AB}) = \frac{1}{4}$，而不是 $\mathrm{P}(F_{AB} \mid C)$。
- 只在部分分支中漏掉母亲的因子 $\frac{1}{2}$：这个因子最终会约去，但前提是每一个分支都一致地使用它。

STEP 之外

人口中的比例 81%、18%、1% 正是 p^2、$2pq$、q^2，其中 $p = 0.9$、$q = 0.1$：这就是哈迪－温伯格平衡（Hardy–Weinberg equilibrium），在随机交配之下，它会一代一代地保持不变。从观察到的结果推断隐藏的原因，是贝叶斯统计学（Bayesian statistics）的核心，而同样的计算也是遗传咨询和亲子鉴定的基础。Blitzstein and Hwang, *Introduction to Probability*（概率论入门）对条件概率和贝叶斯定理有详尽的论述，并附有许多与本题类似的例子。

详细解答

STEP I 2000，第 14 题

概率与统计 • 均匀分布的矩

STEP I 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-s1-q14

关键思路

这里的一切都是均匀分布（uniform distribution）的矩（moment），而 X 在 $[-1,1]$ 上的矩很容易计算：奇数阶的矩因对称性而等于零，偶数阶的矩则为 $1/(k+1)$。至于 $Z=Y-X$，我们根本不需要知道 Z 的分布。把 $(Y-X)^2$ 和 $(Y-X)^4$ 按二项式展开，再利用独立性（independence），每个期望值（expectation）都化为我们已知的矩的乘积。

解答

X 的矩

X 在 $[-1,1]$ 上的密度（density）为 $\frac{1}{2}$，所以对非负整数 k，

$$\mathrm{E}(X^k)=\frac{1}{2}\int_{-1}^{1}x^k\,\mathrm{d}x=\begin{cases}0, & k\text{ 为奇数（被积函数是奇函数），}\\ \dfrac{1}{k+1}, & k\text{ 为偶数.}\end{cases}$$

特别地，$\mathrm{E}(X)=\mathrm{E}(X^3)=0$，$\mathrm{E}(X^2)=\frac{1}{3}$，而 $\mathrm{E}(X^4)=\frac{1}{5}$。因此

$$\mathrm{E}(X^2)=\frac{1}{3},\qquad \mathrm{Var}(X^2)=\mathrm{E}(X^4)-\big(\mathrm{E}(X^2)\big)^2=\frac{1}{5}-\frac{1}{9}=\frac{4}{45}.$$

Y 的分布相同，所以它的矩也相同。

Z^2 的平均值

展开得 $Z^2=Y^2-2XY+X^2$。由于 X 和 Y 独立，$\mathrm{E}(XY)=\mathrm{E}(X)\,\mathrm{E}(Y)=0$，所以

$$\mathrm{E}(Z^2)=\frac{1}{3}-0+\frac{1}{3}=\frac{2}{3}.$$

Z^2 的方差

我们需要 $\mathrm{E}(Z^4)$。根据二项式定理（binomial theorem），

$$Z^4 = (Y-X)^4 = Y^4 - 4Y^3X + 6Y^2X^2 - 4YX^3 + X^4.$$

> **想一想** 我们真的可以把 Y^3X 的期望值写成 $\mathrm{E}(Y^3)\,\mathrm{E}(X)$ 吗?

可以：因为 X 和 Y 独立，所以它们的任何函数（例如 Y^3 和 X）也独立，而独立随机变量（random variable）之积的期望值等于各自期望值之积。中间两个交叉项各含有 X 或 Y 的奇数次幂，所以它们的期望值都是零。因此

$$\begin{aligned}\mathrm{E}(Z^4) &= \mathrm{E}(Y^4) + 6\,\mathrm{E}(Y^2)\,\mathrm{E}(X^2) + \mathrm{E}(X^4)\\ &= \frac{1}{5} + 6\cdot\frac{1}{3}\cdot\frac{1}{3} + \frac{1}{5} = \frac{2}{5} + \frac{2}{3} = \frac{16}{15}.\end{aligned}$$

所以

$$\mathrm{Var}(Z^2) = \mathrm{E}(Z^4) - \big(\mathrm{E}(Z^2)\big)^2 = \frac{16}{15} - \frac{4}{9} = \frac{48-20}{45} = \frac{28}{45}.$$

由于 $\frac{28}{45} = 7\times\frac{4}{45}$，我们得到 $\mathrm{Var}(Z^2) = 7\,\mathrm{Var}(X^2)$，这正是所要证明的。

验算

- Z 在 $[-2,2]$ 上有三角形密度 $f(z) = \frac{1}{4}(2-|z|)$。直接计算，$\mathrm{E}(Z^4) = 2\int_0^2 \frac{1}{4}z^4(2-z)\,\mathrm{d}z = \frac{1}{2}\left(\frac{64}{5} - \frac{32}{3}\right) = \frac{16}{15}$，与展开所得的结果一致。
- $\mathrm{E}(Z^2) = \mathrm{Var}(Y-X) = \mathrm{Var}(Y) + \mathrm{Var}(X) = \frac{1}{3} + \frac{1}{3}$，因为 $\mathrm{E}(Z) = 0$。
- 模拟一百万对 (X,Y)，得出 $\mathrm{Var}(X^2) \approx 0.0889$ 和 $\mathrm{Var}(Z^2) \approx 0.623$，而理论值为 $4/45 \approx 0.0889$ 和 $28/45 \approx 0.622$。

常见错误

- 写成 $\mathrm{Var}(Z^2) = \mathrm{Var}(Y^2) + \mathrm{Var}(X^2)$。这样做忽略了 Z^2 内的 $-2XY$ 一项，它的平均值虽然是零，却贡献了 $4\,\mathrm{Var}(XY) = \frac{4}{9} = \frac{20}{45}$；这个错误会得出 $\frac{8}{45}$，而不是 $\frac{28}{45}$。
- 忘记了 Y^2X^2 一项的二项式系数 6。
- 在 $[-1,1]$ 上把密度当作 1 而不是 $\frac{1}{2}$，结果每个矩都变成两倍。

STEP 之外

$\mathrm{E}(X^k)$ 这些量就是 X 的矩，而处理独立变量之和的简洁方法是矩母函数（moment generating function）$M_X(t) = \mathrm{E}(\mathrm{e}^{tX})$，它满足 $M_{Y-X}(t) = M_Y(t)M_X(-t)$。把这个乘积展开成 t 的幂级数（power series），就可以一次过得到 Z 的所有矩。矩母函数和特征函数（characteristic function）是大学第一门概率课程的核心内容：它们是证明中心极限定理（central limit theorem）的主要工具。Grimmett 与 Welsh 所著的 *Probability: An Introduction*（概率论入门教材）对这些内容有清晰的阐述。

详细解答

STEP II 2000，第 1 题

纯数学 · 埃及分数

> **STEP II 2000，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/00-s2-q1

关键思路

两个例子有共同的规律：第一个分母比 N 大一，第二个分母则是 N 与它的乘积。至于唯一性的部分，先去分母，再“配成矩形”，便可把 $\frac{1}{a}+\frac{1}{b}=\frac{1}{N}$ 化为两个整数之积等于 N^2 的形式。质数（prime）的平方只有极少数的因数分解（factorisation）方式，所以 a 和 b 的可能取值寥寥可数，我们只需逐一列出即可。

解答

一般结果

在 $\frac{1}{2}=\frac{1}{3}+\frac{1}{6}$ 和 $\frac{1}{3}=\frac{1}{4}+\frac{1}{12}$ 中，分母分别是 $N+1$ 和 $N(N+1)$（因为 $6=2\times 3$ 而 $12=3\times 4$）。于是我们猜想

$$\frac{1}{N}=\frac{1}{N+1}+\frac{1}{N(N+1)}.$$

要证明这条式，只需把右边通分：

$$\frac{1}{N+1}+\frac{1}{N(N+1)}=\frac{N+1}{N(N+1)}=\frac{1}{N}.$$

两个分母都是大于 1 的整数，所以两项都是单位分数（unit fraction）。又因为当 $N>1$ 时 $N(N+1)>N+1$，所以两者相异。因此，每个单位分数都可以写成两个相异单位分数之和。

N 为质数时的唯一性

把 $\frac{1}{N}=\frac{1}{a}+\frac{1}{b}$ 两边乘以 abN，得 $ab=Nb+Na$，即

$$ab-Na-Nb=0 \quad\Longleftrightarrow\quad ab-Na-Nb+N^2=N^2 \quad\Longleftrightarrow\quad (a-N)(b-N)=N^2.$$

每一步都可以逆推（因为 a、b、N 均不为零），所以 $(*)$ 的解恰好就是 $(a-N)(b-N)=N^2$ 的解。

> **想一想** $a-N$ 和 $b-N$ 有没有可能同时为负数?

不可能。由于 $\frac{1}{b}>0$，我们有 $\frac{1}{a}<\frac{1}{N}$，所以 $a>N$；同理 $b>N$。因此 $a-N$ 和 $b-N$ 是乘积为 N^2 的正整数。（这一点值得留意：若没有这一步，$(-1,-N^2)$ 这一对便会是另一个需要排除的因数分解。）

现在设 $N=p$ 为质数。p^2 的正因数（divisor）只有 1、p 和 p^2，所以 $(a-p,\,b-p)$ 这一对必定是以下其中之一：

$$(1,\,p^2),\qquad (p,\,p),\qquad (p^2,\,1).$$

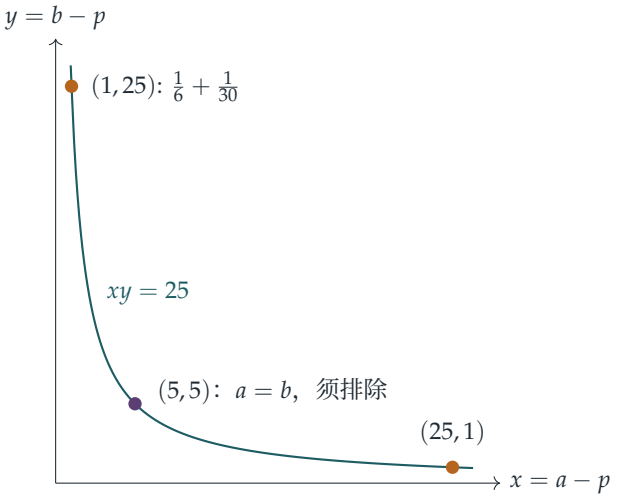
- (p,p) 给出 $a=b=2p$，即 $\frac{1}{p}=\frac{1}{2p}+\frac{1}{2p}$：两个单位分数并不相异，所以这个情况须排除。
- $(1,p^2)$ 给出 $a=p+1$，$b=p^2+p=p(p+1)$，这正是上面找到的表示式；$(p^2,1)$ 则是同一对数，只是次序对调。

所以，若不计两项的次序，唯一的写法是

$$\frac{1}{p}=\frac{1}{p+1}+\frac{1}{p(p+1)},$$

这正是所要证明的。

下图说明了个中原因：这些解就是双曲线（hyperbola）$xy=p^2$ 在正象限内的格点（lattice point），其中 $x=a-p$，$y=b-p$。图中 $p=5$。



形如 $2/N$ 的分数

现在假设 $\frac{2}{N} = \frac{1}{a} + \frac{1}{b}$，其中 $N = p$ 是奇质数。两边乘以 abp，得 $2ab = pa + pb$。要配成矩形，我们先乘以 2：

$$4ab - 2pa - 2pb = 0 \iff 4ab - 2pa - 2pb + p^2 = p^2 \iff (2a - p)(2b - p) = p^2.$$

> ? **想一想** 为甚么要先乘以 2 才作因式分解?

若直接把 $2ab - pa - pb$ 因式分解，便会出现 $(a - \frac{p}{2})$ 这类项，而它们并不是整数。乘以 2 之后，一切都保持为整数，所以因式 $2a - p$ 和 $2b - p$ 都是整数，我们便可以运用整除性（divisibility）。

与之前一样，$\frac{1}{a} < \frac{2}{p}$，所以 $2a > p$，即 $2a - p > 0$；同理 $2b - p > 0$。p^2 的正因数是 1、p、p^2，所以 $(2a - p, 2b - p)$ 是以下其中之一：

$$(1, p^2), \qquad (p, p), \qquad (p^2, 1).$$

- (p, p) 给出 $a = b = p$，即 $\frac{2}{p} = \frac{1}{p} + \frac{1}{p}$：两者并不相异，须排除。
- $(1, p^2)$ 给出

  $$a = \frac{p+1}{2}, \qquad b = \frac{p^2 + p}{2} = \frac{p(p+1)}{2}.$$

  由于 p 是奇数，$p + 1$ 是偶数，所以两者都是整数。由于 $p \geq 3$，$a \geq 2$，所以两者都大于 1；而 $b = pa > a$，所以两者相异。$(p^2, 1)$ 的情况是同一对数，只是次序对调。

直接验证：

$$\frac{2}{p+1} + \frac{2}{p(p+1)} = \frac{2(p+1)}{p(p+1)} = \frac{2}{p}.$$

所以 $\frac{2}{p}$ 可以写成两个相异单位分数之和，而且（不计次序）唯一的写法是

$$\frac{2}{p} = \frac{1}{\frac{1}{2}(p+1)} + \frac{1}{\frac{1}{2}p(p+1)},$$

这就证明了最后的结果。

验算

- $p=3$：$\frac{2}{3} = \frac{1}{2} + \frac{1}{6}$；$p=5$：$\frac{2}{5} = \frac{1}{3} + \frac{1}{15}$；$p=7$：$\frac{2}{7} = \frac{1}{4} + \frac{1}{28}$。
- 以电脑搜索所有 $N < 300$，结果发现：对每个质数 N，$\frac{1}{N}$ 恰好有一种表示法；对每个奇质数 N，$\frac{2}{N}$ 也恰好有一种表示法，与上面的公式吻合。
- 质数这个条件确实重要：$N=6$ 有四种表示法，例如 $\frac{1}{6} = \frac{1}{7} + \frac{1}{42} = \frac{1}{8} + \frac{1}{24} = \frac{1}{9} + \frac{1}{18} = \frac{1}{10} + \frac{1}{15}$，每一种对应一个因数分解 $36 = 1\times 36 = 2\times 18 = 3\times 12 = 4\times 9$（$6\times 6$ 这一对则给出 $a=b$）。
- 为甚么要 $p>2$：当 $p=2$ 时，$\frac{2}{2}=1$ 不能写成 $\frac{1}{a}+\frac{1}{b}$ 的形式（其中 $a,b>1$ 且相异），而 $(p+1)/2$ 也确实不是整数。

常见错误

- 忘记证明 $a-N>0$ 和 $b-N>0$，因而忽略了 N^2 的负因数分解。
- 没有排除 $N\times N$ 这个因数分解，它给出的是两个相等的单位分数。
- 在 $2/N$ 的情况中，没有检查 $\frac{1}{2}(p+1)$ 是否整数（这正是用到 $p\neq 2$ 的地方）。

STEP 之外

把分数写成相异单位分数之和，可追溯至古埃及人，而“埃及分数”（Egyptian fractions）至今仍有未解决的问题：艾狄胥–施特劳斯猜想（Erdős–Straus conjecture）指出，对所有 $N\geq 2$，$\frac{4}{N}$ 总可以写成三个单位分数之和，但至今无人能够证明。把 $xy - Nx - Ny = 0$ 化为 $(x-N)(y-N) = N^2$ 的技巧，是处理丢番图方程（Diophantine equation）的标准工具，而数算解的个数就变成数算因数的个数。Davenport 的 *The Higher Arithmetic*（一本数论入门书）是进一步认识因数与质数的上佳读物。

详细解答

STEP II 2000，第 2 题

纯数学 · 多项式的重根

STEP II 2000，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/00-s2-q2

关键思路

重复因式（repeated factor）经过求导后仍然保留，只是每求导一次，幂次便减一。因此，四次幂的因式 $(x-a)^4$ 会迫使 p 及其首三阶导数（derivative）在 a 处全部等于零。对于这个六次多项式（sextic），三阶导数只是一个三次式，它的根给出一张很短的 a 的候选名单。我们逐一检验这些候选值，最后才决定 k，因为一求导，k 便会消失。

解答

第一个结果

假设 $(x-a)^2$ 是因式，即对某个多项式（polynomial）q 有 $\mathrm{p}(x)=(x-a)^2\,\mathrm{q}(x)$。由积法则（product rule），

$$\mathrm{p}'(x)=2(x-a)\,\mathrm{q}(x)+(x-a)^2\,\mathrm{q}'(x)=(x-a)\big[\,2\mathrm{q}(x)+(x-a)\mathrm{q}'(x)\,\big].$$

代入 $x=a$ 便得 $\mathrm{p}'(a)=0$，正是所求。

相应的结果

同样的计算适用于任何幂次。若 $\mathrm{f}(x)=(x-a)^m\,\mathrm{g}(x)$，其中 $m\geq 1$，而 g 是多项式，则

$$\mathrm{f}'(x)=(x-a)^{m-1}\big[\,m\,\mathrm{g}(x)+(x-a)\,\mathrm{g}'(x)\,\big],$$

而方括号内的式子仍然是多项式。所以若 $(x-a)^m$ 整除 f，则 $(x-a)^{m-1}$ 整除 f'。

现设 $(x-a)^4$ 是 $\mathrm{p}(x)$ 的因式。把上述事实连用三次，

$$(x-a)^3 \text{ 整除}\mathrm{p}'(x),\qquad (x-a)^2 \text{ 整除}\mathrm{p}''(x),\qquad (x-a) \text{ 整除}\mathrm{p}'''(x).$$

再加上 $(x-a)^4$ 整除 $\mathrm{p}(x)$ 本身，代入 $x=a$ 便得

$$\mathrm{p}(a)=\mathrm{p}'(a)=\mathrm{p}''(a)=\mathrm{p}'''(a)=0.$$

求 k

设 $\mathrm{p}(x)=x^6+4x^5-5x^4-40x^3-40x^2+32x+k$。求导三次，

$$\begin{aligned}\mathrm{p}'(x)&=6x^5+20x^4-20x^3-120x^2-80x+32,\\ \mathrm{p}''(x)&=30x^4+80x^3-60x^2-240x-80=10\,(3x^4+8x^3-6x^2-24x-8),\\ \mathrm{p}'''(x)&=120x^3+240x^2-120x-240=120\,(x^3+2x^2-x-2).\end{aligned}$$

? **想一想** 四个条件之中，应该先用哪一个?

先用最容易解的那一个。常数 k 已从 p''' 中消失，而且 p''' 只是三次式，可用分组法因式分解（factorise）：

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

所以根据相应的结果，a 必定是 -2、1 或 -1。我们把它们逐一代入 p'' 检验：

$$\begin{aligned}\tfrac{1}{10}\mathrm{p}''(1)&=3+8-6-24-8=-27\neq 0,\\ \tfrac{1}{10}\mathrm{p}''(-1)&=3-8-6+24-8=5\neq 0,\\ \tfrac{1}{10}\mathrm{p}''(-2)&=48-64-24+48-8=0.\end{aligned}$$

只有 $a=-2$ 过关。它同样通过 p' 的检验：

$$\mathrm{p}'(-2)=-192+320+160-480+160+32=0.$$

最后，由 $\mathrm{p}(-2)=0$ 便可确定 k：

$$\mathrm{p}(-2)=64-128-80+320-160-64+k=k-48=0,\qquad \text{所以}\qquad \boxed{k=48}.$$

? **想一想** 我们只用了*必要*条件。怎样知道当 $k=48$ 时，$(x+2)^4$ 真的是因式?

事实上，$\mathrm{p}(a)=\mathrm{p}'(a)=\mathrm{p}''(a)=\mathrm{p}'''(a)=0$ 这四个条件确实蕴含 $(x-a)^4$ 是因式（由多项式的泰勒定理（Taylor's theorem）可得）。不过，最快捷的确认方法是直接乘开。由于 $(x+2)^4=x^4+8x^3+24x^2+32x+16$，商式必定是首项系数为一（monic）的二次式，其常数项为 $48/16=3$。比较 x^5 的系数，得出 x^2-4x+3，而且确实有

$$(x+2)^4(x^2-4x+3)=x^6+4x^5-5x^4-40x^3-40x^2+32x+48.$$

所以，当 $k = 48$ 时，

$$x^6 + 4x^5 - 5x^4 - 40x^3 - 40x^2 + 32x + 48 = (x+2)^4(x-1)(x-3),$$

而这个因式符合所要求的形式，其中 $a = -2$。

验算

- 把 $(x+2)^4(x-1)(x-3)$ 完全展开：每个系数都与题目给出的多项式相符，而常数项是 $16 \times 3 = 48$。
- 检验系数之和：$\mathrm{p}(1) = 1 + 4 - 5 - 40 - 40 + 32 + 48 = 0$，这是必然的，因为 $(x-1)$ 是因式。
- 我们用电脑在随机选取的多项式 $(x-a)^4\mathrm{q}(x)$ 上验证了一般结果，求得 p''' 的根为 $-2, \pm 1$，而 p'' 只在 -2 处等于零，并得出 $k = 48$。

常见错误

- 从 $\mathrm{p}(a) = 0$ 或 $\mathrm{p}'(a) = 0$ 入手。这两个条件牵涉五次式或 k 本身，难以迅速得出结果。应从次数最低的导数开始。
- 找到 p''' 的根便停下来。候选值 $a = \pm 1$ 是 p''' 的根，却不是 p'' 的根。
- 只把相应的结果写成“$\mathrm{p}'''(a) = 0$”。应写出完整的一组条件 $\mathrm{p}(a) = \mathrm{p}'(a) = \mathrm{p}''(a) = \mathrm{p}'''(a) = 0$，并加以证明。

STEP 之外

逆命题（converse）同样成立：$(x-a)^m$ 整除 p，当且仅当 p 及其首 $m-1$ 阶导数在 a 处都等于零。这是因为多项式等于它自身在 a 处的泰勒展开式。在大学一年级的代数课程中，同一想法用来定义根的*重数*（multiplicity）。它还提供了一个判别重根的方法：用欧几里得算法（Euclid's algorithm）计算 $\gcd(\mathrm{p}, \mathrm{p}')$。在某些域（field）上，例如模质数的整数，形式导数（formal derivative）会有出人意表的表现。Allenby 的 *Rings, Fields and Groups*（环、域与群）对此有详细阐述。

详细解答

STEP II 2000，第 3 题

纯数学 • 误差传递

STEP II 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-s2-q3

关键思路

余弦定理（cosine rule）把 a^2 准确地用 b、c 和 A 表示出来，所以整道题要问的，就是当这三个量各自稍作变动时，a 会怎样随之改变。把每个组成部分都按微小量展开至一阶（first order），舍去微小量之间的乘积，最后利用 $\sqrt{1+x} \approx 1 + \frac{1}{2}x$ 开平方。至于取值范围，留意 η 是若干个独立误差的线性组合（linear combination），所以当每一项都取到极端值，而且正负号都取得有利时，它便达到最大。

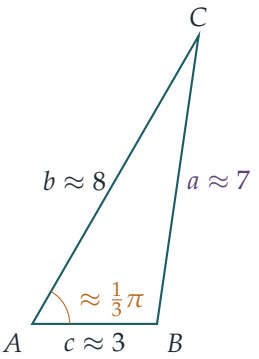
解答

a 的近似式

取未受扰动的数值 $b = 8$、$c = 3$、$A = \frac{1}{3}\pi$，余弦定理给出

$$a^2 = 64 + 9 - 2 \cdot 8 \cdot 3 \cdot \tfrac{1}{2} = 49,$$

所以 $a = 7$。这就解释了首项的来历。



现在加入扰动。以下“≈”一律表示只保留 $\epsilon_1, \epsilon_2, \epsilon_3$ 的一阶项，而舍去所有二次或更高次的项，例如 ϵ_1^2 或 $\epsilon_2\epsilon_3$。于是

$$\begin{aligned} b^2 &= 64 + 16\epsilon_1 + \epsilon_1^2 \approx 64 + 16\epsilon_1, \\ c^2 &= 9 + 6\epsilon_2 + \epsilon_2^2 \approx 9 + 6\epsilon_2, \\ bc &= 24 + 3\epsilon_1 + 8\epsilon_2 + \epsilon_1\epsilon_2 \approx 24 + 3\epsilon_1 + 8\epsilon_2. \end{aligned}$$

至于角度，由复角公式（compound-angle formula）和小角近似（small-angle approximation）$\cos\epsilon_3 \approx 1$、$\sin\epsilon_3 \approx \epsilon_3$，可得

$$\cos A = \cos\tfrac{1}{3}\pi\cos\epsilon_3 - \sin\tfrac{1}{3}\pi\sin\epsilon_3 \approx \tfrac{1}{2} - \tfrac{\sqrt{3}}{2}\epsilon_3.$$

（$\cos\epsilon_3 \approx 1$ 的误差属于 ϵ_3^2 的数量级，所以在这一阶的精确度下确实可以忽略。）把括号乘开，并再次舍去微小量之间的乘积，

$$2bc\cos A \approx 2\big(24 + 3\epsilon_1 + 8\epsilon_2\big)\Big(\tfrac{1}{2} - \tfrac{\sqrt{3}}{2}\epsilon_3\Big) \approx 24 + 3\epsilon_1 + 8\epsilon_2 - 24\sqrt{3}\,\epsilon_3.$$

把各部分代入余弦定理，

$$\begin{aligned} a^2 &\approx (64 + 16\epsilon_1) + (9 + 6\epsilon_2) - (24 + 3\epsilon_1 + 8\epsilon_2 - 24\sqrt{3}\,\epsilon_3) \\ &= 49 + 13\epsilon_1 - 2\epsilon_2 + 24\sqrt{3}\,\epsilon_3. \end{aligned}$$

设 $\delta = 13\epsilon_1 - 2\epsilon_2 + 24\sqrt{3}\,\epsilon_3$，这是一个微小量。由于 $a > 0$，

$$a = \sqrt{49 + \delta} = 7\Big(1 + \frac{\delta}{49}\Big)^{1/2} \approx 7\Big(1 + \frac{\delta}{98}\Big) = 7 + \frac{\delta}{14},$$

这里用了二项式展开（binomial expansion）$(1 + x)^{1/2} \approx 1 + \frac{1}{2}x$，适用于微小的 x。因此 $a \approx 7 + \eta$，其中

$$\eta = \frac{13\epsilon_1 - 2\epsilon_2 + 24\sqrt{3}\,\epsilon_3}{14},$$

这正是所求的结果。

想一想 我们能否避免逐项手算这些乘积的展开?

可以：对 $a^2 = b^2 + c^2 - 2bc\cos A$ 求微分，得到 $2a\,\delta a = (2b - 2c\cos A)\,\delta b + (2c - 2b\cos A)\,\delta c + 2bc\sin A\,\delta A$。在 $a = 7$、$b = 8$、$c = 3$、$A = \frac{1}{3}\pi$ 时，上式变成 $14\,\delta a = 13\,\delta b - 2\,\delta c + 24\sqrt{3}\,\delta A$，一行便得出相同的答案。两种方法其实都只是一阶泰勒展开（Taylor expansion）。

η 的取值范围

试卷上 ϵ_2 的界限印作 $4\cdot 9\times 10^{-2}$，用的是旧式的居中小数点，所以它的意思是 $|\epsilon_2| \le 4.9\times 10^{-2}$。

三个误差各自独立地变化，所以由三角不等式（triangle inequality），

$$|\eta| \le \frac{13|\epsilon_1| + 2|\epsilon_2| + 24\sqrt{3}\,|\epsilon_3|}{14} \le \frac{13\times 2\times 10^{-3} + 2\times 4.9\times 10^{-2} + 24\sqrt{3}\times\sqrt{3}\times 10^{-3}}{14}.$$

分子是 $0.026 + 0.098 + 0.072 = 0.196$，所以

$$|\eta| \le \frac{0.196}{14} = 0.014.$$

这个界限是可以达到的：取 $\epsilon_1 = 2\times 10^{-3}$、$\epsilon_2 = -4.9\times 10^{-2}$、$\epsilon_3 = \sqrt{3}\times 10^{-3}$，便令 η 的每一项都是正数，而且尽可能大，得出 $\eta = 0.014$；把三个正负号全部反过来，则得出 $\eta = -0.014$。由于 η 是各误差的连续函数（continuous function）（其实更是线性函数），而容许的误差所组成的集合是一个长方体，所以两者之间的每一个值也都能取到。因此取值范围是

$$-0.014 \le \eta \le 0.014,$$

也就是说，a 大约落在 7 ± 0.014 之内。

验算

- 当所有误差都是零时，三角形的三边是 8、3、7，并有一个 60° 的角，而确实有 $64 + 9 - 24 = 49$。
- 取上述的极端误差，准确的余弦定理给出 $a \approx 7.0141$ 和 $a \approx 6.9861$，与 7 ± 0.014 非常接近；当中的微小差异来自被舍去的二阶项。
- 以数值计算，a 对 A 的偏导数（partial derivative）在基准值处是 $2.9692\ldots = 24\sqrt{3}/14$，与预期相符。

常见错误

- 忘记最后要开平方：$a^2 \approx 49 + \delta$ 给出的是 $a \approx 7 + \delta/14$，而不是 $7 + \delta$。
- 求最大值时把三个误差都取正数：ϵ_2 前面带负号，所以它必须取负数才能把 η 推高。
- 把 $4 \cdot 9 \times 10^{-2}$ 读成 36×10^{-2}；对于一条长度为 3 的边来说，这根本算不上是微小的误差。

STEP 之外

这就是通过全微分（total differential）进行的误差传播（error propagation）：对于多变量函数 f，有 $\delta f \approx \sum(\partial f/\partial x_i)\,\delta x_i$，而最坏的情况是 $\sum |\partial f/\partial x_i|\,|\delta x_i|$。在大学一年级的课程中，这会发展成多变量的泰勒定理（Taylor's theorem），并附有二阶余项（remainder）的明确界限，让你确切知道“≈”到底有多准确。实验科学家则往往改用适用于独立随机误差的方和根法则（root-sum-of-squares rule）。Bartle 与 Sherbert 的 *Introduction to Real Analysis*（实分析入门）对带余项的泰勒定理有严谨细致的讲解。

详细解答

STEP II 2000，第 4 题

纯数学 · 辐角与反正切恒等式

STEP II 2000，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/00-s2-q4

关键思路

头两个结果说明：复数（complex number）相乘时，它们的辐角（argument）相加。因此，形如“若干个反正切（arctangent）之和 $= \pi/4$”的恒等式，其实是关于 $5-\mathrm{i}$ 和 $1+\mathrm{i}$ 这类高斯整数（Gaussian integer）之乘积的命题：每个满足 $a>0$ 的因子 $a+b\mathrm{i}$ 贡献一个角 $\arctan(b/a)$，而直线 $y=x$ 上任何一点的辐角都是 $\pi/4$。要做的工作是小心地把乘积展开，然后验证两个角确实相等，而不仅仅是相差 2π 的某个倍数。

解答

乘积公式与棣美弗定理

展开并利用 $\mathrm{i}^2=-1$，得

$$\begin{aligned}(\cos\theta+\mathrm{i}\sin\theta)(\cos\phi+\mathrm{i}\sin\phi) &= (\cos\theta\cos\phi-\sin\theta\sin\phi)+\mathrm{i}(\sin\theta\cos\phi+\cos\theta\sin\phi)\\ &= \cos(\theta+\phi)+\mathrm{i}\sin(\theta+\phi),\end{aligned}$$

其中用到 $\cos(A+B)$ 和 $\sin(A+B)$ 的复角公式（compound-angle formulae）。

第二个结果（即棣美弗定理，de Moivre's theorem）我们对 n 用数学归纳法（induction）证明。当 $n=1$ 时，两边都是 $\cos\theta+\mathrm{i}\sin\theta$。假设结果对 $n=k$ 成立，则

$$\begin{aligned}(\cos\theta+\mathrm{i}\sin\theta)^{k+1} &= (\cos k\theta+\mathrm{i}\sin k\theta)(\cos\theta+\mathrm{i}\sin\theta)\\ &= \cos(k\theta+\theta)+\mathrm{i}\sin(k\theta+\theta),\end{aligned}$$

这里先用归纳假设（induction hypothesis），再用取 $\phi=k\theta$ 的乘积公式。所以结果对 $n=k+1$ 成立，因而对每个正整数 n 都成立。

为方便起见，我们把 $\cos\theta + \mathrm{i}\sin\theta$ 记作 $\operatorname{cis}\theta$。于是我们已证明 $\operatorname{cis}\theta\operatorname{cis}\phi = \operatorname{cis}(\theta+\phi)$ 和 $(\operatorname{cis}\theta)^n = \operatorname{cis} n\theta$。

关于角的一个事实

若 $a > 0$ 而 b 为实数，则

$$a + b\mathrm{i} = \sqrt{a^2+b^2}\,\operatorname{cis}\big(\arctan(b/a)\big),$$

这是因为角 $\alpha = \arctan(b/a)$ 位于 $(-\frac{\pi}{2}, \frac{\pi}{2})$ 之内，所以 $\cos\alpha > 0$，而且 $\tan\alpha = b/a$；这两个条件决定了 $\cos\alpha = a/\sqrt{a^2+b^2}$ 及 $\sin\alpha = b/\sqrt{a^2+b^2}$。

此外，若 $R\operatorname{cis} S = R'\operatorname{cis} T$ 且 $R, R' > 0$，则取模（modulus）可得 $R = R'$，所以 $\cos S = \cos T$ 及 $\sin S = \sin T$，因此 $S - T$ 是 2π 的倍数。

第一个恒等式

设 $\alpha = \arctan\frac{1}{5}$。则 $5 - \mathrm{i} = \sqrt{26}\operatorname{cis}(-\alpha)$ 及 $1 + \mathrm{i} = \sqrt{2}\operatorname{cis}\frac{\pi}{4}$，所以由上面的结果得

$$(5-\mathrm{i})^2(1+\mathrm{i}) = 26\sqrt{2}\,\operatorname{cis}\big(\tfrac{\pi}{4} - 2\alpha\big).$$

另一方面，直接展开可得

$$(5-\mathrm{i})^2 = 24 - 10\mathrm{i}, \qquad (24-10\mathrm{i})(1+\mathrm{i}) = 24 + 24\mathrm{i} - 10\mathrm{i} + 10 = 34 + 14\mathrm{i},$$

而 $34 + 14\mathrm{i} = \sqrt{1352}\operatorname{cis}\beta$，其中 $\beta = \arctan\frac{14}{34} = \arctan\frac{7}{17}$。两相比较，$\frac{\pi}{4} - 2\alpha$ 与 β 相差 2π 的某个倍数。

> **想一想** 我们怎样知道这个 2π 的倍数是零？

我们把两个角的范围都确定下来。由于 $0 < \frac{1}{5} < 1$，我们有 $0 < \alpha < \frac{\pi}{4}$，所以 $-\frac{\pi}{4} < \frac{\pi}{4} - 2\alpha < \frac{\pi}{4}$。另外 $0 < \beta < \frac{\pi}{2}$。$(-\frac{\pi}{4}, \frac{\pi}{2})$ 内的两个角相差必定小于 2π，所以该倍数是零，即 $\beta = \frac{\pi}{4} - 2\alpha$。也就是说，

$$\arctan\tfrac{7}{17} + 2\arctan\tfrac{1}{5} = \frac{\pi}{4},$$

这正是所要证明的。

第二个恒等式

同样的想法提示我们考虑 $(4+\mathrm{i})^3(20+\mathrm{i})(1985+\mathrm{i})$，它的辐角应该就是等式的左边。逐步相乘：

$$\begin{aligned}(4+\mathrm{i})^2 &= 15 + 8\mathrm{i},\\ (4+\mathrm{i})^3 &= (15+8\mathrm{i})(4+\mathrm{i}) = 52 + 47\mathrm{i},\\ (4+\mathrm{i})^3(20+\mathrm{i}) &= (52+47\mathrm{i})(20+\mathrm{i}) = 993 + 992\mathrm{i}.\end{aligned}$$

这已经非常接近直线 $y = x$，而最后一个因子正好完成余下的工作：

$$\begin{aligned}(993 + 992\mathrm{i})(1985 + \mathrm{i}) &= (993 \times 1985 - 992) + (993 + 992 \times 1985)\mathrm{i} \\ &= 1970113 + 1970113\,\mathrm{i} = 1970113\sqrt{2}\ \mathrm{cis}\ \tfrac{\pi}{4}.\end{aligned}$$

（如果你好奇 1985 从何而来：要求 $(993 + 992\mathrm{i})(a + \mathrm{i})$ 的实部与虚部相等，便得 $993a - 992 = 992a + 993$，所以 $a = 1985$。）

现在设 $p = \arctan\frac{1}{4}$、$q = \arctan\frac{1}{20}$ 及 $r = \arctan\frac{1}{1985}$。根据上面关于角的事实，$4 + \mathrm{i}$、$20 + \mathrm{i}$、$1985 + \mathrm{i}$ 分别是 $\mathrm{cis}\,p$、$\mathrm{cis}\,q$、$\mathrm{cis}\,r$ 的正实数倍，所以它们的乘积是 $\mathrm{cis}(3p + q + r)$ 的正实数倍。与上一行比较，$3p + q + r - \frac{\pi}{4}$ 是 2π 的倍数。

p, q, r 各自位于 $(0, \frac{\pi}{4})$ 之内，因为它们的正切值都位于 $(0, 1)$ 之内，所以 $0 < 3p + q + r < \frac{5\pi}{4}$。因此 $3p + q + r$ 与 $\frac{\pi}{4}$ 相差小于 2π，该倍数是零，于是

$$3\arctan\tfrac{1}{4} + \arctan\tfrac{1}{20} + \arctan\tfrac{1}{1985} = \frac{\pi}{4}.$$

验算

模必须吻合：$|34 + 14\mathrm{i}|^2 = 1156 + 196 = 1352 = 26^2 \times 2$，正如所料。至于第二个乘积，$17^3 \times 401 \times (1985^2 + 1)$ 应该等于 2×1970113^2，而事实确是如此。在数值上，用计算机计算，两个等式的左边都等于 $0.785398163\ldots = \pi/4$；例如 $\arctan\frac{7}{17} \approx 0.390607$ 而 $\frac{\pi}{4} - 2\arctan\frac{1}{5} \approx 0.390607$。

常见错误

- 由复数相等直接推断角相等，而没有排除两者相差 $2\pi k$ 的可能。关于范围的论证是证明中不可或缺的一环。
- 计算 $(5 - \mathrm{i})^2$ 时出现正负号错误：它等于 $24 - 10\mathrm{i}$，所以 $5 - \mathrm{i}$ 贡献的是 $-\arctan\frac{1}{5}$，这就是为甚么该反正切移到另一边后带着正号。
- 计算大数乘积时出现算术错误；可如上文所述，用模来验算。

STEP 之外

这类恒等式称为梅钦类公式（Machin-type formulae）；梅钦（John Machin）的 $\frac{\pi}{4} = 4\arctan\frac{1}{5} - \arctan\frac{1}{239}$ 是用同样的方法由 $(5 + \mathrm{i})^4(239 - \mathrm{i})$ 得出的，并在 1706 年用来把 π 计算到小数点后一百位。其背后的结构是高斯整数 $\mathbb{Z}[\mathrm{i}]$ 的算术，当中分解为高斯质数（Gaussian primes）的唯一分解性（unique factorisation）解释了哪些这类恒等式有可能

存在。想了解乘法作为旋转的几何意义，可阅读 Needham 的 *Visual Complex Analysis*（以图像方式讲解复分析）；想了解高斯整数，可阅读 Davenport 的 *The Higher Arithmetic*（数论入门）。

详细解答

STEP II 2000，第 5 题

纯数学 · 最小二乘近似

STEP II 2000，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/00-s2-q5

关键思路

把平方展开之后，要求最小值的量其实只是 λ 的二次式（quadratic），其系数都是固定的数（涉及 f 的积分）。只要配方（completing the square），就能一次过求出最佳的 λ 和最小的误差。至于正弦函数的例子，要点在于当 n 很大时，$\sin(\pi x/n)$ 几乎就是一条通过原点的直线，而小角近似（small-angle approximations）正好把这个说法精确化。

解答

λ 的最佳值

设

$$J = \int_0^1 x f(x)\,\mathrm{d}x, \qquad K = \int_0^1 \big(f(x)\big)^2\,\mathrm{d}x.$$

这两个都是常数，与 λ 无关。把平方展开，并利用 $\int_0^1 x^2\,\mathrm{d}x = \frac{1}{3}$，得

$$\begin{aligned} I(\lambda) = \int_0^1 \big(f(x) - \lambda x\big)^2\,\mathrm{d}x &= K - 2\lambda J + \tfrac{1}{3}\lambda^2 \\ &= \tfrac{1}{3}\big(\lambda - 3J\big)^2 + K - 3J^2. \end{aligned}$$

平方项永不为负，而且只在 $\lambda = 3J$ 时等于零，所以 $I(\lambda)$ 恰好在

$$\lambda = 3\int_0^1 x f(x)\,\mathrm{d}x,$$

时取得最小值，这正是所求。（等价地，由 $I'(\lambda) = -2J + \frac{2}{3}\lambda = 0$ 可得 $\lambda = 3J$，而 $I''(\lambda) = \frac{2}{3} > 0$ 确认这是极小值。）

残差

R^2 是 I 的最小值，即残差（residual error）的平方；由配方后的式子可见，这个最小值是 $K-3J^2$。由于 $J=\lambda/3$，我们有 $3J^2=\frac{1}{3}\lambda^2$，所以

$$R^2=\int_0^1 \big(f(x)\big)^2\,\mathrm{d}x-\tfrac{1}{3}\lambda^2,$$

这正是所求。（如果你是用求导的方法找出 λ，可把 $J=\lambda/3$ 代入 $K-2\lambda J+\frac{1}{3}\lambda^2$，得到 $K-\frac{2}{3}\lambda^2+\frac{1}{3}\lambda^2$，结果相同。）

$f(x)=\sin(\pi x/n)$ 的情况

为了令代数运算较为简洁，设 $a=\pi/n$，于是 $f(x)=\sin ax$，而且当 $n\to\infty$ 时 $a\to 0$。取 $u=x$ 和 $\mathrm{d}v=\sin ax\,\mathrm{d}x$ 作分部积分（integration by parts），

$$\begin{aligned}\int_0^1 x\sin ax\,\mathrm{d}x&=\Big[-\frac{x\cos ax}{a}\Big]_0^1+\frac{1}{a}\int_0^1\cos ax\,\mathrm{d}x\\&=-\frac{\cos a}{a}+\frac{\sin a}{a^2},\end{aligned}$$

所以

$$\lambda=\frac{3(\sin a-a\cos a)}{a^2}.$$

想一想 为甚么 $\sin\theta$ 的近似式中需要保留 θ^3 项，而不能只用 $\sin\theta\approx\theta$?

如果只用 $\sin a\approx a$ 和 $\cos a\approx 1$，分子就变成 $a-a=0$，这甚么也没有告诉我们：首项互相抵消了，答案藏在下一阶的项之中。把提示中的两个近似式都用上，

$$\sin a-a\cos a\approx\Big(a-\tfrac{1}{6}a^3\Big)-a\Big(1-\tfrac{1}{2}a^2\Big)=\tfrac{1}{3}a^3,$$

因此当 n 很大时，

$$\lambda\approx\frac{3\cdot\frac{1}{3}a^3}{a^2}=a=\frac{\pi}{n}$$

这就是结果 (i)。

至于 (ii)，我们还需要 K。利用 $\sin^2 ax=\frac{1}{2}(1-\cos 2ax)$，

$$K=\int_0^1\sin^2 ax\,\mathrm{d}x=\frac{1}{2}-\frac{\sin 2a}{4a}.$$

当 $n\to\infty$ 时，$a\to 0$ 而 $\dfrac{\sin 2a}{2a}\to 1$，所以 $K\to\frac{1}{2}-\frac{1}{2}=0$。同时 $\lambda\to 0$，因为 $\lambda\approx\pi/n$。故此

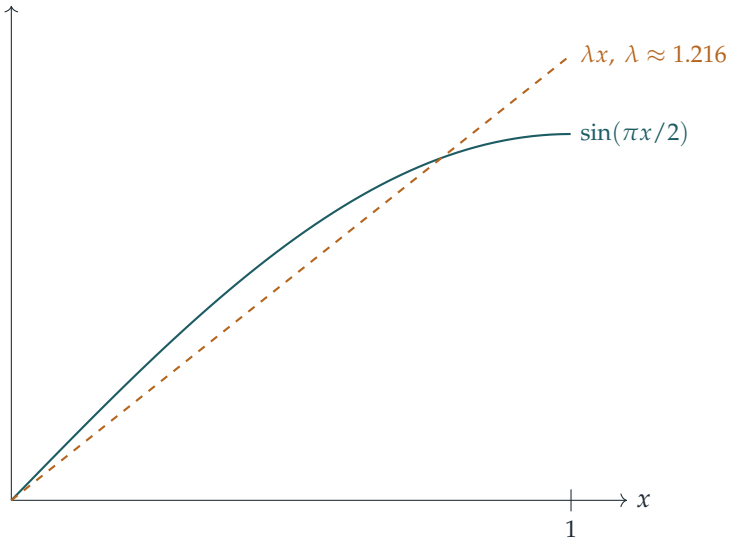
$$R^2=K-\tfrac{1}{3}\lambda^2\to 0-0=0,$$

又因为 $R\ge 0$，我们得到 $\lim_{n\to\infty}R=0$。这就是结果 (ii)。

事实上，这里的抵消远比上述的更彻底。利用 $\sin 2a \approx 2a - \frac{4}{3}a^3$，我们得到 $K \approx \frac{1}{2} - \left(\frac{1}{2} - \frac{1}{3}a^2\right) = \frac{1}{3}a^2$，而 $\frac{1}{3}\lambda^2 \approx \frac{1}{3}a^2$ 亦然。所以 R^2 不单是小：与 K（即 f 本身的大小）相比，它也是小的。这条直线几乎完全捕捉了 f。

为何这些结果在意料之中

在 $0 \le x \le 1$ 上，角 $\pi x/n$ 介乎 0 与 π/n 之间，当 n 很大时这个角很小。所以在整个区间上 $\sin(\pi x/n) \approx \pi x/n$：这个函数本身就非常接近线性函数 λx，其中 $\lambda = \pi/n$。因此，最佳拟合直线的斜率（gradient）应该接近 π/n，而拟合应该变得近乎完美，所以残差应该趋向零。下图显示，即使 $n = 2$，曲线与直线也相差不远。



验算

- 如果本来就有 $f(x) = cx$，则 $\lambda = 3c\int_0^1 x^2\,\mathrm{d}x = c$ 而 $R^2 = \frac{1}{3}c^2 - \frac{1}{3}c^2 = 0$，正如所料。
- 以数值计算，当 $n = 10$ 时，精确公式给出 $\lambda = 0.31107$，而 $\pi/10 = 0.31416$；当 $n = 100$ 时，比值 $\lambda/(\pi/n)$ 为 0.99990。
- 当 $n = 10$ 时，数值积分（quadrature）给出 $R^2 \approx 6.0 \times 10^{-7}$，而 $K \approx 0.032$，所以与 f 相比，残差确实微不足道。

常见错误

- 停在 $\sin a \approx a$、$\cos a \approx 1$ 而得到 $\lambda \approx 0$：这是对的，但毫无用处。要保留三次项。

▶ 忘记了 $\lambda = 3\int_0^1 xf(x)\,\mathrm{d}x$ 中的 3 是来自 $\int_0^1 x^2\,\mathrm{d}x = \frac{1}{3}$。

STEP 之外

这就是最小二乘逼近（least-squares approximation），而它本质上是几何。把 $[0,1]$ 上的函数看成向量，并以 $\langle f,g\rangle = \int_0^1 fg\,\mathrm{d}x$ 作为“点积”（dot product）。那么 λx 就是 f 在 x 所张成的直线上的正交投影（orthogonal projection），而 $R^2 = K - \frac{1}{3}\lambda^2$ 就是毕氏定理（Pythagoras' theorem）。把展开式再推进一步，可得 $R^2 \approx a^6/1575$，这来自把 x^3 项中沿 x 方向的分量投影掉。Strang 的 *Introduction to Linear Algebra*（线性代数入门）详细阐述了投影与最小二乘法，而同样的想法更可引伸至傅里叶级数（Fourier series）。

详细解答

STEP II 2000，第 6 题

纯数学 · 半角代换

STEP II 2000，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/00-s2-q6

关键思路

代换（substitution）$t = \tan\frac{1}{2}\theta$ 可以把任何由 $\sin\theta$ 和 $\cos\theta$ 组成的有理函数（rational function）化为 t 的有理函数，而在这一题，结果是 2 除以一个二次式。配方（completing the square）之后得到反正切（arctangent），而第一部分的第三个恒等式（identity）正好用来化简上限处的反正切。第二个积分则可凭对称性（symmetry）由第一个积分直接得出，毋须再做新的积分。

解答

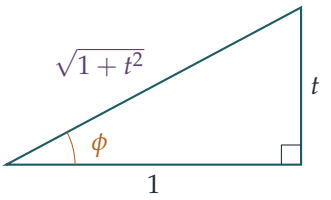
三个恒等式

设 $\phi = \frac{1}{2}\theta$，则 $t = \tan\phi$ 且 $1 + t^2 = \sec^2\phi$。由二倍角公式（double-angle formulae），

$$\sin\theta = 2\sin\phi\cos\phi = 2\tan\phi\cos^2\phi = \frac{2\tan\phi}{\sec^2\phi} = \frac{2t}{1+t^2},$$

$$\cos\theta = \cos^2\phi - \sin^2\phi = \cos^2\phi\,(1 - \tan^2\phi) = \frac{1-t^2}{1+t^2}.$$

（只要 $\tan\frac{1}{2}\theta$ 有定义，即 $\cos\frac{1}{2}\theta \neq 0$，这两条式子便成立。）下图的直角三角形，两条直角边分别为 1 和 t，夹角为 ϕ，是记住这两条公式的好方法。



至于第三个恒等式，当 $\sin\theta \neq 0$ 时，利用 $1+\cos\theta = 2\cos^2\phi$ 和 $\sin\theta = 2\sin\phi\cos\phi$：

$$\frac{1+\cos\theta}{\sin\theta} = \frac{2\cos^2\phi}{2\sin\phi\cos\phi} = \cot\phi = \tan\left(\tfrac{1}{2}\pi - \phi\right) = \tan\left(\tfrac{1}{2}\pi - \tfrac{1}{2}\theta\right),$$

因为对余角（complementary angles）而言，$\cot\phi = \tan(\frac{1}{2}\pi - \phi)$。至此三个恒等式均已证明。

第一个积分

设 $t = \tan\frac{1}{2}\theta$，则 $\theta = 2\arctan t$，所以

$$\mathrm{d}\theta = \frac{2}{1+t^2}\,\mathrm{d}t.$$

当 $\theta = 0$ 时，$t = 0$；当 $\theta = \frac{1}{2}\pi$ 时，$t = \tan\frac{1}{4}\pi = 1$。设 $c = \cos\alpha$ 和 $s = \sin\alpha$；由于 $0 < \alpha < \frac{1}{2}\pi$，两者皆为正数。于是

$$\int_0^{\frac{1}{2}\pi} \frac{\mathrm{d}\theta}{1+c\sin\theta} = \int_0^1 \frac{1}{1+\dfrac{2ct}{1+t^2}} \cdot \frac{2}{1+t^2}\,\mathrm{d}t = \int_0^1 \frac{2}{t^2+2ct+1}\,\mathrm{d}t.$$

配方，并利用 $1-c^2 = s^2$，得

$$t^2 + 2ct + 1 = (t+c)^2 + s^2.$$

由于 $s > 0$，由标准积分 $\int \frac{\mathrm{d}u}{u^2+s^2} = \frac{1}{s}\arctan\frac{u}{s}$ 可得

$$\int_0^1 \frac{2}{(t+c)^2+s^2}\,\mathrm{d}t = \frac{2}{s}\left[\arctan\frac{t+c}{s}\right]_0^1 = \frac{2}{s}\left(\arctan\frac{1+\cos\alpha}{\sin\alpha} - \arctan\frac{\cos\alpha}{\sin\alpha}\right).$$

想一想 怎样化简 $\arctan\dfrac{1+\cos\alpha}{\sin\alpha}$？

第一部分的功夫在这里派上用场。由第三个恒等式（以 α 代替 θ），$\frac{1+\cos\alpha}{\sin\alpha} = \tan(\frac{1}{2}\pi - \frac{1}{2}\alpha)$。不过，在把 arctan 与 tan 互相抵消之前，必须先检查这个角是否位于 arctan 的值域（range）$(-\frac{1}{2}\pi, \frac{1}{2}\pi)$ 之内。由于 $0 < \alpha < \frac{1}{2}\pi$，我们有 $\frac{1}{4}\pi < \frac{1}{2}\pi - \frac{1}{2}\alpha < \frac{1}{2}\pi$，所以确实有

$$\arctan\frac{1+\cos\alpha}{\sin\alpha} = \tfrac{1}{2}\pi - \tfrac{1}{2}\alpha.$$

同理，$\frac{\cos\alpha}{\sin\alpha} = \cot\alpha = \tan(\frac{1}{2}\pi - \alpha)$，而 $0 < \frac{1}{2}\pi - \alpha < \frac{1}{2}\pi$，所以 $\arctan\frac{\cos\alpha}{\sin\alpha} = \frac{1}{2}\pi - \alpha$。因此

$$\int_0^{\frac{1}{2}\pi} \frac{\mathrm{d}\theta}{1+\cos\alpha\sin\theta} = \frac{2}{\sin\alpha}\left(\left(\tfrac{1}{2}\pi - \tfrac{1}{2}\alpha\right) - \left(\tfrac{1}{2}\pi - \alpha\right)\right) = \frac{2}{\sin\alpha}\cdot\frac{\alpha}{2} = \frac{\alpha}{\sin\alpha},$$

此即所求。

第二个积分

代入 $\theta = \frac{1}{2}\pi - u$。于是 $\cos\theta = \sin u$，$\mathrm{d}\theta = -\mathrm{d}u$，而上下限 $0, \frac{1}{2}\pi$ 互换，刚好抵消负号：

$$\int_0^{\frac{1}{2}\pi} \frac{\mathrm{d}\theta}{1+\sin\alpha\cos\theta} = \int_0^{\frac{1}{2}\pi} \frac{\mathrm{d}u}{1+\sin\alpha\sin u}.$$

现在设 $\beta = \frac{1}{2}\pi - \alpha$。于是 $\sin\alpha = \cos\beta$，且 $0 < \beta < \frac{1}{2}\pi$，所以刚才证明的结果可以套用，只需以 β 代替 α：

$$\int_0^{\frac{1}{2}\pi} \frac{\mathrm{d}u}{1+\cos\beta\sin u} = \frac{\beta}{\sin\beta}.$$

由于 $\sin\beta = \cos\alpha$，我们得出结论：对于 $0 < \alpha < \frac{1}{2}\pi$，

$$\int_0^{\frac{1}{2}\pi} \frac{\mathrm{d}\theta}{1+\sin\alpha\cos\theta} = \frac{\frac{1}{2}\pi - \alpha}{\cos\alpha}.$$

验算

- 当 $\alpha \to 0$ 时，$\cos\alpha \to 1$，第一个积分应趋向 $\int_0^{\pi/2} \frac{\mathrm{d}\theta}{1+\sin\theta} = 1$（作代换后得 $\int_0^1 \frac{2}{(1+t)^2}\,\mathrm{d}t = 1$）；而 $\alpha/\sin\alpha \to 1$，果然吻合。
- 当 $\alpha \to \frac{1}{2}\pi$ 时，被积函数（integrand）趋向 1，所以第一个积分应趋向 $\frac{1}{2}\pi$，与 $\frac{\pi/2}{1}$ 相符。同样，当 $\alpha \to \frac{1}{2}\pi$ 时，第二个积分趋向 1，与 $\frac{\frac{1}{2}\pi-\alpha}{\cos\alpha} \to 1$ 一致。
- 数值上，当 $\alpha = \frac{1}{4}\pi$ 时，用辛普森法则（Simpson's rule）算得第一个积分为 1.110721，与 $\frac{\pi\sqrt{2}}{4} \approx 1.110721$ 相符；当 $\alpha = \frac{1}{3}\pi$ 时，第二个积分算得 $1.047198 \approx \frac{1}{3}\pi$，正如公式所预测。

常见错误

- 忘记 $\mathrm{d}\theta$ 所带来的因子 $\frac{2}{1+t^2}$，或忘记把上下限改为 0 和 1。
- 未检查 x 是否位于 $(-\frac{1}{2}\pi, \frac{1}{2}\pi)$ 之内，便写下 $\arctan(\tan x) = x$；在这一题，正是限制条件 $0 < \alpha < \frac{1}{2}\pi$ 令这一步成立。
- 为第二个积分把整个积分过程重做一遍，而不利用 $\theta \mapsto \frac{1}{2}\pi - \theta$ 和 $\alpha \mapsto \frac{1}{2}\pi - \alpha$。

STEP 之外

代换 $t = \tan\frac{1}{2}\theta$（有时称为魏尔斯特拉斯代换，Weierstrass substitution）之所以行得通，是因为点 $(\cos\theta, \sin\theta) = \left(\frac{1-t^2}{1+t^2}, \frac{2t}{1+t^2}\right)$ 是单位圆的一个有理参数化（rational parametrisation）；取 t 为有理数，便可得出所有毕氏三元数组（Pythagorean triple）。在大学里，形

如 $\int_0^{2\pi} R(\cos\theta, \sin\theta)\,\mathrm{d}\theta$ 的积分，通常改为设 $z = e^{i\theta}$ 并运用留数定理（residue theorem）来计算，这样就把问题转化为找出单位圆内的极点（poles）。Needham 的 *Visual Complex Analysis*（《复分析：可视化方法》）是这个想法的一本精彩入门书。

详细解答

STEP II 2000，第 7 题

纯数学 · 直线与圆锥

STEP II 2000，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/00-s2-q7

关键思路

动手做任何事之前，先求出 $\mathbf{s}$ 的长度。$\cos\theta \pm \sqrt{3}$ 这些项是刻意安排的：交叉项会互相抵消，并出现 $\cos^2\theta + \sin^2\theta$，所以 $|\mathbf{s}|$ 完全与 θ 无关。之后的一切都来自纯量积（scalar product）。$s_x - s_z$ 这个组合同样不含 θ，这直接指出了 m 的方向，也解释了为何最后一部分会出现平面 $x - z =$ 常数。

解答

l 与一般直线的夹角

为简便起见，记 $c_\theta = \cos\theta$。$\mathbf{s}$ 的长度为

$$\begin{aligned}|\mathbf{s}|^2 &= (c_\theta + \sqrt{3})^2 + 2\sin^2\theta + (c_\theta - \sqrt{3})^2 \\ &= 2c_\theta^2 + 6 + 2\sin^2\theta = 8,\end{aligned}$$

因为交叉项 $\pm 2\sqrt{3}\,c_\theta$ 互相抵消。所以对任何 θ，都有 $|\mathbf{s}| = 2\sqrt{2}$。

设 $\mathbf{d} = a\mathbf{i} + b\mathbf{j} + c\mathbf{k}$。则

$$\mathbf{s}\cdot\mathbf{d} = a(c_\theta + \sqrt{3}) + \sqrt{2}\,b\sin\theta + c(c_\theta - \sqrt{3}) = (a+c)\cos\theta + \sqrt{2}\,b\sin\theta + \sqrt{3}\,(a-c).$$

因此两条直线的夹角 ϕ 满足

$$\cos\phi = \frac{(a+c)\cos\theta + \sqrt{2}\,b\sin\theta + \sqrt{3}\,(a-c)}{2\sqrt{2}\,\sqrt{a^2+b^2+c^2}}.$$

直线本身没有指定的方向，所以如果你要的是锐角（acute angle），就取右边的绝对值（modulus）。

直线 m

我们要找一个方向 $\mathbf{d}$，使 $\cos\phi$ 与 θ 无关。观察分子可知，恰好当 $a+c=0$ 且 $b=0$ 时，对 θ 的依赖便会消失。所以试取 $\mathbf{d}=\mathbf{i}-\mathbf{k}$（$a=1$，$b=0$，$c=-1$）：

$$\cos\phi=\frac{\sqrt{3}\cdot 2}{2\sqrt{2}\cdot\sqrt{2}}=\frac{\sqrt{3}}{2}.$$

这个值是正数，所以 ϕ 本身已是锐角，而且对任何 θ 都有 $\phi=\pi/6$。因此，通过原点的直线

$$m:\quad \mathbf{r}=\mu(\mathbf{i}-\mathbf{k})$$

符合要求。当 θ 变化时，l 扫出一个圆锥（cone），其轴为 m，半顶角（semi-vertical angle）为 $\pi/6$。

P 描出的圆

在 l 上，$x-z=\lambda(s_x-s_z)=\lambda\big((c_\theta+\sqrt{3})-(c_\theta-\sqrt{3})\big)=2\sqrt{3}\,\lambda$。令此式等于 $4\sqrt{3}$，便得 $\lambda=2$，不论 θ 是甚么，所以

$$\overrightarrow{OP}=2\mathbf{s}=(2\cos\theta+2\sqrt{3})\,\mathbf{i}+2\sqrt{2}\sin\theta\,\mathbf{j}+(2\cos\theta-2\sqrt{3})\,\mathbf{k}.$$

直线 m 与该平面相交于 $\mu-(-\mu)=4\sqrt{3}$ 之处，即 $\mu=2\sqrt{3}$，交点为 C，其中

$$\overrightarrow{OC}=2\sqrt{3}\,\mathbf{i}-2\sqrt{3}\,\mathbf{k}.$$

想一想 为甚么 C 是圆心的自然人选?

平面 $x-z=4\sqrt{3}$ 的法线（normal）是 $\mathbf{i}-\mathbf{k}$，正是 m 的方向。所以 m 是圆锥的轴，它垂直于该平面，而任何圆锥被垂直于其轴的平面所截，都会得到一个圆，圆心就在轴与该平面相交之处。现在我们直接验证这一点：

$$\overrightarrow{CP}=2\cos\theta\,\mathbf{i}+2\sqrt{2}\sin\theta\,\mathbf{j}+2\cos\theta\,\mathbf{k},\qquad |\overrightarrow{CP}|^2=4\cos^2\theta+8\sin^2\theta+4\cos^2\theta=8.$$

所以 P 的每一个位置都在该平面上，并且与平面上的定点 C 相距 $2\sqrt{2}$。换句话说，P 位于该平面上以 C 为圆心、半径为 $2\sqrt{2}$ 的圆上，而 C 在 m 上。

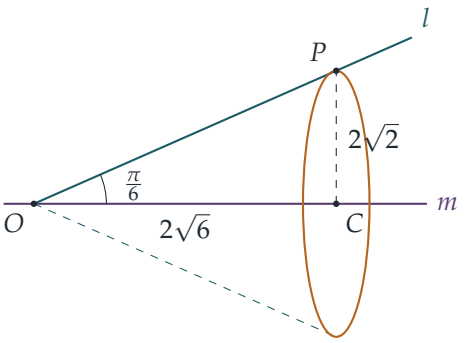
要看出 P 确实描出整个圆，可写成

$$\overrightarrow{CP}=2\sqrt{2}\,(\cos\theta\,\mathbf{u}+\sin\theta\,\mathbf{v}),\qquad \mathbf{u}=\tfrac{1}{\sqrt{2}}(\mathbf{i}+\mathbf{k}),\quad \mathbf{v}=\mathbf{j}.$$

这里 $\mathbf{u}$ 和 $\mathbf{v}$ 是互相垂直的单位向量（unit vector），两者都垂直于 $\mathbf{i}-\mathbf{k}$，所以它们都平行于该平面。当 θ 由 0 变到 2π 时，P 沿圆走了刚好一圈。半径为

$$\boxed{2\sqrt{2}}.$$

求半径还有一个较快的方法：$|OP| = 2|\mathbf{s}| = 4\sqrt{2}$ 而 $|OC| = 2\sqrt{6}$，且 OC 垂直于该平面，所以由毕氏定理（Pythagoras）得 $CP = \sqrt{32-24} = 2\sqrt{2}$。等价地，$CP = OC\tan\frac{\pi}{6} = 2\sqrt{6}/\sqrt{3}$。



验算

- 当 $\theta = 0$ 时，$\mathbf{s} = (1+\sqrt{3}, 0, 1-\sqrt{3})$，其中 $|\mathbf{s}|^2 = (4+2\sqrt{3}) + (4-2\sqrt{3}) = 8$，而 $\mathbf{s}\cdot(\mathbf{i}-\mathbf{k}) = 2\sqrt{3}$，得 $\cos\phi = \sqrt{3}/2$。
- 求半径的两种方法结果一致：$\sqrt{(4\sqrt{2})^2 - (2\sqrt{6})^2} = \sqrt{8}$ 及 $2\sqrt{6}\tan\frac{\pi}{6} = \sqrt{8} \approx 2.828$。
- 数值上，取 20 000 个随机的 θ（一般公式则另取随机的 a, b, c），与 m 的夹角等于 $\pi/6$，P 在该平面上，且 $|CP| = 2.8284\ldots$，误差全部在 10^{-15} 以内。

常见错误

- 忘记除以 $|\mathbf{s}|$，或假设 $\mathbf{s}$ 是单位向量。它的长度是 $2\sqrt{2}$。
- 只证明了 P 在以 O 为球心的球面（sphere）上（即 $|OP| = 4\sqrt{2}$）。球面被平面所截固然得到一个圆，但你仍须说明为何圆心是 m 上的 C：这是因为 OC 垂直于该平面。

STEP 之外

直线 l 扫出一个圆锥，而用平面去截圆锥，正是圆锥曲线（conic sections）名称的由来。垂直于轴的平面截出一个圆；把平面倾斜，便会得到椭圆（ellipse）、抛物线（parabola）或双曲线（hyperbola）。在大学一年级的课程中，你会把圆锥描述为二次曲面（quadric surface）$(\mathbf{r}\cdot\hat{\mathbf{n}})^2 = |\mathbf{r}|^2\cos^2\alpha$，并利用对称矩阵（symmetric matrix）的特征值（eigenvalues）为这类曲面分类。Strang 的 *Introduction to Linear Algebra*（线性代数入门）涵盖了这方面的内容。

详细解答

STEP II 2000，第 8 题

纯数学 · 可分离变量微分方程

STEP II 2000，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/00-s2-q8

关键思路

两条方程都是可分离变量（separable）的：含 x 的部分与含 y 的部分以乘积形式出现，所以可以把所有涉及 y 的项放在一边，所有涉及 x 的项放在另一边。两题中 x 的一边都具有 $x\,\mathrm{e}^{ax^2}$ 的形式，由于 x 是 x^2 的导数（derivative）（的倍数），所以可以立即积分。在第 (ii) 部分，$(y+3)$ 的幂次由我们自行选择，而我们选择的幂次要令 y 的增长恰好与 e^{3x^2} 相符。

解答

第 (i) 部分

由于在 $x=0$ 时 $y=6$，起初附近有 $y+3>0$，因此可以除以 $(y+3)^{1/2}$：

$$(y+3)^{-1/2}\frac{\mathrm{d}y}{\mathrm{d}x}=-4x\,\mathrm{e}^{-x^2}.$$

两边对 x 积分，并利用 $\frac{\mathrm{d}}{\mathrm{d}x}\mathrm{e}^{-x^2}=-2x\,\mathrm{e}^{-x^2}$，得

$$2(y+3)^{1/2}=2\mathrm{e}^{-x^2}+C.$$

代入 $x=0$、$y=6$ 得 $2\cdot 3=2+C$，所以 $C=4$，而

$$(y+3)^{1/2}=\mathrm{e}^{-x^2}+2.$$

右边恒不小于 2，所以它确实是一个合理的（正）平方根，而且 $y+3$ 永远不会达到 0；因此上面的除法对所有 $x\geq 0$ 都成立。两边平方，得

$$\boxed{y=\left(2+\mathrm{e}^{-x^2}\right)^2-3=1+4\mathrm{e}^{-x^2}+\mathrm{e}^{-2x^2}}.$$

当 $x \to \infty$ 时，$\mathrm{e}^{-x^2} \to 0$ 且 $\mathrm{e}^{-2x^2} \to 0$，所以 $y \to 1$，正如题目所求。

第 (ii) 部分

这里的方程是

$$(y+3)^{k-1}\frac{\mathrm{d}y}{\mathrm{d}x} = x\,\mathrm{e}^{6x^2}.$$

要令一般的（可能不是整数的）幂次有意义，我们需要 $y+3>0$；又由于当 $x>0$ 时 $\frac{\mathrm{d}y}{\mathrm{d}x}>0$，$y$ 是递增的，所以只要 $y+3$ 起初是正数，它就会一直保持正数。当 $k \neq 0$ 时，积分得

$$\frac{(y+3)^k}{k} = \frac{\mathrm{e}^{6x^2}}{12} + C, \qquad \text{所以} \qquad (y+3)^k = \frac{k}{12}\,\mathrm{e}^{6x^2} + kC.$$

? **想一想** 我们应该尝试哪一个 k?

若 $k>0$，则当 x 很大时，常数可以忽略，于是 $y+3 \approx (k/12)^{1/k}\,\mathrm{e}^{6x^2/k}$。我们希望 y 的增长与 e^{3x^2} 相若，所以需要 $6/k=3$，即 $k=2$。（任何其他 $k>0$ 都会令 $\mathrm{e}^{-3x^2}y$ 趋向 0 或趋向 ∞。）

因此取 $k=2$。这时方程是 $\frac{\mathrm{d}y}{\mathrm{d}x} = x\,\mathrm{e}^{6x^2}/(y+3)$，其解为

$$(y+3)^2 = \frac{\mathrm{e}^{6x^2}}{6} + D,$$

其中 $D = (y_0+3)^2 - \frac{1}{6}$，而 y_0 是 y 在 $x=0$ 时的值。右边由 $(y_0+3)^2>0$ 开始递增，所以一直保持正数，而解对每一个 $x \geq 0$ 都存在：

$$y = \sqrt{\tfrac{1}{6}\mathrm{e}^{6x^2} + D} \; - 3.$$

乘以 e^{-3x^2}，并把 $\mathrm{e}^{-3x^2} = \sqrt{\mathrm{e}^{-6x^2}}$ 放入根号内，得

$$\mathrm{e}^{-3x^2}y = \sqrt{\tfrac{1}{6} + D\,\mathrm{e}^{-6x^2}} \; - \; 3\mathrm{e}^{-3x^2} \;\longrightarrow\; \sqrt{\tfrac{1}{6}} = \frac{1}{\sqrt{6}}$$

当 $x \to \infty$ 时成立，因为两个指数项都趋向 0。所以当 $k=2$ 时，极限（limit）是

$$\boxed{\lim_{x\to\infty} \mathrm{e}^{-3x^2}y = \frac{1}{\sqrt{6}}},$$

不论初始值是甚么：常数 D 已从答案中消失，这正是题目可以说“任何解”的原因。

当 $k=2$ 时，幂 $(y+3)^{-1}$ 对 $y<-3$ 也有意义。在该分支上 $y+3 = -\sqrt{\frac{1}{6}\mathrm{e}^{6x^2}+D}$，而同样的论证给出极限 $-1/\sqrt{6}$。题目预期的答案是 $1/\sqrt{6}$，对应 $y>-3$ 的解。

验算

- 在 (i) 中，$x = 0$ 给出 $y = (2+1)^2 - 3 = 6$，而对 $y = (2 + \mathrm{e}^{-x^2})^2 - 3$ 求导得 $-4x\mathrm{e}^{-x^2}(2 + \mathrm{e}^{-x^2}) = -4x\mathrm{e}^{-x^2}(y+3)^{1/2}$。
- 以数值方法（龙格—库塔法，Runge–Kutta）逐步求解两条方程，结果与公式吻合；例如在 (i) 中 $y(2) \approx 1.07360$，而在 (ii) 中，由 $y_0 = 6$ 开始，$\mathrm{e}^{-12}y(2) \approx 0.408230$，已经很接近 $1/\sqrt{6} \approx 0.408248$。

常见错误

- 到平方之后才记起要加积分常数（constant of integration），或者把 $y = 6$ 代入了错误的式子。
- 把 $x\,\mathrm{e}^{6x^2}$ 积分成 $\mathrm{e}^{6x^2}/6$ 而不是 $\mathrm{e}^{6x^2}/12$；这不会改变 k，但会改变极限。

STEP 之外

第 (ii) 部分是渐近分析（asymptotic analysis）的一个小例子：我们无须理会初始值，就找出了解在 x 很大时的表现，因为积分常数被不断增长的项所淹没。探究解的哪些特征与初始条件（initial conditions）无关，正是微分方程（differential equations）定性理论的核心问题。Boyce 与 DiPrima 的 *Elementary Differential Equations*（微分方程入门教科书）涵盖可分离变量方程以及这类长期表现的讨论。

详细解答

STEP II 2000，第 9 题

力学 · 有空气阻力的运动

STEP II 2000，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：

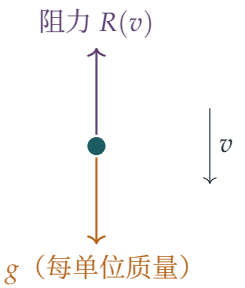
step.maths.org/questions/00-s2-q9

关键思路

每位跳伞者都遵从牛顿第二定律（Newton's second law），而阻力只取决于速率，因此我们得到一条关于 v 和 t 的一阶微分方程（first-order differential equation），而且可以分离变量（separation of variables）。“速率永远达不到”这类结果源自终端速率（terminal speed）：即阻力刚好与重力平衡时的速率。解会慢慢向它靠近，但在有限时间内永远到不了。至于 Karen，只要用无因次（dimensionless）速率 $u = kv/g$ 来表示，那个二次式便可以干净俐落地因式分解。

解答

我们把每位跳伞者看成一个由静止开始垂直下落的质点（particle），重力加速度 g 为常数，方向向下。取 $v \geq 0$ 为向下的速率，并由起跳一刻开始量度 t。以每单位质量计，所受的力是向下的重力 g 和向上的空气阻力（air resistance），所以由牛顿第二定律得 $\dfrac{\mathrm{d}v}{\mathrm{d}t} = g -$ (每单位质量的阻力)。



Jane 的速率上界

Jane 的运动方程（equation of motion）是

$$\frac{\mathrm{d}v}{\mathrm{d}t} = g - kv = k\Big(\frac{g}{k} - v\Big).$$

起初 $v = 0 < g/k$。只要 $v < g/k$，我们便可以除以 $g/k - v > 0$，然后分离变量：

$$\int_0^v \frac{\mathrm{d}w}{g/k - w} = \int_0^t k\,\mathrm{d}s \quad\Longrightarrow\quad -\ln\Big(1 - \frac{kv}{g}\Big) = kt,$$

所以

$$v = \frac{g}{k}\big(1 - \mathrm{e}^{-kt}\big).$$

这个解对所有 $t \geq 0$ 都成立，因为对任何有限的 t，右边都小于 g/k，而这正是我们做除法时所用的条件。由于 $\mathrm{e}^{-kt} > 0$，对所有 t 都有 $v < g/k$。所以当 $t \to \infty$ 时，Jane 的速率趋向 g/k，但永远达不到这个值。

Karen 的速率上界

Karen 的运动方程是

$$\frac{\mathrm{d}v}{\mathrm{d}t} = g - kv - \frac{2k^2}{g}v^2.$$

设 $u = kv/g$，这是一个无因次的速率。于是右边变成

$$g\big(1 - u - 2u^2\big) = g(1 - 2u)(1 + u), \qquad \text{及} \qquad \frac{\mathrm{d}u}{\mathrm{d}t} = \frac{k}{g}\frac{\mathrm{d}v}{\mathrm{d}t} = k(1 - 2u)(1 + u).$$

当 $u = \frac{1}{2}$ 时加速度为零（另一个根 $u = -1$ 并不是物理上合理的速率），即 $v = g/(2k)$。这就是 Karen 的终端速率。

起初 $u = 0 < \frac{1}{2}$。只要 $0 \leq u < \frac{1}{2}$，两个因式都是正数，所以可以分离变量。化成部分分式（partial fractions），

$$\frac{1}{(1 - 2u)(1 + u)} = \frac{2/3}{1 - 2u} + \frac{1/3}{1 + u},$$

这可以在 $1 = A(1 + u) + B(1 - 2u)$ 中代入 $u = -1$ 和 $u = \frac{1}{2}$ 来验证。因此

$$kt = \int_0^u \Big(\frac{2/3}{1 - 2w} + \frac{1/3}{1 + w}\Big)\mathrm{d}w = \frac{1}{3}\Big[-\ln(1 - 2w) + \ln(1 + w)\Big]_0^u = \frac{1}{3}\ln\frac{1 + u}{1 - 2u}.$$

把 $\dfrac{1 + u}{1 - 2u} = \mathrm{e}^{3kt}$ 重新整理，得

$$u = \frac{\mathrm{e}^{3kt} - 1}{2\mathrm{e}^{3kt} + 1} = \frac{1}{2} - \frac{3/2}{2\mathrm{e}^{3kt} + 1} < \frac{1}{2}.$$

基于与之前相同的理由，这个解对所有 $t \geq 0$ 都成立，所以 Karen 的速率满足

$$v < \frac{g}{2k} \quad \text{对所有} t.$$

这就是对应的结果：她的速率趋近 $g/(2k)$，但永远达不到这个值。

Jane 达到 $g/(3k)$ 所需的时间

在 Jane 的解中令 $v = g/(3k)$，得 $1 - \mathrm{e}^{-kt} = \frac{1}{3}$，所以 $\mathrm{e}^{-kt} = \frac{2}{3}$，而

$$t = \frac{1}{k} \ln \frac{3}{2},$$

正如题目所求。

Karen 达到 $g/(3k)$ 所需的时间

> **想一想** 我们需要以 t 表示 u 的公式吗？还是有更快的途径？

更快的途径是利用以 u 表示 t 的公式，而这个公式我们已经有了。速率为 $g/(3k)$ 即 $u = \frac{1}{3}$，它小于 $\frac{1}{2}$，所以她确实会达到这个速率。于是

$$\frac{1+u}{1-2u} = \frac{4/3}{1/3} = 4, \qquad \text{所以} \qquad t = \frac{1}{3k} \ln 4 = \frac{2}{3k} \ln 2.$$

验算

- 在任何正的速率下，Karen 所受的阻力都比 Jane 大，所以她应该要用更长时间才达到相同的速率。果然，$\frac{2}{3} \ln 2 \approx 0.462$ 大于 $\ln \frac{3}{2} \approx 0.405$。
- 当 t 很小时，两条公式都给出 $v \approx gt$，即没有阻力的自由下落（free fall），这正是应有的结果。把 $u = (\mathrm{e}^{3kt} - 1)/(2\mathrm{e}^{3kt} + 1) \approx 3kt/3 = kt$ 展开，便得 $v \approx gt$。
- 我们由静止开始，用数值方法逐步求解两条方程（龙格－库塔法（Runge–Kutta），取 $g = 9.8$，$k = 0.7$）。速率一直保持在 14 和 7 以下，这两个数正是 g/k 和 $g/(2k)$；而速率 $g/(3k)$ 分别在 $t \approx 0.5792$ 和 $t \approx 0.6601$ 时达到。这些数值与 $k^{-1} \ln \frac{3}{2}$ 和 $\frac{2}{3k} \ln 2$ 相符，误差约为 10^{-9}。

常见错误

- 单凭加速度为正便断言“速率永远达不到 g/k”。这只说明速率在增加，却不能说明它一直低于 g/k。你需要明确的解，或者论证达到终端速率所需的时间是无限的。
- 弄错 $\ln(1-2u)$ 一项的正负号。对 $\frac{1}{1-2u}$ 积分，得 $-\frac{1}{2} \ln(1-2u)$。

STEP 之外

两条方程都是自治（autonomous）方程：$\mathrm{d}v/\mathrm{d}t = f(v)$。终端速率是这类方程的稳定平衡点（stable equilibrium）。常微分方程（ODE）的唯一性定理（uniqueness theorem）表

明, 当 f 光滑时, 解永远不能在有限时间内到达平衡点, 这就是两个上界背后的一般原因。Karen 的方程是一条里卡蒂方程（Riccati equation）, 而同样的逻辑斯蒂型（logistic-type）行为亦见于人口模型。Boyce 与 DiPrima 的 *Elementary Differential Equations* 涵盖自治方程及其平衡点，而 Kleppner 与 Kolenkow 的 *An Introduction to Mechanics* 则讨论有阻力的运动。

详细解答

STEP II 2000，第 10 题

力学 · 连接滑轮系统

STEP II 2000，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：

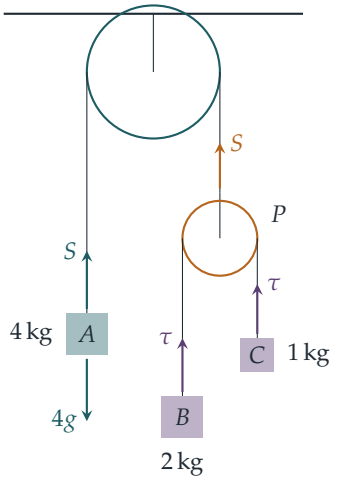
step.maths.org/questions/00-s2-q10

关键思路

这是一部双重阿特伍德机（double Atwood machine）。当 A 被固定时，第二个滑轮（pulley）不动，B、C 组成一部普通的阿特伍德机。一旦放开 A，第二个滑轮也会移动，所以最干净利落的入手方法，是量度 B 和 C 相对于该滑轮的运动，并利用“轻滑轮所受合力（net force）为零”这个事实：滑轮上方的绳的拉力，恰好是下方那条绳的张力（tension）的两倍。

解答

我们把各质点（particle）视为点质量，把绳视为轻（light）而不可伸长（inextensible），并把两个滑轮都视为光滑（smooth）而轻。称第二个滑轮为 P。所有加速度（acceleration）一律取向下为正，以 S 表示穿过 A 的那条绳的张力，以 τ 表示 BC 的张力，而 g 的单位为 $\mathrm{m\,s^{-2}}$。



由于 P 是轻的，它所受的合力必定为零（否则它的加速度会是无限大）。它被 S 向上拉，两边又各被 τ 向下拉，所以在任何时刻都有

$$S = 2\tau$$

。

第一阶段：A 被固定

A 被固定不动时，绕过固定滑轮的绳不会移动，所以 P 固定不动。这时 B 和 C 就是一部普通的阿特伍德机：若 B 以 a 向下加速，则 C 以 a 向上加速。由牛顿第二定律（Newton's second law）得

$$2a = 2g - \tau, \qquad 1 \cdot a = \tau - g.$$

两式相加，得 $3a = g$，所以

$$a = \tfrac{1}{3}g, \qquad \tau = \tfrac{4}{3}g.$$

因此 B 向下加速，C 向上加速，大小各为 $\frac{1}{3}g\ \mathrm{m\,s^{-2}}$。

在这个阶段结束时（$t = T$），B 正向下运动，C 正向上运动，速率各为 $\frac{1}{3}gT$，而 B 已下降了 $\frac{1}{2} \cdot \frac{1}{3}g\,T^2 = \frac{1}{6}gT^2$。在整个阶段中，$A$ 和 P 都静止不动。

第二阶段：全部自由运动

设 A 的加速度为 a_A（向下）。绕过固定滑轮的绳长度不变，所以 P 的加速度为 $-a_A$：A 下降得多快，它就上升得多快。

? **想一想** 当 B 和 C 所悬挂的滑轮本身也在加速时，我们该怎样描述它们的加速度？

相对于 P 而言，绳 BC 的表现与固定的阿特伍德机完全一样：若 B 相对于 P 以加速度 r 向下运动，则 C 相对于 P 以加速度 r 向上运动。再加上 P 本身的加速度，得

$$a_B = -a_A + r, \qquad a_C = -a_A - r.$$

现在对每个质点应用牛顿第二定律（向下为正），并利用 $S = 2\tau$：

$$\begin{aligned} A:&\quad 4a_A = 4g - 2\tau,\\ B:&\quad 2(-a_A + r) = 2g - \tau,\\ C:&\quad 1(-a_A - r) = g - \tau. \end{aligned}$$

由 A 的方程得 $\tau = 2g - 2a_A$。把它代入 B 的方程，得 $-2a_A + 2r = 2a_A$，所以 $r = 2a_A$。把两者都代入 C 的方程：

$$-3a_A = g - (2g - 2a_A) = -g + 2a_A \quad\Longrightarrow\quad a_A = \tfrac{1}{5}g.$$

于是 $r = \frac{2}{5}g$，$\tau = \frac{8}{5}g$，$S = \frac{16}{5}g$，并且

$$a_A = \tfrac{1}{5}g\,(\text{向下}), \qquad a_B = -\tfrac{1}{5}g + \tfrac{2}{5}g = \tfrac{1}{5}g\,(\text{向下}), \qquad a_C = -\tfrac{1}{5}g - \tfrac{2}{5}g = -\tfrac{3}{5}g,$$

所以 C 以 $\frac{3}{5}g\ \mathrm{m\,s^{-2}}$ 向上加速。

这些加速度都是常数，而且在 A 被释放的一刻，没有任何量突然改变：A 和 P 由静止开始运动，而 B、C 则保持原有的速率 $\frac{1}{3}gT$。

A 的速率

在第一阶段中 B 移动了 $\frac{1}{6}gT^2$，所以它还要再走

$$0.6\,gT^2 - \tfrac{1}{6}gT^2 = \tfrac{3}{5}gT^2 - \tfrac{1}{6}gT^2 = \tfrac{13}{30}gT^2$$

。在第二阶段中，B 开始时的向下速度为 $\frac{1}{3}gT$，向下加速度为 $\frac{1}{5}g$，所以它的速度一直为正，所走的距离等于它的位移。若这段路程需时 t，则由 $s = ut + \frac{1}{2}at^2$ 得

$$\tfrac{1}{3}gT\,t + \tfrac{1}{10}gt^2 = \tfrac{13}{30}gT^2.$$

设 $t = kT$，并把全式乘以 $30/(gT^2)$：

$$3k^2 + 10k - 13 = 0 \quad\Longrightarrow\quad (k-1)(3k+13) = 0.$$

由于 $t > 0$，故 $k = 1$：第二阶段恰好历时 T 秒。A 在这个阶段由静止开始，加速度为 $\frac{1}{5}g$，所以它这时的速率为

$$v_A = \tfrac{1}{5}g \cdot T = \tfrac{gT}{5}\ \mathrm{m\,s^{-1}}.$$

验算

- 把求得的数值代入，每条运动方程都成立：对 A 而言，$4\cdot\frac{1}{5}g = 4g - \frac{16}{5}g$；对 C 而言，$-\frac{3}{5}g = g - \frac{8}{5}g$。
- 一个独立的验算：两条绳迫使 $y_C = \text{常数} - 2y_A - y_B$（各位置均以向下量度），而以两个自由坐标 y_A、y_B 运用能量方法，同样得出 $a_A = a_B = \frac{1}{5}g$，$a_C = -\frac{3}{5}g$。
- 取 $g = 9.8$，$T = 1.7$，以数值方法逐步计算运动，B 在 $t = 3.400$（即 $2T$）时走了 $0.6gT^2$，此时 A 的速率为 $3.332 = gT/5$。
- 总重量为 $7g$，但固定滑轮只承受 $2S = \frac{32}{5}g$，小于 $7g$，所以在第二阶段中，系统的质心（centre of mass）向下加速，理应如此。

常见错误

- 在第二阶段仍把 P 当作固定，并对 B 和 C 沿用 $\pm\frac{1}{3}g$。
- 忘记了轻滑轮有 $S = 2\tau$，或者为滑轮加上质量项。
- 把整段 $0.6gT^2$ 当作第二阶段的距离，而没有减去已经走过的 $\frac{1}{6}gT^2$。

STEP 之外

像第二项验算那样，选取能自动满足绳的约束条件的位置坐标，正是拉格朗日力学（Lagrangian mechanics）的起点：把动能（kinetic energy）和势能（potential energy）用自由坐标表示出来，运动方程便随之而得，完全不必理会张力。Kleppner 与 Kolenkow 所著的 *An Introduction to Mechanics*（力学入门教材）对滑轮这类受约束系统有细致的处理，是很好的下一步读物。

详细解答

STEP II 2000，第 11 题

力学 · 弹性绳与简谐运动

STEP II 2000，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/00-s2-q11

关键思路

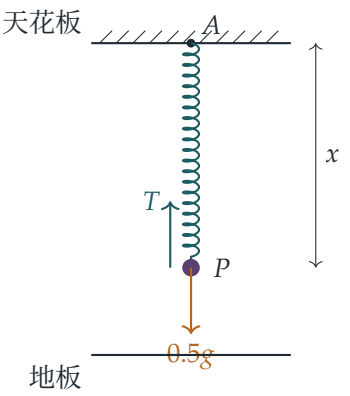
用能量可以很快解决两个定性的结论：把释放时储存的弹性势能（elastic potential energy），与到达自然长度（natural length）及到达天花板所需的重力势能作比较。至于时间，则把运动分成两个阶段：弹性绳拉紧时，质点绕其平衡位置（equilibrium position）作简谐运动（simple harmonic motion）；弹性绳一旦松弛，质点便只在重力作用下自由运动。回程就是把上升的过程倒过来走一遍，所以我们把上升所需的时间加倍。

解答

我们把 P 视为系于一条轻弹性绳上的质点（particle），弹性绳服从胡克定律（Hooke's law），并且没有空气阻力。设 x 为 P 在 A 之下的距离，以向下为正，单位为米。当 $x > 1.5$ 时弹性绳拉紧，此时其伸长量（extension）为 $x - 1.5$，张力（tension）为

$$T = \frac{\lambda}{l}(x - 1.5) = \frac{5g}{1.5}(x - 1.5) = \frac{10g}{3}(x - 1.5).$$

房间高 2.5 m，所以 P 由 $x = 2$ 出发，伸长量为 0.5 m。



弹性绳变得松弛

释放时，储存的弹性势能为

$$\frac{\lambda e^2}{2l} = \frac{5g\,(0.5)^2}{2(1.5)} = \frac{5g}{12}\,\text{J}.$$

弹性绳要回复到自然长度，P 必须上升 0.5 m，所增加的势能为 $0.5g \times 0.5 = \frac{g}{4}$ J。由于 $\frac{5g}{12} > \frac{g}{4}$，$P$ 到达 $x = 1.5$ 时仍在向上运动，其动能（kinetic energy）为

$$\tfrac{1}{2}(0.5)v^2 = \frac{5g}{12} - \frac{g}{4} = \frac{g}{6}, \qquad \text{所以} \qquad v^2 = \frac{2g}{3}.$$

因此它会越过弹性绳处于自然长度的那一点，继续向上，弹性绳于是变得松弛。

P 不会到达天花板

要到达天花板，P 须由起点上升 2 m，势能要增加 $0.5g \times 2 = g$ J。可用的能量就只有释放时储存在弹性绳中的 $\frac{5g}{12}$ J（弹性绳一旦松弛便不储存任何能量），而 $\frac{5g}{12} < g$。所以 P 不可能到达天花板。

更准确地说，弹性绳松弛之后，P 是一个以速率 v 沿铅垂方向运动的抛体（projectile），所以它会再上升 $\dfrac{v^2}{2g} = \dfrac{1}{3}$ m。它的最高点位于 $x = 1.5 - \frac{1}{3} = \frac{7}{6}$，仍在天花板之下 $\frac{7}{6}$ m。

弹性绳拉紧时的简谐运动

当 $x > 1.5$ 时，由牛顿第二定律（Newton's second law）（以向下为正）得

$$0.5\,\ddot{x} = 0.5g - \frac{10g}{3}(x - 1.5).$$

平衡位置就是右边等于零之处：$x - 1.5 = 0.15$，所以 $x = 1.65$。设 $y = x - 1.65$，即在平衡位置之下的位移（displacement）。于是

$$\ddot{y} = 2g - \frac{20g}{3}(y + 0.15) = -\frac{20g}{3}\,y,$$

这是绕 $y=0$ 的简谐运动，其中 $\omega^2=\dfrac{20g}{3}$。只要弹性绳处于拉紧状态，这结果便成立，正如题目所要求。

返回所需的时间

P 由静止出发，起点为 $y=2-1.65=0.35$，所以在弹性绳拉紧期间

$$y=0.35\cos\omega t.$$

弹性绳在 $x=1.5$ 时变得松弛，即 $y=-0.15$。所以首次出现这情况的时刻 t_1 满足

$$\cos\omega t_1=-\frac{0.15}{0.35}=-\frac{3}{7}.$$

想一想 $\cos\omega t_1=-\frac{3}{7}$ 的各个解之中，我们要的是哪一个？又怎样用 arccos 把它写出来？

当 t 由 0 开始增加时，ωt 亦由 0 开始增加，而 $\cos\omega t$ 首次达到 $-\frac{3}{7}$ 时，ωt 位于 $(\frac{\pi}{2},\pi)$ 之内。由于 $\arccos\frac{3}{7}\in(0,\frac{\pi}{2})$ 且 $\cos(\pi-\theta)=-\cos\theta$，这个值就是 $\omega t_1=\pi-\arccos\frac{3}{7}$。因此

$$t_1=\frac{1}{\omega}\left(\pi-\arccos\tfrac{3}{7}\right)=\left(\frac{3}{20g}\right)^{1/2}\left(\pi-\arccos\tfrac{3}{7}\right).$$

（作为验算，此时的速率为 $\omega\sqrt{0.35^2-0.15^2}=\sqrt{\frac{20g}{3}\times0.1}=\sqrt{\frac{2g}{3}}$，与能量计算的结果吻合。）

弹性绳松弛期间，P 只在重力作用下运动。它以速率 $v=\sqrt{2g/3}$ 向上离开 $x=1.5$，然后以相同的速率向下回到 $x=1.5$，所需时间为

$$t_2=\frac{2v}{g}=\frac{2}{g}\sqrt{\frac{2g}{3}}=\left(\frac{8}{3g}\right)^{1/2}.$$

若作用力只取决于位置，运动便是可逆的。P 在 $x=1.5$ 重新进入拉紧区域时，速率与离开时相同，于是它把简谐运动阶段倒过来重走一遍，再用 t_1 的时间在 $y=0.35$（即原来的位置）静止下来。总时间为

$$\begin{aligned}2t_1+t_2&=2\left(\frac{3}{20g}\right)^{1/2}\left(\pi-\arccos\tfrac{3}{7}\right)+\left(\frac{8}{3g}\right)^{1/2}\\&=\left(\frac{8}{3g}\right)^{1/2}+\left(\frac{3}{5g}\right)^{1/2}\left(\pi-\arccos\tfrac{3}{7}\right),\end{aligned}$$

其中用到 $2\sqrt{3/(20g)}=\sqrt{12/(20g)}=\sqrt{3/(5g)}$。这就是所要求的结果。

验算

- 求自然长度处速率的两种方法（能量法，以及简谐运动的 $v^2 = \omega^2(a^2 - y^2)$）同样得出 $v^2 = \frac{2g}{3}$。
- 取 $g = 9.8$，公式给出约 1.0199 s。以数值方法逐步求解运动方程（当 $x < 1.5$ 时把张力关掉），得出返回起点并静止所需的时间相同，而最高点亦在天花板之下 $\frac{7}{6}$ m。
- 若没有松弛阶段，简谐运动的周期（period）会是 $2\pi/\omega \approx 0.77$ s。我们的 $2t_1 \approx 0.50$ s 比这个值小，这是应有的结果，因为只完成了振动的一部分。

常见错误

- 把简谐运动的中心写成自然长度处，而不是平衡位置。重力把中心移到伸长量为 0.15 m 之处。
- 取 $\omega t_1 = \arccos(-\frac{3}{7})$，却没有认出这就是 $\pi - \arccos\frac{3}{7}$。
- 忘记回程：所求的时间包括上升以及返回，所以两个阶段都要加倍。

STEP 之外

这是一个分段定义的，或称“非光滑”（non-smooth）的动力系统（dynamical system）：不同的区域适用不同的运动方程，而各段的解则在边界处通过位置和速度的配对拼接起来。同样的想法也支配着弹跳的球、碰撞和开关电路，而我们用过的可逆性论证，正是初次窥见保守系统（conservative system）中的时间反演对称性（time-reversal symmetry）。想了解更多，可参阅 Kleppner 与 Kolenkow 的 *An Introduction to Mechanics*（力学入门教科书）。

详细解答

STEP II 2000，第 12 题

概率与统计 · 正态变量之和

> **STEP II 2000，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/00-s2-q12

关键思路

Henry 迟到，正正就是他的总行程时间超过他在 0900 之前所余时间的时候，所以对每条路线，我们把两段独立的正态行程时间相加，得出单一个正态变量（normal variable）。标准化（standardising）之后，“迟到”就变成“一个标准正态变量超过某个数”，而由于 Φ 是递增的，两条路线毋须查表便可比较。余下的工作就是全概率定律（law of total probability）和贝叶斯定理（Bayes' theorem）。

解答

比较两条路线

我们假设在每条路线上，火车行程与其后的行程互相独立（independent）。

King's Cross。设 $X_1 \sim N(55, 25)$ 为火车行程时间，$Y_1 \sim N(30, 144)$ 为其后的行程时间。由于它们是独立的正态变量，它们的和也是正态的，而平均值（mean）与方差（variance）各自相加：

$$T_1 = X_1 + Y_1 \sim N(85,\ 169), \qquad \text{标准差13.}$$

火车在 0715 开出，所以 Henry 有 105 分钟，而当 $T_1 > 105$ 时他便会迟到。设 $Z \sim N(0, 1)$，则

$$\mathrm{P}(\text{迟到} \mid \text{King's Cross}) = \mathrm{P}\left(Z > \frac{105 - 85}{13}\right) = 1 - \Phi\left(\tfrac{20}{13}\right) = A.$$

Liverpool Street。同样地，$T_2 = X_2 + Y_2$，其中 $X_2 \sim N(65, 16)$ 与 $Y_2 \sim N(25, 9)$ 互相独立，所以

$$T_2 \sim N(90,\ 25), \qquad \text{标准差5.}$$

这班火车在 0720 开出，余下 100 分钟，所以

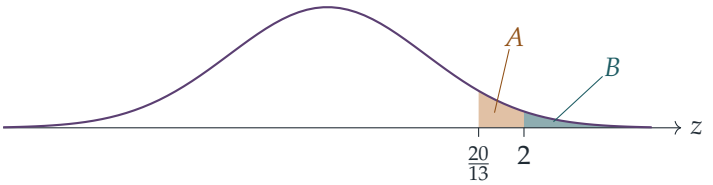
$$\mathrm{P}(\text{迟到} \mid \text{Liverpool Street}) = \mathrm{P}\left(Z > \frac{100-90}{5}\right) = 1 - \Phi(2) = B.$$

想一想 不查表，怎样比较 A 与 B?

Φ 是一个累积分布函数（cumulative distribution function），其密度为正，所以它是严格递增的。由于 $\frac{20}{13} < \frac{26}{13} = 2$，我们有 $\Phi(\frac{20}{13}) < \Phi(2)$，因此

$$A = 1 - \Phi\left(\tfrac{20}{13}\right) > 1 - \Phi(2) = B.$$

所以 Henry 若乘搭往 King's Cross 的火车，迟到的机会较大，这正是所要证明的。（有趣的是，这条路线平均而言反而较快；它之所以落败，是因为由 King's Cross 出发的那段地铁行程变化太大。）



标准正态曲线下涂上阴影的尾部：A 是整个阴影部分的面积，B 是 2 以外颜色较深的部分。

迟到次数的期望值

在单一次行程中，设 K 为 Henry 乘搭往 King's Cross 的火车这一事件，故 $\mathrm{P}(K) = p$，并设 D 为他迟到这一事件。根据全概率定律，

$$\mathrm{P}(D) = \mathrm{P}(D \mid K)\,\mathrm{P}(K) + \mathrm{P}(D \mid K')\,\mathrm{P}(K') = pA + (1-p)B.$$

假设 M 次行程互相独立，迟到次数便服从参数为 M 和 $\mathrm{P}(D)$ 的二项分布（binomial distribution），所以

$$L = M(pA + (1-p)B) = M(B + p(A-B)).$$

由于 $A - B > 0$ 且 $0 \le p \le 1$，表达式 $B + p(A-B)$ 随 p 递增，由 $p = 0$ 时的 B 增至 $p = 1$ 时的 A。因此对 p 的每一个可能值，

$$BM \le L \le AM,$$

左边的等号仅在 $p = 0$ 时成立，右边的等号仅在 $p = 1$ 时成立。用文字来说：L/M 是 A 与 B 的加权平均（weighted average），所以它必定介乎两者之间。

估计 p

在迟到的日子当中，他乘搭了往 King's Cross 火车的比例，是 $\mathrm{P}(K \mid D)$ 的估计值；而因为 M 很大，我们把观察所得的 $\frac{3}{5}$ 视为等于这个概率。根据贝叶斯定理，

$$\mathrm{P}(K \mid D) = \frac{\mathrm{P}(D \mid K)\,\mathrm{P}(K)}{\mathrm{P}(D)} = \frac{pA}{pA + (1-p)B}.$$

令此式等于 $\frac{3}{5}$：

$$5pA = 3pA + 3(1-p)B$$
$$2pA + 3pB = 3B,$$

所以估计值是

$$\hat{p} = \frac{3B}{2A + 3B}.$$

验算

- 以数值计算，$A \approx 0.0620$ 而 $B \approx 0.0228$，所以经 King's Cross 约有 6% 的时候迟到，经 Liverpool Street 则约 2%。对每条路线各模拟 4×10^5 次行程，结果吻合至小数点后三位。
- 代入这些数值，$\hat{p} \approx 0.355$。代回原式，$pA/(pA + (1-p)B) = 0.6$ 准确成立；而以这个 p 模拟整个过程，所得的迟到个案中约有 60% 来自 King's Cross。
- 对最后答案作合理性检查：假如两条路线同样可靠（$A = B$），迟到的情况便只反映他选乘哪班火车，所以当中有 $\frac{3}{5}$ 经 King's Cross，就表示 $p = \frac{3}{5}$。公式给出 $3B/5B = \frac{3}{5}$。由于 King's Cross 是较不可靠的路线，它在迟到的日子中所占的比例偏高，所以 p 必定小于 $\frac{3}{5}$，而事实正是如此。

常见错误

- 把标准差（standard deviation）相加，而不是把方差相加：King's Cross 路线的总时间方差为 $25 + 144 = 169$，所以其标准差是 13，而不是 $5 + 12 = 17$。
- 两班火车都用同一个 105 分钟的时限：往 Liverpool Street 的火车迟五分钟开出。
- 直接令 $p = \frac{3}{5}$。这个 $\frac{3}{5}$ 是 $\mathrm{P}(K \mid D)$，而不是 $\mathrm{P}(K)$；两者不同，是因为两条路线并非同样可靠。

STEP 之外

这里有两个想法，会在大学第一个概率课程中反复出现。独立正态变量之和仍是正态的这个事实，用矩母函数（moment generating function）来证明最为简洁；它也是正态分布族的稳定性（stability）的一个特例，而这种稳定性正是中心极限定理（central limit theorem）的基础。最后一部分则是统计推论（statistical inference）的缩影：我们观察到一个结果（迟到），然后运用贝叶斯定理推断隐藏的原因（选乘哪班火车）。Grimmett 与 Welsh 的 *Probability: An Introduction*（概率入门）涵盖了这两个课题，而 Blitzstein 与 Hwang 的 *Introduction to Probability*（概率导论）则有更多正是这种风格的贝叶斯题目。

详细解答

STEP II 2000，第 13 题

概率与统计 · 捕捉再捕捉估计

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

本题在 *Beyond STEP* 第 13.2 节计数田鼠中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch13/voles.html

关键思路

这是捉放法（capture–recapture）：第二次抽样是从 N 只田鼠中随机选出 200 只，而当中有 200 只带有标记，所以样本中带标记田鼠的数目服从超几何分布（hypergeometric distribution）。要找出 p_N 在何处最大，不要求导（N 的阶乘（factorial）并没有简单意义下的导数）；应该改为通过比值 p_N/p_{N-1} 比较相邻两项，这时几乎所有东西都会约去。最后的恒等式则不费吹灰之力：第三次抽样的各个概率之和必定是 1。

解答

概率 p_N

把第二次抽样看成从 N 只田鼠中选出 200 只，每个大小为 200 的子集出现的机会均等。这样的子集共有 $\binom{N}{200}$ 个。在 N 只田鼠中，200 只带有标记，$N-200$ 只没有。一个恰好含 11 只带标记田鼠的样本，就是选出 11 只带标记的和 189 只没有标记的，所以

$$p_N = \frac{\binom{200}{11}\binom{N-200}{189}}{\binom{N}{200}}.$$

把含有 N 的两个二项式系数（binomial coefficient）写成阶乘，

$$p_N = \binom{200}{11} \cdot \frac{(N-200)!}{189!\,(N-389)!} \cdot \frac{200!\,(N-200)!}{N!}$$
$$= k\,\frac{((N-200)!)^2}{N!\,(N-389)!}, \qquad \text{其中} k = \binom{200}{11}\frac{200!}{189!},$$

而这个常数与 N 无关。这正是所要求的形式。（这里需要 $N \geq 389$，因为两次抽样合共包含 $200 + 189 = 389$ 只不同的田鼠。）

求 p_N 的最大值

对于 $N \geq 390$，把 p_N 除以 p_{N-1}。利用 $(N-200)! = (N-200)\,(N-201)!$、$N! = N\,(N-1)!$ 和 $(N-389)! = (N-389)\,(N-390)!$，所有东西都会约去，只余下

$$\frac{p_N}{p_{N-1}} = \frac{(N-200)^2}{N(N-389)}.$$

分子和分母都是正数，所以

$$\begin{aligned} p_N > p_{N-1} &\iff (N-200)^2 > N(N-389) \\ &\iff N^2 - 400N + 40000 > N^2 - 389N \\ &\iff 40000 > 11N \\ &\iff N < 3636.\overline{36}. \end{aligned}$$

想一想 这告诉我们数列（sequence）$p_{389}, p_{390}, \ldots$ 的形状是怎样的?

当 $390 \leq N \leq 3636$ 时，比值大于 1，所以数列严格递增，直至 p_{3636}。当 $N \geq 3637$ 时，比值小于 1（等号成立需要 $11N = 40000$，而这没有整数解），所以数列自此严格递减。因此 p_N 有唯一的最大值，位于

$$N = 3636.$$

这就是 $\frac{200\times200}{11}$ 的整数部分。这也正是你会猜到的答案：只要认为样本中带标记的比例 $\frac{11}{200}$ 应该等于总体中带标记的比例 $\frac{200}{N}$ 便可。

第三次抽样

取总体大小为 3636，即刚才求得的估计值。第二次放回之后，带标记的田鼠是原来的 200 只加上新标记的 189 只，所以有 389 只带有标记，$3636 - 389 = 3247$ 只没有标记。按照与之前完全相同的推理，一个大小为 200 的随机样本中恰好有 j 只带标记田鼠的概率是

$$\frac{\binom{389}{j}\binom{3247}{200-j}}{\binom{3636}{200}}, \qquad j = 0, 1, \ldots, 200.$$

由于 $389 \geq 200$ 且 $3247 \geq 200$，由 0 至 200 的每一个 j 都有可能出现，而这些就是全部可能的结果。它们互斥（mutually exclusive），所以它们的概率之和是 1：

$$\sum_{j=0}^{200} \frac{\binom{389}{j}\binom{3247}{200-j}}{\binom{3636}{200}} = 1.$$

乘以 $\binom{3636}{200}$ 便得

$$\sum_{j=0}^{200} \binom{389}{j}\binom{3247}{200-j} = \binom{3636}{200},$$

这正是所要证明的。

验算

- 用整数二项式系数准确计算 p_N：对于若干个 N 的值（例如 389、1000 和 5000），结果与阶乘公式一致；而在 $389 \leq N < 20000$ 的范围内搜索，最大值位于 $N = 3636$。
- 在转折点处，$p_{3636}/p_{3635} = 3436^2/(3636 \times 3246) \approx 1.0003 > 1$，但 $p_{3637}/p_{3636} < 1$。
- 最后的恒等式两边以整数准确计算，结果相等。留意 $389 + 3247 = 3636$，这是必然的。

常见错误

- 把 N 当作连续变量来求导。应该改用相邻两项的比值，并准确地判定不等式。
- 忘记了第二次放回之后，带标记的田鼠是 389 只，而不是 200 只。
- 说“最大值位于 $N = 3636.36$”：总体大小必须是整数，而且你应该检查相邻的两个整数中哪一个胜出。

STEP 之外

选取使观测数据出现的概率最大的参数值，就是*最大似然法*（maximum likelihood），这是大学一年级统计学的核心思想。在这里，它给出捉放法的 Lincoln–Petersen 估计量（estimator）。最后的恒等式是范德蒙德恒等式（Vandermonde's identity），而所用的论证方法，即用两种方法数算同一样东西，是组合数学（combinatorics）中一种强而有力的技巧。Grimmett and Welsh 的 *Probability: An Introduction*（概率入门教材）涵盖了超几何分布和这类论证方法。

详细解答

STEP II 2000，第 14 题

概率与统计 • 中位数的分布

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

本题在 *Beyond STEP* 第 13.4 节平均数还是中位数？中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch13/mean-or-median.html

关键思路

中位数（median）的概率密度函数（probability density function）是形如 $y^r(1-y)^s$ 的多项式，所以题目给出的积分已包办所有工作：先用一次求归一化常数（normalising constant），再把 y 的幂次分别提高一和提高二，求出首两个矩（moments）。至于最后一部分，Y 和 $\bar{X}$ 都以 $\frac{1}{2}$ 为中心，而对正态分布（normal distribution）来说，落在距离中心某固定距离之内的概率只取决于标准差（standard deviation）。离散程度较小的那个变量胜出。

解答

常数 k

密度函数的积分必须等于 1。在题目给出的结果中取 $r = s = n$，

$$1 = \int_0^1 ky^n(1-y)^n \,\mathrm{d}y = k\,\frac{n!\,n!}{(2n+1)!}, \qquad \text{所以} \qquad k = \frac{(2n+1)!}{(n!)^2},$$

正如所求。

平均值

密度函数关于 $y=\frac{1}{2}$ 对称（把 y 换成 $1-y$，函数不变），所以 $\mathrm{E}(Y)=\frac{1}{2}$。要用积分来确认这一点，取 $r=n+1$、$s=n$：

$$\mathrm{E}(Y)=k\int_0^1 y^{n+1}(1-y)^n\,\mathrm{d}y=\frac{(2n+1)!}{(n!)^2}\cdot\frac{(n+1)!\,n!}{(2n+2)!}=\frac{n+1}{2n+2}=\frac{1}{2}.$$

方差

现在取 $r=n+2$、$s=n$：

$$\mathrm{E}(Y^2)=\frac{(2n+1)!}{(n!)^2}\cdot\frac{(n+2)!\,n!}{(2n+3)!}=\frac{(n+2)(n+1)}{(2n+3)(2n+2)}=\frac{n+2}{2(2n+3)}.$$

因此

$$\mathrm{Var}(Y)=\frac{n+2}{2(2n+3)}-\frac{1}{4}=\frac{2(n+2)-(2n+3)}{4(2n+3)}=\frac{1}{4(2n+3)}.$$

比较中位数与平均值

每个 X_i 都在 $[0,1]$ 上服从均匀分布（uniform distribution），所以 $\mathrm{E}(X_i)=\frac{1}{2}$，而 $\mathrm{Var}(X_i)=\frac{1}{12}$。由于各 X_i 互相独立（independent），

$$\mathrm{E}(\bar{X})=\tfrac{1}{2},\qquad \mathrm{Var}(\bar{X})=\frac{1}{(2n+1)^2}\cdot(2n+1)\cdot\frac{1}{12}=\frac{1}{12(2n+1)}.$$

想一想 两个变量都以 $\frac{1}{2}$ 为中心。哪一个数值决定了谁较有可能落在距离中心 $d/\sqrt{n}$ 之内?

答案是标准差。当 n 很大时，我们可以取 $Y\sim\mathrm{N}(\frac{1}{2},\sigma_Y^2)$ 和 $\bar{X}\sim\mathrm{N}(\frac{1}{2},\sigma_{\bar{X}}^2)$。若 $W\sim\mathrm{N}(\frac{1}{2},\sigma^2)$ 且 $c>0$，则以 Φ 表示标准正态分布函数（standard normal distribution function），有

$$\mathrm{P}\big(|W-\tfrac{1}{2}|<c\big)=\mathrm{P}\Big(|Z|<\frac{c}{\sigma}\Big)=2\Phi\Big(\frac{c}{\sigma}\Big)-1,$$

其中 $Z\sim\mathrm{N}(0,1)$。这是 σ 的严格递减函数，因为 Φ 严格递增。所以取 $c=d/\sqrt{n}$ 时，所求的不等式成立，当且仅当 $\sigma_Y>\sigma_{\bar{X}}$，即 $\mathrm{Var}(Y)>\mathrm{Var}(\bar{X})$：

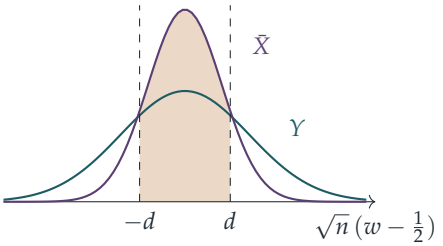
$$\begin{aligned}\frac{1}{4(2n+3)}>\frac{1}{12(2n+1)}&\iff 12(2n+1)>4(2n+3)\\&\iff 24n+12>8n+12\\&\iff n>0.\end{aligned}$$

这对每个 $n\geq 1$ 都成立。所以只要 n 大得足以令正态近似（normal approximation）有效，则不论正数 d 是甚么，Y 的概率都严格小于 $\bar{X}$ 的概率。证明至此完成。

要看看差距有多大，可以把变量标准化：

$$\frac{d/\sqrt{n}}{\sigma_Y} = 2d\sqrt{\frac{2n+3}{n}} \to 2\sqrt{2}\,d, \qquad \frac{d/\sqrt{n}}{\sigma_X} = d\sqrt{\frac{12(2n+1)}{n}} \to 2\sqrt{6}\,d.$$

渐近而言，平均值的标准差是中位数的 $\sqrt{3}$ 分之一。



图中以 $\sqrt{n}\,(w-\frac{1}{2})$ 为标度，画出两个极限密度函数。较窄的曲线（平均值）在任何一条中央带内都有较大的概率。

验算

- $n=0$: 一个均匀变量的“中位数”就是 X_1 本身, 而我们的公式给出 $k=1$ 和 $\mathrm{Var}(Y)=\frac{1}{12}$，正是均匀分布的方差，与预期相符。
- 当 $n=0$ 时两个方差都等于 $\frac{1}{12}$，而当 $n\geq 1$ 时两者便分开，这与上面的不等式 $n>0$ 吻合。
- 以 $n=50$（即 101 个均匀变量）进行模拟, 得到 $\mathrm{Var}(Y)\approx 0.00243$, 而 $1/412\approx 0.00243$。取 $d=0.3$ 时，模拟得到 $\mathrm{P}(|Y-\frac{1}{2}|<d/\sqrt{n})\approx 0.607$ 和 $\mathrm{P}(|\bar{X}-\frac{1}{2}|<d/\sqrt{n})\approx 0.859$, 与正态分布给出的数值 0.611 和 0.860 相当接近。

常见错误

- 写成 $\mathrm{Var}(\bar{X})=\frac{1}{12}$，或者只除以 $2n+1$ 一次，而不是先在分母得到 $(2n+1)^2$，再乘以 $2n+1$。
- 弄错方向：方差较大，落在中心附近的概率反而较小。
- 忘记了计算 $\mathrm{E}(Y^2)$ 时用的是 $r=n+2$，而不是 $r=2n$。

STEP 之外

中位数的密度函数是一个贝塔分布（Beta distribution）$\mathrm{Beta}(n+1,n+1)$，而题目给出的积分就是贝塔函数（Beta function）。你刚才所作的比较，正是渐近相对效率（asymptotic

relative efficiency）的概念：对均匀分布的数据，用样本平均值（sample mean）估计中心，其方差只有中位数的三分之一；但对厚尾（heavy-tailed）的数据，中位数却可以胜出。顺序统计量（order statistics）和贝塔函数在 Blitzstein 与 Hwang 合著的 *Introduction to Probability*（概率论入门）中有很好的讲解。

详细解答

STEP III 2000，第 1 题

纯数学 · 双曲线的切线

STEP III 2000，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/00-s3-q1

关键思路

两条曲线都是以原点为中心的等轴双曲线（rectangular hyperbola），而且在半周旋转（half-turn）$(x, y) \mapsto (-x, -y)$ 之下都保持不变。这个对称性（symmetry）让我们不费吹灰之力就得到 Q。至于切线（tangent），巧妙的做法不是硬解联立方程，而是利用 $ab = 1$ 和 $a^2 - b^2 = 2$ 这两个事实，验证题目给出的点同时位于两条直线上。求得四个点之后，绕原点旋转四分之一周就能解释为何得出正方形。

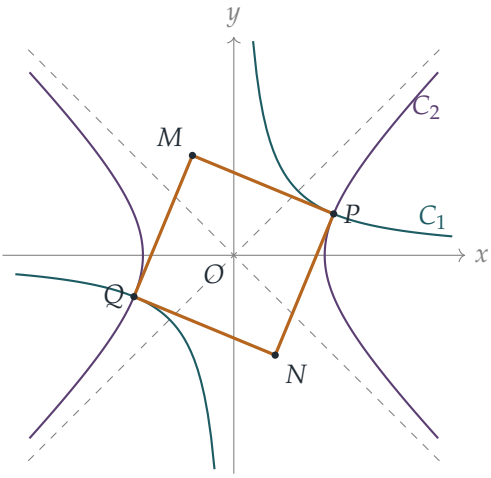
解答

略图

C_1，即 $xy = 1$，是大家熟悉的等轴双曲线，位于第一和第三象限，以两条坐标轴为渐近线（asymptote）。

C_2，即 $x^2 - y^2 = 2$，与 x 轴相交于 $(\pm\sqrt{2}, 0)$，且永不与 y 轴相交（否则需要 $-y^2 = 2$）。写成 $y^2 = x^2 - 2$，可见 $|y| < |x|$，而且当 $|x|$ 很大时 $y \approx \pm x$，所以它的渐近线是 $y = \pm x$。它就是把 C_1 绕原点旋转 45°（并按 $\sqrt{2}$ 的比例放大）所得的曲线，两支分别向左、右开口。

两条曲线在哪里相交？由于 $xy = 1 > 0$，任何公共点都位于第一或第三象限。把 $y = 1/x$ 代入 C_2，得 $x^4 - 2x^2 - 1 = 0$，所以 $x^2 = 1 + \sqrt{2}$（另一个根 $1 - \sqrt{2}$ 是负数）。因此恰好有两个交点，在这两个象限中各有一个。



图中蓝绿色的是 C_1，紫红色的是 C_2，虚线是 C_2 的渐近线 $y = \pm x$。琥珀色的正方形 $PMQN$ 就是我们即将求出的图形。

点 Q

把 (x, y) 换成 $(-x, -y)$，xy 和 $x^2 - y^2$ 都保持不变。所以若 (a, b) 同时位于两条曲线上，则 $(-a, -b)$ 亦然。由于只有两个交点，

$$Q = (-a, -b).$$

以下我们会一直用到 P 同时位于两条曲线上所带来的两个事实：

$$ab = 1, \qquad a^2 - b^2 = 2.$$

切线

对于 C_1，把 $xy = 1$ 作隐函数求导（implicit differentiation），得 $y + x\frac{\mathrm{d}y}{\mathrm{d}x} = 0$，所以 C_1 在其上一点 (p, q) 的斜率（gradient）是 $-q/p$。该点的切线是 $p(y - q) = -q(x - p)$，即

$$qx + py = 2pq = 2.$$

对于 C_2，把 $x^2 - y^2 = 2$ 求导，得 $2x - 2y\frac{\mathrm{d}y}{\mathrm{d}x} = 0$，所以在 (p, q) 的斜率是 p/q。切线是 $q(y - q) = p(x - p)$，即

$$px - qy = p^2 - q^2 = 2.$$

（即使在切线垂直的点，这两个形式仍然成立；它们其实就是曲线的梯度向量（gradient vector）与 (x, y) 的点积。）代入 $P = (a, b)$ 和 $Q = (-a, -b)$：

$$\begin{aligned}
\text{曲线}C_1\text{ 的切线，切点为}P:&\quad bx + ay = 2,\\
\text{曲线}C_2\text{ 的切线，切点为}Q:&\quad -ax + by = 2,\\
\text{曲线}C_2\text{ 的切线，切点为}P:&\quad ax - by = 2,\\
\text{曲线}C_1\text{ 的切线，切点为}Q:&\quad -bx - ay = 2.
\end{aligned}$$

点 M

M 是头两条直线的交点。试代入 $(x, y) = (-b, a)$：

$$\begin{aligned}
bx + ay &= -b^2 + a^2 = 2,\\
-ax + by &= ab + ab = 2ab = 2.
\end{aligned}$$

所以 $(-b, a)$ 同时位于两条直线上。这两条直线并不平行，因为系数的行列式（determinant）是 $b \cdot b - a \cdot (-a) = a^2 + b^2 \neq 0$，所以它们恰好相交于一点。因此

$$M = (-b, a),$$

正如题目所求。

> **想一想** 验证一个已给出的答案，不是比直接求解来得弱吗?

在这里并非如此。两条不平行的直线恰好相交于一点，所以任何同时位于两条直线上的点就是交点。直接求解当然也可以，但到头来我们还是要用 $ab = 1$ 和 $a^2 - b^2 = 2$ 来整理答案。

点 N

同样地，N 是 $ax - by = 2$ 与 $-bx - ay = 2$ 的交点。代入点 $(b, -a)$ 得

$$ab + ab = 2, \qquad -b^2 + a^2 = 2,$$

而行列式是 $a(-a) - (-b)(-b) = -(a^2 + b^2) \neq 0$。所以

$$N = (b, -a).$$

$PMQN$ 是正方形

看看这四个位置向量（position vector）：

$$P = (a, b), \quad M = (-b, a), \quad Q = (-a, -b), \quad N = (b, -a).$$

绕 O 逆时针旋转 90° 会把 (x,y) 送到 $(-y,x)$。所以它把 $P \to M \to Q \to N \to P$。这已经暗示了答案；以下是一个直接的论证。

- 对角线 PQ 和 MN 的中点都是 O，所以它们互相平分，$PMQN$ 是平行四边形（parallelogram）。
- 两条对角线的长度都是 $2\sqrt{a^2+b^2}$，所以它们相等，这个平行四边形是长方形。
- $\overrightarrow{OP}\cdot\overrightarrow{OM} = a(-b)+b(a) = 0$，所以对角线互相垂直，这个长方形是菱形（rhombus）。

既是长方形又是菱形的四边形就是正方形。顶点的次序也正确：P、M、Q、N 绕 O 依次相隔四分之一周，所以 $PMQN$ 是沿着正方形绕一圈，而不是自我交叉。因此 $PMQN$ 是正方形。 ■

它的边长的平方是 $|PM|^2 = (a+b)^2+(b-a)^2 = 2(a^2+b^2)$。由于 $a^2 = 1+\sqrt{2}$ 及 $b^2 = 1/a^2 = \sqrt{2}-1$，这个值是 $4\sqrt{2}$。

验算

- 以数值计算，$a = \sqrt{1+\sqrt{2}} \approx 1.5538$ 及 $b \approx 0.6436$。用数值方法求切线的交点，得 $M \approx (-0.6436,\ 1.5538)$ 和 $N \approx (0.6436,\ -1.5538)$，与所述相符。四条边的长度都是 $2.3784 \approx \sqrt{4\sqrt{2}}$，而且相邻的边互相垂直。
- P 的方向角是 $\arctan(b/a) = \arctan(\sqrt{2}-1) = 22.5^\circ$，刚好在 C_1 的渐近线 $y=0$ 与 C_2 的渐近线 $y=x$ 的正中间，所以正方形倾斜了 22.5°，正如略图所示。

常见错误

- 把 C_2 的切线写成 $px+qy=2$。负号十分重要，正是它令整个运算行得通。
- 凭空猜测有反射对称，于是取 $Q=(b,a)$ 或 $(-b,-a)$。对 $y=x$ 的反射会保持 C_1 不变，但不会保持 C_2 不变；只有绕 O 的半周旋转才同时保持两者不变。
- 只证明四边相等。这样只得到菱形；你还需要一个直角（或对角线相等）。

STEP 之外

我们用过的切线公式 $qx+py=2$ 和 $px-qy=2$，都是以下一般法则的例子：圆锥曲线（conic）$\mathbf{x}^\mathsf{T}A\mathbf{x}=c$ 在 $\mathbf{p}$ 的切线是 $\mathbf{p}^\mathsf{T}A\mathbf{x}=c$。在大学一年级关于二次型（quadratic form）的课程中，你会学到如何用旋转把 A 对角化（diagonalise），这正是 $xy=1$ 和 $x^2-y^2=2$ 其实是同一条曲线转了 45° 的原因。Strang 的 *Introduction to Linear Algebra* 清楚地讲解

了二次型和旋转，而 Siklos 的 *Advanced Problems in Mathematics* 则有更多 STEP 风格的坐标几何题目。

详细解答

STEP III 2000，第 2 题

纯数学 · 代换积分法

STEP III 2000，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/00-s3-q2

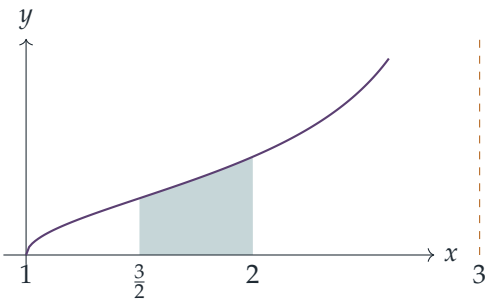
关键思路

像 $(x-1)/(3-x)$ 这类商的平方根，最适合用半角公式（half-angle formulae）处理：设 $x = 2-\cos\theta$，分子变成 $1-\cos\theta$，分母变成 $1+\cos\theta$，而两者之比是一个完全平方，即 $\tan^2\frac{\theta}{2}$。至于一般情况的结果，我们不必把运算重做一遍。只要作一个线性变量代换（linear change of variable），把由 a 到 b 的区间映到由 1 到 3 的区间，一般的积分就会变成第一个积分的倍数。

解答

第一个积分

被积函数（integrand）在 $1 \le x < 3$ 上有定义，而我们的积分范围 $\frac{3}{2} \le x \le 2$ 正好在其中。下图的阴影面积就是所求的积分。



设 $x = 2-\cos\theta$，则 $\mathrm{d}x = \sin\theta\,\mathrm{d}\theta$。当 $x = \frac{3}{2}$ 时，$\cos\theta = \frac{1}{2}$；当 $x = 2$ 时，$\cos\theta = 0$，所

以我们取 θ 由 $\frac{\pi}{3}$ 到 $\frac{\pi}{2}$。于是

$$x-1=1-\cos\theta=2\sin^2\tfrac{\theta}{2},\qquad 3-x=1+\cos\theta=2\cos^2\tfrac{\theta}{2},$$

所以

$$\left(\frac{x-1}{3-x}\right)^{\frac12}=\left|\tan\tfrac{\theta}{2}\right|=\tan\tfrac{\theta}{2},$$

这是因为在我们的范围内 $0<\frac{\theta}{2}<\frac{\pi}{2}$，所以正切值为正。

想一想 当被积函数变成 $\tan\frac{\theta}{2}\sin\theta$ 之后，怎样才可以避免对正切函数求积分?

写出 $\sin\theta=2\sin\frac{\theta}{2}\cos\frac{\theta}{2}$。余弦项会互相抵消：

$$\tan\tfrac{\theta}{2}\,\sin\theta=\frac{\sin\frac{\theta}{2}}{\cos\frac{\theta}{2}}\cdot 2\sin\tfrac{\theta}{2}\cos\tfrac{\theta}{2}=2\sin^2\tfrac{\theta}{2}=1-\cos\theta.$$

因此

$$\begin{aligned}\int_{3/2}^{2}\left(\frac{x-1}{3-x}\right)^{\frac12}\mathrm{d}x&=\int_{\pi/3}^{\pi/2}(1-\cos\theta)\,\mathrm{d}\theta=\Big[\theta-\sin\theta\Big]_{\pi/3}^{\pi/2}\\&=\left(\frac{\pi}{2}-1\right)-\left(\frac{\pi}{3}-\frac{\sqrt3}{2}\right)=\frac{\pi}{6}-1+\frac{\sqrt3}{2}\\&=\frac{\pi+3\sqrt3-6}{6}.\end{aligned}$$

一般的积分

第一个积分就是 $a=1$、$b=3$ 的情况，此时 $p=\frac{3+3}{4}=\frac32$，$q=\frac{1+3}{2}=2$。所以我们要找一个线性代换，把 $u=1$ 送到 $x=a$，把 $u=3$ 送到 $x=b$：

$$x=a+\tfrac12(b-a)(u-1),\qquad \mathrm{d}x=\tfrac12(b-a)\,\mathrm{d}u.$$

于是

$$x-a=\tfrac12(b-a)(u-1),\qquad b-x=(b-a)-\tfrac12(b-a)(u-1)=\tfrac12(b-a)(3-u),$$

又因为 $b-a\neq0$，公因子可以约去：

$$\frac{x-a}{b-x}=\frac{u-1}{3-u}.$$

至于积分上下限：$u=\frac32$ 给出 $x=a+\frac14(b-a)=\frac{3a+b}{4}=p$，而 $u=2$ 给出 $x=a+\frac12(b-a)=\frac{a+b}{2}=q$。由于 $b-a>0$，这个代换是递增的，所以上下限的次序保持不变。因此

$$\begin{aligned}\int_p^q\left(\frac{x-a}{b-x}\right)^{\frac12}\mathrm{d}x&=\frac{b-a}{2}\int_{3/2}^{2}\left(\frac{u-1}{3-u}\right)^{\frac12}\mathrm{d}u\\&=\frac{b-a}{2}\cdot\frac{\pi+3\sqrt3-6}{6}=\frac{(b-a)(\pi+3\sqrt3-6)}{12},\end{aligned}$$

其中用到了第一个结果。这正是我们要证明的。

（另一个等价的做法，是直接代入 $x = \frac{a+b}{2} - \frac{b-a}{2}\cos\theta$。上下限同样是 $\frac{\pi}{3}$ 和 $\frac{\pi}{2}$，运算过程也一样，只是多了一个因子 $\frac{b-a}{2}$。）

验算

- 数值上，$\frac{1}{6}(\pi + 3\sqrt{3} - 6) \approx 0.38962$。把辛普森法则（Simpson's rule）用于原来的被积函数，在 $[\frac{3}{2}, 2]$ 上计算，所得的值准确至小数点后十位都相同。
- 合理性：被积函数在长度为 $\frac{1}{2}$ 的区间上由 $\sqrt{1/3} \approx 0.577$ 上升至 1，所以面积必定介乎约 0.29 与 0.5 之间。结果确实如此。
- 随机选取 20 组 $a < b$（包括 a 为负数的情况），一般公式都与辛普森法则的结果吻合。
- 缩放：把区间 $[a, b]$ 按因子 k 拉伸，被积函数的取值不变，但 $[p, q]$ 会拉伸 k 倍。所以答案必定与 $b - a$ 成正比，而事实正是这样。

常见错误

- 忘记了 $\mathrm{d}x = \sin\theta\,\mathrm{d}\theta$，试图只对 $\tan\frac{\theta}{2}$ 关于 θ 求积分。
- 把上下限取为 $\theta = -\frac{\pi}{3}$ 或 $\frac{5\pi}{3}$。无论怎样选取，都必须使 θ 由 $\frac{\pi}{3}$ 递增至 $\frac{\pi}{2}$，并且途经的范围内 $\sin\theta > 0$ 和 $\tan\frac{\theta}{2} > 0$ 都成立。
- 在一般情况中漏掉 $\mathrm{d}x$ 带来的因子 $\frac{1}{2}(b-a)$，结果分母变成 6 而不是 12。

STEP 之外

第二部分说明，定积分（definite integral）在仿射变量代换（affine change of variable）之下的变化是有规律可循的。这就是为甚么像这样的量只通过长度 $b-a$ 依赖于 a 和 b。被积函数本身属于贝塔函数（beta function）类型：在整个区间上，$\int_a^b \left(\frac{x-a}{b-x}\right)^{1/2}\mathrm{d}x = \frac{\pi}{2}(b-a)$，而函数 $\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,\mathrm{d}t$ 在概率和分析中反复出现。Blitzstein 与 Hwang 的 *Introduction to Probability*（概率导论）把它们作为贝塔分布（Beta distribution）的归一化常数（normalising constants）来介绍。Bartle 与 Sherbert 的 *Introduction to Real Analysis*（实分析导论）则严谨地证明了代换法则（substitution rule）。

详细解答

STEP III 2000，第 3 题

纯数学 · 复数与几何

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

本题在 *Beyond STEP* 第 5.6 节费马点中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch05/fermat-point.html

关键思路

在阿根图（Argand plane）上，把一个向量旋转 60° 只不过是乘以 $\alpha = e^{i\pi/3}$，所以在给定边上作等边三角形（equilateral triangle），其第三个顶点一行就可以写出来。其后的一切都来自恒等式（identity）$1 + \alpha^2 = \alpha$ 及其相关的式子，例如 $\alpha^3 = -1$。它们把各顶点的表达式化成非常简单的形式，一旦看出这些形式，最后一部分几乎立即可得。

解答

恒等式

由于 $\alpha^2 = e^{2i\pi/3} = -\frac{1}{2} + \frac{\sqrt{3}}{2}i$,

$$1 + \alpha^2 = \tfrac{1}{2} + \tfrac{\sqrt{3}}{2}i = \cos\tfrac{\pi}{3} + i\sin\tfrac{\pi}{3} = \alpha.$$

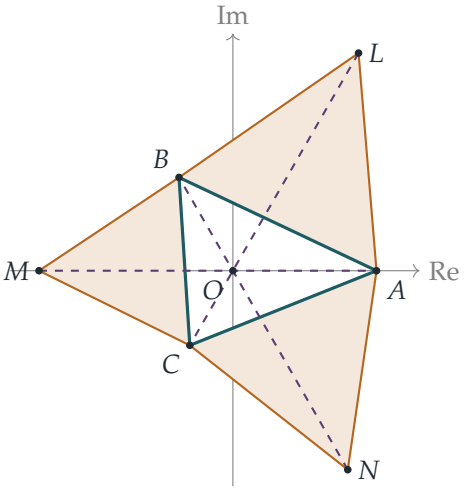
我们还会用到以下几个推论：

- $\alpha^3 = e^{i\pi} = -1$;
- $1 - \alpha = -\alpha^2$（这是把恒等式移项所得）;
- $\alpha^2 - \alpha = -1$（同样是移项所得）;
- $1 - \alpha = \bar{\alpha} = e^{-i\pi/3}$，因为 $\alpha + \bar{\alpha} = 2\cos\frac{\pi}{3} = 1$。

最后一项说明，乘以 $1-\alpha$ 相当于顺时针旋转 $60°$。

三角形 ABC

$A=p$ 位于正实轴上，$B=q\alpha^2=q\,\mathrm{e}^{2\mathrm{i}\pi/3}$ 位于辐角（argument）为 $120°$ 的射线上，而 $C=-r\alpha=r\,\mathrm{e}^{4\mathrm{i}\pi/3}$ 位于辐角为 $240°$ 的射线上。因此三个顶点分别落在由 O 出发、彼此相隔 $120°$ 的三条射线上，与 O 的距离分别为 p,q,r。原点在三角形之内，而 $A\to B\to C$ 是逆时针方向。下图取 $p=2$、$q=1.5$、$r=1.2$，并同时画出在各边上所作的三角形。



顶点 N、L 和 M

若 XYZ 是一个按顺时针方向标记顶点的等边三角形，那么从 X 出发来看，边 XZ 就是边 XY 绕 X 顺时针旋转 $60°$ 所得。因此

$$Z-X=(Y-X)\,\mathrm{e}^{-\mathrm{i}\pi/3}=(Y-X)(1-\alpha),\qquad \text{即}\qquad Z=\alpha X+(1-\alpha)Y.$$

? **想一想** 为甚么“顺时针”会令新作的三角形落在 ABC 的外面？

由于 $A\to B\to C$ 是逆时针方向，C 位于有向边 AB 的左方。把 AB 绕 A 顺时针旋转，它便摆向右方，所以 L 与 C 分别位于 AB 的两侧。其余两条边的情况相同，这与图中所见吻合。

N（三角形 CAN，故 $X=C$、$Y=A$）：

$$N=\alpha(-r\alpha)+(1-\alpha)p=(1-\alpha)p-\alpha^2 r,$$

正如所求。利用 $1-\alpha=-\alpha^2$，可把它整理成 $N=-\alpha^2(p+r)$。

L（三角形 ABL，故 $X=A$、$Y=B$）：

$$L=\alpha p+(1-\alpha)q\alpha^2=\alpha p+q(\alpha^2-\alpha^3)=\alpha p+q(1+\alpha^2)=(p+q)\alpha,$$

当中先用 $\alpha^3=-1$，再用 $1+\alpha^2=\alpha$。

M（三角形 BCM，故 $X=B$、$Y=C$）：

$$M=\alpha\cdot q\alpha^2+(1-\alpha)(-r\alpha)=q\alpha^3+r(\alpha^2-\alpha)=-q-r=-(q+r),$$

当中用到 $\alpha^3=-1$ 和 $\alpha^2-\alpha=-1$。

因此，仿照题目所给 N 的形式，

$$L=\alpha p+\alpha q,\qquad M=-q-r,\qquad N=(1-\alpha)p-\alpha^2 r.$$

三条直线

现在把每一对点都看成某个单位复数的正数倍：

$$\begin{aligned} L&=(p+q)\,\mathrm{e}^{\mathrm{i}\pi/3}, & C&=-r\alpha=r\,\mathrm{e}^{\mathrm{i}\pi/3}\cdot(-1),\\ M&=(q+r)\cdot(-1), & A&=p\cdot 1,\\ N&=(p+r)\cdot(-\alpha^2), & B&=q\,\alpha^2. \end{aligned}$$

在每一对之中，两点分别是 u 和 $-u$ 的正实数倍，其中 u 是一个单位复数（分别为 $u=\alpha$、$u=1$ 和 $u=\alpha^2$）。所以这两点位于同一条通过 O 的直线上，并且分处 O 的两侧。由此可知，直线 LC、MA、NB 各自都通过原点，而且原点其实位于每一条线段上。又因为两点分处 O 的两侧，每条线段的长度就是两点与 O 的距离之和：

$$LC=(p+q)+r,\qquad MA=(q+r)+p,\qquad NB=(p+r)+q,$$

全部等于 $p+q+r$，正如所求。

验算

取 $p=1$、$q=2$、$r=3$，由公式得 $L=3\alpha\approx 1.5+2.598\mathrm{i}$、$M=-5$ 和 $N=-4\alpha^2\approx 2-3.464\mathrm{i}$。计算得 $|L-A|=|L-B|=|A-B|=\sqrt{7}$，确认 ABL 是等边三角形。我们又用电脑直接由 $\pm60^\circ$ 的两个旋转作出每个三角形，保留顺时针的那一个，并就 10 000 组随机选取的 p,q,r 确认了三条公式、三线共点（concurrence）于 O，以及共同长度 $p+q+r$。一个漂亮的特殊情况：当 $p=q=r$ 时，三角形 ABC 是以 O 为中心的等边三角形，而整个图形具有预期中的三重旋转对称。

常见错误

- 旋转方向弄错，乘了 α 而不是 $\bar{\alpha} = 1 - \alpha$。这样 L、M、N 会落在三角形之内，题目所给 N 的形式也推不出来。
- 只证明了 O 在每条直线上，却忘记要得出长度，还需要 O 位于两个端点之间。各实数倍的正负号正好解决这一点。

STEP 之外

这是拿破仑构形（Napoleon's configuration）的一个特殊情况：对任意三角形，在各边向外作等边三角形，直线 AM、BN、CL 都交于同一点（费马点，Fermat point），而且三条线段长度相等。费马点使它到三个顶点的距离总和达到最小；在本题中它就是 O，从这一点看每条边所张的角都是 $120°$。复数把旋转变成乘法，这正是此类证明可以化为简短代数运算的原因。Needham 的 *Visual Complex Analysis*（《复分析：可视化方法》）把这种几何观点发挥得十分精彩。

详细解答

STEP III 2000，第 4 题

纯数学 · 有理函数的描绘

> **STEP III 2000，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/00-s3-q4

关键思路

令 $f(x) = x + c$ 并消去分母后，三次项互相抵消，余下的是一个关于 x 的二次方程（quadratic）。于是直线与曲线是否相交，就取决于一个判别式（discriminant）的正负；而这个判别式本身又是 c 的二次式，所以我们要求“判别式的判别式”。至于描绘图像，把 f 拆成线性部分加上部分分式（partial fractions），就能得知一切：各条渐近线（asymptote）、曲线位于渐近线的哪一侧，以及曲线在何处穿过斜渐近线（oblique asymptote）。

解答

直线与曲线在何处相交

当 $x \neq \pm 1$ 时，方程 $f(x) = x + c$ 等价于

$$x(x-2)(x-a) = (x+c)(x^2-1).$$

展开得 $x^3 - (a+2)x^2 + 2ax = x^3 + cx^2 - x - c$。两边的 x^3 项互相抵消，余下

$$q(x) \equiv (a+2+c)x^2 - (2a+1)x - c = 0. \qquad (*)$$

因此，直线与曲线相交，当且仅当 $(*)$ 有一个不等于 $x = \pm 1$ 的实根（real root）。

> ? **想一想** $(*)$ 的根会不会恰好是被排除的值 $x = \pm 1$ 之一？

代入可得 $q(1) = (a+2+c) - (2a+1) - c = 1 - a$ 及 $q(-1) = (a+2+c) + (2a+1) - c = 3(a+1)$。两者都与 c 无关，所以只有当 $a = 1$ 时 $x = 1$ 才是根，只有当 $a = -1$ 时 $x = -1$ 才是根。当 $|a| \neq 1$ 时，$(*)$ 的每一个实根都对应一个真正的交点。

$a+2+c=0$ 的情况。这时 $(*)$ 是线性方程：$-(2a+1)x+(a+2)=0$。若 $|a|>1$，则 $2a+1\neq 0$，所以有根 $x=\frac{a+2}{2a+1}$。

$a+2+c\neq 0$ 的情况。这时 $(*)$ 是真正的二次方程，其判别式为

$$D(c)=(2a+1)^2+4c(a+2+c)=4c^2+4(a+2)c+(2a+1)^2.$$

把 D 看成 c 的二次式，其首项系数（leading coefficient）为正。它自身的判别式是

$$\begin{aligned}16(a+2)^2-16(2a+1)^2&=16\big[(a+2)-(2a+1)\big]\big[(a+2)+(2a+1)\big]\\&=16(1-a)\cdot 3(1+a)=48(1-a^2).\end{aligned}$$

若 $|a|>1$，上式为负，所以对任何 c 都有 $D(c)>0$；于是 $(*)$ 有两个相异实根，而且都不等于 ± 1。连同线性的情况，可知对任何 c，直线 $y=x+c$ 都与曲线相交。

若 $|a|=1$，对任何 c 都有 $D(c)\geq 0$，所以 $(*)$ 仍然总有实根，这就是该命题在代数上的内容。不过这时曲线本身会退化：分子和分母约去一个公因式，曲线变成一条缺了一点的双曲线（hyperbola）。例如，取 $a=1$ 时，我们得到 $f(x)=x-3+\frac{3}{x+1}$ $(x\neq 1)$，而直线 $y=x-3$（即它的渐近线）与它并不相交；在 $(*)$ 中这对应 $c=-3$，此时唯一的根正是被排除的 $x=1$。所以当 $|a|=1$ 时，除了这些例外的 c 值之外，命题都成立；而当 $|a|>1$ 时，命题毫无例外地成立。

若 $|a|<1$，二次式 $D(c)$ 有两个实根，并且在两根之间严格为负。一个明确的选择是 $c=-\frac{1}{2}(a+2)$，即两根的中点：

$$D\big(-\tfrac{1}{2}(a+2)\big)=(a+2)^2-2(a+2)^2+(2a+1)^2=3(a^2-1)<0.$$

对于这个 c，$a+2+c=\frac{1}{2}(a+2)\neq 0$（因为 $a>-1$），所以 $(*)$ 是一个判别式为负的真正二次方程，没有实根。因此直线 $y=x-\frac{1}{2}(a+2)$ 与曲线不相交，这正是所要证明的。

斜渐近线

以 x^2-1 除 $x^3-(a+2)x^2+2ax$，得

$$f(x)=x-(a+2)+\frac{(2a+1)x-(a+2)}{x^2-1}.$$

当 $x\to\pm\infty$ 时，该分式趋向 0，所以斜渐近线是

$$y=x-a-2.$$

为了描绘图像，把该分式拆成部分分式会很有帮助：

$$f(x)=x-(a+2)+\frac{a-1}{2(x-1)}+\frac{3(a+1)}{2(x+1)}.$$

由此我们得知一切：

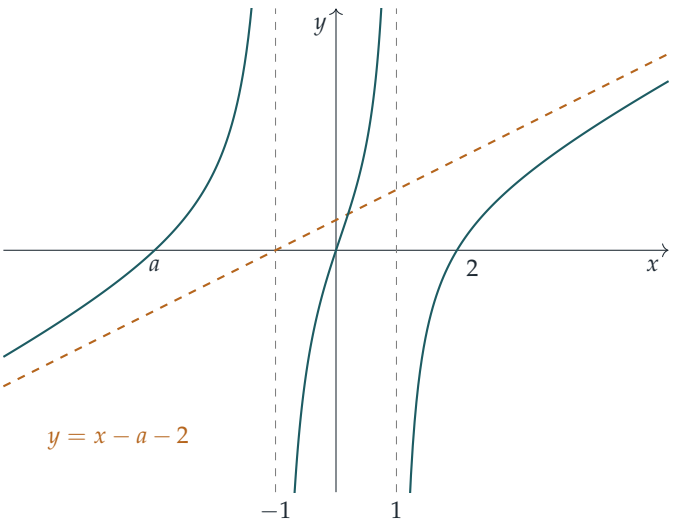
- 零点位于 $x = 0,\ 2,\ a$，铅垂渐近线（vertical asymptote）位于 $x = \pm 1$；
- 曲线只在 $(2a+1)x = a+2$ 之处与斜渐近线相交，即在 $x_0 = \frac{a+2}{2a+1}$（这又是 $(*)$ 的线性情况）；
- 当 $|x|$ 很大时，差距 $f(x) - (x-a-2)$ 的表现有如 $\frac{2a+1}{x}$。在两种情况中都有 $2a+1 < 0$，所以当 $x \to +\infty$ 时曲线位于渐近线下方，而当 $x \to -\infty$ 时则位于其上方；
- 系数 $\frac{a-1}{2}$ 和 $\frac{3(a+1)}{2}$ 决定了曲线在 $x = 1$ 和 $x = -1$ 两侧的表现。

情况 (i)： $a < -1$

两个系数都是负数：$\frac{a-1}{2} < 0$ 及 $\frac{3(a+1)}{2} < 0$。所以当 $x \to 1^-$ 或 $x \to -1^-$ 时 $f(x) \to +\infty$，而当 $x \to 1^+$ 或 $x \to -1^+$ 时 $f(x) \to -\infty$。此外

$$f'(x) = 1 - \frac{a-1}{2(x-1)^2} - \frac{3(a+1)}{2(x+1)^2} > 0,$$

因为被减去的两项都是负数，所以 f 在它的三个分支上都是递增的，完全没有转向点（turning point）。交点满足 $x_0 + 1 = \frac{3(a+1)}{2a+1} > 0$ 及 $x_0 - 1 = \frac{1-a}{2a+1} < 0$，所以 $-1 < x_0 < 1$：曲线在中间的分支上穿过它的渐近线。零点分别位于 $a < -1$（左分支）、0（中间分支）和 2（右分支）。



（图中取 $a = -3$，这时渐近线是 $y = x + 1$，而 $x_0 = \frac{1}{5}$。）

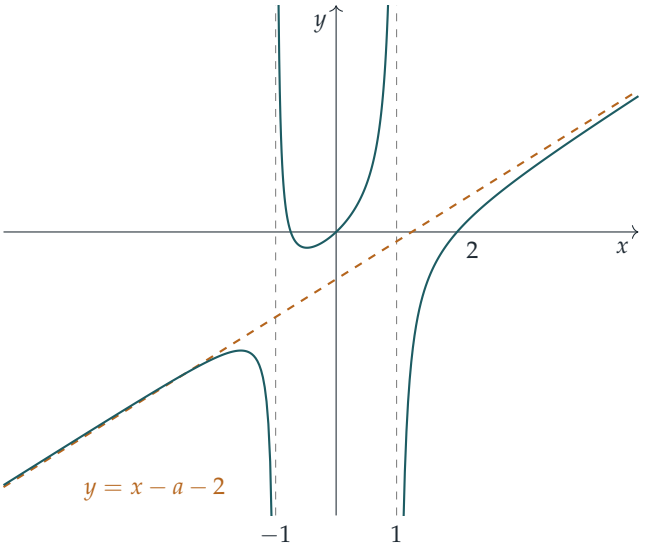
情况 (ii)： $-1 < a < -\frac{1}{2}$

现在 $\frac{a-1}{2} < 0$，但 $\frac{3(a+1)}{2} > 0$。在 $x = 1$ 附近，表现与之前一样（从左边趋向 $+\infty$，从右边趋向 $-\infty$），但在 $x = -1$ 附近则反转过来：当 $x \to -1^-$ 时 $f(x) \to -\infty$，而当 $x \to -1^+$

时 $f(x) \to +\infty$。现在 $x_0 + 1 = \frac{3(a+1)}{2a+1} < 0$，所以 $x_0 < -1$：交点在左分支上。零点 a 和 0 都位于 $(-1, 0)$ 之内。

> **想一想** 虽然题目没有要求我们求出转向点，但它们必定在哪里？

在左分支上，曲线在渐近线上方从 $-\infty$ 升上来，在 x_0 处穿过渐近线，然后当 $x \to -1^-$ 时跌向 $-\infty$；所以它有一个极大点（maximum）。曲线在这个分支上没有零点（三个零点全都 > -1），所以整个分支都位于 x 轴下方。在中间分支上，$f \to +\infty$ 于两端成立，曲线在 a 与 0 之间降到轴的下方，在其他地方则为正，所以在 a 与 0 之间有一个极小点（minimum）。在右分支上，当 $x > 1$ 时 $\frac{3(a+1)}{2(x+1)^2} < \frac{3}{4} \cdot \frac{1}{4} < 1$，所以 $f'(x) > 0$：曲线从 $-\infty$ 上升，经过 $x = 2$，并从下方趋近渐近线。



（图中取 $a = -\frac{3}{4}$：渐近线为 $y = x - \frac{5}{4}$，交点在 $x_0 = -\frac{5}{2}$，而曲线在 $-\frac{3}{4}$ 与 0 之间略为降到轴的下方。）

验算

- 取 $a = 0$ 及 $c = -1$（即我们所选的 $c = -\frac{1}{2}(a+2)$），$(*)$ 变成 $x^2 - x + 1 = 0$，其判别式为 -3：没有交点，与预期相符。
- 取 $a = -3$：$f(-1.1) \approx 30.9$ 及 $f(-0.9) \approx -28.8$，与情况 (i) 中就 $x \to -1^{\mp}$ 所得的正负号吻合。取 $a = -\frac{3}{4}$，则 $f(-1.1) \approx -5.7$ 及 $f(-0.9) \approx 2.1$，正是情况 (ii) 中反转了的表现。
- 在每个范围内取若干个 a 值，以数值方法数算 f' 的变号次数，可以确认情况 (i) 中没

有转向点，而情况 (ii) 中左分支和中间分支各恰有一个转向点。

常见错误

- 忘记 $(*)$ 的根 $x = \pm 1$ 并不是交点；当 $|a| = 1$ 时出错的正是这一点。
- 忽略 $a + 2 + c = 0$ 的情况，这时 $(*)$ 是线性方程，判别式的论证并不适用。
- 在情况 (ii) 中弄错曲线在铅垂渐近线 $x = -1$ 两侧的走向：$a + 1$ 的正负号会改变那里的图像。

STEP 之外

当 $|a| < 1$ 时，$D(c)$ 的两个根给出与曲线相切的直线 $y = x + c$，即切线（tangent），而介乎两者之间的直线则与曲线不相交。当 a 越过 ± 1 时，这些切线会出现或消失，曲线的形状也随之改变：一个零点穿过一条铅垂渐近线，渐近线旁边的表现随即反转。研究一族曲线的定性形状如何随参数（parameter）而改变，正是分歧理论（bifurcation theory）的起点，在大学一年级的微分方程（differential equations）课程中就会遇到；可参阅 Boyce 与 DiPrima 的 *Elementary Differential Equations*（微分方程入门教科书）。若想多练习这类图像描绘，Siklos 的 *Advanced Problems in Mathematics*（进阶数学难题集）是极佳的选择。

详细解答

STEP III 2000，第 5 题

纯数学 · 线性变换

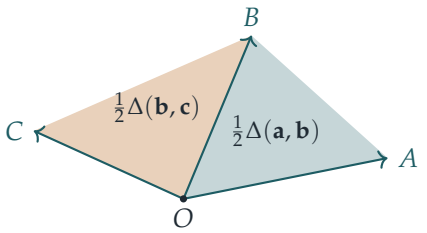
STEP III 2000，第 5 题

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/00-s3-q5

关键思路

$\Delta(\mathbf{a},\mathbf{b})$ 是以 $\mathbf{a}$ 和 $\mathbf{b}$ 为列的矩阵（matrix）的行列式（determinant），所以它是三角形 OAB 有向面积（signed area）的两倍，而且恰好在 $\mathbf{a}$ 与 $\mathbf{b}$ 不平行时才不为零。线性映射（linear map）会把所有这类面积乘以同一个因子，即它的行列式。单凭这一点，必要性便立刻得出。至于充分性，(ii) 中的恒等式（identity）只用各个 Δ 就把 $\mathbf{c}$ 以 $\mathbf{a}$ 和 $\mathbf{b}$ 表示出来，所以一旦 $\mathbf{M}$ 由 A 和 B 确定下来，C 被送到哪里便再无选择余地。



解答

以下一律以 $[\mathbf{u}\ \mathbf{v}]$ 表示以 $\mathbf{u}$ 和 $\mathbf{v}$ 为列的 2×2 矩阵。于是

$$\Delta(\mathbf{u},\mathbf{v}) = u_1v_2 - u_2v_1 = \det[\mathbf{u}\ \mathbf{v}].$$

对于非零的 $\mathbf{u}$、$\mathbf{v}$，$\Delta(\mathbf{u},\mathbf{v}) = 0$ 恰好在 $\mathbf{u}$ 与 $\mathbf{v}$ 平行时成立：行列式为零，当且仅当其中一列是另一列的倍数。因此 $\Delta(\mathbf{a},\mathbf{b})$、$\Delta(\mathbf{b},\mathbf{c})$、$\Delta(\mathbf{c},\mathbf{a})$、$\Delta(\mathbf{p},\mathbf{q})$、$\Delta(\mathbf{q},\mathbf{r})$、$\Delta(\mathbf{r},\mathbf{p})$ 这六个数全部不为零。

还有两个事实会反复用到。由于 $\mathbf{M}[\mathbf{u}\ \mathbf{v}] = [\mathbf{Mu}\ \ \mathbf{Mv}]$（逐列计算），而且乘积的行列式等

于行列式的乘积，

$$\Delta(\mathbf{Mu},\mathbf{Mv}) = \det\mathbf{M}\,\Delta(\mathbf{u},\mathbf{v}). \tag{$*$}$$

另外，$\Delta(\mathbf{v},\mathbf{u}) = -\Delta(\mathbf{u},\mathbf{v})$。

第 (i) 部分

我们需要 $\mathbf{Ma} = \mathbf{p}$ 和 $\mathbf{Mb} = \mathbf{q}$，两者合起来就是

$$\mathbf{M}\,[\mathbf{a}\ \mathbf{b}] = [\mathbf{p}\ \mathbf{q}].$$

由于 $\mathbf{a}$ 与 $\mathbf{b}$ 不平行，$\det[\mathbf{a}\ \mathbf{b}] = \Delta(\mathbf{a},\mathbf{b}) \neq 0$，所以 $[\mathbf{a}\ \mathbf{b}]$ 可逆（invertible），而

$$\mathbf{M} = [\mathbf{p}\ \mathbf{q}]\,[\mathbf{a}\ \mathbf{b}]^{-1} = \frac{1}{\Delta(\mathbf{a},\mathbf{b})}\begin{pmatrix} p_1 & q_1 \\ p_2 & q_2 \end{pmatrix}\begin{pmatrix} b_2 & -b_1 \\ -a_2 & a_1 \end{pmatrix}$$

正符合要求。事实上，这个 $\mathbf{M}$ 是唯一符合要求的矩阵，因为我们可以在右边乘以逆矩阵。取行列式，得 $\det\mathbf{M} = \Delta(\mathbf{p},\mathbf{q})/\Delta(\mathbf{a},\mathbf{b})$，此数不为零。

第 (ii) 部分：恒等式

取左方的第一个分量：

$$\begin{aligned}&(a_1b_2 - a_2b_1)c_1 + (c_1a_2 - c_2a_1)b_1 + (b_1c_2 - b_2c_1)a_1\\ &\quad = a_1b_2c_1 - a_2b_1c_1 + a_2b_1c_1 - a_1b_1c_2 + a_1b_1c_2 - a_1b_2c_1 = 0,\end{aligned}$$

因为六项两两相消。第二个分量的计算完全相同，只是把各个 Δ 以外的 $\mathbf{a}$、$\mathbf{b}$、$\mathbf{c}$ 的下标全部由 1 改为 2：

$$\begin{aligned}&(a_1b_2 - a_2b_1)c_2 + (c_1a_2 - c_2a_1)b_2 + (b_1c_2 - b_2c_1)a_2\\ &\quad = a_1b_2c_2 - a_2b_1c_2 + a_2b_2c_1 - a_1b_2c_2 + a_2b_1c_2 - a_2b_2c_1 = 0.\end{aligned}$$

所以

$$\Delta(\mathbf{a},\mathbf{b})\,\mathbf{c} + \Delta(\mathbf{c},\mathbf{a})\,\mathbf{b} + \Delta(\mathbf{b},\mathbf{c})\,\mathbf{a} = \mathbf{0}, \tag{1}$$

这正是所要证明的。这个恒等式对任意三个向量（vector）都成立，所以对 $\mathbf{p}$、$\mathbf{q}$、$\mathbf{r}$ 同样成立：

$$\Delta(\mathbf{p},\mathbf{q})\,\mathbf{r} + \Delta(\mathbf{r},\mathbf{p})\,\mathbf{q} + \Delta(\mathbf{q},\mathbf{r})\,\mathbf{p} = \mathbf{0}. \tag{2}$$

> ? **想一想** 这个比例条件写成等式，确切是在说甚么？

所有 Δ 都不为零，所以条件 $\Delta(\mathbf{a},\mathbf{b}) : \Delta(\mathbf{b},\mathbf{c}) : \Delta(\mathbf{c},\mathbf{a}) = \Delta(\mathbf{p},\mathbf{q}) : \Delta(\mathbf{q},\mathbf{r}) : \Delta(\mathbf{r},\mathbf{p})$ 的意思是：存在同一个数 $k \neq 0$，使得

$$\Delta(\mathbf{p},\mathbf{q}) = k\,\Delta(\mathbf{a},\mathbf{b}), \qquad \Delta(\mathbf{q},\mathbf{r}) = k\,\Delta(\mathbf{b},\mathbf{c}), \qquad \Delta(\mathbf{r},\mathbf{p}) = k\,\Delta(\mathbf{c},\mathbf{a}). \tag{3}$$

用文字来说：三角形 OPQ、OQR、ORP 的三个有向面积，就是 OAB、OBC、OCA 的面积全部按同一个因子缩放的结果。

第 (ii) 部分：必要性

假设 $\mathbf{Ma}=\mathbf{p}$、$\mathbf{Mb}=\mathbf{q}$ 及 $\mathbf{Mc}=\mathbf{r}$。由 $(*)$，

$$\Delta(\mathbf{p},\mathbf{q})=\det\mathbf{M}\,\Delta(\mathbf{a},\mathbf{b}),\quad \Delta(\mathbf{q},\mathbf{r})=\det\mathbf{M}\,\Delta(\mathbf{b},\mathbf{c}),\quad \Delta(\mathbf{r},\mathbf{p})=\det\mathbf{M}\,\Delta(\mathbf{c},\mathbf{a}).$$

这就是 $k=\det\mathbf{M}$ 时的 (3)，而且因为 $\Delta(\mathbf{p},\mathbf{q})\neq 0$，所以 $k\neq 0$。因此两组比相等。

第 (ii) 部分：充分性

现在假设 (3) 对某个 $k\neq 0$ 成立。由第 (i) 部分，存在矩阵 $\mathbf{M}$ 使得 $\mathbf{Ma}=\mathbf{p}$ 及 $\mathbf{Mb}=\mathbf{q}$；我们必须证明这同一个 $\mathbf{M}$ 也把 $\mathbf{c}$ 送到 $\mathbf{r}$。

把 $\mathbf{M}$ 作用于 (1)。矩阵乘法是线性的，而各个 Δ 只是数，所以

$$\Delta(\mathbf{a},\mathbf{b})\,\mathbf{Mc}=-\Delta(\mathbf{c},\mathbf{a})\,\mathbf{q}-\Delta(\mathbf{b},\mathbf{c})\,\mathbf{p}.$$

两边乘以 k，并对每一项使用 (3)：

$$\Delta(\mathbf{p},\mathbf{q})\,\mathbf{Mc}=-\Delta(\mathbf{r},\mathbf{p})\,\mathbf{q}-\Delta(\mathbf{q},\mathbf{r})\,\mathbf{p}.$$

由 (2)，右方等于 $\Delta(\mathbf{p},\mathbf{q})\,\mathbf{r}$。所以

$$\Delta(\mathbf{p},\mathbf{q})\,\mathbf{Mc}=\Delta(\mathbf{p},\mathbf{q})\,\mathbf{r},$$

又因为 $\Delta(\mathbf{p},\mathbf{q})\neq 0$（$\mathbf{p}$、$\mathbf{q}$ 不平行），故 $\mathbf{Mc}=\mathbf{r}$。因此 P、Q、R 就是 A、B、C 在 $\mathbf{M}$ 之下的像（image）。

必要性与充分性合起来，便证明了这个比例条件是充分必要条件（necessary and sufficient）。

验算

- 试取 $\mathbf{a}=\binom{1}{0}$、$\mathbf{b}=\binom{0}{1}$、$\mathbf{c}=\binom{1}{1}$。这时 $\Delta(\mathbf{a},\mathbf{b})=1$、$\Delta(\mathbf{b},\mathbf{c})=-1$、$\Delta(\mathbf{c},\mathbf{a})=-1$，而 (1) 就是 $\binom{1}{1}-\binom{0}{1}-\binom{1}{0}=\mathbf{0}$。取 $\mathbf{M}=\begin{pmatrix}2&1\\0&3\end{pmatrix}$，得 $\mathbf{p}=\binom{2}{0}$、$\mathbf{q}=\binom{1}{3}$、$\mathbf{r}=\binom{3}{3}$，三个 Δ 分别是 6、-6、-6：比例仍然是 $1:-1:-1$，其中 $k=6=\det\mathbf{M}$。
- 在数值方面，我们用精确分数运算，以 20,000 组随机整数向量三元组检验了这个恒等式。在其中约 17,700 组没有任何两个向量平行的三元组中，我们按第 (i) 部分构造 $\mathbf{M}$，确认了必要性，并确认 $\mathbf{Mc}=\mathbf{r}$ 恰好在 (3) 成立时才成立。

常见错误

- 把比例条件写成一条分数等式，却没有说明各个 Δ 不为零。题目假设任何两个向量都不平行，正是为了保证每一次除法都合法。
- 只证明必要性。真正的内容在于充分性：矩阵由它对两个线性独立（independent）向量的作用所确定，所以这个条件必须能迫使第三个像符合要求。
- 因 Δ 内的次序而弄错正负号：循环次序 $\mathbf{b},\mathbf{c}$ 然后 $\mathbf{c},\mathbf{a}$ 是有关系的，因为 $\Delta(\mathbf{a},\mathbf{c}) = -\Delta(\mathbf{c},\mathbf{a})$。

STEP 之外

恒等式 (1) 其实是克莱姆法则（Cramer's rule）的另一种面貌：它以行列式之比给出 $\mathbf{c}$ 在基（basis）$\mathbf{a}$、$\mathbf{b}$ 之下的坐标。事实 $(*)$，即线性映射把每个有向面积都按 $\det \mathbf{M}$ 缩放，正是行列式的几何定义，而二重积分（double integral）换元时出现的雅可比（Jacobian）因子，背后也是这个道理。Strang 的 *Introduction to Linear Algebra*（《线性代数导论》）对这两个想法都有详细阐述。

详细解答

STEP III 2000，第 6 题

纯数学 · 四次多项式的因式分解

STEP III 2000，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/00-s3-q6

关键思路

比较系数（comparing coefficients）可把这个因式分解（factorisation）化为三条把 p,q,r 与 a,b,c 联系起来的方程。要得到只含 a 的方程，就必须消去 b 和 c，而恒等式（identity）$4bc=(b+c)^2-(b-c)^2$ 正好做到这一点，因为 $b+c$ 和 $b-c$ 正是自然出现的组合。至于最后的四次式（quartic），须先消去 y^3 项，才能用上那条三次方程（cubic），而平移 $y=x+k$ 便可做到。

解答

求 p、q 和 r

展开右边，得

$$(x^2-ax+b)(x^2+ax+c)=x^4+(b+c-a^2)x^2+a(b-c)x+bc,$$

其中 x^3 项 ax^3-ax^3 互相抵消。比较系数，得

$$p=b+c-a^2,\qquad q=a(b-c),\qquad r=bc.$$

a^2 所满足的三次方程

设 $u=a^2$。由第一条方程，$b+c=p+u$。

若 $a\neq 0$，第二条方程给出 $b-c=q/a$，所以 $(b-c)^2=q^2/u$。于是

$$4bc=(b+c)^2-(b-c)^2 \quad\Longrightarrow\quad 4r=(p+u)^2-\frac{q^2}{u}.$$

两边乘以 $u \neq 0$，

$$4ru = u(u^2 + 2pu + p^2) - q^2 \quad \Longrightarrow \quad u^3 + 2pu^2 + (p^2 - 4r)u - q^2 = 0,$$

所以 $u = a^2$ 是一个根（root）。

若 $a = 0$，则 $q = a(b - c) = 0$，而 $u = a^2 = 0$ 也是一个根，因为这时三次式的每一项在 $u = 0$ 时都等于零。因此在任何情况下 a^2 都是一个根，正如题目所求。

> ? **想一想** 为甚么不论 p、q 和 r 是甚么，都必定有一个 $u \geq 0$ 的根?

把这个三次式记为 $f(u)$。在 $u = 0$ 处有 $f(0) = -q^2 \leq 0$，而由于首项是 u^3，当 $u \to +\infty$ 时 $f(u) \to +\infty$。若 $q = 0$，则 $u = 0$ 本身就是一个根。否则 $f(0) < 0$，而对某个足够大的 U 有 $f(U) > 0$；由于 f 是连续（continuous）的，根据介值定理（intermediate value theorem），在 0 与 U 之间（不含两端）必有一个根。无论哪种情况，都有一个非负根。

这一点很重要，因为我们需要 $a = \sqrt{u}$ 是实数。非负根必定给出实数 a，所以每个不含 x^3 项的实系数四次式，都可以分解为两个符合题目形式的实系数二次式（quadratic）。（若 $u > 0$，我们可以由 $b + c = p + u$ 和 $b - c = q/\sqrt{u}$ 求回 b 和 c。）

$p = -1$、$q = -6$、$r = 15$ 的情况

这里 $p^2 - 4r = 1 - 60 = -59$，所以三次方程是

$$u^3 - 2u^2 - 59u - 36 = 0.$$

在 $u = 9$ 处：$729 - 162 - 531 - 36 = 0$，所以 $u = 9$ 是一个根。（除去这个因式后得 $(u - 9)(u^2 + 7u + 4)$，其中二次因式有两个负根，所以 $u = 9$ 是唯一的非负根。）

分解四次式

题目给出的四次式含有 y^3 项，所以我们先作平移把它消去。各根之和为 8，故其平均值为 2，于是设 $y = x + 2$。逐项展开：

$$\begin{aligned}
(x+2)^4 &= x^4 + 8x^3 + 24x^2 + 32x + 16,\\
-8(x+2)^3 &= -8x^3 - 48x^2 - 96x - 64,\\
23(x+2)^2 &= 23x^2 + 92x + 92,\\
-34(x+2) &= -34x - 68,
\end{aligned}$$

把以上各式与 39 相加，得

$$y^4 - 8y^3 + 23y^2 - 34y + 39 = x^4 - x^2 - 6x + 15.$$

这正是 $p = -1$、$q = -6$、$r = 15$ 的情况，这就是题目先考虑该情况的原因。

取 $u = a^2 = 9$，$a = 3$。则

$$b + c = p + a^2 = 8, \qquad b - c = \frac{q}{a} = -2,$$

所以 $b = 3$，$c = 5$。验算一下：$bc = 15 = r$。因此

$$x^4 - x^2 - 6x + 15 = (x^2 - 3x + 3)(x^2 + 3x + 5).$$

最后代入 $x = y - 2$：

$$x^2 - 3x + 3 = (y^2 - 4y + 4) - (3y - 6) + 3 = y^2 - 7y + 13,$$
$$x^2 + 3x + 5 = (y^2 - 4y + 4) + (3y - 6) + 5 = y^2 - y + 3.$$

所以

$$y^4 - 8y^3 + 23y^2 - 34y + 39 = (y^2 - 7y + 13)(y^2 - y + 3).$$

若改取 $a = -3$，则 b 和 c 互换，得出同样的两个因式，只是次序相反。

验算

- 展开答案：$(y^2 - 7y + 13)(y^2 - y + 3) = y^4 - 8y^3 + (3 + 7 + 13)y^2 - (21 + 13)y + 39$，与原式相符。
- 代入 $y = 1$：四次式的值是 $1 - 8 + 23 - 34 + 39 = 21$，而两个因式给出 $7 \times 3 = 21$。
- 两个二次因式的判别式（discriminant）都是负数（分别为 -3 和 -11），所以两者在实数范围内都不能再分解，而这个四次式没有实根。这与 $y = 1$ 时取正值相符。

我们也随机选取了 2000 组有理数 a, b, c 来检验 p, q, r 的公式和那条三次方程，并对 2000 组随机的 (p, q, r) 用二分法（bisection）找到了非负根。

常见错误

- 除以 a 得出 $b - c = q/a$，却没有处理 $a = 0$ 的情况。
- 以“三次方程必有实根”作为论据：这句话虽然正确，却没有证明该根是非负的。关键在于 $f(0) = -q^2$ 的正负号。
- 最后忘记由 x 代回 y。

STEP 之外

这基本上就是笛卡儿（Descartes）（以及费拉里（Ferrari））解一般四次方程的方法：先用平移消去三次项，再通过解一条辅助的三次方程，把四次式分解为两个二次式；这条

辅助方程称为预解三次方程（resolvent cubic）。配合卡尔达诺（Cardano）的三次方程公式，这个方法说明了每一条四次方程都可以用根式求解（solvable by radicals）。为甚么这种做法到四次为止，一般的五次方程（quintic）再没有这样的公式，正是伽罗瓦理论（Galois theory）所研究的课题。Allenby 的 *Rings, Fields and Groups*（环、域与群）是很好的入门读物。

详细解答

STEP III 2000，第 7 题

纯数学 · 不等式与极限

STEP III 2000，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/00-s3-q7

关键思路

两个不等式都是把一个乘积与我们已经熟悉的东西作比较而得出的。用二项式定理（binomial theorem）展开 $\left(1+\frac{1}{n}\right)^n$，所得各项与 $\frac{1}{r!}$ 十分相似，只是略小一点。至于乘积 $\mathrm{P}(n)$，算术平均–几何平均不等式（AM–GM inequality）可以把 n 个因子的乘积化为它们平均值的 n 次方。这些因子是 $1+\frac{1}{2^k}$，它们的平均值刚好略小于 $1+\frac{1}{n}$，于是便直接连回第一部分。

解答

上界 $\left(1+\frac{1}{n}\right)^n < \mathrm{e}$

由二项式定理，

$$\left(1+\frac{1}{n}\right)^n = \sum_{r=0}^{n}\binom{n}{r}\frac{1}{n^r} = \sum_{r=0}^{n}\frac{1}{r!}\cdot\frac{n(n-1)(n-2)\cdots(n-r+1)}{n^r}.$$

第二个分数是 r 个因子的乘积，

$$\frac{n(n-1)\cdots(n-r+1)}{n^r} = 1\cdot\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{r-1}{n}\right),$$

由于 $0 \le j \le r-1 \le n-1$，每个因子都在 $(0,1]$ 之内。因此这个乘积不超过 1，而展开式中每一项都不超过 $\frac{1}{r!}$。所以

$$\left(1+\frac{1}{n}\right)^n \le \sum_{r=0}^{n}\frac{1}{r!} < \sum_{r=0}^{\infty}\frac{1}{r!} = \mathrm{e},$$

其中最后一个不等式是严格的，因为被略去的各项 $\frac{1}{(n+1)!}, \frac{1}{(n+2)!}, \ldots$ 全部为正数。这对每一个正整数 n 都成立。

上界 $\mathrm{P}(n) < \mathrm{e}$

把乘积写成 $\mathrm{P}(n) = a_1 a_2 \cdots a_n$，其中 $a_k = \dfrac{2^k+1}{2^k} = 1 + \dfrac{1}{2^k}$，它们全部为正数，所以 AM–GM 不等式适用。算术平均（arithmetic mean）是

$$\frac{a_1 + \cdots + a_n}{n} = \frac{1}{n}\left(n + \frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^n}\right) = 1 + \frac{1 - 2^{-n}}{n},$$

这里用了等比级数（geometric series）求和 $\frac{1}{2} + \cdots + \frac{1}{2^n} = 1 - 2^{-n}$。由于 $1 - 2^{-n} < 1$，这个平均值严格小于 $1 + \frac{1}{n}$。于是 AM–GM 不等式给出

$$\mathrm{P}(n)^{1/n} \le 1 + \frac{1 - 2^{-n}}{n} < 1 + \frac{1}{n}.$$

两边都是正数，所以取 n 次方后不等式仍然成立：

$$\mathrm{P}(n) < \left(1 + \frac{1}{n}\right)^n < \mathrm{e},$$

其中第二个不等式用了第一部分的结果。这对每一个正整数 n 都成立。

> ? **想一想** 严格不等式是从哪里来的？AM–GM 不等式只给出 $\le$。

严格性来自 $1 - 2^{-n} < 1$，而不是来自 AM–GM 不等式。事实上，当 $n \ge 2$ 时，各因子 a_k 互不相同，所以 AM–GM 不等式本身也是严格的。但当 $n = 1$ 时只有一个因子，AM–GM 不等式成为等式（$\frac{3}{2} = \frac{3}{2}$）。而 $1 + \frac{1-2^{-n}}{n} < 1 + \frac{1}{n}$ 这一步在任何情况下都是严格的，所以这个论证涵盖所有 n。

极限

每个因子 $a_{n+1} = 1 + 2^{-(n+1)}$ 都大于 1，所以 $\mathrm{P}(n+1) = a_{n+1}\mathrm{P}(n) > \mathrm{P}(n)$。因此数列 $\mathrm{P}(n)$ 严格递增（strictly increasing），并且有上界（bounded above）e。有上界的递增数列必定收敛（converge），所以 $\mathrm{P}(n)$ 趋向某个极限（limit）L。

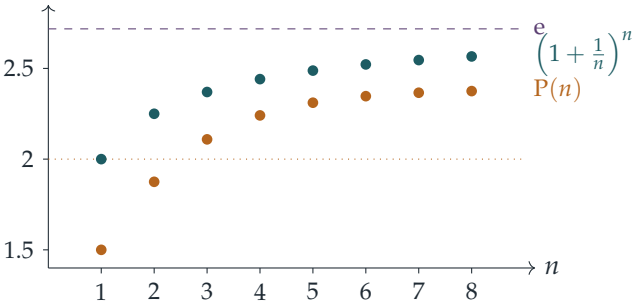
先看上界：每一个 $\mathrm{P}(n) < \mathrm{e}$，所以极限满足 $L \le \mathrm{e}$。（取极限只能保留非严格不等式，这就是为甚么我们得到的是 $\le$ 而不是 $<$。）

再看下界，先计算几项：

$$\mathrm{P}(1) = \frac{3}{2}, \qquad \mathrm{P}(2) = \frac{3}{2}\cdot\frac{5}{4} = \frac{15}{8}, \qquad \mathrm{P}(3) = \frac{15}{8}\cdot\frac{9}{8} = \frac{135}{64} > \frac{128}{64} = 2.$$

由于数列递增，$L \ge \mathrm{P}(4) > \mathrm{P}(3) > 2$。综合起来，

$$2 < L \le \mathrm{e}.$$



验算

- $n=1$：$(1+1)^1=2<\mathrm{e}$，而 $\mathrm{P}(1)=1.5<2$。当 $n=4$：$\mathrm{P}(4)=\frac{2295}{1024}\approx 2.241<\left(\frac{5}{4}\right)^4\approx 2.441$。
- 以数值方法计算这个乘积，$\mathrm{P}(n)$ 稳定在 $L\approx 2.3842$（到 $n=60$ 时，它在小数点后很多个位都已不再变动），这个数值稳妥地落在 2 与 $\mathrm{e}\approx 2.7183$ 之间。

常见错误

- 仅仅因为每个 $\mathrm{P}(n)<\mathrm{e}$ 就声称 $L<\mathrm{e}$。严格不等式在取极限后可能变成等式；能够推出的只有 $L\le \mathrm{e}$。（这里 $L<\mathrm{e}$ 其实是对的，但需要另作论证。）
- 证明了 $\mathrm{P}(n)>2$ 对某个 n 成立，却忘记说明数列是递增的，而正是这一点才把这个界传给 L。
- 使用 AM–GM 不等式时没有检查各 a_k 是否为正数。

STEP 之外

“递增且有上界，所以收敛”这一步就是单调收敛定理（monotone convergence theorem），它与实数的完备性（completeness）等价。从基本原理出发证明这个定理，是大学分析课程最先做的事情之一。取对数后，$\mathrm{P}(n)$ 变成级数 $\sum \ln\left(1+2^{-k}\right)$，把它与 $\sum 2^{-k}$ 比较便可知它收敛。这是研究无穷乘积（infinite product）的标准方法。Abbott 的 *Understanding Analysis*（一本分析学入门书）把这两个想法都讲得十分精彩。

详细解答

STEP III 2000，第 8 题

纯数学 · 递推关系

STEP III 2000，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/00-s3-q8

关键思路

定义数列的规则是非线性的，而且涉及除法，直接处理并不方便。窍门是把它改写成乘积的形式：$a_n a_{n-2} - a_{n-1}^2 = 1$。这是说，连续三项的某个组合恒等于 1。把同一事实往后推一步再写一次，然后将两个式子相减，非线性的规则便化为线性的规则。递推关系（recurrence）一旦变成线性，求通项公式（closed form）就只需用上二阶递推关系的标准方法：先求辅助二次方程（auxiliary quadratic）的根，再配合两个初始值。

解答

预备工作：数列有明确定义

每一步都要除以 a_{n-2}，所以我们应先确定没有任何一项是零。事实上每一项都是正数。我们有 $a_0 = a_1 = 1 > 0$。若 a_{n-1} 和 a_{n-2} 均为正数，则 $1 + a_{n-1}^2 > 0$ 且 $a_{n-2} > 0$，所以 $a_n > 0$。根据归纳法（induction），对所有 n 均有 $a_n > 0$。最初几项是

$$a_0 = 1, \quad a_1 = 1, \quad a_2 = \frac{1+1}{1} = 2, \quad a_3 = \frac{1+4}{1} = 5, \quad a_4 = \frac{1+25}{2} = 13.$$

线性递推关系

把定义式乘以 a_{n-2}，可得对每个 $n \geq 2$，

$$a_n a_{n-2} = 1 + a_{n-1}^2. \tag{$*$}$$

设 $P(k)$ 为命题 $a_k = 3a_{k-1} - a_{k-2}$。我们对 k 用归纳法，证明 $P(k)$ 对所有 $k \geq 2$ 成立。

基础步骤。$a_2 = 2$，而 $3a_1 - a_0 = 3 - 1 = 2$，所以 $P(2)$ 成立。

归纳步骤。假设 $P(k)$ 对某个 $k \geq 2$ 成立，即

$$a_k + a_{k-2} = 3a_{k-1}.$$

我们需要证明 $a_{k+1} = 3a_k - a_{k-1}$。根据定义，$a_{k+1} = (1 + a_k^2)/a_{k-1}$。做法是利用 $n = k$ 时的 $(*)$，把分子中那个碍事的 1 换掉；该式说明 $1 = a_k a_{k-2} - a_{k-1}^2$：

$$\begin{aligned} a_{k+1} &= \frac{a_k a_{k-2} - a_{k-1}^2 + a_k^2}{a_{k-1}} \\ &= \frac{a_k\,(a_k + a_{k-2}) - a_{k-1}^2}{a_{k-1}} \\ &= \frac{a_k \cdot 3a_{k-1} - a_{k-1}^2}{a_{k-1}} && \text{根据} P(k) \\ &= 3a_k - a_{k-1} && \text{因为} a_{k-1} > 0. \end{aligned}$$

所以 $P(k+1)$ 成立。根据归纳法，对所有 $n \geq 2$ 均有 $a_n = 3a_{n-1} - a_{n-2}$。

想一想 归纳假设（inductive hypothesis）在哪里用上了？为甚么可以放心使用 $n = k$ 时的 $(*)$？

归纳假设只用了一次，就是把 $a_k + a_{k-2}$ 换成 $3a_{k-1}$。这个括号之所以出现，全因我们把 1 换成了 $a_k a_{k-2} - a_{k-1}^2$。使用 $(*)$ 则完全不需要任何假设：它不过是把 a_k 的定义乘开，对每个 $k \geq 2$ 都成立。换句话说，$a_n a_{n-2} - a_{n-1}^2$ 这个量在每一步都相同，正是它把线性关系一步一步传递下去。

通项公式

记 $\beta = \alpha^{-1}$。由于 $\alpha = \frac{1+\sqrt{5}}{2}$，

$$\beta = \frac{2}{1+\sqrt{5}} = \frac{2(\sqrt{5}-1)}{5-1} = \frac{\sqrt{5}-1}{2}, \qquad \text{所以} \qquad \alpha + \beta = \sqrt{5}, \quad \alpha\beta = 1.$$

因此

$$\alpha^2 + \beta^2 = (\alpha+\beta)^2 - 2\alpha\beta = 5 - 2 = 3, \qquad \alpha^2\beta^2 = 1,$$

所以 α^2 和 β^2 就是我们这个递推关系的辅助方程 $x^2 - 3x + 1 = 0$ 的两个根。（具体来说，它们是 $\frac{3\pm\sqrt{5}}{2}$。）

现在对 $n \geq 1$ 定义

$$b_n = \frac{\alpha^{2n-1} + \beta^{2n-1}}{\sqrt{5}}.$$

我们只要验证 b_n 与 a_n 的首两项相同，并且满足同一个递推关系，便可证明 $b_n = a_n$。

它满足递推关系。若 x 是 $x^2 = 3x - 1$ 的根，则乘以 x^{n-2} 可得 $x^n = 3x^{n-1} - x^{n-2}$。由于 $\alpha^{2n-1} = \beta\,(\alpha^2)^n$（这里用了 $\alpha\beta = 1$），取 $x = \alpha^2$ 并乘以 β，得

$$\alpha^{2n-1} = 3\alpha^{2n-3} - \alpha^{2n-5}.$$

同样地，取 $x = \beta^2$ 并乘以 α，可得 $\beta^{2n-1} = 3\beta^{2n-3} - \beta^{2n-5}$。两式相加再除以 $\sqrt{5}$：

$$b_n = 3b_{n-1} - b_{n-2} \qquad (n \geq 3).$$

初始值相符。

$$b_1 = \frac{\alpha + \beta}{\sqrt{5}} = \frac{\sqrt{5}}{\sqrt{5}} = 1 = a_1,$$
$$b_2 = \frac{\alpha^3 + \beta^3}{\sqrt{5}} = \frac{(\alpha+\beta)^3 - 3\alpha\beta(\alpha+\beta)}{\sqrt{5}} = \frac{5\sqrt{5} - 3\sqrt{5}}{\sqrt{5}} = 2 = a_2.$$

结论。(a_n) 和 (b_n) 都对 $n \geq 3$ 满足 $u_n = 3u_{n-1} - u_{n-2}$，而且在 $n = 1$ 和 $n = 2$ 时相等。这类递推关系由前两项决定每一项，所以根据（强）归纳法（strong induction），对每个 $n \geq 1$ 均有 $a_n = b_n$。也就是说，

$$a_n = \frac{\alpha^{2n-1} + \alpha^{-(2n-1)}}{\sqrt{5}} \qquad (n \geq 1),$$

这正是所要证明的。

> ? **想一想** 为甚么从 $n = 1$ 而不是 $n = 0$ 开始比较?

a_n 的线性递推关系在 $n = 3$ 时涉及 a_2 和 a_1，所以要处理 $n \geq 1$ 的情况，只需核对 b_1 和 b_2 便已足够。其实公式在 $n = 0$ 时同样成立：$b_0 = (\alpha^{-1} + \alpha)/\sqrt{5} = 1 = a_0$。所以无论从哪里开始，都没有损失。

验算

- 各项为 $1, 1, 2, 5, 13, 34, 89, 233, \ldots$，即斐波那契数（Fibonacci numbers）$F_1, F_3, F_5, \ldots$（其中 $a_0 = 1 = F_{-1}$）。每一项都等于前一项的三倍减去再前一项：$3 \cdot 13 - 5 = 34$。
- 直接验证 $(*)$：$34 \cdot 5 - 13^2 = 170 - 169 = 1$。
- 数值上，$a_{10} = 4181$，而 $(\alpha^{19} + \alpha^{-19})/\sqrt{5} \approx 4181.000$。$\alpha^{-19}/\sqrt{5}$ 这一项极小，所以 a_n 就是最接近 $\alpha^{2n-1}/\sqrt{5}$ 的整数。
- 按原来的规则精确计算数列直至 $n = 39$，所得各项都是整数，每次都满足线性递推关系，并与通项公式吻合。

常见错误

- 试图把 $a_k = 3a_{k-1} - a_{k-2}$ 代入 $(1 + a_k^2)/a_{k-1}$ 再展开，以证明 $a_{k+1} = 3a_k - a_{k-1}$。这样只会弄得一团糟。那个 1 必须用前一步的定义来替换。
- 忘记定义中要除以 a_{n-2}。完整的答案应指出各项均为正数，因此这个除法以及最后约去 a_{k-1} 的一步都是合理的。
- 只验证 b_n 满足递推关系，却没有核对两个初始值。二阶递推关系需要两个初始值才能确定。

STEP 之外

常数 $a_n a_{n-2} - a_{n-1}^2$ 是斐波那契数的卡西尼恒等式（Cassini's identity）$F_{m+1}F_{m-1} - F_m^2 = (-1)^m$ 的一个例子。这里只出现奇数下标，所以符号永远是 $+1$。更一般地，对任何二阶线性递推关系，$u_{n+1}u_{n-1} - u_n^2$ 每一步都乘以一个固定的因子（它是一个 2×2 行列式（determinant），而行列式具有可乘性）。像本题这样的有理递推关系，竟然出人意料地总是产生整数，这与“Somos 数列”（Somos sequences）和洛朗现象（Laurent phenomenon）的研究有关。至于斐波那契型数列和连分数（continued fractions）的数论（留意 α 就是黄金比（golden ratio）），Davenport 的 *The Higher Arithmetic*（《高等算术》）是一本十分值得细读的好书。

详细解答

STEP III 2000，第 9 题

力学 · 弹性斜碰撞

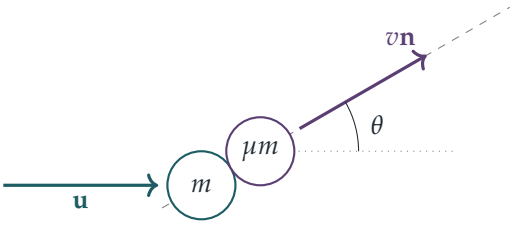
> **STEP III 2000，第 9 题**
>
> 继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：
>
> step.maths.org/questions/00-s3-q9

关键思路

两个圆盘都是光滑的，所以碰撞期间它们之间唯一的力沿连心线（line of centres）作用。原本静止的圆盘因此会沿这条线离开，所以 $\mathbf{n}$ 就是连心线的方向。把它的速度写成 $v\mathbf{n}$，用动量守恒（conservation of momentum，这是一条向量方程）求出另一个圆盘的速度，再让动能守恒（conservation of kinetic energy）定出 v。第 (ii) 部分只是多用一条能量方程，而不等式则来自余弦值不能超过 1 这个要求。

解答

我们把两个圆盘当作光滑水平面上的光滑质点（particle），所以圆盘之间以及圆盘与平面之间都没有摩擦力。设 θ 为 $\mathbf{u}$ 与 $\mathbf{n}$ 之间的夹角，并记 $u = |\mathbf{u}|$，于是 $\mathbf{u}\cdot\mathbf{n} = u\cos\theta$。



在碰撞的一刻，连心线（虚线）指向 $\mathbf{n}$ 的方向，与 $\mathbf{u}$ 的方向（点线）成 θ 角。这幅图只是示意图，代数运算并不需要用到它。

第 (i) 部分

质量为 μm 的圆盘原本静止，碰撞期间它所受的唯一水平力，是另一个圆盘施加的接触力，而这个力沿连心线作用。由于它沿 $\mathbf{n}$ 的方向离开，它碰撞后的速度是 $v\mathbf{n}$，其中 $v > 0$ 是某个速率。

设质量为 m 的圆盘碰撞后的速度为 $\mathbf{w}$。由动量守恒得

$$m\mathbf{u} = m\mathbf{w} + \mu m v\,\mathbf{n} \qquad \Longrightarrow \qquad \mathbf{w} = \mathbf{u} - \mu v\,\mathbf{n}.$$

这次碰撞是完全弹性（perfectly elastic）的，所以动能守恒：

$$\tfrac{1}{2}m\,\mathbf{u}\cdot\mathbf{u} = \tfrac{1}{2}m\,\mathbf{w}\cdot\mathbf{w} + \tfrac{1}{2}\mu m v^2.$$

利用 $\mathbf{n}\cdot\mathbf{n} = 1$，

$$\mathbf{w}\cdot\mathbf{w} = (\mathbf{u} - \mu v\mathbf{n})\cdot(\mathbf{u} - \mu v\mathbf{n}) = u^2 - 2\mu v\,\mathbf{u}\cdot\mathbf{n} + \mu^2 v^2.$$

代入并消去 $\frac{1}{2}m$ 和 u^2，

$$0 = -2\mu v\,\mathbf{u}\cdot\mathbf{n} + \mu^2 v^2 + \mu v^2 = \mu v\big((1+\mu)v - 2\,\mathbf{u}\cdot\mathbf{n}\big).$$

现在 $\mu > 0$，而且 $v \neq 0$，因为被撞的圆盘确实动了。所以我们可以除以 μv，得到

$$v = \frac{2\,\mathbf{u}\cdot\mathbf{n}}{1+\mu},$$

这就是所求的速率。顺带一提，$v > 0$ 告诉我们 $\mathbf{u}\cdot\mathbf{n} > 0$，所以 $0 \le \theta < \frac{1}{2}\pi$。

想一想 为甚么这里不需要用牛顿实验定律（Newton's experimental law）?

其实也可以用：取 $e = 1$，相对速度沿 $\mathbf{n}$ 的分量会反转，而垂直于 $\mathbf{n}$ 的分量则因为圆盘光滑而保持不变。但“完全弹性”的意思就是动能守恒，而这一条纯量方程（scalar equation）连同动量守恒，已经足以确定 v。上面的做法与之完全等价，而且不必分解成分量。

第 (ii) 部分

碰撞后的总动能等于最初的 $\frac{1}{2}mu^2$。如果两个圆盘平分这些动能，每个就有 $\frac{1}{4}mu^2$，所以对被撞的圆盘而言

$$\tfrac{1}{2}\mu m v^2 = \tfrac{1}{4}mu^2 \qquad \Longrightarrow \qquad \mu v^2 = \tfrac{1}{2}u^2.$$

利用第 (i) 部分的结果和 $\mathbf{u}\cdot\mathbf{n} = u\cos\theta$，我们有 $v = \dfrac{2u\cos\theta}{1+\mu}$，所以

$$\frac{4\mu u^2\cos^2\theta}{(1+\mu)^2} = \frac{u^2}{2} \qquad \Longrightarrow \qquad \cos^2\theta = \frac{(1+\mu)^2}{8\mu}.$$

由于 $\cos\theta > 0$（见第 (i) 部分末段），我们取正根：

$$\cos\theta = \frac{1+\mu}{2\sqrt{2\mu}}.$$

要这样的角存在，我们需要 $\cos\theta \le 1$，即 $\cos^2\theta \le 1$：

$$\frac{(1+\mu)^2}{8\mu} \le 1 \iff (1+\mu)^2 \le 8\mu \qquad \text{（两边乘以}8\mu > 0\text{）}$$
$$\iff \mu^2 - 6\mu + 1 \le 0.$$

二次式 $\mu^2 - 6\mu + 1$ 的根是 $\mu = 3 \pm \sqrt{9-1} = 3 \pm \sqrt{8}$，而它恰好在两根之间（包括两根）为负数或零。因此

$$3 - \sqrt{8} \le \mu \le 3 + \sqrt{8},$$

这正是所要证明的结果。（下界是正数，$3 - \sqrt{8} \approx 0.172$，所以与 $\mu > 0$ 并无矛盾。）反过来说，对这个范围内的每一个 μ，公式都给出一个在 $(0, 1]$ 内的有效余弦值，所以平分动能确实可以做到；在范围的两端 $\theta = 0$，即正面碰撞（head-on collision）。

验算

- 正面碰撞的情况：如果 $\mathbf{n}$ 与 $\mathbf{u}$ 平行，第 (i) 部分给出 $v = 2u/(1+\mu)$，这是大家熟悉的一维弹性碰撞结果。当 $\mu = 1$ 时就是 $v = u$：运动中的圆盘停下来，把全部运动传给对方，就像桌球中的情形一样。
- 质量相等，$\mu = 1$：第 (ii) 部分给出 $\cos\theta = 2/(2\sqrt{2}) = 1/\sqrt{2}$，所以 $\theta = 45°$。这符合一个经典事实：两个质量相等的光滑圆盘作弹性碰撞后，会沿互相垂直的方向离开；动能相等又迫使两者速率相等，所以它们各与 $\mathbf{u}$ 成 $45°$ 角。
- 数值方面，我们模拟了 10 000 次随机的光滑弹性碰撞（冲量沿 $\mathbf{n}$ 方向），确认了动量守恒、能量守恒和第 (i) 部分的速率，并确认了当 μ 介乎 0.172 与 5.83 之间时，在第 (ii) 部分求得的角度下两者动能相等。举例来说，$\mu = 0.1$ 给出 $\cos^2\theta = 1.51 > 1$，这是不可能的。

常见错误

- 把 $\mathbf{u}$ 和 $\mathbf{n}$ 当作平行。这是斜碰（oblique collision）；只有沿连心线的分量 $\mathbf{u}\cdot\mathbf{n}$ 会传递过去。
- 除以 v 却没有说明它为何不等于零（$v = 0$ 是能量方程中“没有发生碰撞”的那个解）。
- 忘记排除 $\cos\theta$ 的负平方根。

STEP 之外

把被撞质点的速度写成沿冲量（impulse）方向，然后运用动量和能量，这个技巧正是物理学中弹性散射（elastic scattering）的核心。在质心参考系（centre-of-mass frame）中，情况最为简洁：弹性碰撞只是把两个速度旋转，而不改变它们的大小；第 (ii) 部分的角度范围，于是变成一个关于实验室参考系（lab frame）中哪些角度可以达到的命题。Kleppner 与 Kolenkow 合著的 *An Introduction to Mechanics*（力学入门教材）把两个参考系中的碰撞都讲解得非常清楚。

详细解答

STEP III 2000，第 10 题

力学 · 极限摩擦力平衡

STEP III 2000，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/00-s3-q10

关键思路

先从几何入手：题目选用 $\sqrt{3}a$ 和 a 这两个距离，是为了令梁与地面成一个特殊角，而这个角决定了题中所有方向。然后对两个能令未知力消失的点取力矩（moment）。对铰链（hinge）X 取力矩时，梁的方程只含球体给予的法向反作用力（normal reaction），因为球体给予的摩擦力（friction）正好沿著梁本身的方向。对球心取力矩时，除了两个摩擦力之外，其余各力都通过球心，所以两个摩擦力必定相等。

解答

模型与几何

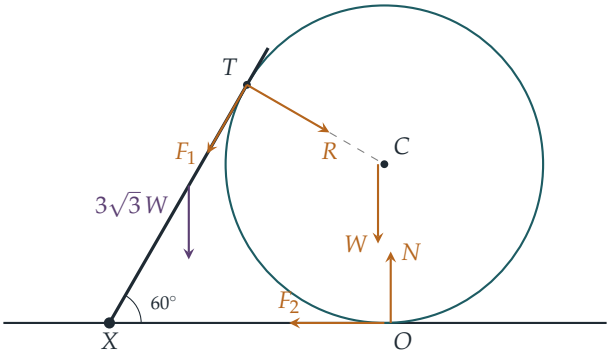
我们把梁视为均匀（uniform）杆，把球体视为均匀刚体（rigid body），其重量作用于球心 C。两个接触点都是粗糙的。设 O 为球体与地面的接触点，则 $OX = \sqrt{3}a$，$OC = a$，而 OC 是铅垂的。

梁与球体在点 T 接触，并在该处与球体相切。地面则是由 X 所作的另一条切线（tangent）。在直角三角形 XOC 中，

$$\tan\angle OXC = \frac{a}{\sqrt{3}a} = \frac{1}{\sqrt{3}}, \qquad \text{所以} \qquad \angle OXC = 30^\circ.$$

由 X 所作的两条切线关于直线 XC 对称，所以梁与地面的夹角是 $2\times 30^\circ = 60^\circ$。由同一点所作的切线长度相等，所以 $XT = XO = \sqrt{3}a$。由于 $\sqrt{3}a < 2a$，接触点确实在梁上，位置稍稍超过梁的中点。

半径 CT 垂直于梁。因此，在 T 的法向反作用力与水平方向成 30°，而在 T 的摩擦力沿著梁的方向，与水平方向成 60°。



琥珀色的箭头表示作用于球体的力。梁所受的 R 和 F_1 方向与此相反，此外还有作用于梁中点的重量 $3\sqrt{3}\,W$（与 X 相距 a），以及铰链处一个未知的反作用力。

摩擦力的方向是我们猜测的：对球体而言，F_1 沿著梁向下指向 X，而 F_2 沿著地面指向 X。假如猜错了，该力算出来便会是负数。

梁：对 X 取力矩

铰链的反作用力作用于 X，所以它对 X 没有力矩。球体给予的摩擦力沿著梁的方向，而梁所在的直线通过 X，所以它也没有力矩。法向反作用力 R 垂直于梁，作用点的距离为 $XT = \sqrt{3}\,a$。重量的作用线与 X 的水平距离为 $a\cos 60^\circ = \frac{1}{2}a$。因此

$$R \cdot \sqrt{3}\,a = 3\sqrt{3}\,W \cdot \tfrac{1}{2}a, \qquad \text{所以} \qquad R = \tfrac{3}{2}W.$$

球体：对 C 取力矩

R、W 和 N 这三个力都通过 C。（法向反作用力沿半径方向作用。）两个摩擦力都沿切线方向，各与 C 相距 a。在 T 的 F_1 使球体绕 C 向一个方向转动，而在 O 的 F_2 则使它向相反方向转动，所以

$$F_1 a = F_2 a, \qquad \text{所以} \qquad F_1 = F_2.$$

把这个共同的值记为 F。

> ? **想一想** 在对两个摩擦力还一无所知的时候，为甚么我们已经可以肯定它们相等?

球体处于平衡（equilibrium）状态，所以对任何一点的总力矩都是零。选球心最为巧妙：重量和两个法向反作用力都通过球心，于是只剩下两个摩擦力，而它们的力臂（lever arm）同样是 a。

球体：分解力

水平方向（resolving，即把力分解），取远离 X 的方向为正：R 的分量是 $R\cos 30^\circ$，F_1 的分量是 $-F\cos 60^\circ$，而 F_2 的贡献是 $-F$。所以

$$\frac{\sqrt{3}}{2}R = \frac{1}{2}F + F = \frac{3}{2}F, \qquad \text{所以} \qquad F = \frac{R}{\sqrt{3}} = \frac{\sqrt{3}}{2}W.$$

这个值是正数，可见先前猜测的摩擦力方向是对的。铅垂方向，R 向下推，分量为 $R\sin 30^\circ$；F_1 向下拉，分量为 $F\sin 60^\circ$。所以

$$N = W + \tfrac{1}{2}R + \tfrac{\sqrt{3}}{2}F = W + \tfrac{3}{4}W + \tfrac{3}{4}W = \tfrac{5}{2}W.$$

摩擦系数

球体在两个接触点都即将滑动，所以两处的摩擦都达到极限（limiting friction）：$F_1 = \mu_1 R$ 及 $F_2 = \mu_2 N$。因此，两个摩擦系数（coefficient of friction）为

$$\mu_1 = \frac{F}{R} = \frac{\frac{\sqrt{3}}{2}W}{\frac{3}{2}W} = \frac{1}{\sqrt{3}}, \qquad \mu_2 = \frac{F}{N} = \frac{\frac{\sqrt{3}}{2}W}{\frac{5}{2}W} = \frac{\sqrt{3}}{5}.$$

验算

- 用坐标作数值验算（$a = 1.3$，$W = 2$），确认了 T 在球面上，而且 CT 垂直于梁。直接解球体的三条平衡方程，得出 $R = 1.5W$、$F_1 = F_2 \approx 0.8660W$ 和 $N = 2.5W$，所以 $\mu_1 \approx 0.5774 = 1/\sqrt{3}$，$\mu_2 \approx 0.3464 = \sqrt{3}/5$。
- 一条独立的方程：对球体以 O 为支点取力矩。这时 N、F_2 和 W 都没有力矩。R 的作用线通过 C，所以它的力矩等于其水平分量在高度 a 处的力矩，即 $\frac{\sqrt{3}}{2}R\,a$。F_1 对 O 的力矩算出来是 $\frac{3}{2}aF$。令两者相等，便再次得到 $F = R/\sqrt{3}$，而无须用到对 C 的力矩方程。
- 两个系数都小于 1，对一般的粗糙接触面来说，这在物理上是合理的。

常见错误

- 把梁的倾角当作 30°（即 XC 的倾角），而不是 60°。
- 求 N 时忘记了 F_1 的铅垂分量。这样会得出 $N = \frac{7}{4}W$，以及错误的 μ_2。
- 在梁对 X 的力矩方程中计入球体给予的摩擦力。这个力沿著梁的方向，所以它对 X 的力矩是零。

STEP 之外

平面上的刚体有三个自由度（degrees of freedom），所以可给出三条独立的平衡方程。本题中，球体有五个未知的接触力，而梁提供了 R，减少了未知数的数目。再加上两个“极限摩擦”条件，问题便成为一个可解的方程组。像这样把方程的数目与未知数的数目互相比较，并辨认出一个系统何时属于静不定（statically indeterminate），正是结构力学（structural mechanics）的起点。Kleppner 与 Kolenkow 所著的 *An Introduction to Mechanics*（力学入门教科书）对刚体静力学和力矩（torque）有细致的论述。

详细解答

STEP III 2000，第 11 题

力学 • 绕粗糙支点的转动

STEP III 2000，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/00-s3-q11

关键思路

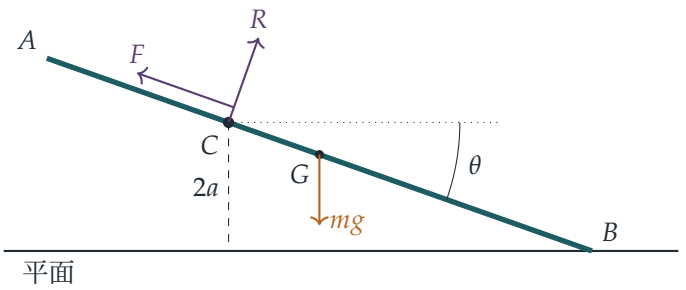
只要没有滑动,棒就只是绕着横梁上的点 C 转动,因此它是一个绕固定轴转动的刚体(rigid body)。由能量可求得角速率（angular speed），由力矩方程可求得角加速度（angular acceleration)。然后对质心（centre of mass）应用牛顿第二定律（Newton's second law)，分别沿棒的方向和垂直于棒的方向分解，便可知道横梁必须提供的摩擦力（friction）和法向反作用力（normal reaction)。只要摩擦力不超过法向反作用力的 μ 倍，棒就不会滑动，而我们需要这个条件一直成立，直至 B 到达平面为止。

解答

模型与几何

我们把棒视为均匀刚棒，把横梁视为一条固定而粗糙的细线，因此横梁只在 C 处施加单一个接触力：垂直于棒的法向反作用力 R 和沿棒方向的摩擦力 F。

质心 G 是棒的中点，与 A 相距 $4.5a$。由于 $AC = 3a$，G 位于 C 靠 B 的一侧，且 $CG = \frac{3}{2}a$，$CB = 6a$。所以重力对 C 有力矩（moment），棒会倾侧，B 端向下。设 θ 为棒转过的角，即棒与水平线所成的角为 θ。



当 B 下降了 $2a$ 时便到达平面，即当 $6a\sin\theta = 2a$ 时，亦即

$$\sin\theta = \tfrac{1}{3}, \qquad \tan\theta = \frac{1}{2\sqrt{2}}.$$

绕 C 的转动

根据平行轴定理（parallel axes theorem），对 C 的转动惯量（moment of inertia）为

$$I_C = \tfrac{1}{12}m(9a)^2 + m\left(\tfrac{3}{2}a\right)^2 = \tfrac{27}{4}ma^2 + \tfrac{9}{4}ma^2 = 9ma^2.$$

在没有滑动的期间，C 固定不动，接触力不作功。G 下降了 $\frac{3}{2}a\sin\theta$，所以由能量守恒（conservation of energy）（从静止开始）得

$$\tfrac{1}{2}(9ma^2)\dot{\theta}^2 = mg\cdot\tfrac{3}{2}a\sin\theta \quad\Longrightarrow\quad \dot{\theta}^2 = \frac{g\sin\theta}{3a}.$$

对 C 取力矩（只有重力有力矩）：

$$9ma^2\,\ddot{\theta} = mg\cdot\tfrac{3}{2}a\cos\theta \quad\Longrightarrow\quad \ddot{\theta} = \frac{g\cos\theta}{6a}.$$

（把能量方程对 t 求导，会得到相同的结果，这是一个有用的验算。）

接触力

G 在以 C 为圆心、半径为 $\frac{3}{2}a$ 的圆上运动，所以它的加速度有一个沿棒指向 C 的分量 $\frac{3}{2}a\dot{\theta}^2$，以及一个垂直于棒的分量 $\frac{3}{2}a\ddot{\theta}$，方向为 θ 增加的方向（即垂直于棒向下）。

> ? **想一想** 摩擦力的方向是哪一边？为甚么它不单单是 $mg\sin\theta$？

重力沿棒向下的分量 $mg\sin\theta$ 会把棒朝 B 的方向拉离横梁；此外，G 还需要一个指向 C 的向心力（centripetal force）。两者都必须由摩擦力提供，因此摩擦力沿棒向上，指向 A。沿棒的方向分解，并以指向 C 的方向作为 G 的加速度的正方向：

$$F - mg\sin\theta = m\cdot\tfrac{3}{2}a\dot{\theta}^2 = m\cdot\tfrac{3}{2}a\cdot\frac{g\sin\theta}{3a} = \tfrac{1}{2}mg\sin\theta,$$

所以

$$F = \tfrac{3}{2} mg \sin\theta.$$

垂直于棒的方向分解，取 G 加速的方向：

$$mg\cos\theta - R = m \cdot \tfrac{3}{2} a\ddot{\theta} = m \cdot \tfrac{3}{2} a \cdot \frac{g\cos\theta}{6a} = \tfrac{1}{4} mg\cos\theta,$$

所以

$$R = \tfrac{3}{4} mg\cos\theta,$$

这在整个过程中都是正数，所以棒一直与横梁保持接触。

不滑动的条件

只要 $F \le \mu R$，棒就不会在 C 处滑动，即

$$\frac{F}{R} = \frac{\frac{3}{2} mg\sin\theta}{\frac{3}{4} mg\cos\theta} = 2\tan\theta \le \mu.$$

比值 $2\tan\theta$ 随 θ 递增，在释放时由 0 开始，所以这个条件在最后一刻最难满足，亦即 B 刚到达平面之时。该处 $\tan\theta = 1/(2\sqrt{2})$，所以我们需要

$$\mu \ge 2 \cdot \frac{1}{2\sqrt{2}} = \frac{1}{\sqrt{2}}.$$

因此 μ 的最小值为

$$\boxed{\mu = \frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2} \approx 0.707.}$$

若恰好取这个值，摩擦力只在 B 撞到平面的一瞬间才达到极限；若 μ 再小一点，棒便会在此之前开始滑动。

验算

- 释放时 $\dot{\theta} = 0$，所以 $F = 0$ 而 $R = \frac{3}{4} mg$：横梁所承受的力小于整个重量，因为 G 开始以 $\frac{3}{2} a\ddot{\theta} = \frac{1}{4} g$ 向下加速。这就是常见的“放手后的木板感觉变轻”效应。
- 单位：F 和 R 都是 mg 的倍数，而 μ 是无量纲（dimensionless）的，理应如此。
- 我们由静止开始，用数值方法逐步求解方程 $\ddot{\theta} = g\cos\theta/(6a)$，以有限差分（finite differences）计算 G 的加速度，再由牛顿第二定律还原出 F 和 R。结果与 $\frac{3}{2} mg\sin\theta$ 和 $\frac{3}{4} mg\cos\theta$ 吻合，而 F/R 稳步上升，在 $\sin\theta = \frac{1}{3}$ 时达到 $0.7071\ldots = 1/\sqrt{2}$。

常见错误

- 对棒的中心取力矩，或忘记平行轴定理中的附加项：$I_C = 9ma^2$，而不是 $\frac{27}{4}ma^2$。
- 把问题当作静力学（statics）来处理，取 $F = mg\sin\theta$、$R = mg\cos\theta$，从而得出 $\mu \geq \tan\theta$。其实 G 的向心加速度和切向加速度都会改变这些力。
- 只在释放的一刻检查摩擦条件（那时 $F = 0$），而没有在条件最严苛的位置检查。

STEP 之外

这道题让我们初尝带有约束力（constraint forces）的刚体动力学：我们先由能量和力矩求出运动，然后才对质心应用牛顿定律，倒推出约束必须提供的力。同样的模式（先解出运动，再检查约束是否真的能提供所需的力）可以用来判断梯子何时开始滑动、倾倒的铅笔何时离开桌面，或者珠子何时离开铁线。Kleppner 与 Kolenkow 合著的 *An Introduction to Mechanics*（力学入门）仔细讲解了定轴转动，以及力矩与质心运动之间的关系。

详细解答

STEP III 2000，第 12 题

概率与统计 · 泊松近似

> **STEP III 2000，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/00-s3-q12

关键思路

一旦知道自己的彩票中奖，余下的随机性就只在其他玩家的彩票上。他们每人各自独立地以概率 $\frac{1}{N}$ 选中中奖号码，所以其他中奖者的人数服从二项分布（binomial distribution）。当玩家众多而成功概率很小时，这个二项分布接近泊松分布（Poisson distribution），而题目给出的极限 $\left(1-\frac{a}{n}\right)^n \approx \mathrm{e}^{-a}$ 正是令这个近似成立的关键。

解答

精确概率

假设你持有一张中奖彩票。设 X 为其余 $2N-1$ 位玩家中同样持有中奖号码的人数，则中奖者总数为 $W = 1 + X$。

机器为每位玩家选号时，是均匀（uniform）地抽选，与其他人的号码互相独立（independent），亦与开奖结果独立。因此，给定中奖号码（不论是哪一个）及你的彩票后，其余 $2N-1$ 位玩家每人以概率 $\frac{1}{N}$ 选中该号码，且彼此独立。所以

$$X \sim \mathrm{B}\left(2N-1,\ \tfrac{1}{N}\right).$$

中奖者不多于三人即 $X \le 2$，故精确概率为

$$\mathrm{P}(W \le 3) = \sum_{k=0}^{2} \binom{2N-1}{k} \left(\frac{1}{N}\right)^k \left(1-\frac{1}{N}\right)^{2N-1-k}.$$

完整写出来就是

$$\left(1-\frac{1}{N}\right)^{2N-1} + \frac{2N-1}{N}\left(1-\frac{1}{N}\right)^{2N-2} + \frac{(2N-1)(2N-2)}{2N^2}\left(1-\frac{1}{N}\right)^{2N-3}.$$

近似值

对每个固定的 $k \in \{0,1,2\}$，把幂拆开：

$$\left(1-\frac{1}{N}\right)^{2N-1-k} = \left[\left(1-\frac{1}{N}\right)^{N}\right]^{2}\left(1-\frac{1}{N}\right)^{-(1+k)}.$$

在题目给出的结果中取 $a=1$ 及 $n=N$，方括号内的式子约为 e^{-1}。最后一个因子是一个非常接近 1 的数的固定小次幂，所以约为 1。因此每个幂都约为 e^{-2}。

至于二项式系数（binomial coefficient），k 最大为 2 而 N 很大，所以

$$\binom{2N-1}{k}\frac{1}{N^k} = \frac{(2N-1)(2N-2)\cdots(2N-k)}{k!\,N^k} \approx \frac{(2N)^k}{k!\,N^k} = \frac{2^k}{k!}.$$

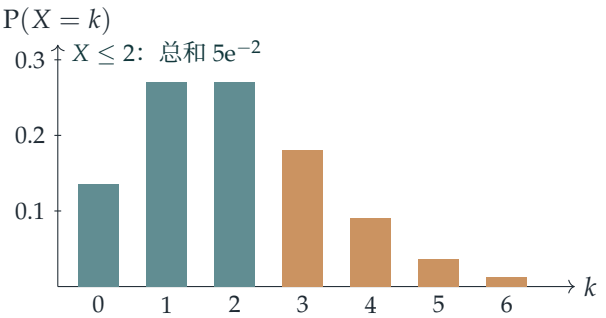
综合以上各项，$k=0,1,2$ 的项分别约为 e^{-2}、$2\mathrm{e}^{-2}$ 和 $2\mathrm{e}^{-2}$，所以

$$\mathrm{P}(W \le 3) \approx \mathrm{e}^{-2}\left(1+2+\frac{2^2}{2!}\right) = 5\mathrm{e}^{-2} \approx 0.677 \approx \frac{2}{3},$$

其中用了 $\mathrm{e}^2 \approx 7.39$。这就是所求的结果。

> **想一想** 为甚么 X 的表现有如一个平均值为 2 的泊松变量?

上面的计算就是二项分布的标准泊松极限。试验次数 $2N-1$ 很大，成功概率 $\frac{1}{N}$ 很小，而两者之积始终接近 2。在这种情况下，$\binom{n}{k}p^k(1-p)^{n-k} \to \mathrm{e}^{-\lambda}\lambda^k/k!$，其中 $\lambda = np$，而这里 $\lambda \to 2$。



X 的极限分布是平均值为 2 的泊松分布。蓝绿色的长条（$X \le 2$，即总共最多三位中奖者）约占总和的三分之二。

玩家自行选号

如果由玩家自行选号，号码就不再均匀分布：人们偏爱生日（不超过 31 的号码）、幸运号码和有规律的号码。假设每位玩家以概率 p_i 选择号码 i，其中 $\sum p_i = 1$，而中奖号码

仍是均匀抽出。那么，在你中奖的条件下，你的号码是 i 的概率与 $p_i \cdot \frac{1}{N}$ 成正比，即概率为 p_i。所以热门号码在中奖彩票中所占的比例偏高。若你的号码是 i，其他中奖者的人数服从 $\mathrm{B}(2N-1, p_i)$，平均值为 $(2N-1)p_i$。对 i 取平均，其他中奖者人数的期望值（expectation）为

$$(2N-1)\sum_i p_i^2 \;\geq\; (2N-1)\cdot\frac{1}{N},$$

这是根据柯西－施瓦兹不等式（Cauchy–Schwarz inequality）$(\sum p_i)^2 \leq N\sum p_i^2$ 得出的，等号仅在每个 $p_i = \frac{1}{N}$ 时成立。所以选号不均匀时，平均而言会有更多人与你分享奖金，而中奖者不多于三人的概率通常会下降。（个别玩家若刻意选择冷门号码，表现会比平均好，但玩家整体而言则较差。）

中奖者人数的期望值

由于 $X \sim \mathrm{B}\left(2N-1, \frac{1}{N}\right)$，其平均值为 $np = (2N-1)\cdot\frac{1}{N}$。（等价地，可把 X 写成 $2N-1$ 个指示变量（indicator variable）之和，每位其他玩家对应一个，各自的平均值为 $\frac{1}{N}$。）因此

$$\mathrm{E}(W) = 1 + \mathrm{E}(X) = 1 + \frac{2N-1}{N} = 3 - \frac{1}{N} = 3 - N^{-1},$$

正如所求。注意这是精确结果，并非近似值。

验算

- 当 $N = 2$ 和 $N = 3$ 时，列举 $2N$ 位玩家所有可能的号码分配及所有可能的中奖号码，所得结果与二项式表达式完全吻合，期望值亦恰为 $3 - \frac{1}{N}$。
- 精确概率在 $N = 10$ 时为 0.7054，$N = 100$ 时为 0.6794，$N = 10\,000$ 时为 0.6767，趋近 $5\mathrm{e}^{-2} = 0.67668$。
- 以 $N = 50$ 进行模拟（10^5 个星期），得出 0.683，而精确值为 0.682；中奖者人数的平均值为 2.981，而 $3 - \frac{1}{50} = 2.98$。
- 当 $N = 10$ 而其中一半号码的受欢迎程度是另一半的两倍时，概率由 0.705 降至 0.641，中奖者人数的期望值则由 2.9 升至 3.11。

常见错误

- 用了 $2N$ 次试验而非 $2N-1$ 次：你自己的彩票已知中奖，所以只有其他玩家是随机的。若用 $2N$ 次试验，期望值会算成 $1+2=3$，而不是 $3 - \frac{1}{N}$。
- 用了无条件平均值：就所有星期而言，中奖者人数的期望值是 $2N\cdot\frac{1}{N} = 2$，但知道你中奖后，便多了一位必然的中奖者，余下 $2N-1$ 位其他玩家。

- 遗漏了 $k = 2$ 的项，或把它写成 2^2e^{-2} 而漏掉 $\frac{1}{2!}$。

STEP 之外

这里的近似就是稀有事件定律（law of rare events）：试验次数多而成功概率小的二项分布会收敛（converge）至泊松分布。最后一部分触及检查悖论（inspection paradox），又称大小偏差抽样（size-biased sampling）。以你的彩票中奖为条件，会令多人选择的号码更有可能出现，这正是条件平均值是 3 而不是 2 的原因。这两个概念在 Grimmett 与 Welsh 的 *Probability: An Introduction*（概率入门教材）以及 Blitzstein 与 Hwang 的 *Introduction to Probability*（概率导论教材）中都有很好的论述。

详细解答

STEP III 2000，第 13 题

概率与统计 · 概率生成函数

> **STEP III 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-s3-q13

关键思路

处理这两个随机变量（random variable），最方便的做法是考虑“尾部”或“累积”事件。第一个六点在第 r 次掷骰之后才出现，当且仅当每一颗骰子在首 r 次掷骰中全部落空；最后一个六点在第 r 次掷骰或之前已经出现，当且仅当每一颗骰子都已经掷出过六点。这两个事件各自都是若干独立事件（independent events）的交集，所以其概率是一个乘积，而把相邻两个值相减便得出分布。一旦认出“第一个六点”这个变量服从几何分布（geometric distribution），与它有关的一切都可以很快求得。

解答

以下一律假设各颗骰子之间互相独立，而每次掷骰之间亦互相独立。对单独一颗骰子，设 X_i 为第 i 颗骰子首次掷出六点时的掷骰次数，则

$$\mathrm{P}(X_i > r) = q^r \qquad (r = 0, 1, 2, \dots),$$

因为 $X_i > r$ 表示连续 r 次都掷不出六点。于是 $R = \min_i X_i$，$S = \max_i X_i$。

第一个六点

事件 $R > r$ 表示 n 颗骰子在首 r 次掷骰中都没有掷出六点，即全部 nr 次个别投掷的结果都不是六点。由独立性，

$$\mathrm{P}(R > r) = q^{nr}.$$

因此，对 $r \geq 1$，

$$\mathrm{P}(R = r) = \mathrm{P}(R > r-1) - \mathrm{P}(R > r) = q^{n(r-1)} - q^{nr} = q^{nr}\left(q^{-n} - 1\right),$$

这正是所求的结果。

记 $Q = q^n$，上式即 $\mathrm{P}(R = r) = Q^{r-1}(1-Q)$：$R$ 服从几何分布，每次把整组骰子掷一次，“成功”（至少出现一个六点）的概率为 $1 - Q$。所以它的概率生成函数（probability generating function，简称 PGF）是一个等比级数（geometric series）：

$$\begin{aligned} G_R(t) = \sum_{r=1}^{\infty} t^r Q^{r-1}(1-Q) &= (1-Q)\, t \sum_{r=1}^{\infty} (Qt)^{r-1} \\ &= \frac{(1-q^n)\, t}{1 - q^n t}, \end{aligned}$$

对 $|t| < q^{-n}$ 成立（特别是在 $t = 1$ 附近）。

要求平均值，用商法则（quotient rule）求导：

$$G_R'(t) = (1-q^n)\,\frac{(1-q^n t) + q^n t}{(1-q^n t)^2} = \frac{1-q^n}{(1-q^n t)^2},$$

所以

$$\mathrm{E}(R) = G_R'(1) = \frac{1}{1-q^n}.$$

当 $n = 2$ 且 $p = \frac{1}{6}$ 时，$q^2 = \frac{25}{36}$，故 $\mathrm{E}(R) = 1 \big/ \frac{11}{36} = \frac{36}{11}$，正如所求。

最后一个六点

事件 $S \le r$ 表示每一颗骰子都在首 r 次掷骰中的某一次掷出过六点。对一颗骰子而言，这个概率是 $1 - q^r$，而各颗骰子互相独立，所以

$$\mathrm{P}(S \le r) = (1-q^r)^n.$$

把相邻两个值相减，对 $r \ge 1$，

$$\mathrm{P}(S = r) = \mathrm{P}(S \le r) - \mathrm{P}(S \le r-1) = \left(1-q^r\right)^n - \left(1-q^{r-1}\right)^n,$$

正如所求。（当 $r = 1$ 时，第二项是 $0^n = 0$，结果正确。）

> **想一想** 要对 $t^r\left[(1-q^r)^2 - (1-q^{r-1})^2\right]$ 求和，看来相当繁复。怎样才可以把它化成等比级数？

当 $n = 2$ 时，把平方展开：

$$\begin{aligned} \mathrm{P}(S = r) &= \left(1 - 2q^r + q^{2r}\right) - \left(1 - 2q^{r-1} + q^{2r-2}\right) \\ &= 2q^{r-1}(1-q) - q^{2(r-1)}(1-q^2) \\ &= 2p\, q^{r-1} - (1-q^2)\,(q^2)^{r-1}. \end{aligned}$$

每一项都是一个几何分布（的倍数），所以

$$G_S(t) = \sum_{r=1}^{\infty} t^r \,\mathrm{P}(S=r) = \frac{2pt}{1-qt} - \frac{(1-q^2)\,t}{1-q^2t}.$$

这个结果合乎情理：由于 $S = X_1 + X_2 - R$，它的分布就是“参数为 p 的几何分布的两倍，减去 R 的分布”。要化简，可把各项通分，以 $(1-qt)(1-q^2t)$ 为公分母，并利用 $1-q^2 = p(1+q)$：

$$\begin{aligned} 2p(1-q^2t) - p(1+q)(1-qt) &= p\big[\,2 - 2q^2t - 1 - q + qt + q^2t\,\big] \\ &= p\big[(1-q) + qt(1-q)\big] = p^2(1+qt). \end{aligned}$$

因此

$$G_S(t) = \frac{p^2\,t\,(1+qt)}{(1-qt)(1-q^2t)}.$$

作为简单的核对，$G_S(1) = \dfrac{p^2(1+q)}{p\cdot p(1+q)} = 1$。

要求平均值，以拆开的形式求导最为容易，因为每一部分都是一个几何分布的 PGF，它在 1 处的导数就是其平均值乘以相应的系数。对 $\dfrac{(1-a)t}{1-at}$ 而言，在 $t=1$ 处的导数是 $\dfrac{1}{1-a}$（与上面 G_R 的情况完全一样），所以

$$\mathrm{E}(S) = G_S'(1) = 2\cdot\frac{1}{p} - \frac{1}{1-q^2} = \frac{2(1+q)-1}{p(1+q)} = \frac{1+2q}{p(1+q)}.$$

当 $p=\frac{1}{6}$，$q=\frac{5}{6}$ 时：

$$\mathrm{E}(S) = 12 - \frac{36}{11} = \frac{96}{11} \approx 8.73.$$

验算

- $\mathrm{E}(R)+\mathrm{E}(S) = \frac{36}{11}+\frac{96}{11} = 12 = \mathrm{E}(X_1)+\mathrm{E}(X_2)$，这是必然的，因为 $\min+\max = X_1+X_2$。
- 当 $n=1$ 时，第一个六点与最后一个六点重合，而 G_R 化为单颗骰子的几何分布 PGF $pt/(1-qt)$。
- 以蒙地卡罗模拟（Monte Carlo simulation）掷 2×10^5 对骰子，所得样本平均值为 3.274 和 8.709，而理论值为 $\frac{36}{11}\approx3.273$ 和 $\frac{96}{11}\approx8.727$；另外，以数值方法对 G_S 的级数求和，在 t 的若干个取值上都与闭合形式相符。

常见错误

- 把 $\mathrm{P}(S=r)$ 写成“某一颗骰子在第 r 次掷骰时首次掷出六点，而其余骰子在此之前已掷出”，却忘记了两颗骰子在同一次掷骰中首次掷出六点的情况。改用 $\mathrm{P}(S\le r)$ 来处

理，便可完全避免这个问题。

- 对化简后的 G_S 求导，而不是对拆开的形式求导：做法正确，但代数运算多得多，出错的机会也多。

STEP 之外

R 和 S 分别是若干独立随机变量的最小值和最大值，是*顺序统计量*（order statistics）最简单的例子；利用 $\mathrm{P}(\min > r)$ 和 $\mathrm{P}(\max \leq r)$ 这个技巧，对任何分布都适用。具有无记忆性（memoryless）的几何分布，是指数分布（exponential distribution）在离散情况下的近亲；在连续极限之下，n 个指数分布变量的最大值的期望值变成调和和（harmonic sum）$1 + \frac{1}{2} + \cdots + \frac{1}{n}$（再乘以平均值），这是优惠券收集问题（coupon collector's problem）中的关键一步。Grimmett 与 Welsh 所著的 *Probability: An Introduction*（概率论入门）对生成函数及上述各种想法有深入的论述。

详细解答

STEP III 2000，第 14 题

概率与统计 · 协方差与独立性

STEP III 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-s3-q14

关键思路

先补上欠缺的第四个概率（probability），然后把 $\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$ 逐项写出。所有项会合并成单一因子 $a - pq$ 乘以一个不可能为零的量，而 $a = pq$ 正好就是说左上角那一格可以分解成边际概率之积。在 2×2 的表中，只要有一格可以这样分解，边际概率就会迫使其余三格也一样。若变量可取三个值，回旋的空间就多了，一个对称的表便足以打破这种联系。

解答

恒等式

四个概率之和为 1，所以

$$\mathrm{P}(X = x_2,\, Y = y_2) = 1 - a - (q - a) - (p - a) = 1 - p - q + a.$$

沿各行各列相加，便得到边际分布（marginal distribution）：

$$\mathrm{P}(X = x_1) = a + (q - a) = q, \qquad \mathrm{P}(Y = y_1) = a + (p - a) = p,$$

所以 $\mathrm{P}(X = x_2) = 1 - q$ 及 $\mathrm{P}(Y = y_2) = 1 - p$。这就是 p 和 q 的真正意义。

	y_1	y_2	
x_1	a	$q - a$	q
x_2	$p - a$	$1 - p - q + a$	$1 - q$
	p	$1 - p$	

现在

$$\begin{aligned}\mathrm{E}(XY) &= a\,x_1y_1 + (q-a)\,x_1y_2 + (p-a)\,x_2y_1 + (1-p-q+a)\,x_2y_2,\\ \mathrm{E}(X)\mathrm{E}(Y) &= \big(qx_1 + (1-q)x_2\big)\big(py_1 + (1-p)y_2\big)\\ &= pq\,x_1y_1 + q(1-p)\,x_1y_2 + p(1-q)\,x_2y_1 + (1-p)(1-q)\,x_2y_2.\end{aligned}$$

逐个系数相减：

$$\begin{aligned}x_1y_1:&\quad a-pq,\\ x_1y_2:&\quad (q-a)-(q-pq) = -(a-pq),\\ x_2y_1:&\quad (p-a)-(p-pq) = -(a-pq),\\ x_2y_2:&\quad (1-p-q+a)-(1-p-q+pq) = a-pq.\end{aligned}$$

因此

$$\mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y) = (a-pq)(x_1y_1-x_1y_2-x_2y_1+x_2y_2) = (a-pq)(x_1-x_2)(y_1-y_2).$$

所以若 $\mathrm{E}(XY)=\mathrm{E}(X)\mathrm{E}(Y)$，则左方为零，即 $(a-pq)(x_1-x_2)(y_1-y_2)=0$，这正是所求。

独立性

任何一对各取两个相异值的随机变量（random variable），其联合概率表都具有上述形式，只要选取适当的 a、p、q 便可（替各格命名即可）。若 $\mathrm{E}(XY)=\mathrm{E}(X)\mathrm{E}(Y)$，则上面的乘积为零。由于 $x_1\neq x_2$ 及 $y_1\neq y_2$，最后两个因子均不为零，所以 $a=pq$。

> ? **想一想** 我们现在知道有一格可以分解。为甚么这样就能确定其余三格?

因为其余每一格都等于一个边际概率减去一个已知的格：

$$\begin{aligned}&\mathrm{P}(X=x_1,Y=y_2) = q-pq = q(1-p) = \mathrm{P}(X=x_1)\,\mathrm{P}(Y=y_2),\\ &\mathrm{P}(X=x_2,Y=y_1) = p-pq = p(1-q) = \mathrm{P}(X=x_2)\,\mathrm{P}(Y=y_1),\\ &\mathrm{P}(X=x_2,Y=y_2) = 1-p-q+pq = (1-p)(1-q) = \mathrm{P}(X=x_2)\,\mathrm{P}(Y=y_2).\end{aligned}$$

每一格都是其边际概率之积，所以 X 与 Y 独立（independent）。（反过来，独立性可推出 $\mathrm{E}(XY)=\mathrm{E}(X)\mathrm{E}(Y)$，这是标准结果，所以对于只取两个值的变量，这两个条件是等价的。）

取三个值的例子

由于 A 和 B 各自在 $\{-1,0,1\}$ 上均匀分布，我们有 $\mathrm{E}(A)=\mathrm{E}(B)=0$，所以我们需要一个表，其每行每列之和均为 $\frac{1}{3}$，且 $\mathrm{E}(AB)=0$，但并非每一格都等于 $\frac{1}{9}$。一个对称的选择

便可做到：

	$B=-1$	$B=0$	$B=1$
$A=-1$	$\frac{1}{6}$	0	$\frac{1}{6}$
$A=0$	0	$\frac{1}{3}$	0
$A=1$	$\frac{1}{6}$	0	$\frac{1}{6}$

每行及每列之和均为 $\frac{1}{3}$，所以 A 和 B 各自取 -1、0、1 的概率均为 $\frac{1}{3}$。只有四个角落的格才给出非零的乘积 ab：

$$\mathrm{E}(AB)=\tfrac{1}{6}(1)+\tfrac{1}{6}(-1)+\tfrac{1}{6}(-1)+\tfrac{1}{6}(1)=0=\mathrm{E}(A)\mathrm{E}(B).$$

但 $\mathrm{P}(A=0,B=0)=\frac{1}{3}\neq\frac{1}{9}=\mathrm{P}(A=0)\,\mathrm{P}(B=0)$，所以 A 与 B 并不独立。用文字来说：$B=0$ 当且仅当 $A=0$，否则无论 A 是甚么，B 取 ±1 的机会均等。

验算

- 试取 $X=Y$，其中 $x_1=y_1=1$，$x_2=y_2=0$：此时 $a=p=q$，而恒等式给出 $\mathrm{Var}(X)=p-p^2$，这对伯努利变量（Bernoulli variable）而言是正确的。
- 这个恒等式是一个多项式恒等式：我们随机选取了 2000 组 x_i、y_i、a、p、q，以有理数运算精确地验证了它。同一次验算亦检查了取三个值的表：边际概率全为 $\frac{1}{3}$，$\mathrm{E}(AB)=0$，以及 $\mathrm{P}(A=0,B=0)=\frac{1}{3}$。

常见错误

- 证出 $a=pq$ 便停笔。独立性要求每一格都可以分解，所以其余三格也需要检查（很快便可完成）。
- 忘记说明为何可以约去 $(x_1-x_2)(y_1-y_2)$：因为所取的值相异，所以它不为零。
- 在最后一部分提出 $B=A^2$ 或 $B=|A|$。这些例子确实是不独立而协方差为零，但 B 永不等于 -1，所以它们没有题目要求的边际分布。

STEP 之外

$\mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y)$ 这个量就是协方差（covariance），而本题说明“不相关”（uncorrelated）比“独立”为弱，只有取两个值的（类似指示变量的）变量例外：对这类变量，恒等式把一切归结为两个指示变量（indicator）的协方差。你日后会遇到一个相关的事实：对于联合正态（jointly normal）的随机变量，协方差为零确实可推出独立。想进一步了解，可参阅 Grimmett 与 Welsh 的 *Probability: An Introduction*（概率论入门），或 Blitzstein 与 Hwang 的 *Introduction to Probability*（概率论导论）。

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

2000 年详细解答. 全书各章及
所有详细解答均可免费阅读、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。

