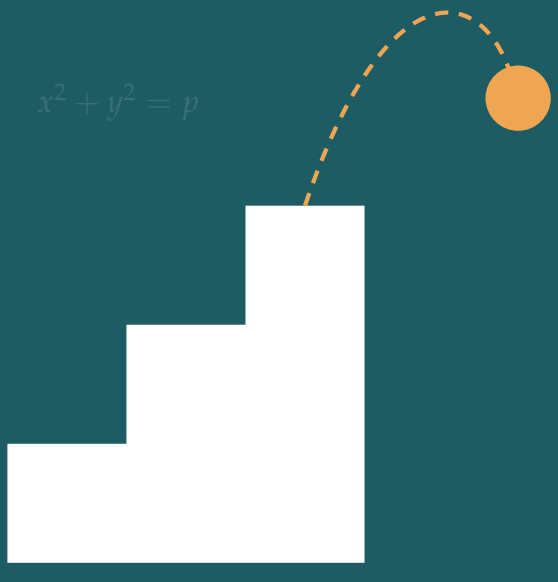
STEP 备试 · 大学数学

Beyond STEP

STEP I, II 及 III

2004 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

2004 年详细解答

Keith McNulty

Beyond STEP：以历届 STEP 试题为门，走进大学数学

2004 年详细解答

2004 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 根式与立方根 纯数学 . . . 3

第 2 题 下取整函数的积分 纯数学 . . . 6

第 3 题 二元多项式的因式分解 纯数学 . . . 9

第 4 题 代换积分法 纯数学 . . . 12

第 5 题 模 5 的剩余 纯数学 . . . 16

第 6 题 三角形的向量几何 纯数学 . . . 20

第 7 题 生成函数 纯数学 . . . 24

第 8 题 单调数列与数学归纳法 纯数学 . . . 27

第 9 题 抛体运动 力学 . . . 31

第 10 题 匀加速运动学 力学 . . . 35

第 11 题 摩擦力与铰接梯子 力学 . . . 39

第 12 题 贝叶斯定理 概率与统计 . . . 43

第 13 题 均匀分布变量的最大值 概率与统计 . . . 46

第 14 题 排列的计数 概率与统计 . . . 50

STEP II

第 1 题 含根式的方程 纯数学 . . . 53

第 2 题 二次不等式 纯数学 . . . 56

第 3 题 曲线描绘 纯数学 . . . 60

第 4 题 几何中的最优化 纯数学 . . . 65

第 5 题 积分方程 纯数学 . . . 69

第 6 题 向量分解 纯数学 . . . 72

第 7 题 二分法 纯数学 . . . 75

第 8 题 微分方程的比较 纯数学 . . . 79

第 9 题 重心 力学 . . . 83
第 10 题 斜面上的摩擦力 力学 . . . 87
第 11 题 功率与阻力 力学 . . . 91
第 12 题 连续分布 概率与统计 . . . 95
第 13 题 最优停止 概率与统计 . . . 99
第 14 题 容斥原理 概率与统计 . . . 103

STEP III

第 1 题 双曲函数积分 纯数学 . . . 107
第 2 题 渐近线与曲线描绘 纯数学 . . . 111
第 3 题 凸性与级数的界限 纯数学 . . . 115
第 4 题 相切圆与级数 纯数学 . . . 119
第 5 题 三角方程 纯数学 . . . 123
第 6 题 递推关系 纯数学 . . . 127
第 7 题 归约公式 纯数学 . . . 131
第 8 题 微分方程 纯数学 . . . 134
第 9 题 力矩与摩擦力 力学 . . . 138
第 10 题 弹性绳与简谐运动 力学 . . . 142
第 11 题 体排斥 力学 . . . 146
第 12 题 指示变量与错排 概率与统计 . . . 150
第 13 题 几何分布 概率与统计 . . . 154
第 14 题 假设检验的错误 概率与统计 . . . 157

详细解答

STEP I 2004，第 1 题

纯数学 · 根式与立方根

STEP I 2004，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/04-s1-q1

关键思路

形如 $a+b\sqrt{n}$（其中 a、b 为整数）的数，本身就像一个小小的数系：两个这样的数相乘，结果仍是同一形式；而两个这样的数相等，当且仅当它们的有理部分（rational part）相同，且 $\sqrt{n}$ 部分亦相同。第 (i) 部分是热身，把这样的一个数立方；第 (ii) 部分则通过比较各部分，把这个过程倒过来做。第 (iii) 部分是一条关于 x^3 的二次方程，而它的根原来正是第 (ii) 部分求得的立方。

解答

第 (i) 部分

由二项式定理（binomial theorem），并利用 $(2\sqrt{5})^2 = 20$ 及 $(2\sqrt{5})^3 = 40\sqrt{5}$，

$$\begin{aligned}\left(3+2\sqrt{5}\right)^3 &= 3^3 + 3\cdot 3^2\cdot 2\sqrt{5} + 3\cdot 3\cdot 20 + 40\sqrt{5}\\ &= 27 + 54\sqrt{5} + 180 + 40\sqrt{5}\\ &= 207 + 94\sqrt{5}.\end{aligned}$$

所以 $a = 207$，$b = 94$。

第 (ii) 部分

把 $\sqrt{5}$ 换成 $\sqrt{2}$，同样展开，对任意 c 和 d 可得

$$\begin{aligned}\left(c-d\sqrt{2}\right)^3 &= c^3 - 3c^2d\sqrt{2} + 3c\cdot 2d^2 - 2d^3\sqrt{2}\\ &= \left(c^3+6cd^2\right) - \left(3c^2d+2d^3\right)\sqrt{2}.\end{aligned}$$

由于实立方根（real cube root）是唯一的，我们需要 $(c-d\sqrt{2})^3 = 99-70\sqrt{2}$。

想一想 为甚么可以把有理部分和 $\sqrt{2}$ 部分分开比较?

若 $p+q\sqrt{2} = r+s\sqrt{2}$，其中 p,q,r,s 为整数且 $q \neq s$，则 $\sqrt{2} = (p-r)/(s-q)$ 便是有理数，但事实并非如此。所以 $q=s$，从而 $p=r$。因此我们需要

$$c^3 + 6cd^2 = 99, \qquad 3c^2d + 2d^3 = 70.$$

由于 c 和 d 均为正数,第一条方程迫使 $c^3 < 99$,故 $c \leq 4$。把每个值逐一代入 $6cd^2 = 99 - c^3$:

- $c=1$：$6d^2 = 98$，没有整数解；
- $c=2$：$12d^2 = 91$，没有；
- $c=3$：$18d^2 = 72$，故 $d=2$；
- $c=4$：$24d^2 = 35$，没有。

唯一的可能是 $c=3$，$d=2$，而它也满足第二条方程：$3\cdot 9\cdot 2 + 2\cdot 8 = 54+16 = 70$。因此

$$\sqrt[3]{99-70\sqrt{2}} = 3-2\sqrt{2}, \qquad c=3,\ d=2.$$

（顺带检查一下正负号：$2\sqrt{2} \approx 2.83 < 3$，所以 $3-2\sqrt{2}$ 是正数，这与 $99-70\sqrt{2} \approx 0.005$ 是正数相符。）

第 (iii) 部分

设 $y = x^3$。方程变成 $y^2 - 198y + 1 = 0$，所以

$$y = \frac{198 \pm \sqrt{198^2-4}}{2} = 99 \pm \sqrt{99^2-1} = 99 \pm \sqrt{9800} = 99 \pm 70\sqrt{2},$$

当中用到 $9800 = 4900\cdot 2$。两个值都是正实数。

由第 (ii) 部分，$99-70\sqrt{2} = (3-2\sqrt{2})^3$。至于另一个根，留意该二次方程两根之积为 1，所以

$$99+70\sqrt{2} = \frac{1}{99-70\sqrt{2}} = \frac{1}{(3-2\sqrt{2})^3} = \left(3+2\sqrt{2}\right)^3,$$

因为 $(3-2\sqrt{2})(3+2\sqrt{2}) = 9-8 = 1$。（同样地，在第 (ii) 部分的展开式中把 d 的正负号改变，亦可直接得出 $(3+2\sqrt{2})^3 = 99+70\sqrt{2}$。）

每个实数恰好有一个实立方根，所以每个 y 值恰好给出一个实数 x。两个实数解是

$$x = 3-2\sqrt{2} \qquad \text{和} \qquad x = 3+2\sqrt{2}.$$

这条六次方程（sextic）其余四个解都是复数：即每个 y 余下的立方根。

验算

- 以数值计算，$(3+2\sqrt{5})^3 \approx 417.19$，而 $207+94\sqrt{5} \approx 417.19$。
- $3+2\sqrt{2} \approx 5.8284$，而 x^6-198x^3+1 在此值处等于零（在舍入误差范围内）；在 $3-2\sqrt{2} \approx 0.1716$ 处亦然。
- 第 (iii) 部分两个答案之积为 1，理应如此：若 x 是根，则 $1/x$ 也是根，因为把方程除以 x^6，便得到关于 $1/x$ 的同一条方程。

常见错误

- 展开时忘记 $(d\sqrt{2})^3 = 2\sqrt{2}\,d^3$，而不是 $\sqrt{2}\,d^3$。
- 由第一条方程求得 $c=3$，$d=2$，却没有验证第二条方程。
- 在第 (iii) 部分给出六个“解”：每个正数只有一个立方根是实数。

STEP 之外

所有形如 $a+b\sqrt{2}$（a,b 为整数）的数组成的集合是一个环（ring），记作 $\mathbb{Z}[\sqrt{2}]$，而 $3+2\sqrt{2}$ 是其中的一个单位（unit）：它在同一个环内有逆元（inverse）$3-2\sqrt{2}$。这类单位与佩尔方程（Pell's equation）$a^2-2b^2=\pm 1$ 息息相关（留意 $99^2-2\cdot 70^2=1$），而 $\mathbb{Z}[\sqrt{2}]$ 的每个单位都是 $\pm(1+\sqrt{2})^n$。想进一步了解，可参阅 Davenport 的 *The Higher Arithmetic*，或 Allenby 的 *Rings, Fields and Groups*。

详细解答

STEP I 2004，第 2 题

纯数学 · 下取整函数的积分

> **STEP I 2004，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/04-s1-q2

关键思路

函数 $[x]$ 在每个区间 $r \le x < r+1$ 上都是常数，所以凡是由它构成的函数都是阶梯函数（step function）。对阶梯函数求积分，不过是把一个个长方形的面积加起来：在各整数处把 $[0, a]$ 分段，在每一段单位长度的区间上，被积函数（integrand）都是常数。当 a 不是整数时，唯一的新情况是最后多了一个较窄的长方形。

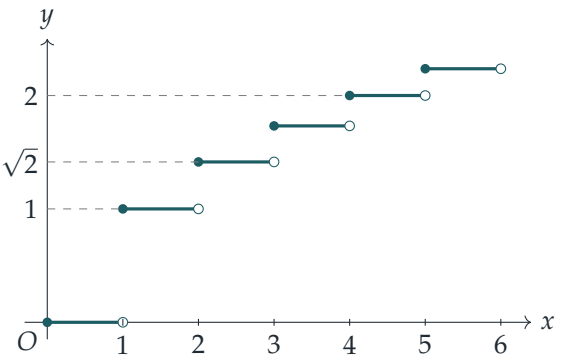
解答

第 (i) 部分

对于 $x \ge 0$，设 $r = [x]$。那么 r 是非负整数，且 $r \le x < r+1$，所以在该区间上

$$y = \sqrt{[x]} = \sqrt{r}.$$

因此图像呈阶梯状：在 $[0,1)$ 上 $y = 0$，在 $[1,2)$ 上 $y = 1$，在 $[2,3)$ 上 $y = \sqrt{2}$，如此类推。每一级都包含左端点（实心点），但不包含右端点（空心点），因为 $[x]$ 恰好在每个整数处向上跳升。由于 $\sqrt{r+1} - \sqrt{r} = 1/(\sqrt{r+1} + \sqrt{r})$ 递减，各级升高的幅度愈来愈小。对于 $x < 0$，我们有 $[x] \le -1$，所以 $\sqrt{[x]}$ 在该处没有定义，图像只在 $x \ge 0$ 的范围内出现。



现设 a 为正整数。在整数 $1, 2, \ldots, a-1$ 处把积分范围分段：

$$\int_0^a \sqrt{[x]}\,\mathrm{d}x = \sum_{r=0}^{a-1} \int_r^{r+1} \sqrt{[x]}\,\mathrm{d}x.$$

在 $r \le x < r+1$ 上，被积函数等于常数 $\sqrt{r}$。（在 $x = r+1$ 这一点上它取另一个值，但改变函数在一点上的值并不会改变其积分。）所以每一段都是一个阔度为 1、高度为 $\sqrt{r}$ 的长方形的面积：

$$\int_r^{r+1} \sqrt{[x]}\,\mathrm{d}x = \sqrt{r}, \qquad \text{因此} \qquad \int_0^a \sqrt{[x]}\,\mathrm{d}x = \sum_{r=0}^{a-1} \sqrt{r},$$

这正是所要证明的结果。

第 (ii) 部分

完全相同的论证也适用于 $2^{[x]}$，它在 $r \le x < r+1$ 上等于 2^r。对于正整数 a，

$$\int_0^a 2^{[x]}\,\mathrm{d}x = \sum_{r=0}^{a-1} \int_r^{r+1} 2^r\,\mathrm{d}x = \sum_{r=0}^{a-1} 2^r = 1 + 2 + 4 + \cdots + 2^{a-1}.$$

这是一个等比级数（geometric series），首项为 1，公比为 2，共有 a 项，所以其和为

$$\frac{2^a - 1}{2 - 1} = 2^a - 1,$$

这就是所求的结果。

第 (iii) 部分

现设 $a > 0$ 不是整数，并令 $n = [a]$，那么 n 是整数，且 $n < a < n+1$ 及 $n \ge 0$。

? **想一想** 积分到了哪里，就不再是若干个完整单位长方形之和?

直到 $x = n$ 为止，一切都与第 (ii) 部分相同；余下的一段是 $n \le x \le a$，这是一条阔度为 $a - n$ 的窄条，完全位于同一级之内，其上 $[x] = n$。（若 $n = 0$，第一段是空的，其贡献为 $2^0 - 1 = 0$，所以同一条公式仍然适用。）因此

$$\begin{aligned}\int_0^a 2^{[x]}\,\mathrm{d}x &= \int_0^n 2^{[x]}\,\mathrm{d}x + \int_n^a 2^n\,\mathrm{d}x \\ &= \left(2^n - 1\right) + (a - n)\,2^n \\ &= 2^{[a]}\left(1 + a - [a]\right) - 1.\end{aligned}$$

验算

- 若 $0 < a < 1$，则 $[a] = 0$，公式给出 $1 \cdot (1 + a) - 1 = a$：正确，因为被积函数在 $[0, a]$ 上等于 1。
- 当 a 从下方趋近整数 m 时，$[a] = m - 1$，公式趋向 $2^{m-1} \cdot 2 - 1 = 2^m - 1$，与第 (ii) 部分吻合。所以这个积分是 a 的连续函数（continuous function）；任何有界函数（bounded function）的积分，作为其积分上限的函数，都必定如此。
- 以数值验算，取 $a = 2.5$，公式给出 $4 \times 1.5 - 1 = 5$，而用中点法则（midpoint rule）计算该积分，同样得出 5.000。至于第 (i) 部分，取 $a = 5$，两边都是 $0 + 1 + \sqrt{2} + \sqrt{3} + 2 \approx 6.1463$。

常见错误

- 由 $r = 1$ 加到 a，而不是由 $r = 0$ 加到 $a - 1$：在 $[r, r + 1)$ 上，高度是 $\sqrt{r}$，即左端点处的值。
- 在第 (iii) 部分，遗漏了那条不完整的窄条，或把它的阔度当作 1 而不是 $a - [a]$。
- 用垂直线把各级连起来，或在两端都画上实心点：下取整函数（floor function）在整数处并不连续。

STEP 之外

把长方形面积相加来求阶梯函数的积分，正是大学里建立黎曼积分（Riemann integral）的方法：先为阶梯函数定义积分，然后用阶梯函数从上方和下方把一般的函数夹住。第 (i) 部分也是比较求和与积分的第一步，由此可得出诸如 $\sum_{r=1}^{n} \sqrt{r} \approx \frac{2}{3} n^{3/2}$ 的估计。Abbott 的 *Understanding Analysis*（一本实分析入门教科书）就是以这种方式建立黎曼积分的。

详细解答

STEP I 2004，第 3 题

纯数学 · 二元多项式的因式分解

STEP I 2004，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/04-s1-q3

关键思路

二元多项式（polynomial in two variables）可以看成只含其中一个变量的多项式，而系数则含有另一个变量。把它看成 x 的多项式时，因式定理（factor theorem）依然适用，长除法（long division）依然适用，而 x 的二次式只要判别式（discriminant）是完全平方（perfect square），就依然可以因式分解（factorise）。在第 (ii) 部分，x 的首项系数（leading coefficient）本身是 y 的一次式，这暗示其中一个因式只含 y。

解答

第 (i) 部分

把式子 $(*)$ 按 x 的幂次整理：

$$P(x) = x^3 + (2y-5)x^2 + (y^2 - 8y + 6)x + (6y - 3y^2).$$

代入 $x = 3$，

$$\begin{aligned} P(3) &= 27 + 9(2y-5) + 3(y^2 - 8y + 6) + 6y - 3y^2 \\ &= (27 - 45 + 18) + (18y - 24y + 6y) + (3y^2 - 3y^2) = 0 \end{aligned}$$

对任何 y 都成立，所以根据因式定理，$x - 3$ 是一个因式，正如题目所求。

要找出另一个因式，设 $P(x) = (x-3)(x^2 + px + q)$，其中 p 和 q 可以与 y 有关。展开右边得 $x^3 + (p-3)x^2 + (q-3p)x - 3q$，再比较系数（comparing coefficients）：

$$\begin{aligned} x^2: &\quad p - 3 = 2y - 5, && \text{所以} p = 2y - 2, \\ x^0: &\quad -3q = 6y - 3y^2, && \text{所以} q = y^2 - 2y, \end{aligned}$$

而 x 的系数亦吻合：$q-3p=y^2-8y+6$，与应有的结果一致。所以二次因式是

$$x^2+(2y-2)x+y^2-2y=(x^2+2xy+y^2)-2(x+y)=(x+y)(x+y-2).$$

想一想 我们是怎样看出这个二次式可以因式分解的?

$x^2+2xy+y^2$ 这几项合起来是完全平方 $(x+y)^2$，而余下的项 $-2x-2y$ 是 $x+y$ 的倍数。（如果一时看不出这个分组方法，可以计算 x 的判别式：$(2y-2)^2-4(y^2-2y)=4$，是完全平方，所以根是 $x=-y+1\pm1$，即 $x=-y$ 和 $x=2-y$。）因此

$$(*)=(x-3)(x+y)(x+y-2),$$

所以 $a=1$, $b=0$, $c=1$, $d=-2$（或者把两个因式对调：$a=1$, $b=-2$, $c=1$, $d=0$）。

第 (ii) 部分

把这个式子称为 Q，同样按 x 的幂次整理：

$$Q=(y+2)x^2+(12-4y-5y^2)x+(6y^3-y^2-21y+10).$$

x^2 的系数是 $y+2$。如果 Q 是三个一次因式的积，x^2 项就来自其中两个因式含 x 的部分，所以很自然会问：$y+2$ 能否整除整个式子？设 $y=-2$：

$$12+8-20=0,\qquad -48-4+42+10=0,$$

所以余下两个系数在 $y=-2$ 时都等于零，根据（关于 y 的）因式定理，两者都可被 $y+2$ 整除：

$$12-4y-5y^2=-(5y^2+4y-12)=-(y+2)(5y-6),$$
$$6y^3-y^2-21y+10=(y+2)(6y^2-13y+5).$$

另外 $6y^2-13y+5=(2y-1)(3y-5)$。因此

$$Q=(y+2)\big[x^2-(5y-6)x+(2y-1)(3y-5)\big].$$

想一想 方括号内的式子可以因式分解吗？我们怎样判断?

它关于 x 的判别式是

$$\begin{aligned}(5y-6)^2-4(2y-1)(3y-5)&=25y^2-60y+36-(24y^2-52y+20)\\&=y^2-8y+16=(y-4)^2,\end{aligned}$$

是完全平方，所以可以分解。更快捷的做法是留意 $(2y-1)+(3y-5)=5y-6$，所以积为 $(2y-1)(3y-5)$、和为 $5y-6$ 的两个数就是 $2y-1$ 和 $3y-5$，由此得根 $x=2y-1$ 和 $x=3y-5$。因此

$$Q=(y+2)(x-2y+1)(x-3y+5).$$

验算

- 在第 (i) 部分代入 $x = y = 1$：$(*)$ 给出 $1-5+2+1-8-3+6+6 = 0$，而 $(1-3)(2)(0) = 0$。取 $x = 2$，$y = 1$：$(*)$ 给出 $8-20+8+2-16-3+12+6 = -3$，而 $(-1)(3)(1) = -3$。
- 在第 (ii) 部分代入 $x = y = 0$：$Q = 10$，而 $(2)(1)(5) = 10$。取 $x = 1$，$y = 1$：$Q = 6-1-21+2+12-4+1-5+10 = 0$，而的确 $x - 2y + 1 = 0$。
- 我们在 2000 个随机点 (x, y) 上以精确的有理数运算核对了两个因式分解，每次结果都一致。

常见错误

- 把式子看成 x 的多项式，却忘记“常数”项是指所有不含 x 的项，在第 (i) 部分即 $6y - 3y^2$。
- 在第 (ii) 部分，尝试找一个根 $x = k$，其中 k 是一个数：形如 $x - k$ 的因式并不存在。应该先找出的，是不含 x 的因式 $y + 2$。
- 把 $-(5y^2 + 4y - 12)$ 除以 $y + 2$ 时出现正负号错误；应展开验算。

STEP 之外

多元多项式构成一个环（ring），而 $\mathbb{Z}[x, y]$ 可以看成 $(\mathbb{Z}[y])[x]$：即系数是 y 的多项式的 x 的多项式。这正是我们刚才的做法，亦是因式定理和除法依然适用的原因。这些环具有唯一分解性质（unique factorisation）（高斯引理，Gauss's lemma），所以我们找到的因式分解，除了次序和正负号之外是唯一的。Allenby 的 *Rings, Fields and Groups*（《环、域与群》）详细阐述了这些概念。

详细解答

STEP I 2004，第 4 题

纯数学 · 代换积分法

STEP I 2004，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/04-s1-q4

关键思路

每当见到 $\sqrt{x^2-1}$，就应该想起恒等式（identity）$\sec^2 t - 1 = \tan^2 t$：设 $x = \sec t$，平方根便化为 $\tan t$，而由于 $\mathrm{d}x = \sec t \tan t \, \mathrm{d}t$，这个 $\tan t$ 亦会一并约去。第 (ii) 和第 (iii) 部分看似不同，但只要对根号内的式子配方（completing the square），两者都会化成同一个形状 $\sqrt{u^2 - a^2}$，其中 $u = x + 2$，而根号外的因子 $x + 2$ 正是这个代换（substitution）所需要的。

解答

导数

把 $\sec t = (\cos t)^{-1}$ 写出，再用链式法则（chain rule），得

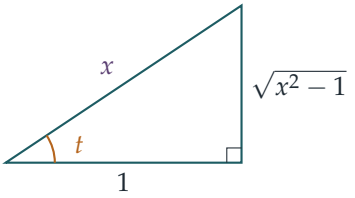
$$\frac{\mathrm{d}}{\mathrm{d}t} \sec t = -(\cos t)^{-2} \cdot (-\sin t) = \frac{\sin t}{\cos^2 t} = \sec t \tan t.$$

第 (i) 部分

设 $x = \sec t$，于是 $\mathrm{d}x = \sec t \tan t \, \mathrm{d}t$。当 $x = \sqrt{2}$ 时，$\cos t = 1/\sqrt{2}$，$t = \frac{\pi}{4}$；当 $x = 2$ 时，$\cos t = \frac{1}{2}$，$t = \frac{\pi}{3}$。对于 $[\frac{\pi}{4}, \frac{\pi}{3}]$ 内的 t，我们有 $\tan t > 0$，所以

$$\sqrt{x^2 - 1} = \sqrt{\sec^2 t - 1} = \sqrt{\tan^2 t} = \tan t$$

（无须加上绝对值符号）。下面的三角形显示的是同一回事：斜边为 x，邻边为 1，所以对边是 $\sqrt{x^2 - 1}$。



积分变成

$$\begin{aligned}\int_{\sqrt{2}}^{2}\frac{1}{x^3\sqrt{x^2-1}}\,\mathrm{d}x &= \int_{\pi/4}^{\pi/3}\frac{\sec t\tan t}{\sec^3 t\,\tan t}\,\mathrm{d}t = \int_{\pi/4}^{\pi/3}\cos^2 t\,\mathrm{d}t\\ &= \int_{\pi/4}^{\pi/3}\tfrac{1}{2}(1+\cos 2t)\,\mathrm{d}t = \left[\tfrac{1}{2}t+\tfrac{1}{4}\sin 2t\right]_{\pi/4}^{\pi/3}\\ &= \left(\frac{\pi}{6}+\frac{\sqrt{3}}{8}\right)-\left(\frac{\pi}{8}+\frac{1}{4}\right) = \frac{\sqrt{3}-2}{8}+\frac{\pi}{24},\end{aligned}$$

其中用到 $\cos^2 t = \frac{1}{2}(1+\cos 2t)$、$\sin\frac{2\pi}{3} = \frac{\sqrt{3}}{2}$ 和 $\sin\frac{\pi}{2} = 1$。这就是所求的结果。

第 (ii) 部分

想一想 第 (ii) 部分里并没有 $\sqrt{x^2-1}$。第 (i) 部分可以怎样帮得上忙?

把根号内的式子展开，然后配方:

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

所以设 $u = x+2$（而 $\mathrm{d}u = \mathrm{d}x$），积分便是

$$\int\frac{1}{u\sqrt{u^2-1}}\,\mathrm{d}u.$$

被积函数（integrand）只在 $u^2 > 1$ 时才有定义。先考虑 $u > 1$（即 $x > -1$），并设 $u = \sec t$，其中 $0 < t < \frac{\pi}{2}$，使得 $\tan t > 0$ 及 $\sqrt{u^2-1} = \tan t$。于是

$$\int\frac{\sec t\tan t}{\sec t\,\tan t}\,\mathrm{d}t = \int 1\,\mathrm{d}t = t+C,$$

而由于 $\sec t = u$，我们有 $t = \operatorname{arcsec} u$（等价地，即 $\arccos(1/u)$）。因此，对于 $x > -1$，

$$\int\frac{1}{(x+2)\sqrt{(x+1)(x+3)}}\,\mathrm{d}x = \operatorname{arcsec}(x+2)+C.$$

为求完整，被积函数在 $x < -3$ 时亦有定义，此时 $u < -1$。在这个情况下，设 $u = -v$，其中 $v > 1$：被积式 $\frac{1}{u\sqrt{u^2-1}}\,\mathrm{d}u$ 变成 $\frac{1}{(-v)\sqrt{v^2-1}}(-\mathrm{d}v) = \frac{1}{v\sqrt{v^2-1}}\,\mathrm{d}v$，其积分为 $\operatorname{arcsec} v$。两种情况都可以用

$$\operatorname{arcsec}|x+2|+C = \arccos\frac{1}{|x+2|}+C,$$

来概括，而两个区间上的常数可以不同。

第 (iii) 部分

再次配方：

$$x^2+4x-5=(x+2)^2-9,$$

所以设 $u=x+2$ 后，我们需要求 $\displaystyle\int \frac{1}{u\sqrt{u^2-9}}\,\mathrm{d}u$。这一次自然的代换是 $u=3\sec t$，这样选是为了令 $u^2-9=9\tan^2 t$。对于 $u>3$（即 $x>1$），取 $0<t<\frac{\pi}{2}$，于是 $\sqrt{u^2-9}=3\tan t$，而 $\mathrm{d}u=3\sec t\tan t\,\mathrm{d}t$。于是

$$\int \frac{3\sec t\tan t}{3\sec t\cdot 3\tan t}\,\mathrm{d}t=\int \frac{1}{3}\,\mathrm{d}t=\frac{t}{3}+C,$$

而 $\sec t=u/3$。因此，对于 $x>1$，

$$\int \frac{1}{(x+2)\sqrt{x^2+4x-5}}\,\mathrm{d}x=\frac{1}{3}\operatorname{arcsec}\frac{x+2}{3}+C.$$

与第 (ii) 部分完全一样，$x<-5$ 的情况（此时 $u<-3$）可用 $u=-v$ 处理，而两个区间都可以用 $\frac{1}{3}\operatorname{arcsec}\left(\frac{1}{3}|x+2|\right)+C$ 来概括。

验算

- 对第 (i) 部分的积分使用辛普森法则（Simpson's rule），得 $0.0974060\ldots$，与 $\frac{1}{8}(\sqrt{3}-2)+\frac{\pi}{24}$ 吻合。
- 对答案求导：当 $u>1$ 时，$\dfrac{\mathrm{d}}{\mathrm{d}u}\operatorname{arcsec}u=\dfrac{1}{u\sqrt{u^2-1}}$（设 $u=\sec t$，并使用开首求得的导数）。例如，第 (ii) 部分的被积函数的 $\int_0^3$ 数值上是 0.322241，而 $\operatorname{arcsec}5-\operatorname{arcsec}2=\arccos\frac{1}{5}-\frac{\pi}{3}$ 与之相符。
- 第 (iii) 部分应该是第 (ii) 部分经过缩放后的版本：设 $u=3w$，则 $\int\frac{\mathrm{d}u}{u\sqrt{u^2-9}}=\frac{1}{3}\int\frac{\mathrm{d}w}{w\sqrt{w^2-1}}$，这就解释了因子 $\frac{1}{3}$ 的由来。

常见错误

- 在第 (i) 部分忘记更改积分上下限，或把上下限弄反：$\sec t=\sqrt{2}$ 给出 $t=\frac{\pi}{4}$，而不是 $\frac{\pi}{3}$。
- 写下 $\sqrt{\tan^2 t}=\tan t$，却没有检查 $\tan t$ 在所用范围内的正负号。
- 在第 (iii) 部分代入 $u=\sec t$ 而不是 $u=3\sec t$，结果剩下 $\sqrt{\sec^2 t-9}$，没有恒等式可以把它化简。

STEP 之外

反正割函数（inverse secant）只是一个家族的其中一员：代换 $x = a\sin t$, $a\tan t$, $a\sec t$（或它们的双曲函数（hyperbolic function）近亲 $a\sinh t$, $a\cosh t$）可以驯服 $\sqrt{a^2 - x^2}$, $\sqrt{a^2 + x^2}$ 和 $\sqrt{x^2 - a^2}$，而配方可以把任何 $\sqrt{\text{二次式}}$ 化为其中一种。在正负号上，以及在原函数（antiderivative）定义于哪个区间上所需要的谨慎，正是初尝分析（analysis）课程如何令积分变得严谨的滋味。Bartle 与 Sherbert 的 *Introduction to Real Analysis*（实分析入门）严格地处理了代换定理（substitution theorem），而 Siklos 的 *Advanced Problems in Mathematics*（高等数学难题集）则有更多同类风格的 STEP 积分题。

详细解答

STEP I 2004，第 5 题

纯数学 • 模 5 的剩余

> **STEP I 2004，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/04-s1-q5

关键思路

这五个等差数列（arithmetic progression）把正整数按除以 5 的余数（remainder）分成五类。只要为每一类写出像 $5n-3$ 这样的通项（general term），就可以把和与平方展开，再抽出因子 5，从而追踪它们落在哪一类。这样便能完全掌握平方数可以落在哪些类别，而两个“不可能”的证明都由此而来：等式右边落在一个左边永远到不了的类别。

解答

通项

每个数列的公差（common difference）都是 5，所以对 $n=1,2,3,\ldots$，第 n 项分别是

$$A:\ 5n-4, \qquad B:\ 5n-3, \qquad C:\ 5n-2, \qquad D:\ 5n-1, \qquad E:\ 5n.$$

（代入 $n=1$，可验证每条式子都给出题目所列的首项。）因此，一个正整数属于 A、B、C、D 或 E，当且仅当它除以 5 的余数分别是 1、2、3、4 或 0。

B 的一项与 C 的一项之和

任取 B 的一项 $5m-3$ 和 C 的一项 $5n-2$，其中 $m,n\geq 1$。那么

$$(5m-3)+(5n-2)=5m+5n-5=5(m+n-1).$$

由于 $m+n-1\geq 1$ 是正整数，这就是 E 的第 $(m+n-1)$ 项，正如所求。

B 和 C 各项的平方

对 B 的一项 $5m-3$,

$$\begin{aligned}(5m-3)^2 &= 25m^2-30m+9\\ &= 5(5m^2-6m+2)-1.\end{aligned}$$

这里 $k=5m^2-6m+2$ 是整数,而且对所有 $m\geq 1$ 都有 $k=5m^2-6m+2=m(5m-6)+2\geq 1$(当 $m=1$ 时它等于 1;当 $m\geq 2$ 时 m 和 $5m-6$ 都是正数)。所以 $(5m-3)^2=5k-1$ 是 D 的第 k 项。B 中每一项的平方都属于 D。

关于 C 的类似命题是:*C 中每一项的平方也是 D 的一项*。事实上,对 C 的一项 $5m-2$,

$$\begin{aligned}(5m-2)^2 &= 25m^2-20m+4\\ &= 5(5m^2-4m+1)-1,\end{aligned}$$

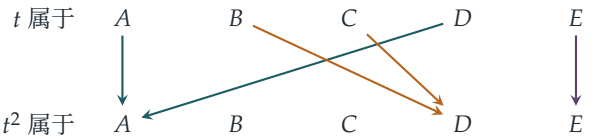
而且当 $m\geq 1$ 时 $k=5m^2-4m+1=m(5m-4)+1\geq 2$。所以 $(5m-2)^2$ 是 D 的第 k 项。

想一想 A、D 和 E 各项的平方又会落在哪里?

第 (i) 和第 (ii) 部分都要用到这个结果,所以值得用同样的方法处理余下三类:

$$\begin{aligned}(5m-4)^2 &= 25m^2-40m+16 = 5(5m^2-8m+4)-4 && \text{(属于 } A\text{)},\\ (5m-1)^2 &= 25m^2-10m+1 = 5(5m^2-2m+1)-4 && \text{(属于 } A\text{)},\\ (5m)^2 &= 5(5m^2) && \text{(属于 } E\text{)}.\end{aligned}$$

当 $m\geq 1$ 时,括号内的都是正整数(例如 $5m^2-8m+4=m(5m-8)+4\geq 1$,以及 $5m^2-2m+1\geq 4$)。所以平方运算把各类对应如下:



特别要留意,*没有完全平方数*(*perfect square*)*属于 B 或 C*:每个正的平方数都属于 A、D 或 E。

第 (i) 部分

假设正整数 x 和 y 满足 $x^2+5y=243\,723$。由于 $243\,723=5\times 48\,744+3$,等式右边是 C 的一项 $5\times 48\,745-2$。于是

$$x^2=243\,723-5y=5(48\,745-y)-2.$$

这个数是正数（它等于 x^2），所以 $48\,745 - y$ 是正整数，而 x^2 是 C 的一项。但我们已证明没有平方数属于 C。这个矛盾（contradiction）说明这样的正整数 x 和 y 并不存在。

第 (ii) 部分

四次方（fourth power）就是平方的平方。平方数 x^2 属于 A、D 或 E；把 A 或 D 的一项平方，得到 A 的一项，而把 E 的一项平方，则得到 E 的一项。所以对任何正整数 x，

$$x^4 \text{ 属于} A \text{ 或} E, \quad \text{也就是说，} \quad x^4 \text{ 的余数是1 或0，这里的除数是5.}$$

y^4 也是如此，所以 $2y^4$ 的余数是 2 或 0。两者相加，$x^4 + 2y^4$ 除以 5 的余数等于以下其中一个数的余数：

$$1 + 2 = 3, \qquad 1 + 0 = 1, \qquad 0 + 2 = 2, \qquad 0 + 0 = 0,$$

所以 $x^4 + 2y^4$ 属于 C、A、B 或 E，而永远不会属于 D。

> ? **想一想** 为甚么可以直接把余数相加?

如果 $x^4 = 5a + r$ 而 $2y^4 = 5b + s$，那么 $x^4 + 2y^4 = 5(a + b) + (r + s)$，所以这个和与 $r + s$ 相差 5 的倍数，与 $r + s$ 属于同一类。这正是上面计算 $B + C$ 时所做的运算，只是写成了一般形式。

现在 $26\,081\,974 = 5 \times 5\,216\,395 - 1$ 的末位数字是 4，所以它是 D 的一项。因此，对任何正整数 x 和 y，它都不可能等于 $x^4 + 2y^4$，正如所求。

验算

- 小数值的例子：$2^2 = 4$、$7^2 = 49$ 和 $12^2 = 144$ 的末位数字都是 4 或 9，所以属于 D；$3^2 = 9$、$8^2 = 64$、$13^2 = 169$ 亦然。
- 四次方：$1, 16, 81, 256, 625, 1296$ 的末位数字都是 1、6 或 5，所以属于 A 或 E。
- 用电脑搜索所有 $x \le \sqrt{243\,723} \approx 493.7$，找不到任何 x 使 $243\,723 - x^2$ 是 5 的正倍数；再搜索所有满足 $x^4 < 26\,081\,974$ 的 x（即 $x \le 71$），也找不到任何 y 使 $2y^4 = 26\,081\,974 - x^4$。

常见错误

- 把 E 的通项写成 $5n + 5$、把 A 的通项写成 $5n + 1$ 并无不可，但这时 n 要由 0 开始；把两种写法混在一起用，项数就会差了一项。

- 在第 (i) 部分，忘记利用 $x^2 > 0$ 来说明 $48\,745 - y$ 是正整数；少了这一步，x^2 就不一定是 C 的一项，因为这些数列只包含正整数。
- 在第 (ii) 部分，只处理 x^4，却忘记因子 2 会把 y^4 的余数由 1 变成 2。

STEP 之外

这五个数列就是模 5 的剩余类（residue classes），而整道题目其实是 $\mathbb{Z}/5\mathbb{Z}$ 中的算术：那里的平方数是 $0, 1, 4$（即二次剩余（quadratic residues）），四次方则是 $0, 1$，这正是费马小定理（Fermat's little theorem）的一个例子：当 $5 \nmid a$ 时 $a^4 \equiv 1 \pmod 5$。要证明一条丢番图方程（Diophantine equation）无解，首选的工具就是选一个合适的数，把方程对它取模。Davenport 的 *The Higher Arithmetic* 对同余（congruences）和二次剩余有十分精彩的阐述。

详细解答

STEP I 2004，第 6 题

纯数学 · 三角形的向量几何

> **STEP I 2004，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/04-s1-q6

关键思路

以位置向量（position vector）代替分开处理的 x 坐标和 y 坐标：设 $\mathbf{a} = (p_1, q_1)$、$\mathbf{b} = (p_2, q_2)$、$\mathbf{c} = (p_3, q_3)$，则 H 的位置向量为 $\mathbf{a} + \mathbf{b} + \mathbf{c}$。这样，题目中每一条直线都有简短的向量描述，而“互相垂直”就变成“纯量积（scalar product）为零”。出现的纯量积是一个平方差（difference of two squares）：$(\mathbf{c} + \mathbf{b}) \cdot (\mathbf{c} - \mathbf{b})$，而题目的后半部分就全在于此。

解答

我们假设 A、B、C 不共线（collinear），因而构成一个真正的三角形（否则第一部分的直线可能重合，便没有唯一的交点）。

中线相交

设 L、M、N 分别为 BC、CA、AB 的中点（midpoint），其位置向量为

$$\mathbf{l} = \tfrac{1}{2}(\mathbf{b} + \mathbf{c}), \qquad \mathbf{m} = \tfrac{1}{2}(\mathbf{a} + \mathbf{c}), \qquad \mathbf{n} = \tfrac{1}{2}(\mathbf{a} + \mathbf{b}).$$

直线 AL 上的一般点为

$$\mathbf{a} + s(\mathbf{l} - \mathbf{a}) = (1 - s)\,\mathbf{a} + \tfrac{s}{2}\,\mathbf{b} + \tfrac{s}{2}\,\mathbf{c},$$

而直线 BM 上的一般点为

$$\mathbf{b} + t(\mathbf{m} - \mathbf{b}) = \tfrac{t}{2}\,\mathbf{a} + (1 - t)\,\mathbf{b} + \tfrac{t}{2}\,\mathbf{c}.$$

只要 $\mathbf{a}$、$\mathbf{b}$ 和 $\mathbf{c}$ 的系数对应相等，这两个表达式便一致：$1 - s = \tfrac{t}{2}$、$\tfrac{s}{2} = 1 - t$ 及 $\tfrac{s}{2} = \tfrac{t}{2}$。最后一式给出 $s = t$，于是由 $1 - s = \tfrac{s}{2}$ 得 $s = t = \tfrac{2}{3}$，这组值同时满足三条等式。因此两

条直线都通过点 G，其位置向量为

$$\mathbf{g} = \tfrac{1}{3}(\mathbf{a}+\mathbf{b}+\mathbf{c}), \qquad \text{即} \qquad G = \left(\tfrac{p_1+p_2+p_3}{3}, \tfrac{q_1+q_2+q_3}{3}\right).$$

? **想一想** 比较系数找到了一个公共点。我们怎样知道它就是唯一的交点?

两条相异直线最多只有一个交点，所以我们只需确定 AL 和 BM 是不同的直线。假如它们是同一条直线，这条直线便包含 A、B 和 L，因而包含整条直线 BL，亦即直线 BC；那么 A 就会在 BC 上，与我们的假设矛盾。所以 G 是唯一的交点。

至于第三条中线（median）CN，取点 $\mathbf{c}+u(\mathbf{n}-\mathbf{c})$，其中 $u=\frac{2}{3}$：

$$\mathbf{c}+\tfrac{2}{3}\left(\tfrac{1}{2}\mathbf{a}+\tfrac{1}{2}\mathbf{b}-\mathbf{c}\right) = \tfrac{1}{3}\mathbf{a}+\tfrac{1}{3}\mathbf{b}+\tfrac{1}{3}\mathbf{c} = \mathbf{g}.$$

所以 G 位于连接 C 与 AB 中点的直线上，正如所求。（每次都得到 $\frac{2}{3}$ 这个值，说明 G（即形心（centroid））位于每条中线上由顶点起计三分之二的位置。）

在 H 处的垂直关系

点 H 的位置向量为 $\mathbf{h}=\mathbf{a}+\mathbf{b}+\mathbf{c}$，所以

$$\overrightarrow{AH} = \mathbf{h}-\mathbf{a} = \mathbf{b}+\mathbf{c}, \qquad \overrightarrow{BC} = \mathbf{c}-\mathbf{b}.$$

若直线 AH 与 BC 垂直相交，则这两个方向向量（direction vector）的纯量积为零：

$$0 = (\mathbf{c}+\mathbf{b})\cdot(\mathbf{c}-\mathbf{b}) = \mathbf{c}\cdot\mathbf{c}-\mathbf{b}\cdot\mathbf{b} = (p_3^2+q_3^2)-(p_2^2+q_2^2),$$

其中交叉项 $\mathbf{c}\cdot\mathbf{b}$ 与 $\mathbf{b}\cdot\mathbf{c}$ 互相抵销，因为纯量积符合交换律（commutative）。因此

$$p_2^2+q_2^2 = p_3^2+q_3^2,$$

这就是所求的结果。用文字来说，B 和 C 与原点 O 的距离相等。

以完全相同的方法，$\overrightarrow{BH}=\mathbf{a}+\mathbf{c}$ 而 $\overrightarrow{AC}=\mathbf{c}-\mathbf{a}$，所以若 BH 与 AC 垂直相交，则 $(\mathbf{c}+\mathbf{a})\cdot(\mathbf{c}-\mathbf{a})=0$，由此得

$$p_1^2+q_1^2 = p_3^2+q_3^2.$$

第三条高线

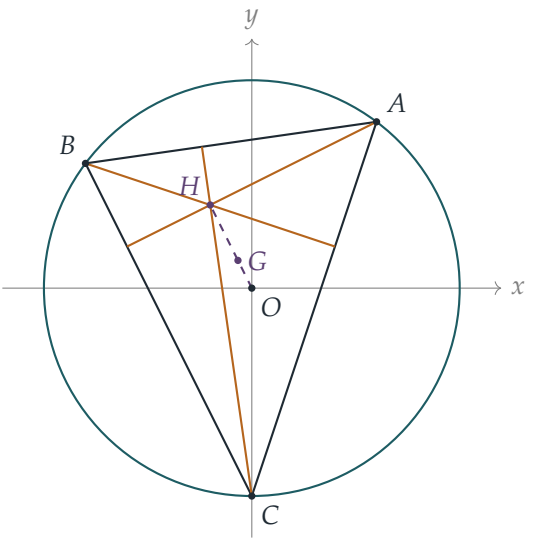
假设两个垂直条件都成立。结合以上两个结果，

$$p_1^2+q_1^2 = p_3^2+q_3^2 = p_2^2+q_2^2, \qquad \text{所以} \qquad \mathbf{a}\cdot\mathbf{a} = \mathbf{b}\cdot\mathbf{b}.$$

现在 $\overrightarrow{CH}=\mathbf{a}+\mathbf{b}$ 而 $\overrightarrow{AB}=\mathbf{b}-\mathbf{a}$，并且

$$\overrightarrow{CH}\cdot\overrightarrow{AB} = (\mathbf{b}+\mathbf{a})\cdot(\mathbf{b}-\mathbf{a}) = \mathbf{b}\cdot\mathbf{b}-\mathbf{a}\cdot\mathbf{a} = 0.$$

所以 CH 垂直于 AB（前提是 $H \neq C$，这样 CH 才是一条真正的直线；若 $H = C$，这个命题便没有内容）。推论至此完成。



上图显示最后一部分的情况：A、B、C 位于以原点 O 为圆心的圆上，三条高线（altitude）（琥珀色）都通过 H，而 O、G、H 共线，且 $\mathbf{h} = 3\mathbf{g}$。

验算

- 取 $A = (3,4)$、$B = (-4,3)$、$C = (0,-5)$，三点与 O 的距离都是 5。于是 $H = (-1,2)$，而 $\overrightarrow{AH} \cdot \overrightarrow{BC} = (-4)(4) + (-2)(-8) = 0$；另外两个纯量积也是 0。形心为 $G = (-\frac{1}{3}, \frac{2}{3})$，即 H 的三分之一。
- 数值方面，对大约 2000 个顶点为整数点的随机三角形，以分数精确计算出的两条中线交点总是 $\frac{1}{3}(\mathbf{a} + \mathbf{b} + \mathbf{c})$，并且位于第三条中线上；另外，对 1000 个内接于以 O 为圆心的圆的随机三角形，关于 H 的三个纯量积全部为零。

常见错误

- 证明了两条中线都通过 G，却没有说明它们为何不可能是同一条直线。只要用一句话指出 A、B、C 不共线，问题便解决了。
- 把 $\overrightarrow{OH} = \mathbf{a} + \mathbf{b} + \mathbf{c}$ 当作 AH 的方向。正确的方向是 $\mathbf{h} - \mathbf{a} = \mathbf{b} + \mathbf{c}$。
- 以为这个结果对所有三角形都成立。题目的假设迫使 O 与 A、B、C 等距；若原点是任意选取的，H 便不是垂心（orthocentre）。

STEP 之外

最后一部分证明了三角形的三条高线共点（concurrent）：把原点放在外心（circumcentre）O（与各顶点等距），则点 $\mathbf{a}+\mathbf{b}+\mathbf{c}$ 同时位于三条高线上。由于 $\mathbf{h}=3\mathbf{g}$，这也证明了外心、形心和垂心位于同一条直线上，这条直线称为欧拉线（Euler line），且 $OG:GH=1:2$。这种向量式的几何，由不依赖坐标的恒等式担起主要工作，正是大学一年级线性代数（linear algebra）处理几何的方式；Strang 的 *Introduction to Linear Algebra*（线性代数入门教材）对此有详细阐述，而 Siklos 的 *Advanced Problems in Mathematics* 则提供更多 STEP 风格的练习。

详细解答

STEP I 2004，第 7 题

纯数学 · 生成函数

STEP I 2004，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/04-s1-q7

关键思路

线性递推关系（linear recurrence）与幂级数（power series）是同一个对象的两种描述。把级数乘以 $1-7x+10x^2$，每个系数 a_n 便会与 a_{n-1} 和 a_{n-2} 对齐，于是递推关系令乘积中几乎所有系数都变为零。剩下来的是一个多项式，所以 f 是有理函数（rational function）。接着利用部分分式（partial fractions）和几何级数（geometric series），便可直接得出 a_n。第 (ii) 部分用的是同一套方法，只需先由已知数值求出 p 和 q。

解答

第 (i) 部分

当 $|x|<\frac{1}{5}$ 时，级数收敛（converge），所以我们可以把级数乘以 x 的幂，并逐项相加：

$$\begin{aligned} \mathrm{f}(x) &= a_0 + a_1 x + \sum_{n=2}^{\infty} a_n x^n, \\ -7x\,\mathrm{f}(x) &= -7a_0 x - \sum_{n=2}^{\infty} 7a_{n-1} x^n, \\ 10x^2\mathrm{f}(x) &= \sum_{n=2}^{\infty} 10a_{n-2} x^n. \end{aligned}$$

（在第二行和第三行，我们重新标记了求和指标，使每个和式都是对 x^n（$n \geq 2$）求和。）相加得

$$\mathrm{f}(x) - 7x\,\mathrm{f}(x) + 10x^2\mathrm{f}(x) = a_0 + (a_1 - 7a_0)x + \sum_{n=2}^{\infty} \left(a_n - 7a_{n-1} + 10a_{n-2}\right)x^n.$$

根据递推关系，和式中每个括号都等于零，而 $a_0 = 2$，$a_1 - 7a_0 = 7 - 14 = -7$。所以

$$(1 - 7x + 10x^2)\,\mathrm{f}(x) = 2 - 7x.$$

现在 $1 - 7x + 10x^2 = (1 - 2x)(1 - 5x)$，当 $|x| < \frac{1}{5}$ 时不等于零，所以可以相除：

$$\mathrm{f}(x) = \frac{2 - 7x}{(1 - 2x)(1 - 5x)}.$$

分子正好是两个因式之和，$(1 - 5x) + (1 - 2x) = 2 - 7x$，所以

$$\mathrm{f}(x) = \frac{(1 - 5x) + (1 - 2x)}{(1 - 2x)(1 - 5x)} = \frac{1}{1 - 2x} + \frac{1}{1 - 5x},$$

正如所求。

当 $|x| < \frac{1}{5}$ 时，我们有 $|2x| < 1$ 和 $|5x| < 1$，所以两个分式都是收敛几何级数之和：

$$\mathrm{f}(x) = \sum_{n=0}^{\infty} (2x)^n + \sum_{n=0}^{\infty} (5x)^n = \sum_{n=0}^{\infty} \left(2^n + 5^n\right) x^n.$$

想一想 为甚么现在可以直接读出 a_n？

我们有两个幂级数，它们对区间 $|x| < \frac{1}{5}$ 内的每个 x 都相等。幂级数由它在 0 附近的值决定：它的系数可以由 $a_n = \mathrm{f}^{(n)}(0)/n!$ 还原出来。所以两者的系数必定相同，于是

$$a_n = 2^n + 5^n \quad \text{对所有} n \geq 0.$$

（简单验算一下：$a_0 = 1 + 1 = 2$，$a_1 = 2 + 5 = 7$。）

第 (ii) 部分

首先要求出 p 和 q。把 $n = 2$ 和 $n = 3$ 代入递推关系：

$$\begin{aligned} b_2 = pb_1 + qb_0: &\quad 40 = 10p + 5q, \quad \text{所以} 2p + q = 8; \\ b_3 = pb_2 + qb_1: &\quad 100 = 40p + 10q, \quad \text{所以} 4p + q = 10. \end{aligned}$$

两式相减，得 $2p = 2$，所以 $p = 1$ 及 $q = 6$。递推关系是 $b_n = b_{n-1} + 6b_{n-2}$。

与第 (i) 部分完全一样，乘以 $1 - x - 6x^2$ 会消去由 x^2 起的所有系数：

$$(1 - x - 6x^2)\,\mathrm{g}(x) = b_0 + (b_1 - b_0)x = 5 + 5x.$$

由于 $1 - x - 6x^2 = (1 - 3x)(1 + 2x)$，当 $|x| < \frac{1}{3}$ 时不等于零，

$$\mathrm{g}(x) = \frac{5 + 5x}{(1 - 3x)(1 + 2x)} = \frac{A}{1 - 3x} + \frac{B}{1 + 2x},$$

其中 $5+5x = A(1+2x)+B(1-3x)$。代入 $x=\frac{1}{3}$ 得 $\frac{20}{3}=\frac{5}{3}A$，所以 $A=4$；代入 $x=-\frac{1}{2}$ 得 $\frac{5}{2}=\frac{5}{2}B$，所以 $B=1$。因此

$$g(x)=\frac{4}{1-3x}+\frac{1}{1+2x}.$$

当 $|x|<\frac{1}{3}$ 时，我们有 $|3x|<1$ 和 $|2x|<1$，所以把两者都展开成几何级数，

$$g(x)=\sum_{n=0}^{\infty}\left(4\cdot 3^n+(-2)^n\right)x^n,$$

再如前比较系数，

$$b_n=4\cdot 3^n+(-2)^n.$$

验算

这些公式与已知数值吻合：$4+1=5$，$12-2=10$，$36+4=40$，$108-8=100$。由递推关系逐项计算 a_n 和 b_n 直至 $n=39$，结果与 2^n+5^n 和 $4\cdot 3^n+(-2)^n$ 完全一致。数值上，在 $x=0.1$ 时，级数的部分和（partial sums）给出 f 的值为 3.25，g 的值为 6.547619…，与 $\frac{1}{0.8}+\frac{1}{0.5}=3.25$ 和 $\frac{4}{0.7}+\frac{1}{1.2}=\frac{275}{42}$ 相符。

常见错误

- 忘记递推关系只由 $n=2$ 起成立：乘积中的 x^0 项和 x^1 项必须另行计算，而它们正好给出分子。
- 在第 (ii) 部分，在 p 和 q 仍然未知时便使用递推关系，或把分子写成 b_0+b_1x 而不是 $b_0+(b_1-pb_0)x$。
- 因式分解 $1-x-6x^2$ 时弄错正负号，结果得出 3^n 和 2^n，而不是 3^n 和 $(-2)^n$。

STEP 之外

级数 f 称为数列 (a_n) 的*生成函数*（generating function），而这个方法适用于任何常系数线性递推关系：分母总是由递推关系构造出来，而它的因式则给出“特征根”（characteristic roots）（这里是 2 和 5）。生成函数是组合数学（combinatorics）和概率论的核心工具；在概率论中，概率生成函数（probability generating functions）把随机变量之和化为乘积。“相等的幂级数有相等的系数”这一步，要在大学第一门分析课程中才有严格的论证；关于幂级数及其唯一性，可参阅 Abbott 的 *Understanding Analysis*。

详细解答

STEP I 2004，第 8 题

纯数学 · 单调数列与数学归纳法

STEP I 2004，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/04-s1-q8

关键思路

要证明一个数列（sequence）递增，可把差 $t_{n+1}-t_n$ 看成 t_n 的函数。在两部分中，这个差都可以因式分解，而所得的因式正好对应递推关系（recurrence）的不动点（fixed point）2 和 3（即令 $t_{n+1}=t_n$ 的值）。于是差的正负只取决于 t_n 相对于 2 和 3 的位置，所以真正要下工夫的，是用归纳法（induction）证明每一项都留在适当的区域内。

解答

第 (i) 部分

对任意 n，

$$x_{n+1}-x_n=\frac{x_n^2+6}{5}-x_n=\frac{x_n^2-5x_n+6}{5}=\frac{(x_n-2)(x_n-3)}{5}.$$

若 $x_n>3$，则两个因式皆为正，所以 $x_{n+1}-x_n>0$，即 $x_{n+1}>x_n$。

余下要验证每一项都大于 3，我们用归纳法来证明。根据假设，命题 $x_n>3$ 在 $n=0$ 时成立。假设 $x_n>3$。由上面的计算，$x_{n+1}>x_n>3$，所以亦有 $x_{n+1}>3$。因此对所有 $n\geq 0$ 都有 $x_n>3$，从而对所有 $n\geq 0$ 都有 $x_{n+1}>x_n$：这个数列严格递增（strictly increasing）。

? **想一想** 为甚么只证明 $x_1>x_0$ 并不足够?

不等式 $x_{n+1}>x_n$ 是由 $x_n>3$ 推导出来的，所以我们在每一步都需要这个条件，而不仅是第一步。在这里，这个条件不费工夫便可得到：从大于 3 的位置再向上走一步，仍然大于 3。（留意若起始值恰好是 $x_0=3$，便会得到常数数列，所以严格不等式是必要的。）

第 (ii) 部分

首先留意，若 $2 < y_n < 3$，则 $y_n > 0$，所以 $y_{n+1} = 5 - 6/y_n$ 有定义。对这样的 y_n，

$$y_{n+1} - y_n = 5 - \frac{6}{y_n} - y_n = -\frac{y_n^2 - 5y_n + 6}{y_n} = \frac{(y_n - 2)(3 - y_n)}{y_n},$$

此式为正，因为 $y_n - 2 > 0$、$3 - y_n > 0$ 及 $y_n > 0$。所以只要 $2 < y_n < 3$，便有 $y_{n+1} > y_n$。

与第 (i) 部分一样，我们必须证明每一项都位于这个区间内。这次它不会自动成立，因为递增的数列有可能越过 3。所以我们考察各项与 3 的距离：

$$3 - y_{n+1} = 3 - 5 + \frac{6}{y_n} = \frac{6 - 2y_n}{y_n} = \frac{2(3 - y_n)}{y_n}.$$

? **想一想** 我们怎样知道这个数列永远不会跳过 3?

上式说明，只要 $y_n > 0$，$3 - y_{n+1}$ 便与 $3 - y_n$ 同号。所以小于 3 的一项之后，紧接的一项必定也小于 3。

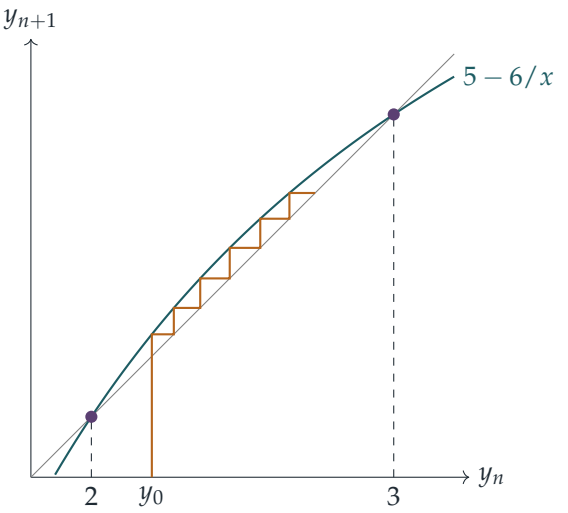
现在进行归纳。断言：对所有 $n \geq 0$，$2 < y_n < 3$。根据假设，这在 $n = 0$ 时成立。假设 $2 < y_n < 3$。那么：

- 由关于差的计算，$y_{n+1} > y_n > 2$；
- $3 - y_{n+1} = 2(3 - y_n)/y_n > 0$，因为 $3 - y_n > 0$ 及 $y_n > 0$，所以 $y_{n+1} < 3$。

因此 $2 < y_{n+1} < 3$，归纳完成。

由于对每个 n 都有 $2 < y_n < 3$，关于差的计算在每一步都适用，所以对所有 $n \geq 0$ 都有 $y_{n+1} > y_n$。这个数列严格递增，但对所有 n 仍有 $y_n < 3$，正如所求。

蛛网图（cobweb diagram）显示了实际的情况：曲线 $y = 5 - 6/x$ 与直线 $y = x$ 相交于 $x = 2$ 和 $x = 3$，在两者之间曲线位于直线上方，而各次迭代的值向着 3 攀升，却始终达不到它。



验算

- 在不动点处，递推关系给出常数数列：$(3^2+6)/5=3$，以及 $5-6/3=3$、$5-6/2=2$，与因式 $(t-2)(t-3)$ 吻合。
- 由 $y_0=2.5$ 开始：$y_1=2.6$，$y_2\approx 2.692$，$y_3\approx 2.771$，递增且小于 3；经过 200 步后，y_n 与 3 相差不超过 10^{-15}。由 $x_0=3.1$ 开始：$x_1=3.122$，$x_2\approx 3.149$，递增。
- 我们用电脑以精确分数对数百个随机起始值检验了两个归纳论证，并确认当 x_0 略小于 3 时会得到 $x_1<x_0$，所以条件 $x_0>3$ 是必需的。

常见错误

- 在假设 $x_n>3$ 之下证明了 $x_{n+1}>x_n$ 便停下来，没有用归纳法证明每一项都大于 3。
- 在第 (ii) 部分，证明了 $y_{n+1}<3$，却忘记要继续归纳还需要 $y_{n+1}>2$（它可由 $y_{n+1}>y_n$ 推出）。
- 把不等式两边同乘以 y_n，却没有指出 $y_n>0$。

STEP 之外

第 (ii) 部分的数列是有界的递增数列，而数学分析中的*单调收敛定理*（monotone convergence theorem）指出，这样的数列必定收敛；任何极限 L 都满足 $L=5-6/L$，所以这里 $y_n\to 3$。在第 (i) 部分，数列递增而且无上界，因为有限的极限必须是一个大于 3 的不

动点，而这样的不动点并不存在。迭代的值究竟被不动点吸引还是排斥，取决于映射在该点的斜率，这正是离散动力系统（discrete dynamical systems）理论的起点。Abbott 的 *Understanding Analysis* 对单调收敛有严谨细致的论述。

详细解答

STEP I 2004，第 9 题

力学 · 抛体运动

STEP I 2004，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/04-s1-q9

关键思路

第一部分是标准的“在最高点垂直速度为零”的计算。在隧道中，题目没有给出速率 u，但它也不是任意的：一旦选定 θ，要求路径通过 P 就确定了 u。因此我们消去 u，把最大高度写成只含 θ 的函数（以 $t = \tan\theta$ 表示最为方便），再问它何时保持在隧道顶之下。答案最后化为一个关于 t 的二次不等式（quadratic inequality）。

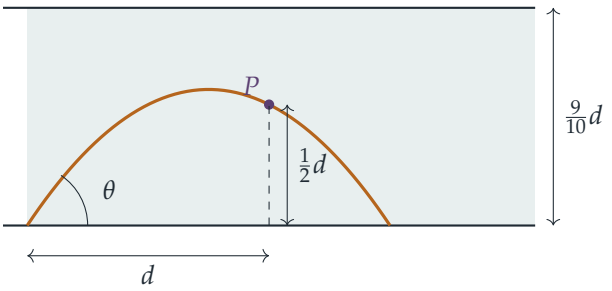
解答

最大高度

把粒子视为在重力作用下自由运动的质点，不计空气阻力（air resistance）。取抛射点为原点建立坐标轴，x 轴水平，y 轴垂直向上。速度的垂直分量（vertical component）起初为 $u\sin\theta$，并有恒定加速度 $-g$。在最高点垂直速度为零，所以在垂直方向应用 $v^2 = v_0^2 + 2as$，得

$$0 = u^2\sin^2\theta - 2gH, \qquad \text{所以} \qquad H = \frac{u^2\sin^2\theta}{2g}.$$

隧道



粒子的路径可由 $x = uT\cos\theta$ 与 $y = uT\sin\theta - \frac{1}{2}gT^2$ 消去时间 T 而得：

$$y = x\tan\theta - \frac{gx^2}{2u^2\cos^2\theta} = xt - \frac{gx^2(1+t^2)}{2u^2},$$

其中 $t = \tan\theta$，并用了 $\sec^2\theta = 1 + t^2$。

通过 P。 代入 $x = d$、$y = \frac{1}{2}d$，再除以 d：

$$\frac{1}{2} = t - \frac{gd(1+t^2)}{2u^2}, \qquad \text{所以} \qquad \frac{gd(1+t^2)}{2u^2} = t - \frac{1}{2}.$$

左边为正数，所以必须有 $t > \frac{1}{2}$。（这是合理的：粒子起初朝 θ 方向运动，而重力只会把路径向下弯，所以只有瞄准的方向高于通往 P 的直线时才能到达 P，而该直线的斜率为 $\frac{1}{2}$。）对每个这样的 θ，恰好有一个速率符合要求：

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

? **想一想** 题目没有给出 u 的值。条件怎么可能只与 θ 有关?

因为速率并不由我们选择：一旦 θ 固定，“通过 P”这个条件就把 u 确定下来。所以最大高度其实只是 θ 的函数。

不碰隧道顶。 把 $\sin^2\theta = \dfrac{t^2}{1+t^2}$ 代入 H 的公式，

$$H = \frac{u^2}{2g}\cdot\frac{t^2}{1+t^2} = \frac{d(1+t^2)}{2(2t-1)}\cdot\frac{t^2}{1+t^2} = \frac{dt^2}{2(2t-1)}.$$

粒子不碰隧道顶，当且仅当 $H < \frac{9}{10}d$。由于 $2t-1 > 0$，我们可以把不等式两边同乘 $10(2t-1)/d$ 而不必改变不等号的方向：

$$\begin{aligned}\frac{dt^2}{2(2t-1)} < \frac{9d}{10} &\iff 5t^2 < 9(2t-1)\\ &\iff 5t^2 - 18t + 9 < 0\\ &\iff (5t-3)(t-3) < 0\\ &\iff \tfrac{3}{5} < t < 3.\end{aligned}$$

这个区间内的每个 t 都满足 $t > \frac{1}{2}$，所以先前的条件没有带来额外限制。若 $t = \frac{3}{5}$ 或 $t = 3$，粒子刚好触及隧道顶，这种情况被排除。由于粒子始终没有碰到隧道顶，它会落回地面，并在隧道内着地。

最后，θ 介乎 0 与 $\frac{\pi}{2}$ 之间，而 tan 在此范围内递增，所以

$$\arctan\tfrac{3}{5} < \theta < \arctan 3,$$

这正是所要证明的结果。

验算

- 当 $t = 3$ 时：$H = \dfrac{9d}{2\cdot 5} = \frac{9}{10}d$；当 $t = \frac{3}{5}$ 时：$H = \dfrac{(9/25)d}{2\cdot(1/5)} = \frac{9}{10}d$。两个端点都刚好擦过隧道顶，正如所料。
- 当 $t \to \frac{1}{2}^{+}$ 时，所需速率 $u \to \infty$，且 $H \to \infty$：粒子几乎沿着通往 P 的直线运动，所以它必须非常快，其后亦会升得非常高。
- 数值方面，取 $d = 3$ 和 $g = 9.8$，我们测试了 10^5 个随机角度。对每个角度，我们求出 u，确认路径通过 P，并发现 $H < \frac{9}{10}d$ 当且仅当 $0.5404 < \theta < 1.2490$（即以弧度表示的 $\arctan 0.6$ 和 $\arctan 3$）。逐步模拟亦确认了 $H = u^2\sin^2\theta/(2g)$。

常见错误

- 把 u 当作另一个独立的未知数，结果做不下去：必须利用在 P 的条件把它消去。
- 把不等式两边同乘 $2t-1$ 之前，忘记检查它的正负。
- 只检查 P 是否在隧道顶之下。路径的最高点通常出现在 P 之后，所以真正重要的是最大高度。

STEP 之外

把轨迹（trajectory）以 $\tan\theta$ 表示，可将抛体（projectile）问题化为关于 t 的二次式。同一技巧可求出以给定速率抛射的所有轨迹的“包络线”（envelope，即作为边界的抛物线），并给出击中目标的两个可能角度。像这样的问题，部分参数由约束条件确定，其余参数则构成一个容许区域，由此可通往最优化（optimisation）和变分法（calculus of variations）。Kleppner 与 Kolenkow 的 *An Introduction to Mechanics*（力学入门）是进一步学习这类力学的极佳读物。

详细解答

STEP I 2004，第 10 题

力学 · 匀加速运动学

STEP I 2004，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：

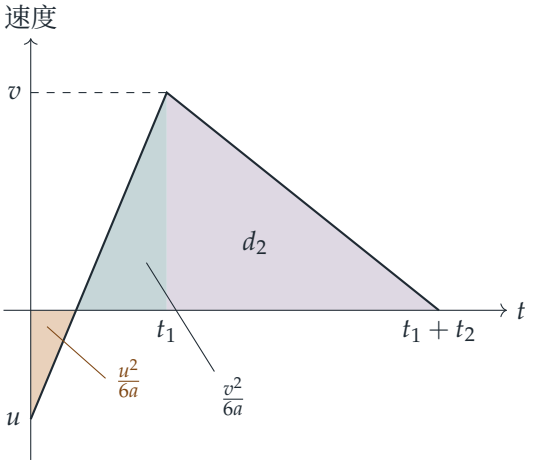
step.maths.org/questions/04-s1-q10

关键思路

这里全部都是匀加速运动（constant acceleration），所以单靠 $v^2 = u^2 + 2as$ 和 $v = u + at$ 便足以解决整条题目。唯一的陷阱是距离（distance）这个词。当 $u < 0$ 时，质点（particle）先向后运动，停下来，然后才向前加速，所以距离 d_1 是两段路程之和，而不是单一个位移（displacement）。画一幅速度—时间图（velocity–time graph）便一目了然：距离就是图线与时间轴之间的总面积，轴下方的面积也当作正数计算。

解答

取最终速度 v 的方向为正方向。在第一阶段，加速度是 $+3a$（它令速度由 u 上升至 $v > u$）；在第二阶段，加速度是 $-a$，直至速度变为 0。全程均有 $a > 0$。



上图所示为 $u < 0$ 的情况：轴下方的小三角形就是向后运动的那段路程。

第二阶段（两种情况皆适用）。 质点以减速度（deceleration）a 由速率 v 减至静止，期间从不掉头，所以

$$0 = v^2 - 2ad_2 \quad \Longrightarrow \quad d_2 = \frac{v^2}{2a}.$$

第 (i) 部分

若 $u > 0$，速度由 u 上升至 v，一直保持正值，所以距离等于位移：

$$v^2 = u^2 + 2(3a)d_1 \quad \Longrightarrow \quad d_1 = \frac{v^2 - u^2}{6a}.$$

因此

$$3d_1 = \frac{v^2 - u^2}{2a} < \frac{v^2}{2a} = d_2,$$

因为 $u^2 > 0$。所以 $3d_1 < d_2$，正如所求。

第 (ii) 部分

现在 $u < 0$，第一阶段要分成两段。

> ? **想一想** 为甚么不能再直接使用 $v^2 = u^2 + 6ad_1$?

这条公式给出的是位移，即 $\frac{v^2-u^2}{6a}$。但质点是先向后运动的。在加速度 $3a$ 之下，它由速率 $|u|$ 减慢至静止，走过 $\frac{u^2}{6a}$（由 $0 = u^2 - 2(3a)s$ 得出）。然后它由静止加速至 v，走过 $\frac{v^2}{6a}$。距离是两者之和：

$$d_1 = \frac{u^2 + v^2}{6a}, \qquad \text{所以} \qquad 3d_1 = \frac{u^2 + v^2}{2a}.$$

与 $d_2 = \frac{v^2}{2a}$ 和 $2d_2 = \frac{2v^2}{2a}$ 比较：

- $3d_1 - d_2 = \dfrac{u^2}{2a} > 0$，因为 $u \neq 0$；
- $2d_2 - 3d_1 = \dfrac{v^2 - u^2}{2a} > 0$，因为 $v > |u|$。

因此 $d_2 < 3d_1 < 2d_2$，正如所求。

平均速率

所需时间与方向无关。在第一阶段，速度以变率 $3a$ 改变了 $v - u$；在第二阶段，速度以变率 a 改变了 v，所以

$$t_1 = \frac{v - u}{3a}, \qquad t_2 = \frac{v}{a}, \qquad t_1 + t_2 = \frac{4v - u}{3a},$$

不论 u 的正负如何。

$u > 0$ 的情况。总距离是 $\frac{v^2-u^2}{6a} + \frac{3v^2}{6a} = \frac{4v^2-u^2}{6a}$，所以平均速率（average speed）是

$$S_+ = \frac{4v^2-u^2}{6a} \cdot \frac{3a}{4v-u} = \frac{4v^2-u^2}{2(4v-u)}.$$

$u < 0$ 的情况。设 $u = -w$，其中 $0 < w < v$。总距离是 $\frac{v^2+w^2}{6a} + \frac{3v^2}{6a} = \frac{4v^2+w^2}{6a}$，所需时间是 $\frac{4v+w}{3a}$，所以

$$S_- = \frac{4v^2+w^2}{2(4v+w)}.$$

要比较这两个数，可把它们分别与 $\frac{1}{2}v$ 比较：

$$S_+ - \tfrac{1}{2}v = \frac{4v^2-u^2-v(4v-u)}{2(4v-u)} = \frac{u(v-u)}{2(4v-u)},$$
$$S_- - \tfrac{1}{2}v = \frac{4v^2+w^2-v(4v+w)}{2(4v+w)} = -\frac{w(v-w)}{2(4v+w)}.$$

在第一种情况中 $0 < u < v$，第一个式子的分子和分母皆为正数，所以 $S_+ > \frac{1}{2}v$。在第二种情况中 $0 < w < v$，第二个式子是负数，所以 $S_- < \frac{1}{2}v$。因此

$$S_- < \tfrac{1}{2}v < S_+,$$

即 $u > 0$ 时平均速率较大，正如所求。留意这个结论对两种情况中任何 $|u|$ 的值都成立，而不限于两者相等的时候：只要 v 和 a 相同，$\frac{1}{2}v$ 便把每一段 $u > 0$ 的旅程与每一段 $u < 0$ 的旅程分隔开。

验算

- 取 $v = 2$，$a = 1$。当 $u = 1$：$d_1 = \frac{1}{2}$，$d_2 = 2$，所以 $3d_1 = 1.5 < 2$，所需时间是 $\frac{7}{3}$，得出 $S_+ = \frac{2.5}{7/3} \approx 1.071 > 1$。当 $u = -1$：$d_1 = \frac{5}{6}$，所以 $2 < 3d_1 = 2.5 < 4$，所需时间是 3，得出 $S_- = \frac{17/6}{3} \approx 0.944 < 1$。以数值方法逐步模拟这个运动，也得出相同的数值。
- 极限（limit）：当 $u \to 0$ 或 $|u| \to v$ 时，两个平均速率都趋向 $\frac{1}{2}v$，这正是由静止起步、最后停下的简单旅程的平均速率。在容许范围的边界上，各个不等式也变成等式：在两个部分中，当 $u \to 0$ 时 $3d_1 \to d_2$；在第 (ii) 部分中，当 $u \to -v$ 时 $3d_1 \to 2d_2$。

常见错误

- 在第 (ii) 部分使用 $d_1 = \frac{v^2-u^2}{6a}$。这是位移，而且会再次得出 $3d_1 < d_2$，与题目要求证明的结果矛盾。

- 以为 $u < 0$ 时，时间也需要分段计算。其实不必：$v = u + 3at_1$ 在经过折返点前后一直成立。

STEP 之外

位移与距离之别，就是 $\int v\,\mathrm{d}t$ 与 $\int |v|\,\mathrm{d}t$ 之别。后者是路径的弧长（arc length），也就是汽车里程表所量度的数值。在大学的力学课程中，弧长成为描述弯曲轨迹的自然参数，由此引出加速度的切向分量（tangential component）和法向分量（normal component）。Kleppner 与 Kolenkow 合著的 *An Introduction to Mechanics*（力学入门）正是以这种方式建立运动学（kinematics）的。

详细解答

STEP I 2004，第 11 题

力学 · 摩擦力与铰接梯子

STEP I 2004，第 11 题

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

step.maths.org/questions/04-s1-q11

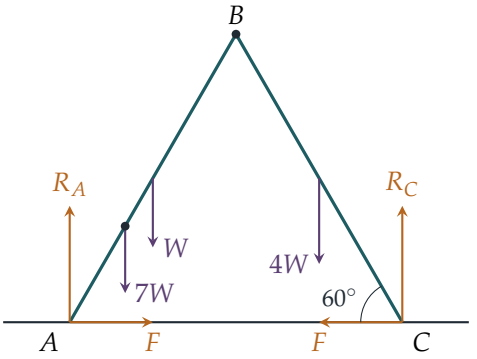
关键思路

在任何一处滑动之前，整个梯架处于平衡（equilibrium），所以我们可以由整个系统求出两个法向反作用力（normal reaction），再单独考虑其中一把梯子来求摩擦力（对铰链取力矩，便可消去未知的铰链力）。水平方向的外力只有两个摩擦力，所以它们大小相等。法向反作用力较小的那只梯脚，所需的比值 F/R 较大，因此最先滑动的就是那只梯脚。

解答

我们把每把梯子视为长度为 $2a$ 的均匀杆（uniform rod），把 B 处的铰链（hinge）视为光滑，并把装修工人视为质点（particle）。由于 $AB = BC$ 且 $\angle ABC = 60°$，三角形 ABC 是等边三角形：$AC = 2a$，而每把梯子与地面成 $60°$ 角。于是 B 与每只梯脚的水平距离为 a，离地面的高度为 $a\sqrt{3}$。

当装修工人沿 AB 爬到三分之一处时，她与 A 的水平距离为 $\frac{1}{3}a$。AB 的中心与 A 的水平距离为 $\frac{1}{2}a$，而 BC 的中心与 A 的水平距离为 $\frac{3}{2}a$。



法向反作用力

设 A 和 C 处的法向反作用力分别为 R_A 和 R_C。对整个系统沿铅垂方向分解力，得

$$R_A + R_C = W + 4W + 7W = 12W.$$

对整个系统以 A 为支点取力矩（moments）（铰链力是内力，所以不会出现）：

$$R_C \cdot 2a = W \cdot \tfrac{1}{2}a + 7W \cdot \tfrac{1}{3}a + 4W \cdot \tfrac{3}{2}a = \tfrac{53}{6}Wa,$$

所以

$$R_C = \tfrac{53}{12}W, \qquad R_A = 12W - \tfrac{53}{12}W = \tfrac{91}{12}W.$$

摩擦力

两只梯脚有向外张开的趋势，所以每只梯脚所受的摩擦力（friction）都指向内侧。对整个系统沿水平方向分解力，摩擦力是仅有的水平外力，所以两者大小相同；记这个大小为 F。

要求出 F，我们需要单独考虑其中一把梯子。取没有装修工人的 BC，并以 B 为支点取力矩，使未知的铰链力消失。以 B 为支点，反作用力 R_C 的力臂（lever arm）为 a，重量 $4W$ 的力臂为 $\frac{1}{2}a$，而摩擦力 F（水平方向，作用于地面的高度）的力臂为 $a\sqrt{3}$。重量和 F 使 BC 向一方转动，而 R_C 使它向另一方转动，所以

$$R_C \cdot a = 4W \cdot \tfrac{1}{2}a + F \cdot a\sqrt{3} \quad \Longrightarrow \quad F = \frac{R_C - 2W}{\sqrt{3}} = \frac{29W}{12\sqrt{3}}.$$

作为验算，对 AB 以 B 为支点作同样的计算，得 $R_A \cdot a = W \cdot \frac{1}{2}a + 7W \cdot \frac{2}{3}a + Fa\sqrt{3}$，所以 $F\sqrt{3} = \frac{91}{12}W - \frac{31}{6}W = \frac{29}{12}W$，与上面的结果一致。

哪一把梯子会滑动?

想一想 两只梯脚所受的摩擦力相同，那么我们怎样判断哪一只会先失守?

当所需的摩擦力等于 μ 乘以该处的法向反作用力时，梯脚便处于即将滑动的状态。两只梯脚所需的 F 相同，所以法向反作用力较小的梯脚所需的比值较大:

$$\frac{F}{R_A}=\frac{29}{91\sqrt{3}}\approx 0.184, \qquad \frac{F}{R_C}=\frac{29}{53\sqrt{3}}\approx 0.316.$$

所以 C 最先达到极限摩擦（limiting friction）：滑动的是**较重的梯子** BC，尽管装修工人是在 AB 上。她的重量把 A 牢牢压在地面上，而 C 所承受的负荷则少得多。

我们还应确定在此之前没有任何一处已经滑动。若她爬到全长的 t 倍处，同样的计算给出 $R_C=(\frac{13}{4}+\frac{7}{2}t)W$ 和 $F\sqrt{3}=R_C-2W=(\frac{5}{4}+\frac{7}{2}t)W$，所以

$$\frac{F}{R_C}=\frac{1}{\sqrt{3}}\cdot\frac{\frac{5}{4}+\frac{7}{2}t}{\frac{13}{4}+\frac{7}{2}t}=\frac{1}{\sqrt{3}}\left(1-\frac{2}{\frac{13}{4}+\frac{7}{2}t}\right),$$

它随 t 递增。同样，F/R_A 也随 t 递增（它的分子上升，而分母 $R_A=12W-R_C$ 下降）。所以随着她向上爬，所需的比值不断增大，而第一次滑动恰好发生在 F/R_C 达到 μ 之时，即 $t=\frac{1}{3}$。因此

$$\mu=\frac{29}{53\sqrt{3}}=\frac{29\sqrt{3}}{159}\approx 0.316.$$

验算

- 对 AB 以 B 为支点取力矩，所得的 F 与对 BC 以 B 为支点取力矩所得的相同（见上文），所以各反作用力互相吻合。
- 完全没有装修工人时，由于 BC 较重，R_C 应大于 R_A；事实上，此时以 A 为支点取力矩得 $R_C=\frac{13}{4}W$ 和 $R_A=\frac{7}{4}W$，两者之和为 $5W$。
- 当 $t=\frac{11}{14}$ 时 $R_A=R_C$；只有超过这一点，滑动的才会是 AB，所以在 $t=\frac{1}{3}$ 时滑动的必定是 BC。
- 数值上，由坐标重新计算每一个力（并对两把梯子检查以 B 为支点的力矩），得 $\mu\approx 0.31591$。

常见错误

- 以为载着装修工人的那把梯子会滑动。她的重量增大了 A 处的法向反作用力，使那只梯脚更加稳固。

- 对单独一把梯子以梯脚为支点取力矩，这样会引入未知的铰链力。对一把梯子以 B 为支点取力矩便可避开它。
- 忘记摩擦力对铰链的力臂为 $a\sqrt{3}$（即 B 的高度）。

STEP 之外

这是处理铰接结构的静力学（statics）基本策略：先把整个结构当作一个整体来求外部反作用力，然后在某个接合处把它切开，求出其中一个构件必须提供的力。哪一个接触点最先失效这类问题，会引向框架的分析，以及虚功原理（principle of virtual work）；在虚功原理中，平衡是通过考虑符合约束条件的微小位移来求得的。Kleppner 与 Kolenkow 的 *An Introduction to Mechanics*（力学入门教科书）对刚体静力学和摩擦有细致的论述。

详细解答

STEP I 2004，第 12 题

概率与统计 · 贝叶斯定理

STEP I 2004，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/04-s1-q12

关键思路

从第 1 组抽出的样本，大致告诉我们：已知一块晶片属于第 1 组，它来自机器 A 的条件概率（conditional probability）是多少。因此，我们利用贝叶斯定理（Bayes′ theorem），以 λ、p 和 q 表示这个条件概率，令它等于观察所得的比例，再解出 λ。分拣出错只改变了每部机器所生产的晶片落入第 1 组的概率；方法则完全相同。

解答

从工厂的产品中随机抽出一块晶片。设 A 为该晶片由机器 A 制造的事件（故 $\mathrm{P}(A) = \lambda$ 及 $\mathrm{P}(A') = 1 - \lambda$），$P$ 为该晶片完美无瑕的事件，G 为该晶片被放入第 1 组的事件。题目给出 $\mathrm{P}(P \mid A) = p$ 及 $\mathrm{P}(P \mid A') = q$。

从第 1 组随机抽出的一块晶片，就是在 G 发生的条件下，从全部产品中随机抽出的一块晶片。由于样本很大，观察所得的比例 $\frac{2}{5}$ 是 $\mathrm{P}(A \mid G)$ 的良好估计，而我们对 λ 的估计值，就是使 $\mathrm{P}(A \mid G)$ 恰好等于 $\frac{2}{5}$ 的那个值。

分拣完全准确

此时 G 与 P 是同一事件。根据贝叶斯定理，

$$\mathrm{P}(A \mid P) = \frac{\mathrm{P}(P \mid A)\,\mathrm{P}(A)}{\mathrm{P}(P \mid A)\,\mathrm{P}(A) + \mathrm{P}(P \mid A')\,\mathrm{P}(A')} = \frac{\lambda p}{\lambda p + (1-\lambda)q}.$$

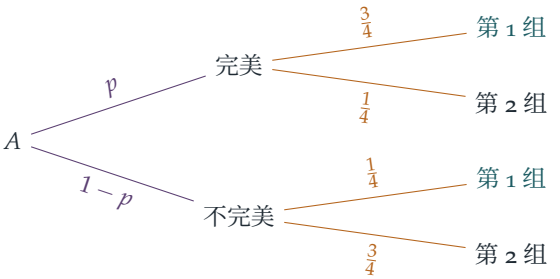
令上式等于 $\frac{2}{5}$，然后交叉相乘（分母为正数）：

$$\begin{aligned} 5\lambda p &= 2\lambda p + 2(1-\lambda)q \\ 3\lambda p + 2\lambda q &= 2q \\ \lambda &= \frac{2q}{3p+2q}, \end{aligned}$$

这就是所求的估计值，因为 $3p + 2q > 0$。

分拣出错

现在，一块晶片进入第 1 组有两种途径：本身完美而被正确分拣，或本身不完美而被错误分拣。我们假设分拣出错与晶片由哪部机器制造互相独立（independent），也就是说，在已知晶片是否完美的情况下，它如何被分拣与它来自哪部机器无关。下面的树形图显示来自机器 A 的晶片的各个分支；机器 B 的情况相同，只需把 p 换成 q。



把通往第 1 组的两个分支相加，

$$\mathrm{P}(G \mid A) = \tfrac{3}{4}p + \tfrac{1}{4}(1-p) = \frac{1+2p}{4}, \qquad \mathrm{P}(G \mid A') = \tfrac{3}{4}q + \tfrac{1}{4}(1-q) = \frac{1+2q}{4}.$$

于是，再次运用贝叶斯定理（因子 $\frac{1}{4}$ 互相约去），

$$\mathrm{P}(A \mid G) = \frac{\lambda(1+2p)}{\lambda(1+2p) + (1-\lambda)(1+2q)}.$$

想一想 证据有没有改变？$\frac{2}{5}$ 还应该沿用吗？

应该沿用：样本确实是从第 1 组抽出的，而其中确实有 $\frac{2}{5}$ 来自机器 A。改变了的，只是我们对晶片如何进入第 1 组所建立的模型。因此，我们令 $\mathrm{P}(A \mid G)$ 的新表达式等于 $\frac{2}{5}$：

$$\begin{aligned} 5\lambda(1+2p) &= 2\lambda(1+2p) + 2(1-\lambda)(1+2q) \\ \lambda\big[3(1+2p) + 2(1+2q)\big] &= 2(1+2q) \\ \lambda &= \frac{2(1+2q)}{5+6p+4q}. \end{aligned}$$

这就是 λ 的新估计值。留意它其实就是第一个答案，只是把 p 和 q 分别换成 $\frac{1+2p}{4}$ 和 $\frac{1+2q}{4}$，即进入第 1 组的新概率。

验算

- 若 $p = q$，两部机器在品质上无从分辨，因此第 1 组的组成应与整体产量的比例一致，而 λ 应为 $\frac{2}{5}$。两条公式都给出这个结果：$\frac{2p}{5p} = \frac{2}{5}$ 及 $\frac{2(1+2p)}{5+10p} = \frac{2}{5}$。
- 两个估计值都严格介乎 0 与 1 之间，因为两者的分子均为正数，且小于分母。
- 以 $\lambda = 0.3$、$p = 0.6$、$q = 0.8$ 及分拣出错的情况模拟 4×10^5 块晶片，结果第 1 组中有 0.2667 来自机器 A，而公式给出的数值是 0.2661；把模拟所得的比例代入公式反过来求解，得回 $\lambda \approx 0.301$。

常见错误

- 令 $\frac{2}{5}$ 等于 $\mathrm{P}(G \mid A)$ 或 λp，而不是等于条件概率 $\mathrm{P}(A \mid G)$。
- 在第二部分，忘记了被错误放入第 1 组的不完美晶片，即只用了 $\frac{3}{4}p$。

STEP 之外

选取能使观察所得数据“完全吻合”的参数值，这种做法称为*矩估计法*（method of moments）；在本题中，它与最大似然估计（maximum likelihood estimation）的结果一致，而后者你会在大学的第一门统计学课程中遇到。出错的分拣机是量度误差（measurement error）的一个简单模型，这种误差又称错误分类（misclassification），在医学检测中是个严重的实际问题：假阳性（false positive）和假阴性（false negative）正是以这种方式扭曲观察所得的比例。想进一步了解贝叶斯定理和条件化，可参阅 Blitzstein 与 Hwang 合著的 *Introduction to Probability*（概率论入门）。

详细解答

STEP I 2004，第 13 题

概率与统计 • 均匀分布变量的最大值

STEP I 2004，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/04-s1-q13

关键思路

若干个独立（independent）的数的最大值不超过 x，当且仅当其中每一个数都不超过 x。这样一来，棘手的“最大值”就化为若干个相同概率的乘积，一行便可写出 X 的分布函数（distribution function）。其余一切都由此而来：求导可得密度（density）和期望值（expectation），而令分布函数等于 0.05，便可求出检验的临界值（critical value）。

解答

第 (i) 部分

设 U_1, U_2, U_3 为该三个数，它们互相独立，且各自在 $[0,1]$ 上服从均匀分布（uniform distribution），所以当 $0 \le x \le 1$ 时，$\mathrm{P}(U_i \le x) = x$。由于 $X = \max(U_1, U_2, U_3)$，

$$X \le x \iff U_1 \le x,\ U_2 \le x \text{ 及 } U_3 \le x.$$

由独立性，

$$F(x) = \mathrm{P}(X \le x) = x^3 \qquad (0 \le x \le 1).$$

特别地，$\mathrm{P}(X \le 0.8) = 0.8^3 = 0.512$，正是所求。

求导后，X 在 $[0,1]$ 上的密度为 $f(x) = F'(x) = 3x^2$，所以

$$\mathrm{E}(X) = \int_0^1 x \cdot 3x^2 \,\mathrm{d}x = \left[\tfrac{3}{4}x^4\right]_0^1 = \frac{3}{4}.$$

这个结果合情合理：三个数中最大的一个，理应远高于单独一个数的平均值 $\frac{1}{2}$。

第 (ii) 部分

现在有 N 个独立的数，各自在 $[0,a]$ 上服从均匀分布，所以当 $0 \le x \le a$ 时，$\mathrm{P}(U_i \le x) = x/a$，而同样的论证给出

$$\mathrm{P}(X \le x) = \left(\frac{x}{a}\right)^N \qquad (0 \le x \le a),$$

而当 $x \ge a$ 时，$\mathrm{P}(X \le x) = 1$。由于 X 是连续的，$\mathrm{P}(X < x)$ 与此相同。

临界值。 当 $x < c$ 时，检验拒绝 H_0；而 5% 的显著性水平（significance level）表示，在 H_0 为真（$a = 1$）时拒绝它的概率为 0.05：

$$\mathrm{P}(X < c \mid a = 1) = c^N = 0.05, \qquad \text{所以} \qquad c = 0.05^{1/N}.$$

x 的值偏小，是支持 $a < 1$ 的证据，这就是拒绝域（rejection region）为 $x < c$ 的原因。

> ? **想一想** “$x \le 0.8$ 必然导致拒绝”怎样化为一个关于 N 的不等式?

每一个 $x \le 0.8$ 都被拒绝，当且仅当每一个这样的 x 都小于 c，即 $0.8 < c$。由于 $t \mapsto t^N$ 在 $t > 0$ 时是递增的，这等价于

$$0.8^N < c^N = 0.05.$$

所以我们要找满足 $0.8^N < 0.05$ 的最小整数 N。（留意当 N 增大时，c 会向 1 递增，所以 N 愈大，检验就愈倾向于拒绝。）

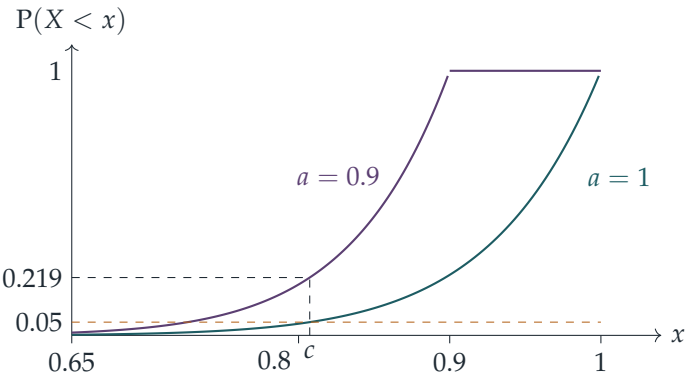
利用 $2^{10} \approx 10^3$。写成 $0.8 = 2^3/10$，于是 $0.8^N = 2^{3N}/10^N$。

- $N = 13$：$0.8^{13} = \dfrac{2^{39}}{10^{13}} = \dfrac{2^9 \cdot 2^{30}}{10^{13}} \approx \dfrac{512 \times 10^9}{10^{13}} = 0.0512 > 0.05$。
- $N = 14$：$0.8^{14} = \dfrac{2^{42}}{10^{14}} = \dfrac{2^2 \cdot 2^{40}}{10^{14}} \approx \dfrac{4 \times 10^{12}}{10^{14}} = 0.04 < 0.05$。

由于 0.8^N 随 N 增大而递减，每一个 $N \le 13$ 同样不符合要求。所以最小的值是

$$N = 14.$$

（这个近似会略为低估 2 的幂，因为 $2^{10} = 1024$；真实数值 $0.8^{13} \approx 0.0550$ 和 $0.8^{14} \approx 0.0440$ 会得出相同的结论。）



当 $N = 14$ 时，临界值为 $c = 0.05^{1/14} \approx 0.807$，仅仅高于 0.8。

若 $a = 0.8$。 所有的数都落在 $[0, 0.8]$ 内,所以必然有 $X \leq 0.8 < c$。零假设(null hypothesis)被拒绝的概率为 1。

若 $a = 0.9$。 此时 $c \approx 0.807 < 0.9$，所以

$$\mathrm{P}(\text{拒绝}) = \mathrm{P}(X < c \mid a = 0.9) = \left(\frac{c}{0.9}\right)^{14} = \frac{c^{14}}{0.9^{14}} = \frac{0.05}{0.9^{14}}.$$

这里 $0.9^2 = 0.81$，$0.9^4 = 0.6561$，$0.9^8 \approx 0.4305$，因此 $0.9^{14} = 0.9^8 \cdot 0.9^4 \cdot 0.9^2 \approx 0.2288$。于是

$$\mathrm{P}(\text{拒绝} \mid a = 0.9) = \frac{0.05}{0.9^{14}} \approx 0.219.$$

所以即使有 14 个观测值，这个检验通常也察觉不到 a 是 0.9 而不是 1。

验算

- 蒙地卡罗模拟（Monte Carlo simulation）（2×10^5 次试验）在第 (i) 部分得出 $\mathrm{P}(X \leq 0.8) \approx 0.512$ 和 $\mathrm{E}(X) \approx 0.750$。
- 一般而言，$[0, 1]$ 上 n 个均匀分布的数的最大值，其平均值为 $\frac{n}{n+1}$；当 $n = 1$ 时这是 $\frac{1}{2}$，理应如此，而当 $n = 3$ 时则是 $\frac{3}{4}$。
- 取 $N = 14$ 和 $c = 0.05^{1/14}$，模拟得出的拒绝率在 $a = 1$ 时约为 0.0505（即检验的大小），在 $a = 0.9$ 时约为 0.218，与 $0.05/0.9^{14} = 0.2186$ 吻合。

常见错误

- 以 $1 - 0.8^N$ 代替 0.8^N：最大值偏小的概率是乘积 0.8^N，而不是它的补数。

- 把 $a = 0.9$ 时的拒绝概率当作 $(0.8/0.9)^{14}$。检验是在 $x < c$ 时拒绝，而 c 并不是 0.8。
- 因为 0.0512“大约是 0.05”而停在 $N = 13$。我们需要的是严格小于 0.05。

STEP 之外

最大值是一个顺序统计量（*order statistic*），而对于 $[0, a]$ 上的均匀分布，它是 a 的充分统计量（*sufficient statistic*）：它包含了样本中关于 a 的全部信息。把拒绝概率看成真实 a 的函数，就是该检验的检定力函数（*power function*）。在这里，当 $a \leq 1$ 时它是 $\min\{1, 0.05/a^N\}$，而统计学家正是通过比较检定力函数来在不同的检验之间作出选择。Grimmett 与 Welsh 的 *Probability: An Introduction* 涵盖了最大值和最小值的分布；Blitzstein 与 Hwang 的 *Introduction to Probability* 则详细讨论了顺序统计量。

详细解答

STEP I 2004，第 14 题

概率与统计 • 排列的计数

STEP I 2004，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/04-s1-q14

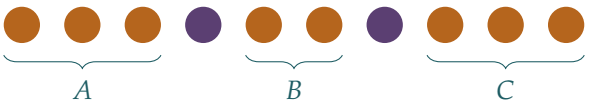
关键思路

想像把整个宝箱里的硬币逐枚取出，排成一行。海盗其实只是在两枚铅币的位置把这一行切开：第一个海盗得到第一枚铅币之前的金币，第二个海盗得到两枚铅币之间的金币，第三个海盗得到第二枚铅币之后的金币。硬币的每一种排列次序出现的可能性都相同，所以整个问题归结为数算两枚铅币可以落在哪些位置。

解答

设想全部 $n+2$ 枚硬币按随机次序逐一抽出，尽管第三个海盗实际上只是把剩下的一手捞走。这并不改变谁得到哪些硬币。抽取是随机的，所以从 $n+2$ 个位置中为铅币选出两个位置的 $\binom{n+2}{2}$ 种选法，每一种的可能性都相同。

设 A、B 和 C 分别为第一、第二和第三个海盗取得的金币数目。那么 $A+B+C=n$，而这一行硬币看起来是这样的：



图中琥珀色的是金币，紫红色的是铅币。铅币的位置决定了 (A,B,C)，而每一个满足 $A+B+C=n$ 的非负整数三元组，都恰好来自一种位置选法。所以这 $\binom{n+2}{2}=\frac{1}{2}(n+1)(n+2)$ 个三元组 (A,B,C) 的可能性全部相同。

第 (i) 部分

第一个海盗得不到金币，当且仅当 $A=0$，即抽出的第一枚硬币就是铅币。由于 $n+2$ 枚硬币中每一枚最先被抽出的可能性都相同，这个事件的概率（probability）是 $\frac{2}{n+2}$。因此

$$\mathrm{P}(A\geq 1)=1-\frac{2}{n+2}=\frac{n}{n+2}.$$

（用数算的方法也得到相同答案：$A=0$ 时余下 $B+C=n$，它有 $n+1$ 个解，而 $(n+1)\big/\frac{1}{2}(n+1)(n+2)=\frac{2}{n+2}$。）

第 (ii) 部分

> **想一想** 第二个海盗是在一个随机的时刻才开始抽的。她的处境真的不比第一个海盗差吗?

是的，她的处境并不较差。我们数算 $B=0$ 的三元组：这时 $A+C=n$，它有 $n+1$ 个解，与第 (i) 部分完全一样。所以

$$\mathrm{P}(B\geq 1)=1-\frac{n+1}{\frac{1}{2}(n+1)(n+2)}=\frac{n}{n+2}.$$

事实上，由于所有三元组 (A,B,C) 的可能性都相同，三人所得的份额是*可交换的*（exchangeable）：每个海盗得到金币的机会都同样是 $\frac{n}{n+2}$，连第三个海盗也不例外。

第 (iii) 部分

我们需要 A、B 和 C 全部至少为 1。设 $A=A'+1$、$B=B'+1$ 和 $C=C'+1$。于是我们想要的三元组就对应于 $A'+B'+C'=n-3$ 的非负整数解。根据隔板法（stars and bars），这样的解共有 $\binom{n-1}{2}$ 个，这一点也可以直接看出：在相邻金币之间的 $n-1$ 个空隙中，选出两个来放铅币。因此，对于 $n\geq 1$，

$$\mathrm{P}(A\geq 1,\ B\geq 1,\ C\geq 1)=\frac{\binom{n-1}{2}}{\binom{n+2}{2}}=\frac{(n-1)(n-2)}{(n+1)(n+2)}.$$

当 $n=1$ 和 $n=2$ 时，这个概率是 0，这是必然的，因为金币不足三枚时，三个海盗不可能人人都有金币。（当 $n=0$ 时，概率显然是 0；公式在这里并不适用，因为 $\binom{n-1}{2}$ 所数算的空隙根本不存在。）

验算

- 当 $n=1$ 时，硬币的排列有 GLL、LGL 和 LLG 三种，可能性全部相同。第一个海盗只在 GLL 中得到金币，其概率为 $\frac{1}{3}=\frac{1}{1+2}$；第二个海盗只在 LGL 中得到金币，概率同样是 $\frac{1}{3}$。

- 当 $n = 3$ 时，唯一合乎要求的排列是 GLGLG，它是 $\binom{5}{2} = 10$ 种排列中的一种，而公式给出 $\frac{2\cdot 1}{4\cdot 5} = \frac{1}{10}$。
- 对 $n = 1, \ldots, 14$ 逐一穷举，结果与三条公式全部吻合。以 $n = 5$ 模拟 2×10^5 个宝箱，得出 0.715、0.714 和 0.286，而理论值分别为 $\frac{5}{7} \approx 0.714$、$\frac{5}{7}$ 和 $\frac{12}{42} = \frac{2}{7} \approx 0.286$。

常见错误

- 把第二个海盗的抽取当作从一个装满的宝箱重新开始。留给她的金币有多少，取决于第一个海盗抽得怎样。
- 在第 (iii) 部分把 $\mathrm{P}(A \geq 1)$、$\mathrm{P}(B \geq 1)$ 和 $\mathrm{P}(C \geq 1)$ 相乘。各人的份额并非独立（independent），因为它们的总和必须是 n。

STEP 之外

铅币位置与 $A + B + C = n$ 的解之间的对应，就是“隔板法”的双射（bijection）。它说明了随机放置的标记之间的间隔是可交换的。它的连续版本，是把点均匀地投在一根棒上，由此引出均匀间距（uniform spacings）理论和狄利克雷分布（Dirichlet distribution），而这类次序统计量（order statistics）在大学一年级的概率论中随处可见。Blitzstein 与 Hwang 合著的 *Introduction to Probability*（概率论入门）对这种数算方法以及相关的对称性论证，都有十分出色的讲解。

详细解答

STEP II 2004，第 1 题

纯数学 · 含根式的方程

STEP II 2004，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/04-s2-q1

关键思路

每一条方程都含有 $\sqrt{x}$ 和 $\sqrt{3x^2+1}$，所以设 $t=\sqrt{x}$（其中 $t\geq 0$），把前者化为多项式（polynomial）。然后把余下的平方根单独移到一边，平方一次。这样会得到一条关于 t 的四次方程（quartic），而且很容易因式分解（factorise），因为 $t=0$ 和 $t=1$ 必定是根。真正要花工夫的是最后一步：平方可能引入增根（false roots），所以我们只保留那些令被平方的一边不是负数的根。

解答

三条方程都需要 $x\geq 0$，$\sqrt{x}$ 才是实数，所以我们写 $x=t^2$，其中 $t\geq 0$。于是 $\sqrt{x}=t$，而 $3x^2+1=3t^4+1$。

平方这一步只在一个方向上成立。对于实数 A 和 B，

$$\sqrt{A}=B \iff A=B^2 \text{ 且} B\geq 0.$$

所以平方之后，我们必须检查右边在每一个根处是否非负。

第 (i) 部分

把根式单独移到一边，得

$$\sqrt{3t^4+1}=2t^2-t+1.$$

右边恒为正数：它的判别式（discriminant）是 $1-8<0$，而首项系数（leading coefficient）是正数。所以在这里平方不会有任何损失，平方后所得方程的每一个根都是真正的解。平

方得

$$3t^4+1=(2t^2-t+1)^2=4t^4-4t^3+5t^2-2t+1,$$
$$0=t^4-4t^3+5t^2-2t=t\,(t^3-4t^2+5t-2).$$

这个三次式在 $t=1$ 时等于零（因为 $1-4+5-2=0$），除去因式 $t-1$ 后得

$$t^3-4t^2+5t-2=(t-1)(t^2-3t+2)=(t-1)^2(t-2).$$

所以 $t\,(t-1)^2(t-2)=0$，得 $t=0,1,2$。三个值都是非负的，所以

$$x=0,\quad x=1,\quad x=4.$$

验算一下：当 $x=4$ 时，我们得到 $\sqrt{49}+2-8-1=0$。

第 (ii) 部分

现在我们有

$$\sqrt{3t^4+1}=-t^2+2t+1.$$

平方得

$$3t^4+1=(-t^2+2t+1)^2=t^4-4t^3+2t^2+4t+1,$$
$$0=2t^4+4t^3-2t^2-4t=2t\,(t^3+2t^2-t-2).$$

把三次式的各项分组为 $t^2(t+2)-(t+2)$，得

$$2t\,(t+2)(t-1)(t+1)=0.$$

由于 $t\geq 0$，只剩下 $t=0$ 和 $t=1$。右边 $-t^2+2t+1$ 在这两处的值分别是 1 和 2，都是正数，所以两个根都是真正的解：

$$x=0,\quad x=1.$$

第 (iii) 部分

这一次

$$\sqrt{3t^4+1}=t^2+2t-1.$$

平方得

$$3t^4+1=(t^2+2t-1)^2=t^4+4t^3+2t^2-4t+1,$$
$$0=2t^4-4t^3-2t^2+4t=2t\,(t^3-2t^2-t+2).$$

把三次式分组为 $t^2(t-2)-(t-2)$，得

$$2t\,(t-2)(t-1)(t+1)=0.$$

非负的根是 $t=0,1,2$。

想一想 这三个值是否全都满足原来的方程?

不是。当 $t=0$ 时，右边 t^2+2t-1 等于 -1，但平方根不可能是负数。所以 $t=0$ 是增根，是由平方引入的。（直接验证：当 $x=0$ 时，(iii) 的左边是 $1-0-0+1=2\neq 0$。）当 $t=1$ 和 $t=2$ 时，右边分别是 2 和 7，都是正数，所以这两个根是真正的解。因此

$$x=1,\quad x=4.$$

(ii) 和 (iii) 之间有一个巧妙的联系。把 t 换成 $-t$，$-t^2+2t+1$ 就变成 $-(t^2+2t-1)$，而平方会消去负号。所以 (iii) 中平方后的方程，就是把 (ii) 中平方后的方程的 t 换成 $-t$ 所得的方程，它的根 $0,\pm1,2$ 正是 (ii) 的根 $0,\pm1,-2$ 的相反数。

验算

把每个答案代入原来的方程。例如，在 (iii) 中代入 $x=4$，得 $\sqrt{49}-4-4+1=0$。在 $0\leq x\leq 50$ 的范围内对每条方程的左边作细密的数值扫描，没有发现其他零点。在 (i) 中，左边在 $x=1$ 处触及零但不变号，这与二重因式 $(t-1)^2$ 吻合。

常见错误

- 忘记检查被平方那一边的正负号，因而在 (iii) 中保留了 $x=0$。
- 除以 t（或除以 $\sqrt{x}$），因而在 (i) 和 (ii) 中漏掉了 $x=0$ 这个根。
- 一边仍有两个平方根时就平方。应先把 $\sqrt{3x^2+1}$ 单独移到一边，这样平方一次便能把它消去。

STEP 之外

映射（map）$A\mapsto A^2$ 在 $\mathbb{R}$ 上不是单射（injective），所以两边平方得到的只是蕴涵关系（implication），而不是等价关系（equivalence）。留意哪些步骤可逆，是每一个严谨解方程过程背后的习惯；到了分析学（analysis），这就变成“函数在甚么时候有反函数（inverse）、在哪个定义域（domain）上有反函数”的问题。Bartle 与 Sherbert 的 *Introduction to Real Analysis*（《实分析导论》）对此有细致的阐述。

详细解答

STEP II 2004，第 2 题

纯数学 · 二次不等式

STEP II 2004，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/04-s2-q2

关键思路

绝对值（modulus）是唯一棘手的地方，而只要设 $t = |x|$，它就会消失。由于 $x^2 = |x|^2$，原不等式便化为一个关于 t 的普通二次不等式（quadratic inequality），并附带条件 $t \geq 0$。接着只要配方（completing the square），或者考察判别式（discriminant），一切便一目了然。t 的每一个容许区间都对应 x 的两个区间，分别位于原点两侧。

解答

以下一律设 $t = |x| \geq 0$。由于 $x^2 = t^2$，不等式 $(*)$ 正好等价于

$$q(t) = t^2 - \alpha t + 2 < 0, \qquad t = |x| \geq 0.$$

当 $|\alpha| < 2\sqrt{2}$ 时无解

配方得

$$q(t) = \left(t - \frac{\alpha}{2}\right)^2 + 2 - \frac{\alpha^2}{4}.$$

若 $|\alpha| < 2\sqrt{2}$，则 $\alpha^2 < 8$，故 $2 - \frac{\alpha^2}{4} > 0$。平方项永不为负，所以对任何实数 t 都有 $q(t) > 0$，对 $t = |x|$ 自然也成立。因此对所有 x 均有 $x^2 - \alpha|x| + 2 > 0$，没有任何 x 的值满足 $(*)$。

$\alpha = 3$ 的情况

此时 $q(t) = t^2 - 3t + 2 = (t-1)(t-2)$，它恰好在 $1 < t < 2$ 时为负。所以 $(*)$ 成立当且仅当 $1 < |x| < 2$，解集为

$$\{x : -2 < x < -1\} \cup \{x : 1 < x < 2\}.$$

长度之和 S

设 $\alpha > 2\sqrt{2}$。则 $\alpha^2 > 8$，故 q 有两个相异实根

$$t_1 = \frac{\alpha - \sqrt{\alpha^2 - 8}}{2}, \qquad t_2 = \frac{\alpha + \sqrt{\alpha^2 - 8}}{2},$$

而且由于 q 的首项系数为正，$q(t) < 0$ 恰好在 $t_1 < t < t_2$ 时成立。

> **想一想** 区间 $t_1 < t < t_2$ 会不会有一部分落在 t 为负的范围？我们需要担心这一点吗？

不需要。两根之积为 $t_1 t_2 = 2 > 0$，两根之和为 $t_1 + t_2 = \alpha > 0$，所以两根皆为正数。因此整个区间 (t_1, t_2) 都位于 $t > 0$ 之内，其中每一个 t 值都确实是 $|x|$ 可取的值。

所以 $(*)$ 成立当且仅当 $t_1 < |x| < t_2$，即

$$-t_2 < x < -t_1 \qquad \text{或} \qquad t_1 < x < t_2.$$

这是两个互不相交的区间（因为 $t_1 > 0$），长度各为 $t_2 - t_1 = \sqrt{\alpha^2 - 8}$。因此

$$S = 2\sqrt{\alpha^2 - 8}.$$

由于 $\alpha^2 - 8 < \alpha^2$ 且 $\alpha > 0$，取正平方根得 $\sqrt{\alpha^2 - 8} < \alpha$，所以

$$S < 2\alpha,$$

这正是所要证明的。

S 对 α 的图像

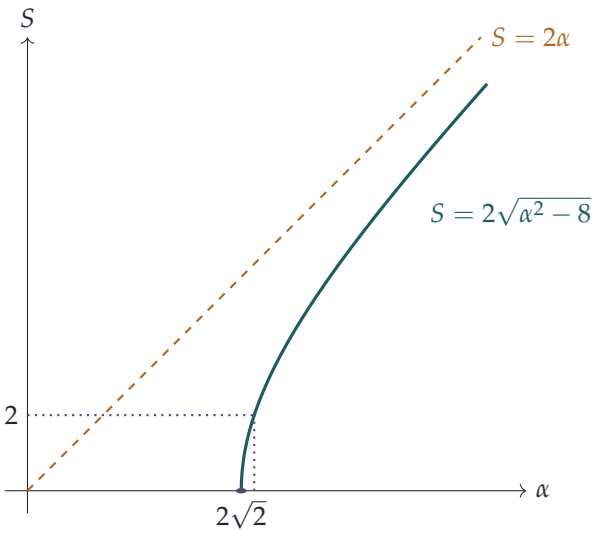
函数 $S = 2\sqrt{\alpha^2 - 8}$ 的定义域为 $\alpha > 2\sqrt{2}$（当 $\alpha \le 2\sqrt{2}$ 时根本没有任何区间：在 $\alpha = 2\sqrt{2}$ 时有 $q(t) = (t - \sqrt{2})^2 \ge 0$，而只要 $\alpha \le 0$，对所有 $t \ge 0$ 都有 $q(t) \ge 2$）。它的主要特征如下：

- 当 $\alpha \to 2\sqrt{2}^{+}$ 时 $S \to 0$，而 $\dfrac{\mathrm{d}S}{\mathrm{d}\alpha} = \dfrac{2\alpha}{\sqrt{\alpha^2 - 8}}$ 在该处趋向无穷大：曲线由 $(2\sqrt{2}, 0)$ 出发，该点的切线是铅垂的。
- S 是递增的，而且 $\dfrac{\mathrm{d}^2 S}{\mathrm{d}\alpha^2} = -\dfrac{16}{(\alpha^2 - 8)^{3/2}} < 0$，所以曲线是凹的（concave）。

- 直线 $S = 2\alpha$ 是渐近线（asymptote），曲线从下方趋近它：由于 $(2\alpha - S)(2\alpha + S) = 4\alpha^2 - 4(\alpha^2 - 8) = 32$,

$$2\alpha - S = \frac{32}{2\alpha + S} \to 0 \quad \text{当} \alpha \to \infty.$$

事实上 $\frac{\alpha^2}{8} - \frac{S^2}{32} = 1$，所以这个图像是双曲线（hyperbola）的四分之一，渐近线为 $S = 2\alpha$。



验算

- 当 $\alpha = 3$ 时，公式给出 $S = 2\sqrt{1} = 2$，与 $(-2, -1)$ 和 $(1, 2)$ 这两个长度各为 1 的区间吻合（即图中以虚点线标示、位于 $\alpha = 3$ 的那一点）。
- 当 $\alpha = 5$ 时，$S = 2\sqrt{17} \approx 8.246$；以数值方法量度 $[-20, 20]$ 内满足 $x^2 - 5|x| + 2 < 0$ 的 x 所组成的集合，所得的值相同，而且明显小于 $2\alpha = 10$。
- 当 α 很大时，$2\alpha - S \approx 8/\alpha$：在 $\alpha = 100$ 时，两者相差 $0.0800\ldots$。

常见错误

- 忘记 $|x|$ 的每一个区间对应 x 的两个区间，因而得出 $S = \sqrt{\alpha^2 - 8}$。
- 把 $\alpha = 3$ 的答案只写成 $1 < x < 2$，漏掉了负的那一半。
- 使用判别式时，没有指出关于 t 的两个根皆为正数；当 $\alpha < 0$ 时，同样的判别式条件却完全没有解，因为两个根都是负数。

STEP 之外

函数 $x \mapsto x^2 - \alpha|x| + 2$ 是偶函数（even function），所以它的解集是对称的，而长度 S 就是该函数取负值的集合的测度（measure）。量度由不等式所定义的集合的大小，正是测度论（measure theory）和勒贝格积分（Lebesgue integration）的起点，大学二年级的分析课程会学到这些内容。若想细读本题用到的极限、导数和不等式的严谨论述，可参阅 Abbott 的 *Understanding Analysis*（一本实分析入门书）。

详细解答

STEP II 2004，第 3 题

纯数学 · 曲线描绘

> **STEP II 2004，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/04-s2-q3

关键思路

一切都源自同一条曲线。对于 C，先把导数（derivative）彻底因式分解：重复因式 $(x-2)$ 在求导之后仍然保留，余下的只是一个二次式。其余两条曲线都是由 C 建构出来的，毋须从头做起：在 (i) 中取 $y=\pm\sqrt{f(x)}$，在 (ii) 中则把 x^2 代入 f。要留意 f 在哪里取正值、在哪里有单根（simple root）、在哪里有重根（repeated root）；曲线的形状正是由这些细节决定的。

解答

设 $f(x)=x(x+1)(x-2)^4$，则 C 就是 $y=f(x)$。

C 的驻点

利用乘积法则（product rule），并抽出公因式 $(x-2)^3$，

$$\begin{aligned} f'(x) &= (2x+1)(x-2)^4 + 4x(x+1)(x-2)^3 \\ &= (x-2)^3\left[(2x+1)(x-2)+4x^2+4x\right] \\ &= (x-2)^3(6x^2+x-2) \\ &= (x-2)^3(3x+2)(2x-1). \end{aligned}$$

所以驻点（stationary point）位于 $x=-\frac{2}{3}$、$x=\frac{1}{2}$ 和 $x=2$，而

$$f\left(-\tfrac{2}{3}\right)=\left(-\tfrac{2}{3}\right)\left(\tfrac{1}{3}\right)\left(-\tfrac{8}{3}\right)^4=-\frac{8192}{729},\qquad f\left(\tfrac{1}{2}\right)=\tfrac{1}{2}\cdot\tfrac{3}{2}\cdot\left(-\tfrac{3}{2}\right)^4=\frac{243}{64},\qquad f(2)=0.$$

要判别这些驻点的类型，可考察 f' 各因式的正负号：

	$x<-\frac{2}{3}$	$-\frac{2}{3}<x<\frac{1}{2}$	$\frac{1}{2}<x<2$	$x>2$
$(x-2)^3$	$-$	$-$	$-$	$+$
$(3x+2)(2x-1)$	$+$	$-$	$+$	$+$
$f'(x)$	$-$	$+$	$-$	$+$

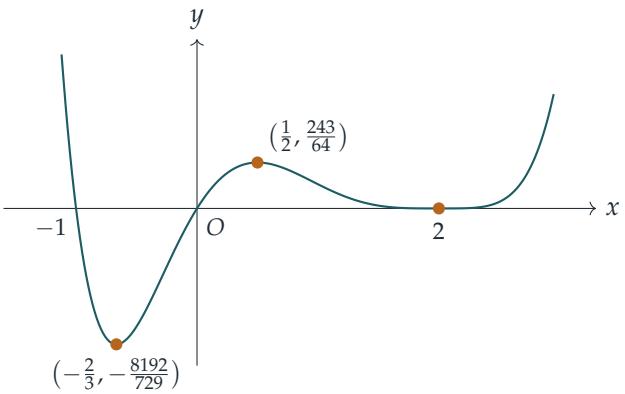
因此

- $\left(-\frac{2}{3},-\frac{8192}{729}\right)$ 是极小点（minimum）（f' 由 $-$ 变为 $+$）；
- $\left(\frac{1}{2},\frac{243}{64}\right)$ 是极大点（maximum）（f' 由 $+$ 变为 $-$）；
- $(2,0)$ 是极小点（f' 由 $-$ 变为 $+$）。

最后一个即使不用微积分也显而易见：在 $x=2$ 附近 $f(x)\geq 0$，因为在那里 $x(x+1)>0$，而 $(x-2)^4\geq 0$。

描绘 C

曲线在 $x=-1$ 和 $x=0$ 处穿过 x 轴（两者都是单根，斜率分别为 $f'(-1)=-81$ 和 $f'(0)=16$），并在 $x=2$ 处与该轴相切。在 $x=2$ 附近，$f(x)\approx 2\cdot 3\,(x-2)^4=6(x-2)^4$，所以曲线在那里非常平坦。由于 f 是一个 6 次多项式，而且首项系数为正，$y\to+\infty$ 当 $x\to\pm\infty$。草图中的纵轴比例经过压缩。



第 (i) 部分

现在 $y^2=f(x)$，所以 $y=\pm\sqrt{f(x)}$。以下三点足以确定曲线的形状。

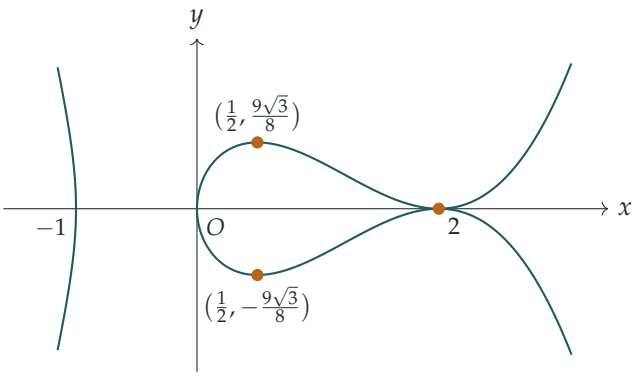
- *曲线存在的范围*。我们需要 $f(x)\geq 0$，由 C 可知这表示 $x\leq -1$ 或 $x\geq 0$。在 $-1<x<0$ 的范围内没有曲线。
- *对称性*。把 y 换成 $-y$，方程不变，所以曲线关于 x 轴对称。只需描绘上半部分 $y=\sqrt{f(x)}=(x-2)^2\sqrt{x(x+1)}$，再作反射即可。

- 驻点。对于 $y=\sqrt{f}$，当 $f>0$ 时有 $\mathrm{d}y/\mathrm{d}x = f'/(2\sqrt{f})$，所以上半部分的驻点恰好出现在 $f'=0$ 且 $f>0$ 之处。这样只剩下 $x=\frac{1}{2}$（$x=-\frac{2}{3}$ 这一点在定义域之外），于是在 $\left(\frac{1}{2}, \sqrt{243/64}\right) = \left(\frac{1}{2}, \frac{9\sqrt{3}}{8}\right)$ 有一个极大点，而在下半部分的 $\left(\frac{1}{2}, -\frac{9\sqrt{3}}{8}\right)$ 有一个极小点。

想一想 在 $f=0$ 的三点 $(-1,0)$、$(0,0)$ 和 $(2,0)$，曲线会怎样?

在单根处，例如 $x=0$，$f(x)\approx 16x$，所以在原点附近 $y^2\approx 16x$：这是一条横卧的抛物线，曲线穿过横轴时的切线是铅垂的。在 $x=-1$ 处情况相同，那里 $y^2\approx 81(-1-x)$。在重根 $x=2$ 处，$y^2\approx 6(x-2)^4$，所以 $y\approx \pm\sqrt{6}\,(x-2)^2$：这是两段形如抛物线的弧，一段在横轴之上，一段在横轴之下，两者都在 $(2,0)$ 与 x 轴相切。因此上下两半在该处相遇，并有共同的水平切线；在 $(2,0)$ 处，上半部分取得极小值，下半部分取得极大值。

当 $|x|$ 很大时，$y\approx\pm|x|^3$。所以曲线包括一个由 $(0,0)$ 到 $(2,0)$ 的闭合环，它在 $(2,0)$ 处收窄成一点，然后延伸为两条向右张开的分支；另外还有独立的一段，经过 $(-1,0)$ 并向左张开。



第 (ii) 部分

这里 $y=x^2(x^2+1)(x^2-2)^4=f(x^2)$。以下两点观察已解决了大部分问题。

- 曲线关于 y 轴对称，因为 x 只以 x^2 的形式出现。
- 当 x 走遍 $[0,\infty)$ 时，$u=x^2$ 亦递增地走遍 $[0,\infty)$，所以对于 $x\geq 0$，这条曲线就是 C 中 $x\geq 0$ 的部分，只是重新参数化而已。特别是处处都有 $y\geq 0$，而 C 中对应负 x 的一段（包括其极小点）完全不起作用。

由链式法则（chain rule），

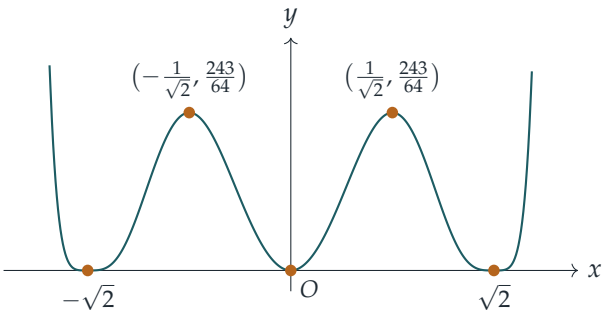
$$\frac{\mathrm{d}y}{\mathrm{d}x} = 2x\,f'(x^2) = 2x\,(x^2-2)^3(3x^2+2)(2x^2-1),$$

当 $x=0$、$x^2=\frac{1}{2}$ 或 $x^2=2$ 时，上式等于零（因式 $3x^2+2$ 永不为零）。在 $u\geq 0$ 上，f

由 0 上升至 $u=\frac{1}{2}$ 处的极大值，再下降至 $u=2$ 处的 0，之后一直上升；而对于 $x\geq 0$，$u=x^2$ 保持这个次序。所以驻点为

- 位于 $(0,0)$ 的极小点，该处 $y\approx 16x^2$；
- 位于 $\left(\pm\frac{1}{\sqrt{2}},\frac{243}{64}\right)$ 的极大点；
- 位于 $(\pm\sqrt{2},0)$ 的极小点，曲线在该处非常平坦地与横轴相切：$y\approx 384\,(x\mp\sqrt{2})^4$。

当 $|x|$ 很大时，$y\approx x^{12}$，上升得极为陡峭。



验算

先对导数作一个简单的合理性检查：在 $x=0$ 处，$f(x)=x(x+1)(x-2)^4\approx 16x$，而确实有 $f'(0)=(-8)(2)(-1)=16$。数值验算方面，在随机选取的点上把 f' 的因式分解与有限差分（finite difference）导数比较，结果吻合；又得出 $f(-\frac{2}{3})\approx -11.237$ 和 $f(\frac{1}{2})=3.796875$，并求得 $\sqrt{f}$ 在 $[0,2]$ 上的最大值为 $1.94856\ldots=\frac{9\sqrt{3}}{8}$，在 $x=0.5$ 处取得。至于 (ii)，$\mathrm{d}y/\mathrm{d}x$ 在 $x\approx\pm 0.707$ 和 ± 1.414（以及 0）处变号，与预期一致。

常见错误

- 先把 f 展开才求导。这样 f' 会是一个看不出任何因式的五次式；把 $(x-2)^3$ 留在括号外，就只需解一个二次方程。
- 在 (i) 中，于 $(0,0)$ 和 $(-1,0)$ 画出尖点（cusp）或折角。在 f 的单根处，$y^2=f(x)$ 的切线是铅垂的。
- 在 (ii) 中，照搬 C 在 $x=-\frac{2}{3}$ 处的极小点。由于 $x^2\geq 0$，只会用到 C 中 $x\geq 0$ 的部分，而 y 永不为负。

STEP 之外

$y^2 = f(x)$ 在 f 的根附近的表现，正是研究代数曲线（algebraic curve）奇点（singular point）的起点：单根给出一个带铅垂切线的光滑点，而重根则给出两条分支相遇之处，如 $(2,0)$ 的情况。当 f 是三次式时，曲线 $y^2 = f(x)$ 称为*椭圆曲线*（*elliptic curves*），是现代数论的核心对象。若想仔细了解曲线描绘背后的微积分（单调性（monotonicity）、导数的正负号和链式法则），可参阅 Abbott 的 *Understanding Analysis*。

详细解答

STEP II 2004，第 4 题

纯数学 · 几何中的最优化

STEP II 2004，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：

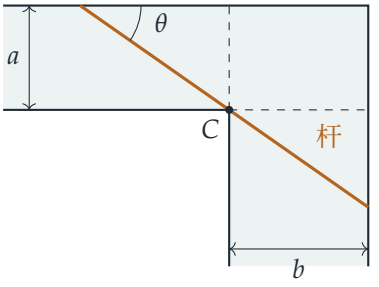
step.maths.org/questions/04-s2-q4

关键思路

无论怎样搬，杆（或桌子）都要由指向一条走廊的方向，转到指向另一条走廊的方向，所以在某一刻，它与墙壁所成的角必会取遍两者之间的每一个角 θ。对固定的角度而言，空间最宽裕的位置就是刚好擦过内角的那条直线，而这段长度用直角三角形（right-angled triangle）很容易算出。因此，物件能够转过弯角，当且仅当它的长度不超过这些长度在所有角度之中的最小者，而微积分（calculus）正好可以求出这个最小值（minimum）。

解答

以下一律把外角放在原点（origin），外墙分别沿 $y = 0$（阔度为 a 的走廊，向左延伸）和 $x = 0$（阔度为 b 的走廊，向下延伸）。于是内角位于 $C = (-b, -a)$。



第 (i) 部分

假设在某一刻，杆与 a 走廊的墙壁成锐角（acute angle）θ，并如图所示由外墙 $y = 0$ 向下斜向外墙 $x = 0$。

*角度为 θ 时最长的线段。*取通过 C 而倾角为 θ 的直线。由 C 向上到墙 $y = 0$，垂直升高了 a，所以这一段的长度是 $a\operatorname{cosec}\theta$；由 C 横向到墙 $x = 0$，水平移动了 b，所以这一段的长度是 $b\sec\theta$。因此，这条直线夹在两道外墙之间的弦（chord）的长度为

$$f(\theta) = a\operatorname{cosec}\theta + b\sec\theta.$$

任何其他倾角为 θ 而由一条走廊伸入另一条走廊的直线，都必须在 C 的外侧经过（否则它会穿过内墙）。这样的直线与两道外墙围出一个较小的相似（similar）直角三角形，所以它在外墙之间的弦比 $f(\theta)$ 短。另外，假如此刻杆完全位于 a 走廊之内，它就容纳在一条阔度为 a 的长带里，所以 $L\sin\theta \le a$，由此得 $L \le a\operatorname{cosec}\theta < f(\theta)$；同理，假如它完全位于 b 走廊之内，则 $L \le b\sec\theta < f(\theta)$。无论哪一种情况，都有

$$L \le f(\theta) \quad \text{对杆所经过的每一个角 } \theta \text{ 均成立。}$$

当杆由沿着第一条走廊转到沿着第二条走廊时，θ 由接近 0 连续地变到接近 $\frac{1}{2}\pi$，所以它会取遍其间的每一个值。因此 $L \le f(\theta)$ 对*每一个* $\theta \in (0, \frac{1}{2}\pi)$ 都成立，也就是说，L 不超过 f 的最小值。

*求 f 的最小值。*求导（differentiate）得

$$f'(\theta) = -\frac{a\cos\theta}{\sin^2\theta} + \frac{b\sin\theta}{\cos^2\theta} = \frac{b\sin^3\theta - a\cos^3\theta}{\sin^2\theta\cos^2\theta}.$$

这个式子等于零，当且仅当 $\tan^3\theta = a/b$；由于 tan 在 $(0, \frac{1}{2}\pi)$ 上递增，这个方程在该区间内只有一个解 α。f' 的正负号与 $b\tan^3\theta - a$ 相同，在 α 之前为负，之后为正，所以 α 给出最小值。（此外，在区间两端均有 $f \to \infty$。）在 $L \le f(\theta)$ 中取 $\theta = \alpha$，得

$$L \le a\operatorname{cosec}\alpha + b\sec\alpha, \qquad \tan^3\alpha = \frac{a}{b},$$

这正是所要证明的。

第 (ii) 部分

现在由桌子的长边担当杆的角色。桌子的内侧长边位于一条倾角为 θ 的直线上，外侧长边则位于与之平行、向外角方向再远 w 的直线上。桌子外侧的两个角都必须位于外墙之内。

*外侧长边可用的空间。*设一条倾角为 θ 的直线与外角 O 的距离为 d，它在 O 处截出一个直角三角形，其斜边（hypotenuse）为 c（即两墙之间的弦），两条直角边为 $c\cos\theta$ 和 $c\sin\theta$，由 O 所作的高为 d。这个三角形面积的两倍既等于 $c\cos\theta \cdot c\sin\theta$，又等于 $c\,d$，所以

$$c = \frac{d}{\sin\theta\cos\theta} = 2d\operatorname{cosec}2\theta.$$

可见弦长与直线到 O 的距离成正比。因此，把直线向内（即向 O 的方向）移动 w，它的弦便会缩短 $2w\operatorname{cosec}2\theta$。

想一想 为甚么外侧长边必须容纳在两道外墙之间，而且是在比通过 C 的直线再靠内 w 的那条直线上?

内侧长边不能越过内角，所以它位于通过 C 的直线上，或在该直线的外侧。于是外侧长边所在的直线至少再向外 w，其弦长至多为 $f(\theta) - 2w\operatorname{cosec} 2\theta$。（两条短边由外侧的两个角向内伸展，与长边垂直，并且留在走廊之内，所以不会带来额外的限制。）假如整张桌子都位于 a 走廊之内，则它横跨长带的高度满足 $l\sin\theta + w\cos\theta \le a$。由此得出的上界 $l \le (a - w\cos\theta)\operatorname{cosec}\theta$ 比下文的 $g(\theta)$ 小了 $(b - w\sin\theta)\sec\theta > 0$，这是因为 $w < b$。b 走廊的情况亦相同。所以，与第 (i) 部分完全一样，要成功搬过弯角，必须有 $l \le g(\theta)$ 对所有 $\theta \in (0, \frac{1}{2}\pi)$ 成立，其中

$$g(\theta) = a\operatorname{cosec}\theta + b\sec\theta - 2w\operatorname{cosec} 2\theta.$$

（桌子必须比两条走廊都窄，即 $w < a$ 且 $w < b$，否则它连沿走廊移动也办不到。）

求 g 的最小值。由于 $\dfrac{\mathrm{d}}{\mathrm{d}\theta}\operatorname{cosec} 2\theta = -\dfrac{2\cos 2\theta}{\sin^2 2\theta}$，

$$g'(\theta) = -\frac{a\cos\theta}{\sin^2\theta} + \frac{b\sin\theta}{\cos^2\theta} + \frac{4w\cos 2\theta}{\sin^2 2\theta}.$$

两边乘以 $\sin^2\theta\cos^2\theta = \frac{1}{4}\sin^2 2\theta > 0$。于是 $g'(\theta) = 0$ 当且仅当

$$-a\cos^3\theta + b\sin^3\theta + w\cos 2\theta = 0, \qquad \text{即} \qquad w\cos 2\theta = \cos^3\theta\,(a - b\tan^3\theta).$$

利用 $\cos 2\theta = \cos^2\theta\,(1 - \tan^2\theta)$ 改写，再除以 $\cos^2\theta$。假如 $\theta = \frac{1}{4}\pi$，方程便会变成 $0 = \frac{1}{2\sqrt{2}}(a - b)$，但 $a \ne b$，所以这不可能。因此在驻点（stationary point）处 $\tan^2\theta \ne 1$，我们可以除以 $1 - \tan^2\theta$，得到

$$w = \left(\frac{a - b\tan^3\theta}{1 - \tan^2\theta}\right)\cos\theta.$$

当 $\theta \to 0^+$ 时，$g(\theta) \approx (a - w)/\theta \to +\infty$，因为 $w < a$；同理，当 $\theta \to \frac{1}{2}\pi^-$ 时 $g \to +\infty$，因为 $w < b$。所以连续（continuous）函数 g 在区间内部的某一点取得最小值，而在该点 $g' = 0$。把这一点记为 β；它满足上面列出的方程。在 $l \le g(\theta)$ 中取 $\theta = \beta$，得

$$l \le a\operatorname{cosec}\beta + b\sec\beta - 2w\operatorname{cosec} 2\beta,$$

这正是所要证明的。（事实上，对方程的任何一个根 β 都有 $l \le \min g \le g(\beta)$，所以题目指的是哪一个根并无影响。）

验算

► 在第 (ii) 部分中取 $w = 0$，便得 $\tan^3\beta = a/b$ 以及第 (i) 部分的上界，与预期相符。

- 在第 (i) 部分中，$\sin\alpha = a^{1/3}/\sqrt{a^{2/3}+b^{2/3}}$ 而 $\cos\alpha = b^{1/3}/\sqrt{a^{2/3}+b^{2/3}}$，所以上界是 $\left(a^{2/3}+b^{2/3}\right)^{3/2}$。当 $a=1$，$b=8$ 时，这个值是 $5^{3/2}\approx 11.18$，而以数值方法搜索 f 的最小值，结果亦与此吻合。
- 我们随机选取了数百组 $a \neq b$ 和 $w < \min(a,b)$，以穷举法求出 g 的最小值，并确认取得最小值的角度满足关于 β 的方程。我们亦直接用坐标核对了各弦的长度。

常见错误

- 求了 f 的最大值而不是最小值。杆必须挨得过最差的角度，所以上界是可用长度的最小值。
- 以为受外墙限制的是桌子的内侧长边。其实被挤压的是外侧长边，它位于距离外角近 w 的那条直线上。
- 除以 $1-\tan^2\beta$ 时，没有留意到这一步需要 $a \neq b$。

STEP 之外

第 (i) 部分是经典的“梯子转弯”问题，其答案 $\left(a^{2/3}+b^{2/3}\right)^{3/2}$ 与星形线（astroid）有关：通过内角的各条直线扫出一条包络线（envelope）。第 (ii) 部分则是搬沙发问题（moving sofa problem）的一个小特例：在阔度为一个单位的直角弯角中，能够搬过去的图形之中，面积最大的是哪一个？这个问题由 Leo Moser 于 1966 年提出，在其后的大部分岁月里一直未能完全解决。如果想做更多这类几何与微积分交汇的题目，可以试试 Siklos 的 *Advanced Problems in Mathematics*。

详细解答

STEP II 2004，第 5 题

纯数学 · 积分方程

STEP II 2004，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/04-s2-q5

关键思路

未知函数藏在积分（integral）之内，看起来很吓人，但只要用复角公式（compound-angle formula）把 $\sin(x+t)$ 展开，t 就可以移到积分之外，而留在积分内的只是两个固定的数。因此 f 必定是 t 加上 $\sin t$ 与 $\cos t$ 的一个组合，整个问题于是归结为由两条线性方程求出两个常数。第一部分的两个积分，正是那条方程所需要的。

解答

两个积分

运用分部积分法（integration by parts），取 $u = x$ 及 $\mathrm{d}v = \sin x\,\mathrm{d}x$：

$$\int_0^{\pi} x\sin x\,\mathrm{d}x = \Big[-x\cos x\Big]_0^{\pi} + \int_0^{\pi}\cos x\,\mathrm{d}x = (-\pi\cos\pi - 0) + \Big[\sin x\Big]_0^{\pi} = \pi.$$

同样地，取 $u = x$ 及 $\mathrm{d}v = \cos x\,\mathrm{d}x$：

$$\int_0^{\pi} x\cos x\,\mathrm{d}x = \Big[x\sin x\Big]_0^{\pi} - \int_0^{\pi}\sin x\,\mathrm{d}x = 0 - \Big[-\cos x\Big]_0^{\pi} = -(1+1) = -2.$$

我们还需要在 $[0,\pi]$ 上的三个标准积分：

$$\int_0^{\pi}\sin^2 x\,\mathrm{d}x = \int_0^{\pi}\cos^2 x\,\mathrm{d}x = \frac{\pi}{2}, \qquad \int_0^{\pi}\sin x\cos x\,\mathrm{d}x = 0.$$

头两个可由 $\sin^2 x = \frac{1}{2}(1-\cos 2x)$ 及 $\cos^2 x = \frac{1}{2}(1+\cos 2x)$ 得出，因为 $\cos 2x$ 在一整个周期上的积分为零；第三个是 $\frac{1}{2}\int_0^{\pi}\sin 2x\,\mathrm{d}x = 0$，理由相同。

f 的形式

展开 $\sin(x+t)=\sin x\cos t+\cos x\sin t$。积分是对 x 进行的，所以 $\sin t$ 和 $\cos t$ 都是常数，可以移到积分之外：

$$\begin{aligned}\mathrm{f}(t)&=t+\int_0^{\pi}\mathrm{f}(x)\big(\cos x\sin t+\sin x\cos t\big)\,\mathrm{d}x\\&=t+\sin t\int_0^{\pi}\mathrm{f}(x)\cos x\,\mathrm{d}x+\cos t\int_0^{\pi}\mathrm{f}(x)\sin x\,\mathrm{d}x\\&=t+A\sin t+B\cos t,\end{aligned}$$

其中 A 和 B 就是题目给出的两个积分。由于它们与 t 无关，所以是真正的常数。这就是所要求的形式。

求 A 和 B

想一想 我们并不知道 f 是甚么，那又怎样求出 A 和 B 呢?

我们其实知道它的形状：$\mathrm{f}(x)=x+A\sin x+B\cos x$。把它代回 $(*)$，积分就变成可以用 A 和 B 表示并计算出来的式子，再比较等式两边，便得到这两个常数所满足的方程。

代入 $(*)$ 的右边，并利用上面 $\sin(x+t)$ 的展开式：

$$t+\sin t\int_0^{\pi}(x+A\sin x+B\cos x)\cos x\,\mathrm{d}x+\cos t\int_0^{\pi}(x+A\sin x+B\cos x)\sin x\,\mathrm{d}x.$$

利用先前求得的积分，

$$\begin{aligned}\int_0^{\pi}(x+A\sin x+B\cos x)\cos x\,\mathrm{d}x&=-2+A\cdot 0+B\cdot\frac{\pi}{2}=-2+\frac{\pi}{2}B,\\\int_0^{\pi}(x+A\sin x+B\cos x)\sin x\,\mathrm{d}x&=\pi+A\cdot\frac{\pi}{2}+B\cdot 0=\pi+\frac{\pi}{2}A.\end{aligned}$$

所以 $(*)$ 表示，对于每一个 t，

$$t+A\sin t+B\cos t=t+\Big(-2+\frac{\pi}{2}B\Big)\sin t+\Big(\pi+\frac{\pi}{2}A\Big)\cos t.$$

比较 $\sin t$ 和 $\cos t$ 的系数（这样做是合理的，因为此恒等式对所有 t 都成立：先代入 $t=\frac{\pi}{2}$，再代入 $t=0$，便可分别抽出每个系数），得

$$A=-2+\frac{\pi}{2}B,\qquad B=\pi+\frac{\pi}{2}A.$$

把第二式代入第一式，

$$A=-2+\frac{\pi^2}{2}+\frac{\pi^2}{4}A\quad\Longrightarrow\quad A\Big(1-\frac{\pi^2}{4}\Big)=\frac{\pi^2}{2}-2,\quad\text{即}\quad A\cdot\frac{4-\pi^2}{4}=-\frac{4-\pi^2}{2}.$$

由于 $\pi^2\neq 4$，我们可以除以 $4-\pi^2$，得 $A=-2$。于是 $B=\pi+\frac{\pi}{2}(-2)=0$。因此

$$A=-2,\qquad B=0,\qquad\text{所以}\qquad \mathrm{f}(t)=t-2\sin t.$$

验算

直接用 A 和 B 的定义来核对答案。取 $\mathrm{f}(x) = x - 2\sin x$,

$$\int_0^\pi \mathrm{f}(x)\cos x\,\mathrm{d}x = -2 - 2\cdot 0 = -2 = A, \qquad \int_0^\pi \mathrm{f}(x)\sin x\,\mathrm{d}x = \pi - 2\cdot\frac{\pi}{2} = 0 = B.$$

数值方面,用辛普森法则(Simpson's rule)可得 $\int_0^\pi x\sin x\,\mathrm{d}x \approx 3.14159$ 及 $\int_0^\pi x\cos x\,\mathrm{d}x \approx -2.00000$。我们又取 $\mathrm{f}(t) = t - 2\sin t$，在五十个随机的 t 值上计算 $(*)$ 的右边，每一次所得的结果与 $\mathrm{f}(t)$ 的差距都在 10^{-9} 之内。

常见错误

- 忘记 $\int_0^\pi x\cos x\,\mathrm{d}x$ 中的负号：边界项为零，而 $-\int_0^\pi \sin x\,\mathrm{d}x = -2$。
- 弄错哪一个积分与 $\sin t$ 相乘。$\sin(x+t)$ 中 $\cos x\sin t$ 的部分给出 $A\sin t$，其中 $A = \int \mathrm{f}(x)\cos x\,\mathrm{d}x$，而不是 $\int \mathrm{f}(x)\sin x\,\mathrm{d}x$。
- 除以 $1 - \frac{\pi^2}{4}$ 时，没有指出它不等于零。

STEP 之外

方程 $(*)$ 是一条*第二类弗雷德霍姆积分方程*（*Fredholm integral equation of the second kind*）。它的核（kernel）$\sin(x+t) = \sin x\cos t + \cos x\sin t$ 是有限个“x 的函数乘以 t 的函数”之和（称为“退化”（degenerate）核或“可分离”（separable）核），这正是问题能够化约为一个 2×2 线性方程组（linear system）的原因。条件 $\pi^2 \neq 4$ 所说的，就是这个方程组的行列式（determinant）不等于零；假如行列式为零，方程便可能无解，或有无限多个解，这就是奇异矩阵（singular matrix）在积分方程中的对应情况。至于这背后的线性代数，可参阅 Strang 的 *Introduction to Linear Algebra*。

详细解答

STEP II 2004，第 6 题

纯数学 · 向量分解

> **STEP II 2004，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/04-s2-q6

关键思路

把一个向量（vector）分拆成沿 $\mathbf{a}$ 方向的部分和垂直（perpendicular）于 $\mathbf{a}$ 的部分，其实就是投影（projection）：由于 $\mathbf{a}$ 是单位向量（unit vector），任何向量 $\mathbf{v}$ 沿它的分量（component）就是 $(\mathbf{v}\cdot\mathbf{a})\,\mathbf{a}$。第二部分是同一个想法，只是高了一个维度。向量 $\mathbf{a}$ 和 $\mathbf{q}$ 是 Π 内两个互相垂直的方向，而 $\mathbf{R}$ 则提供第三个与两者垂直的方向。因此，我们既可以逐一向每个方向投影，也可以先设未知倍数，再用点积（dot product）写出垂直的条件。我们采用第二种做法，并以第一种做法作验算。

解答

向量 p 与 q

由于 $\mathbf{p}$ 平行（parallel）于 $\mathbf{a}$，设 $\mathbf{p}=\lambda\mathbf{a}$。于是

$$\mathbf{q}=\mathbf{a}+\mathbf{b}-\lambda\mathbf{a}=(1-\lambda)\mathbf{a}+\mathbf{b},$$

它自然位于 Π 内，因为 $\mathbf{a}$ 和 $\mathbf{b}$ 都在其中。要使 $\mathbf{q}$ 垂直于 $\mathbf{a}$，我们需要

$$0=\mathbf{q}\cdot\mathbf{a}=(1-\lambda)\,|\mathbf{a}|^2+\mathbf{a}\cdot\mathbf{b}=(1-\lambda)+3,$$

这里用了 $|\mathbf{a}|=1$ 和 $\mathbf{a}\cdot\mathbf{b}=3$。所以 $\lambda=4$，并且

$$\boxed{\mathbf{p}=4\mathbf{a},\qquad \mathbf{q}=\mathbf{b}-3\mathbf{a}.}$$

确实有 $\mathbf{q}\cdot\mathbf{a}=3-3=0$，以及 $\mathbf{p}+\mathbf{q}=\mathbf{a}+\mathbf{b}$。（换一个说法，$\mathbf{p}$ 就是投影 $\big((\mathbf{a}+\mathbf{b})\cdot\mathbf{a}\big)\mathbf{a}=4\mathbf{a}$。）

向量 P、Q 与 R

首先列出几项稍后要用的事实。由 $|\mathbf{b}| = 5$,

$$|\mathbf{q}|^2 = |\mathbf{b}|^2 - 6\,\mathbf{a}\cdot\mathbf{b} + 9|\mathbf{a}|^2 = 25 - 18 + 9 = 16,$$

所以 $|\mathbf{q}| = 4$。特别地，$\mathbf{q} \neq \mathbf{0}$，故 $\mathbf{b}$ 不平行于 $\mathbf{a}$，而 $\mathbf{a}$ 和 $\mathbf{b}$ 张成（span）平面 Π。

设 $\mathbf{P} = \alpha\mathbf{a}$ 及 $\mathbf{Q} = \beta\mathbf{q} = \beta(\mathbf{b} - 3\mathbf{a})$。这样 $\mathbf{R}$ 就必定是

$$\mathbf{R} = \mathbf{a} + \mathbf{b} + \mathbf{c} - \alpha\mathbf{a} - \beta(\mathbf{b} - 3\mathbf{a}) = \mathbf{c} + u\,\mathbf{a} + v\,\mathbf{b}, \qquad \text{其中} u = 1 - \alpha + 3\beta,\ v = 1 - \beta.$$

想一想 怎样把“$\mathbf{R}$ 垂直于 Π”化成方程?

一个向量垂直于一个平面，当且仅当它垂直于该平面内两个不平行的向量。我们刚刚看到 $\mathbf{a}$ 和 $\mathbf{b}$ 正是这样的一对向量。所以条件就是 $\mathbf{R}\cdot\mathbf{a} = 0$ 和 $\mathbf{R}\cdot\mathbf{b} = 0$。利用 $\mathbf{a}\cdot\mathbf{c} = -2$、$\mathbf{b}\cdot\mathbf{c} = 2$、$\mathbf{a}\cdot\mathbf{b} = 3$、$|\mathbf{a}|^2 = 1$ 和 $|\mathbf{b}|^2 = 25$:

$$\mathbf{R}\cdot\mathbf{a} = -2 + u + 3v = 0,$$
$$\mathbf{R}\cdot\mathbf{b} = 2 + 3u + 25v = 0.$$

由第一式得 $u = 2 - 3v$。代入第二式得 $8 + 16v = 0$，所以

$$v = -\tfrac{1}{2}, \qquad u = \tfrac{7}{2}.$$

于是 $\beta = 1 - v = \tfrac{3}{2}$，而 $\alpha = 1 + 3\beta - u = 1 + \tfrac{9}{2} - \tfrac{7}{2} = 2$。因此

$$\boxed{\mathbf{P} = 2\mathbf{a}, \qquad \mathbf{Q} = \tfrac{3}{2}(\mathbf{b} - 3\mathbf{a}), \qquad \mathbf{R} = \mathbf{c} + \tfrac{7}{2}\mathbf{a} - \tfrac{1}{2}\mathbf{b}.}$$

把三者相加，$\mathbf{a}$ 的系数（coefficient）是 $2 - \tfrac{9}{2} + \tfrac{7}{2} = 1$，$\mathbf{b}$ 的系数是 $\tfrac{3}{2} - \tfrac{1}{2} = 1$，$\mathbf{c}$ 的系数是 1，所以 $\mathbf{P} + \mathbf{Q} + \mathbf{R} = \mathbf{a} + \mathbf{b} + \mathbf{c}$，符合要求。

最后，$\mathbf{R} \neq \mathbf{0}$：假如它是零向量，$\mathbf{c} = \tfrac{1}{2}\mathbf{b} - \tfrac{7}{2}\mathbf{a}$ 便会位于 Π 内，与 $\mathbf{c}$ 不平行于 Π 这个假设矛盾。正是这个假设，令 $\mathbf{R}$ 成为一个真正的法向量（normal vector）。

另一个角度：向互相垂直的轴投影

单位向量 $\mathbf{a}$、$\hat{\mathbf{q}} = \tfrac{1}{4}(\mathbf{b} - 3\mathbf{a})$ 和 $\mathbf{n}$（Π 的一个单位法向量）组成一个正交归一组（orthonormal set）。所以 $\mathbf{P}$ 和 $\mathbf{Q}$ 就是 $\mathbf{v} = \mathbf{a} + \mathbf{b} + \mathbf{c}$ 在前两个向量上的投影:

$$\mathbf{v}\cdot\mathbf{a} = 1 + 3 - 2 = 2, \qquad \text{所以} \mathbf{P} = 2\mathbf{a};$$
$$\mathbf{v}\cdot\hat{\mathbf{q}} = \tfrac{1}{4}\big(\mathbf{a}\cdot\mathbf{q} + \mathbf{b}\cdot\mathbf{q} + \mathbf{c}\cdot\mathbf{q}\big) = \tfrac{1}{4}(0 + 16 + 8) = 6, \qquad \text{所以} \mathbf{Q} = 6\hat{\mathbf{q}} = \tfrac{3}{2}(\mathbf{b} - 3\mathbf{a}),$$

这里用了 $\mathbf{b}\cdot\mathbf{q} = 25 - 9 = 16$ 和 $\mathbf{c}\cdot\mathbf{q} = 2 + 6 = 8$。余下的部分 $\mathbf{R}$，就是 $\mathbf{c}$ 沿 Π 法线方向的分量。这与上面的答案一致。

验算

- 直接验算：$\mathbf{R}\cdot\mathbf{a} = -2 + \frac{7}{2} - \frac{3}{2} = 0$，而 $\mathbf{R}\cdot\mathbf{b} = 2 + \frac{21}{2} - \frac{25}{2} = 0$。
- 一个具体例子：取 $\mathbf{a} = \mathbf{i}$、$\mathbf{b} = 3\mathbf{i} + 4\mathbf{j}$ 和 $\mathbf{c} = -2\mathbf{i} + 2\mathbf{j} + \mathbf{k}$，它们满足所有已知条件。这时 $\mathbf{q} = 4\mathbf{j}$、$\mathbf{P} = 2\mathbf{i}$、$\mathbf{Q} = 6\mathbf{j}$ 及 $\mathbf{R} = \mathbf{k}$，而它们的和是 $\mathbf{a} + \mathbf{b} + \mathbf{c} = 2\mathbf{i} + 6\mathbf{j} + \mathbf{k}$。
- 我们亦用 2000 组满足条件、方向随机的向量组合测试过这些公式。每一次都有 $\mathbf{q}\cdot\mathbf{a} = 0$、$|\mathbf{q}| = 4$，$\mathbf{R}$ 是 Π 的法向量，而且三个向量之和等于 $\mathbf{a} + \mathbf{b} + \mathbf{c}$，误差全部在 10^{-9} 以内。

常见错误

- 取 $\mathbf{p} = (\mathbf{a}\cdot\mathbf{b})\mathbf{a} = 3\mathbf{a}$，这只是 $\mathbf{b}$ 本身的投影。我们需要的是 $\mathbf{a} + \mathbf{b}$ 的投影，即 $4\mathbf{a}$。
- 只加上 $\mathbf{R}\cdot\mathbf{a} = 0$ 这一个条件。要垂直于一个平面，需要两个独立的条件。
- 假设 $\mathbf{P} = \mathbf{p}$ 及 $\mathbf{Q} = \mathbf{q}$。加入 $\mathbf{c}$ 之后，平面内的分量也会改变，因为 $\mathbf{c}$ 本身在 Π 内也有分量。

STEP 之外

这是格拉姆－施密特正交化过程（Gram–Schmidt process）的缩影。我们由 $\mathbf{a}$、$\mathbf{b}$、$\mathbf{c}$ 出发，构造出两两垂直的方向 $\mathbf{a}$、$\mathbf{b} - 3\mathbf{a}$ 和 $\mathbf{c} + \frac{7}{2}\mathbf{a} - \frac{1}{2}\mathbf{b}$，然后把一个向量沿这些方向分解。同样的步骤在任何维数都适用。它是正交投影（orthogonal projection）、最小二乘拟合（least-squares fitting）和矩阵的 QR 分解（QR factorisation）的基础。Strang 的 *Introduction to Linear Algebra*（线性代数入门教材）对这些课题都有详细阐述。

详细解答

STEP II 2004，第 7 题

纯数学 · 二分法

STEP II 2004，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/04-s2-q7

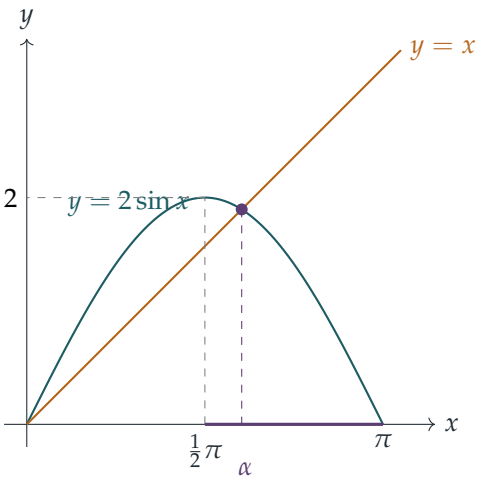
关键思路

方程 $2\sin x - x = 0$ 的根（root），就是曲线 $y = 2\sin x$ 与直线 $y = x$ 相交之处，所以图像部分只涉及一条正弦曲线和一条直线。至于二分法（bisection）的每一步，归根究底都只问同一个问题：在中点（midpoint）处，$2\sin c$ 比 c 大还是小？半角公式（half-angle formula）可把 $\frac{5}{8}\pi$ 和 $\frac{9}{16}\pi$ 的正弦值化为多重根式，而比较两个正数的平方，则令计算只需用到题目给出的近似值。

解答

I_0 内恰有一个根

我们有 $f(x) = 0 \iff 2\sin x = x$，所以各根就是 $y = 2\sin x$ 与 $y = x$ 的交点的 x 坐标。



在 $[\frac{1}{2}\pi, \pi]$ 上：

- 在 $x = \frac{1}{2}\pi$ 处，曲线的高度为 2，直线的高度为 $\frac{1}{2}\pi < 2$（因为 $\pi < 4$），所以曲线在直线之上；
- 在 $x = \pi$ 处，曲线的高度为 0，直线的高度为 $\pi > 0$，所以曲线在直线之下；
- 在这个区间（interval）上，曲线 $y = 2\sin x$ 递减，而直线 $y = x$ 递增。

两个图像都是连续（continuous）的，所以曲线必定在区间内某处穿过直线；又因为一个下降、一个上升，曲线一旦落到直线之下，就不可能再回到直线之上。因此两者恰好相交一次，即 $\mathrm{f}(x) = 0$ 在 I_0 内恰有一个根 α。用 f 来表达：$\mathrm{f}(\frac{1}{2}\pi) = 2 - \frac{1}{2}\pi > 0$，$\mathrm{f}(\pi) = -\pi < 0$，而且 f 在该区间上严格递减（strictly decreasing）。

在以下每一步二分法中，我们都保留 f 变号的那一半。由于 $\mathrm{f}(\frac{1}{2}\pi) > 0$，一切都取决于 $\mathrm{f}(c_n)$ 的正负：若 $\mathrm{f}(c_n) < 0$，根在 c_n 的左边；若 $\mathrm{f}(c_n) > 0$，根就在右边。

求 I_1

这里 $c_0 = \frac{3}{4}\pi$，而

$$\mathrm{f}(\tfrac{3}{4}\pi) = 2 \cdot \frac{1}{\sqrt{2}} - \tfrac{3}{4}\pi \approx 1.4 - \tfrac{3}{4}\sqrt{10}.$$

由于 $\sqrt{10} > 3$，我们有 $\frac{3}{4}\sqrt{10} > 2.25 > 1.4$，所以 $\mathrm{f}(c_0) < 0$。连同 $\mathrm{f}(a_0) > 0$，可得 $\mathrm{f}(a_0)\mathrm{f}(c_0) < 0$，所以

$$I_1 = [\tfrac{1}{2}\pi, \tfrac{3}{4}\pi].$$

求 I_2

现在 $c_1 = \frac{1}{2}(\frac{1}{2}\pi + \frac{3}{4}\pi) = \frac{5}{8}\pi$，我们需要知道 $2\sin\frac{5}{8}\pi - \frac{5}{8}\pi$ 的正负。

想一想 不用计算机，怎样求 $\sin\frac{5}{8}\pi$?

利用 $\sin\frac{5}{8}\pi = \sin(\pi - \frac{5}{8}\pi) = \sin\frac{3}{8}\pi = \cos\frac{1}{8}\pi$，再以 $\theta = \frac{1}{8}\pi$ 代入倍角公式（double-angle formula）$\cos 2\theta = 2\cos^2\theta - 1$：

$$4\cos^2\tfrac{1}{8}\pi = 2\big(1 + \cos\tfrac{1}{4}\pi\big) = 2 + \sqrt{2}.$$

所以 $(2\sin\frac{5}{8}\pi)^2 = 2 + \sqrt{2}$。$2\sin\frac{5}{8}\pi$ 和 $\frac{5}{8}\pi$ 都是正数，所以可以比较它们的平方：

$$(2\sin\tfrac{5}{8}\pi)^2 = 2 + \sqrt{2} \approx 3.4, \qquad (\tfrac{5}{8}\pi)^2 \approx \tfrac{25}{64}\cdot 10 = \tfrac{250}{64} \approx 3.9.$$

因此 $2\sin\frac{5}{8}\pi < \frac{5}{8}\pi$，即 $\mathrm{f}(c_1) < 0$。同样地，$\mathrm{f}(a_1) = \mathrm{f}(\frac{1}{2}\pi) > 0$，所以 $\mathrm{f}(a_1)\mathrm{f}(c_1) < 0$，于是

$$I_2 = [\tfrac{1}{2}\pi, \tfrac{5}{8}\pi],$$

正如题目所求。

求 I_3

现在 $c_2 = \frac{1}{2}(\frac{1}{2}\pi + \frac{5}{8}\pi) = \frac{9}{16}\pi$。与之前一样，$\sin\frac{9}{16}\pi = \sin\frac{7}{16}\pi = \cos\frac{1}{16}\pi$，再以 $\theta = \frac{1}{16}\pi$ 代入倍角公式，得

$$(2\sin\tfrac{9}{16}\pi)^2 = 4\cos^2\tfrac{1}{16}\pi = 2\big(1 + \cos\tfrac{1}{8}\pi\big) = 2 + 2\cos\tfrac{1}{8}\pi = 2 + \sqrt{2 + \sqrt{2}},$$

当中用到上一步得出的 $2\cos\frac{1}{8}\pi = \sqrt{2 + \sqrt{2}}$。取 $\sqrt{2} \approx 1.4$，我们有 $\sqrt{2 + \sqrt{2}} \approx \sqrt{3.4} > 1.8$（因为 $1.8^2 = 3.24$），所以

$$(2\sin\tfrac{9}{16}\pi)^2 > 3.8, \qquad (\tfrac{9}{16}\pi)^2 \approx \tfrac{81}{256}\cdot 10 = \tfrac{810}{256} < 3.2.$$

所以 $2\sin\frac{9}{16}\pi > \frac{9}{16}\pi$，即 $\mathrm{f}(c_2) > 0$。这一次 $\mathrm{f}(a_2)\mathrm{f}(c_2) > 0$，而 $\mathrm{f}(c_2)\mathrm{f}(b_2) = \mathrm{f}(\frac{9}{16}\pi)\,\mathrm{f}(\frac{5}{8}\pi) < 0$。因此

$$I_3 = [\tfrac{9}{16}\pi, \tfrac{5}{8}\pi].$$

这里使用近似值是稳妥的：在每个中点处，所比较的两个量相差约 0.5 或以上，远大于 $\frac{1}{\sqrt{2}} \approx 0.7$ 或 $\pi \approx \sqrt{10}$ 所带来的误差。

验算

- 以数值方法执行二分法，得到的区间相同，而根为 $\alpha \approx 1.8955$，位于 $\frac{9}{16}\pi \approx 1.767$ 与 $\frac{5}{8}\pi \approx 1.963$ 之间。
- 精确值与近似值吻合：$\mathrm{f}(\frac{3}{4}\pi) \approx -0.942$，$\mathrm{f}(\frac{5}{8}\pi) \approx -0.116$，以及 $\mathrm{f}(\frac{9}{16}\pi) \approx 0.194$。
- 每个区间的长度都是前一个的一半：$\frac{1}{2}\pi$、$\frac{1}{4}\pi$、$\frac{1}{8}\pi$、$\frac{1}{16}\pi$。

常见错误

- 只指出函数变号，却没有解释为何只有一个根。单凭变号只能保证至少有一个根；要证明唯一性（uniqueness），还需要“一个递减、一个递增”的论证。
- 以平方来比较 $2\sin c$ 和 c，却没有先指出两者都是正数。

STEP 之外

变号论证其实就是介值定理（intermediate value theorem），它是大学分析课程中最早遇到的重要定理之一；其证明需要用到实数的完备性（completeness），而二分法本质上就是把这个证明当作算法来执行。二分法必定收敛，但速度缓慢（每一步只多得一个二进制位）。牛顿－拉弗森法（Newton–Raphson）等方法收敛得快得多，但有可能失败，所以数值分析（numerical analysis）在很大程度上就是在稳妥与速度之间作取舍。有关理论可参阅 Abbott 的 *Understanding Analysis*。

详细解答

STEP II 2004，第 8 题

纯数学 · 微分方程的比较

STEP II 2004，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/04-s2-q8

关键思路

当 $0 \le x < 1$ 时，右边为正数，所以 x 随 t 递增，我们可以把方程倒转过来看：t 变成 x 的函数，即由 0 爬升到高度 x 所需的时间。在每个高度上斜率都较大，即每个高度都较早到达，所以题目给出的右边各式之间的不等式（inequality）立即定出三条曲线的次序，一条在另一条之上。至于爬升到 $x = 1$ 所需的时间是有限还是无限，就变成一个瑕积分（improper integral）是否收敛（converge）的问题。

解答

以下一律以 $x_n(t)$ 表示指数为 n 时的解。在 $t = 0$ 时，$x = 0$，而对每个 n 斜率都是 $(1-0)^{1/n} = 1$，所以所有曲线都沿直线 $x = t$ 离开原点。当 $0 \le x < 1$ 时，斜率严格为正，所以 x 递增，而

$$\frac{\mathrm{d}t}{\mathrm{d}x} = \frac{1}{(1-x^n)^{1/n}}, \qquad \text{所以} \qquad t_n(x) = \int_0^x \frac{\mathrm{d}\xi}{(1-\xi^n)^{1/n}}$$

就是解到达高度 x 的时刻，其中 $0 \le x < 1$。

第 (i) 部分

取 $n = 1$，方程为 $\dfrac{\mathrm{d}x}{\mathrm{d}t} = 1 - x$。分离变量（separating variables）（这是可行的，因为沿着解有 $1 - x > 0$），

$$t = \int_0^x \frac{\mathrm{d}\xi}{1-\xi} = -\ln(1-x), \qquad \text{所以} \qquad x = 1 - \mathrm{e}^{-t}.$$

图像由原点出发，斜率为 1，一直递增并且向下凹（concave）（其斜率 e^{-t} 递减），当 $t \to \infty$ 时趋近渐近线（asymptote）$x = 1$ 但永远不会到达：当 $x \to 1$ 时，$t_1(x) = -\ln(1-x) \to \infty$。

下图中最低的一条曲线就是它。

第 (ii) 部分

不等式。对 $0 < x < 1$，$1 - x$ 和 $(1 - x^2)^{1/2}$ 都是正数，所以该不等式等价于两边平方后所得的不等式：

$$(1 - x)^2 < 1 - x^2 \iff 1 - 2x + x^2 < 1 - x^2 \iff 2x^2 < 2x \iff x < 1,$$

其中最后一步是除以 $2x > 0$。由于 $x < 1$ 成立，所以 $1 - x < (1 - x^2)^{1/2}$ 也成立。

草图。在 $(0, 1)$ 内的每个高度 x，$n = 2$ 的曲线都比 $n = 1$ 的曲线陡。因此，对 $0 < x < 1$，

$$t_2(x) = \int_0^x \frac{\mathrm{d}\xi}{(1 - \xi^2)^{1/2}} < \int_0^x \frac{\mathrm{d}\xi}{1 - \xi} = t_1(x),$$

因为左边的被积函数（integrand）在 $(0, x]$ 上严格较小。所以 $n = 2$ 的解严格较早到达每一个高度；又因为两个解都是递增的，所以在每个时刻 $t > 0$（只要 $x_2 < 1$），$n = 2$ 的曲线都严格位于 $n = 1$ 的曲线之上。两者都以斜率 1 离开原点，所以在原点附近它们非常接近，然后随着 t 增大而分开。

> **想一想** 两条曲线在原点的斜率相同。为甚么 $n = 1$ 的曲线不能在稍后悄悄升到上面?

要升到上面，它必须在某个高度 x 追上 $n = 2$ 的曲线，即在同一时刻到达 x。但对每个 $x \in (0, 1)$ 都有 $t_2(x) < t_1(x)$：爬得较慢的那个总是严格地迟到。（等价的说法是：凡两条曲线处于同一高度之处，$n = 2$ 的那条都较陡，所以它永远不会被超越。）

求解。设 $x = \sin y$，其中 $0 \le y < \frac{1}{2}\pi$，且在 $t = 0$ 时 $y = 0$。于是 $\cos y > 0$，所以 $(1 - x^2)^{1/2} = (\cos^2 y)^{1/2} = \cos y$，方程变成

$$\cos y \frac{\mathrm{d}y}{\mathrm{d}t} = \cos y \quad \Longrightarrow \quad \frac{\mathrm{d}y}{\mathrm{d}t} = 1 \quad \Longrightarrow \quad y = t,$$

这里用到在 $t = 0$ 时 $y = 0$。所以解为

$$x = \sin t \qquad \text{其中} 0 \le t < \tfrac{1}{2}\pi.$$

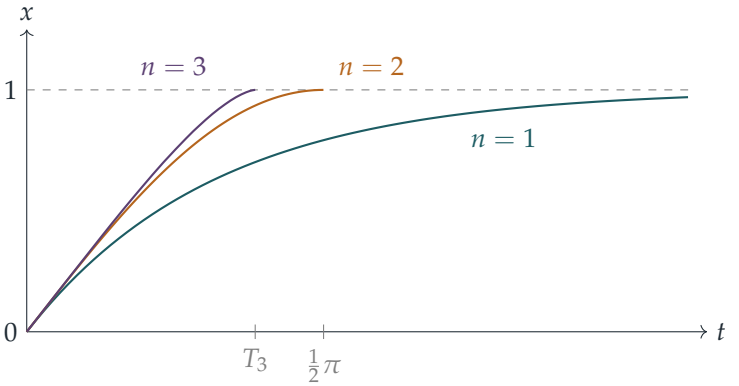
这与 $t_2(x) = \arcsin x$ 吻合。与 $n = 1$ 的情况不同，这个解在有限时间内到达 $x = 1$，时刻为 $t = \frac{1}{2}\pi$，而且到达时斜率为 $\cos \frac{1}{2}\pi = 0$，与直线 $x = 1$ 相切。

第 (iii) 部分

与之前完全一样，在 $(0, 1)$ 上运用题目给出的不等式 $(1 - x^2)^{1/2} < (1 - x^3)^{1/3}$，可知 $n = 3$ 的曲线在每个高度都比 $n = 2$ 的曲线陡，所以

$$t_3(x) = \int_0^x \frac{\mathrm{d}\xi}{(1 - \xi^3)^{1/3}} < \int_0^x \frac{\mathrm{d}\xi}{(1 - \xi^2)^{1/2}} = \arcsin x \qquad (0 < x < 1).$$

所以 $n=3$ 的曲线和其他曲线一样以斜率 1 离开原点，并且在 $t>0$ 时严格位于 $x=\sin t$ 之上。它的斜率 $(1-x^3)^{1/3}$ 在 $x\to 1$ 时降至 0，所以它与直线 $x=1$ 相遇时同样会变得平坦。



在 $t=\frac{1}{2}\pi$ 之前到达 $x=1$。草图显示 $n=3$ 的曲线位于 $\sin t$ 之上，而后者本身在 $\frac{1}{2}\pi$ 到达 1；所以 $n=3$ 的曲线必定不迟于该时刻到达 1。要看出它是严格较早到达，可在积分的比较中令 $x\to 1$。到达 $x=1$ 所需的时间为

$$T_3=\int_0^1 \frac{\mathrm{d}\xi}{(1-\xi^3)^{1/3}}.$$

这个瑕积分收敛：由于在 $[0,1]$ 上 $1-\xi^3=(1-\xi)(1+\xi+\xi^2)\geq 1-\xi$，被积函数至多为 $(1-\xi)^{-1/3}$，而后者在 $[0,1]$ 上的积分为 $\frac{3}{2}$。又由于被积函数在 $(0,1)$ 上严格小于 $(1-\xi^2)^{-1/2}$，

$$T_3<\int_0^1 \frac{\mathrm{d}\xi}{(1-\xi^2)^{1/2}}=\Big[\arcsin\xi\Big]_0^1=\tfrac{1}{2}\pi.$$

所以 $n=3$ 的解在时刻 $T_3<\frac{1}{2}\pi$ 到达 $x=1$，这正是所要证明的。（一旦到达，它就停留在 $x=1$，因为斜率为 0，而且不可能变成负数。）

验算

- 代回原式：$\frac{\mathrm{d}}{\mathrm{d}t}(1-\mathrm{e}^{-t})=\mathrm{e}^{-t}=1-x$，而对 $0\leq t\leq\frac{1}{2}\pi$ 有 $\frac{\mathrm{d}}{\mathrm{d}t}\sin t=\cos t=(1-\sin^2 t)^{1/2}$。以数值方法（龙格—库塔法，Runge–Kutta）解各方程，所得结果与这些公式吻合。
- 数值上，$T_3\approx 1.2092$，明显小于 $\frac{1}{2}\pi\approx 1.5708$。事实上 $T_3=\dfrac{2\pi}{3\sqrt{3}}$ 是准确值（见下文），两者吻合至小数点后七位。
- 在高度 $x=0.9$，三个时间分别为 $-\ln 0.1\approx 2.30$、$\arcsin 0.9\approx 1.12$ 和 $t_3(0.9)\approx 0.98$，次序与预测相符。

常见错误

- 只说“斜率较大，所以曲线较高”，却没有指出所比较的是同一高度上的斜率，而不是同一时刻的斜率。改为把 t 看成 x 的函数，论证才站得住脚。
- 写出 $(1-\sin^2 y)^{1/2}=\cos y$ 而没有说明为何 $\cos y \geq 0$：这需要 $0 \leq y \leq \frac{1}{2}\pi$。
- 把 $x=\sin t$ 延续到 $t=\frac{1}{2}\pi$ 之后：在那里它会递减，但方程规定 $\mathrm{d}x/\mathrm{d}t \geq 0$。

STEP 之外

解在有限时间内到达 $x=1$，然后永远停留在 $x=1$，原因是 $(1-x^n)^{1/n}$ 在 $x=1$ 处不满足利普希茨（Lipschitz）条件；微分方程（differential equation）的唯一性定理（uniqueness theorem）正是在这里失效（当 $n=2$ 时，$x=1$ 和 $x=\sin t$ 都通过 $(\frac{1}{2}\pi,1)$）。到达时间是一个 Beta 函数（Beta function）：代入 $u=\xi^n$ 可得 $T_n=\frac{1}{n}B(\frac{1}{n},1-\frac{1}{n})=\dfrac{\pi/n}{\sin(\pi/n)}$，当 $n=2$ 时为 $\frac{1}{2}\pi$，当 $n=3$ 时为 $\frac{2\pi}{3\sqrt{3}}$，并且当 $n\to\infty$ 时递减趋向 1。Boyce 与 DiPrima 的 *Elementary Differential Equations*（基础微分方程教科书）讨论了解的存在性与唯一性，以及解的比较。

详细解答

STEP II 2004，第 9 题

力学 · 重心

STEP II 2004，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/04-s2-q9

关键思路

半球以曲面着地时，与平面接触的是球面上的一点，而球面的法线必定通过球心。因此反作用力沿着通过底面圆心 O 的铅垂线作用；平衡时，无论放在那里的是甚么物体，其重心（centre of gravity）都必定在同一条铅垂线上。于是题目的每一部分都只是“求出重心，然后倾斜物体，直至重心在 O 的正下方”。采用固定在物体上的坐标，整个计算就轻松得多。

解答

模型与平衡条件

把半球视为刚体（rigid body），它与水平面只在一点 C 接触；把质点（particle）视为固定在边缘上的点质量。作用力只有各个重量和 C 处的接触力。由于 C 在球面上，C 处的法线通过 O；又因为平面是水平的，这条法线是铅垂的。所以 C 在 O 的正下方，而且无论物体怎样滚动，O 的高度始终是 r。

平衡（equilibrium）时，接触力必须与总重量互相抵销，所以它是铅垂的（没有摩擦力），并且与重量沿同一直线作用。因此：

平衡时，系统的重心在通过 O 的铅垂线上。

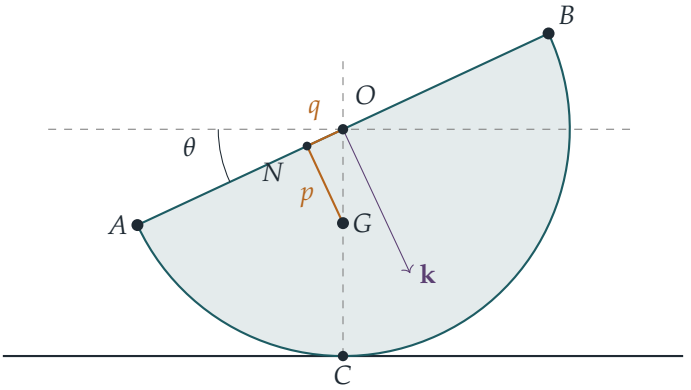
重心在 O 的下方，因为物体位于底面的曲面那一侧。

物体坐标与初始位置

在物体上固定两个互相垂直的单位向量（unit vector）：$\mathbf{k}$ 沿对称轴，由 O 指向曲面；$\mathbf{e}$ 在底面所在的平面内。由于 G 与底面的距离是 p，与 O 的距离是 $\sqrt{p^2+q^2}$，它在底面上的垂足 N 与 O 的距离是 q。选取 $\mathbf{e}$ 由 O 指向 N，便得

$$\overrightarrow{OG} = p\,\mathbf{k} + q\,\mathbf{e}.$$

底面垂直于 $\mathbf{k}$，所以底面与水平面的夹角等于 $\mathbf{k}$ 与铅垂线的夹角。在初始位置，$\overrightarrow{OG}$ 铅垂向下，所以这个角 θ 满足 $\tan\theta = q/p$。



边缘上哪一点最低？边缘上的一点相对于 O 的位置是 $r(\cos\phi\,\mathbf{e} + \sin\phi\,\mathbf{f})$，其中 $\mathbf{f}$ 是第三个单位向量，与前两者都垂直。这一点在 O 以下的深度，就是它沿向下的单位向量 $(p\mathbf{k}+q\mathbf{e})/\sqrt{p^2+q^2}$ 的分量，即

$$\frac{rq\cos\phi}{\sqrt{p^2+q^2}},$$

当 $\phi = 0$ 时最大（这里取 $q > 0$；若 $q = 0$，底面是水平的，没有唯一的最低点）。所以

$$\overrightarrow{OA} = r\,\mathbf{e}, \qquad \overrightarrow{OB} = -r\,\mathbf{e}.$$

用文字说明：A 与 G 在对称轴的同一侧，并且在通过 O 和 G 的铅垂平面内，而整个问题始终留在这个平面内。

质点在 A

半球连同质点的重心 G_A 由下式给出：

$$(M+m)\,\overrightarrow{OG_A} = M(p\,\mathbf{k} + q\,\mathbf{e}) + mr\,\mathbf{e} = Mp\,\mathbf{k} + (Mq + mr)\,\mathbf{e}.$$

在新的平衡位置，$\overrightarrow{OG_A}$ 是铅垂的，所以与求 θ 时完全一样，$\mathbf{k}$ 与铅垂线的夹角（亦即底面与水平面的夹角）就是 $\overrightarrow{OG_A}$ 与 $\mathbf{k}$ 的夹角：

$$\tan\alpha = \frac{Mq+mr}{Mp}, \qquad \text{所以} \qquad \alpha = \tan^{-1}\left(\frac{Mq+mr}{Mp}\right).$$

正如所料，A 处多出的重量把 A 拉得更低，于是 $\alpha > \theta$。

> **想一想** 为甚么底面的倾角就是 $\overrightarrow{OG_A}$ 与对称轴的夹角？

底面垂直于对称轴 $\mathbf{k}$，而水平面垂直于铅垂线。把一对直线各自旋转一个直角，并不改变两者之间的夹角，所以底面与水平面的夹角等于 $\mathbf{k}$ 与铅垂线的夹角。平衡时，通过 O 的铅垂线就是直线 OG_A，它沿 $\mathbf{k}$ 和 $\mathbf{e}$ 的分量分别是 Mp 和 $Mq + mr$（除以 $M + m$ 之后）。它与 $\mathbf{k}$ 所成角的正切就是这两个分量之比。

质点在 B

现在质点位于 $-r\,\mathbf{e}$，所以新的重心 G_B 满足

$$(M+m)\,\overrightarrow{OG_B} = Mp\,\mathbf{k} + (Mq - mr)\,\mathbf{e}, \qquad \text{由此得} \qquad \tan\beta = \frac{Mq - mr}{Mp}.$$

由于 m 很小，我们有 $mr < Mq$，所以 $\overrightarrow{OG_B}$ 仍然与先前一样在对称轴的同一侧，底面向同一方向倾斜，倾角是较小的 β。

两个位置之间的夹角

由复角公式（compound-angle formula），

$$\begin{aligned}\tan(\alpha-\beta) &= \frac{\tan\alpha - \tan\beta}{1 + \tan\alpha\tan\beta} = \frac{\dfrac{2mr}{Mp}}{1 + \dfrac{(Mq+mr)(Mq-mr)}{M^2p^2}} \\ &= \frac{2mMrp}{M^2p^2 + M^2q^2 - m^2r^2} = \frac{2mMrp}{M^2\,(p^2+q^2) - m^2r^2},\end{aligned}$$

其中第三步把分子和分母同乘以 M^2p^2。这就是所要求的结果。

验算

- 若 $q = 0$（例如均匀半球），两个位置互为镜像：$\tan\alpha = mr/(Mp)$ 而 $\beta = -\alpha$。公式给出 $\tan 2\alpha = 2t/(1-t^2)$，其中 $t = mr/(Mp)$，正是二倍角公式，与预期相符。
- 若 $m \to 0$，则 α 和 β 都趋向 θ，而右边趋向 0。
- 取 $M = 1$、$m = 0.1$、$r = 1$、$p = \frac{3}{8}$、$q = \frac{1}{8}$，得 $\tan\alpha = 0.6$ 和 $\tan\beta = \frac{1}{15}$，所以 $\alpha \approx 30.96°$，$\beta \approx 3.81°$。结果的两边都等于 $\frac{20}{39} \approx 0.5128$。我们亦以数值方法求出平衡角：对 2000 组随机数值，把重心的高度减至最小，所得结果与 $\tan\alpha$ 和 $\tan\beta$ 的公式相符，误差在 10^{-13} 以内。

常见错误

- 把 G 放在原来位置的接触点的正下方，而不是 O 的正下方。接触点会随物体滚动而移动；O 才是始终在接触点正上方的那一点。
- 把 A 取在对称轴的错误一侧。边缘的最低点与 G 在同一侧，这正是质点在 A 时倾角增大的原因。
- 当 $mr > Mq$ 时仍把 β 当作正数。这时底面向另一方向倾斜，公式只在把 β 算作负数时才成立；“m 很小”这个条件正是要排除这种情况。

STEP 之外

这里所用的原理是：刚体在重力和单一接触力作用下平衡时，其质心（centre of mass）在该力的作用线上。等价的说法是，质心处于高度的驻值（stationary）位置；而考察物体滚动时这个高度的变化，还可以判定平衡是否稳定。这就是判定稳定性的能量法（energy method）的起点，大学一年级力学会把这套方法发展下去，用来处理摇摆和滚动的物体。Kleppner 与 Kolenkow 的 *An Introduction to Mechanics*（力学入门教科书）正是以这种观点讲述质心和刚体平衡。

详细解答

STEP II 2004，第 10 题

力学 · 斜面上的摩擦力

> **STEP II 2004，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/04-s2-q10

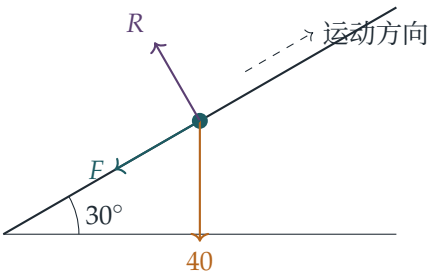
关键思路

摩擦力（friction）永远与运动方向相反，所以质点（particle）沿斜面向上滑与向下滑时的加速度（acceleration）并不相同。先求出这两个加速度：题目的数字经过挑选，结果十分整齐。然后 P 向上滑到停下，再滑回来，与仍在向上滑的 Q 相遇。把两个质点的位置都写成“在最高点以下的距离”，碰撞（collision）方程就成为一个平方差（difference of two squares）。

（印刷试卷的第 (i) 部分有一处误植：碰撞时间中的“b”应为 T。）

解答

我们把每个质点视为粗糙斜面（rough plane）上的质点，并以 x 表示由 A 沿 ℓ 向上量度的位置，取沿斜面向上为正方向。



垂直于斜面的方向没有加速度，所以法向反作用力（normal reaction）为

$$R = 4 \cdot 10 \cos 30^\circ = 20\sqrt{3}\ \text{N},$$

而由于每个质点都在滑动，摩擦力达到极限值（limiting friction）：

$$F = \mu R = \frac{1}{5\sqrt{3}} \cdot 20\sqrt{3} = 4\,\text{N}.$$

重量沿斜面向下的分量是 $40\sin 30^\circ = 20\,\text{N}$。

向上运动时：摩擦力沿斜面向下，所以 $4a = -(20+4)$，减速度为 $6\,\text{m}\,\text{s}^{-2}$。

向下运动时：摩擦力沿斜面向上，所以沿斜面向下的加速度为 $(20-4)/4 = 4\,\text{m}\,\text{s}^{-2}$。

由于 $20 > 4$，质点停下后不会保持静止：极限摩擦力不足以把它留住，所以它会滑回下来。

第 (i) 部分

P 的运动。向上时，$v = 6 - 6t$，所以 P 在 $t = 1$ 停下，此时已走了 $\frac{1}{2} \cdot 6 \cdot 1 = 3\,\text{m}$。当 $0 \le t \le 1$ 时，$x_P = 6t - 3t^2 = 3 - 3(1-t)^2$。其后 P 由 $x = 3$ 从静止开始以加速度 4 向下滑，所以

$$x_P = 3 - 2(t-1)^2 \qquad (t \ge 1),$$

此式一直成立，直至 P 回到 A，那时 $2(t-1)^2 = 3$，即 $t = 1 + \sqrt{3/2}$。

Q 的运动。与上述运动完全相同，只是时间推迟了 T：设 $u = t - T$，则

$$x_Q = 6u - 3u^2 = 3 - 3(1-u)^2 \qquad (0 \le u \le 1).$$

> **想一想** 两个质点会否在都仍向上运动时碰撞?

不会。当两者都向上运动时，Q 所走的正是 P 在 T 秒之前走过的路径，而沿这条路径 x 是递增的，所以 P 严格地在前面。它们只可能在 P 掉头之后相遇，即 $t \ge 1$ 之时。此外，Q 也不可能先到达自己的最高点（$x = 3$，在 $u = 1$ 时），因为那时 P 早已离开 $x = 3$，位于其下方。所以碰撞时 P 正在下滑，而 Q 正在上滑。

碰撞。设 $\tau = t - 1 \ge 0$ 为 P 掉头后所经过的时间。那么 $1 - u = 1 - (t - T) = T - \tau$，而由 $x_P = x_Q$ 得

$$2\tau^2 = 3(T-\tau)^2.$$

τ 和 $T - \tau = 1 - u$ 都是非负的（后者是因为 Q 仍在向上运动），所以可以取正平方根：

$$\sqrt{2}\,\tau = \sqrt{3}\,(T - \tau)$$

$$\tau = \frac{\sqrt{3}\,T}{\sqrt{3}+\sqrt{2}} = \sqrt{3}(\sqrt{3}-\sqrt{2})\,T = (3-\sqrt{6})\,T,$$

其中利用 $(\sqrt{3}+\sqrt{2})(\sqrt{3}-\sqrt{2}) = 1$ 作有理化（rationalising）。因此碰撞发生在 P 抛出后

$$t = (3-\sqrt{6})\,T + 1$$

秒，正如所求。

余下要看看关于 T 的条件在哪里用到。这个时刻必须早于 P 回到 A 的时刻，所以需要 $\tau < \sqrt{3/2}$：

$$(3-\sqrt{6})\,T < \sqrt{\tfrac{3}{2}} \iff T < \frac{\sqrt{3/2}\,(3+\sqrt{6})}{3} = \sqrt{\tfrac{3}{2}} + \tfrac{\sqrt{9}}{3} = 1 + \sqrt{\tfrac{3}{2}},$$

因为 $\sqrt{3/2}\cdot\sqrt{6} = 3$。这正是题目给出的条件。（同一个不等式亦保证 $t \geq T$，即 Q 已经抛出：$(3-\sqrt{6})T + 1 \geq T \iff T \leq 1/(\sqrt{6}-2) = 1+\sqrt{3/2}$。）

第 (ii) 部分

当 $T = 1+\sqrt{2/3}$ 时，

$$\tau = (3-\sqrt{6})\left(1+\sqrt{\tfrac{2}{3}}\right) = 3-\sqrt{6}+3\sqrt{\tfrac{2}{3}}-\sqrt{6}\sqrt{\tfrac{2}{3}} = 3-\sqrt{6}+\sqrt{6}-2 = 1,$$

所以碰撞发生在 $t = 2$。这个 T 值小于 $1+\sqrt{3/2}$，所以第 (i) 部分的结果适用。

因摩擦而损失的能量等于摩擦力乘以滑行的总距离，因为摩擦力的大小恒为 4 N，而且永远与运动方向相反。

- P 向上滑了 3 m，然后向下滑了 $2\tau^2 = 2$ m：合共 5 m。
- Q 尚未掉头，并且与 P 位于同一点，即 $x = 3-2 = 1$：它滑行了 1 m。

所以损失的能量是

$$4\times(5+1) = 24\ \text{J}.$$

验算

用能量收支可以核实这个答案。起初的动能（kinetic energy）是 $2\times\frac{1}{2}\cdot 4\cdot 6^2 = 144$ J。在 $t=2$ 时，P 的速率是 $4\times 1 = 4$（动能 32 J）；Q 有 $u = 1-\sqrt{2/3}$，速率为 $6(1-u) = 6\sqrt{2/3}$，所以 $v^2 = 24$（动能 48 J）。两者都在斜面上 1 m 处，即升高了 $\frac{1}{2}$ m，势能（potential energy）增加了 $2\times 4\cdot 10\cdot\frac{1}{2} = 40$ J。而 $144-32-48-40 = 24$ J。另外，对随机选取的 $T < 1+\sqrt{3/2}$，以数值方法逐步模拟两个质点的运动，所得的碰撞时间与 $(3-\sqrt{6})T+1$ 吻合（例如当 $T \approx 1.816$ 时为 ≈ 2.000），因摩擦而损失的能量为 24.000 J。

常见错误

- P 向下滑时沿用向上滑时的加速度：摩擦力的方向已经逆转，所以应是 4，而不是 6。
- 忘记了 Q 所受的摩擦力：Q 也一直在滑动，它滑行的 1 m 贡献了 4 J。
- 在 $2\tau^2 = 3(T-\tau)^2$ 中取错平方根：正负号由 $\tau \geq 0$ 和 $T-\tau \geq 0$ 决定。

STEP 之外

摩擦力是非保守力（non-conservative force）最简单的例子：它所作的功取决于路径，而不仅取决于起点和终点，这就是为甚么我们需要的是滑行的总距离，而不是位移。这也解释了为甚么机械能（mechanical energy）在某些系统中守恒，在另一些系统中却不守恒；大学一年级的力学课程会用线积分（line integral）和势函数（potential）把这个想法严格地表述出来。Kleppner 与 Kolenkow 的 *An Introduction to Mechanics*（力学入门）对功、保守力和摩擦力有仔细的论述。

详细解答

STEP II 2004，第 11 题

力学 · 功率与阻力

STEP II 2004，第 11 题

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

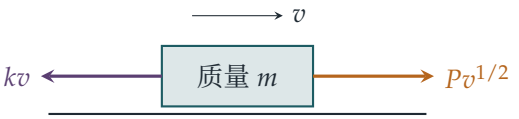
step.maths.org/questions/04-s2-q11

关键思路

功率（power）等于力乘以速率，所以引擎以全功率 $Pv^{3/2}$ 运作时，所施的推力为 $Pv^{1/2}$。由牛顿第二定律（Newton's second law）便得到一条含有 $\sqrt{v}$ 的微分方程（differential equation），而代换（substitution）$u = \sqrt{v}$ 可把它化为一阶线性方程（linear first-order equation），马上就能解出。引擎关掉之后，两列火车的速率各自按指数衰减（exponential decay），而第 (iii) 部分的窍门就是比较这两个衰减率。

解答

我们把每列火车看成沿水平直路轨运动的质点（particle）。引擎以全功率开动时，沿路轨方向的力有驱动力（driving force）和阻力（resistance）kv。



若引擎在速率为 v 时输出的功率为 H，则其驱动力为 H/v。所以在全功率下，A 所受的驱动力为 $Pv^{3/2}/v = Pv^{1/2}$，而 B 所受的驱动力为 $2Pv^{1/2}$。

第 (i) 部分

对于 A，沿路轨方向应用牛顿第二定律，得

$$m\frac{\mathrm{d}v_A}{\mathrm{d}t} = Pv_A^{1/2} - kv_A.$$

设 $u = v_A^{1/2}$，则 $v_A = u^2$，且 $\dfrac{\mathrm{d}v_A}{\mathrm{d}t} = 2u\dfrac{\mathrm{d}u}{\mathrm{d}t}$。于是

$$2mu\frac{\mathrm{d}u}{\mathrm{d}t} = Pu - ku^2, \qquad \text{所以} \qquad 2m\frac{\mathrm{d}u}{\mathrm{d}t} = P - ku,$$

这里两边除以 u；由于火车一直在运动（$u > 0$），这样做是容许的。这是一条线性（而且可分离变量（separable））的方程：

$$\int \frac{2m\,\mathrm{d}u}{P - ku} = \int \mathrm{d}t \quad \Longrightarrow \quad -\frac{2m}{k}\ln(P - ku) = t + \text{常数},$$

所以 $P - ku = \mathrm{C}\mathrm{e}^{-kt/2m}$，其中 C 为某个常数。（开始时 u 极小，故 $P - ku > 0$；而且它一直保持为正，因为解永远不会到达平衡值（equilibrium）$u = P/k$。）

想一想 “以极低的速率缓缓前进”告诉了我们甚么?

它给出了初始条件（initial condition）。在 $t = 0$ 时，两个速率都是正数（所以除以 u 并无问题），但小得可以忽略，因此我们取 $t = 0$ 时 $u = 0$。于是 $C = P$，而

$$u = \frac{P}{k}\left(1 - \mathrm{e}^{-kt/2m}\right), \qquad \text{所以} \qquad v_A = \frac{P^2\left(1 - \mathrm{e}^{-kt/2m}\right)^2}{k^2},$$

正是所要求的结果。（假如火车完全静止，$v \equiv 0$ 也会满足原来的方程，因为那时根本没有驱动力；这就是题目要让火车一开始已在缓慢移动的原因。）

对于 B，方程是 $2m\dfrac{\mathrm{d}v_B}{\mathrm{d}t} = 2Pv_B^{1/2} - kv_B$。用同一个代换可得 $4m\dfrac{\mathrm{d}u}{\mathrm{d}t} = 2P - ku$，这正是 A 的方程，只是把 P 换成 $2P$、m 换成 $2m$。所以

$$v_B = \frac{4P^2\left(1 - \mathrm{e}^{-kt/4m}\right)^2}{k^2}.$$

第 (ii) 部分

两个速率都是正数，所以 $9v_A = 4v_B$ 等价于 $3\sqrt{v_A} = 2\sqrt{v_B}$。设 $x = \mathrm{e}^{-kt/4m}$，则 $\mathrm{e}^{-kt/2m} = x^2$，而且当 $t > 0$ 时 $0 < x < 1$。条件变成

$$\frac{3P}{k}\left(1 - x^2\right) = \frac{4P}{k}(1 - x) \quad \Longrightarrow \quad 3(1 - x)(1 + x) = 4(1 - x).$$

由于 $x \neq 1$，我们可以除以 $1 - x$，得 $3 + 3x = 4$，所以 $x = \frac{1}{3}$（即 $t = \frac{4m}{k}\ln 3$）。于是 $1 - x^2 = \frac{8}{9}$，$1 - x = \frac{2}{3}$，所以

$$v_A = \frac{64P^2}{81k^2}, \qquad v_B = \frac{16P^2}{9k^2}.$$

简单验算一下：$9v_A = \frac{64P^2}{9k^2} = 4v_B$。

第 (iii) 部分

从引擎关掉的一刻起计算时间 τ，并设 $V_A = \frac{64P^2}{81k^2}$ 和 $V_B = \frac{16P^2}{9k^2}$ 为该刻的速率。这时只有阻力起作用：

$$m\frac{\mathrm{d}v_A}{\mathrm{d}\tau} = -kv_A, \qquad 2m\frac{\mathrm{d}v_B}{\mathrm{d}\tau} = -kv_B,$$

所以

$$v_A = V_A\,\mathrm{e}^{-k\tau/m}, \qquad v_B = V_B\,\mathrm{e}^{-k\tau/2m}.$$

较重的火车 B 减速的指数率只有一半，所以把 v_B 平方，两个指数项便一致了：

$$v_B^2 = V_B^2\,\mathrm{e}^{-k\tau/m} = \frac{V_B^2}{V_A}\,v_A.$$

最后

$$\frac{V_B^2}{V_A} = \frac{256P^4}{81k^4}\cdot\frac{81k^2}{64P^2} = \frac{4P^2}{k^2},$$

所以 $v_B^2 = \dfrac{4P^2}{k^2}v_A$，即在其后的所有时刻都有 $k^2v_B^2 = 4P^2v_A$，正是所要求的结果。

验算

- 当 $t \to \infty$ 时，$v_A \to P^2/k^2$，这正是驱动力 $Pv^{1/2}$ 与阻力 kv 相等之处。同样地，$v_B \to 4P^2/k^2$，那里 $2Pv^{1/2} = kv$。
- 当 t 很小时，$v_A \approx P^2t^2/(4m^2)$，$v_B \approx P^2t^2/(4m^2)$：两列火车起步时并驾齐驱，这是应有的结果，因为两者每单位质量的驱动力都是 $Pv^{1/2}/m$。
- 取 $m = 1.7$、$P = 2.3$、$k = 0.9$，我们由一个极小的初速率出发，用数值方法（龙格—库塔法（Runge–Kutta））逐步计算两条运动方程，结果与第 (i) 部分的公式吻合至约 10^{-5}；在 $9v_A = 4v_B$ 的一刻，得 $\mathrm{e}^{-kt/4m} = 0.33333$、$v_A = 5.1602$ 和 $v_B = 11.6104$；关掉引擎后继续逐步计算，$k^2v_B^2$ 一直与 $4P^2v_A$ 相等（例如两秒后两者都是 37.8739）。

常见错误

- 把 $Pv^{3/2}$ 当作力。它是功率；力等于功率除以速率。
- 忘记第 (ii) 部分的条件是一条关于 $x = \mathrm{e}^{-kt/4m}$ 的二次方程，除了 $x = \frac{1}{3}$ 之外，还有根 $x = 1$（即 $t = 0$，此时两个速率皆为零）。

STEP 之外

代换 $u = \sqrt{v}$ 是求解伯努利方程（Bernoulli equation）$y' + p(t)y = q(t)y^n$ 的方法的一个特例；在该方法中，$u = y^{1-n}$ 总能给出一条线性方程。关于火车完全静止的那个备注，让我们初次窥见唯一性理论（uniqueness theory）：由于 $\sqrt{v}$ 在 $v = 0$ 处的斜率为无穷大，方程由 $v = 0$ 出发的解不止一个，皮卡－林德勒夫定理（Picard–Lindelöf theorem）解释了这个现象。Boyce 与 DiPrima 的 *Elementary Differential Equations*（微分方程入门教材）涵盖了这两个课题；至于功率与阻力的力学，可参阅 Kleppner 与 Kolenkow 的 *An Introduction to Mechanics*（力学入门教材）。

详细解答

STEP II 2004，第 12 题

概率与统计 · 连续分布

STEP II 2004，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/04-s2-q12

关键思路

一切都源自一个原函数（antiderivative）：除了一个常数因子之外，$x\mathrm{e}^{-ax^2}$ 就是 e^{-ax^2} 的导数。由此可求出归一化常数（normalising constant），并得到累积分布函数（cumulative distribution function）的闭合形式，所以只需解一条指数方程便可求得中位数（median）。众数（mode）就是曲线达到顶点的位置，前提是该顶点位于 $[0,1]$ 之内；否则概率密度在 $x=1$ 处仍在上升。至于条件概率（conditional probability），只需用到累积分布函数在 m 和 h 的值。

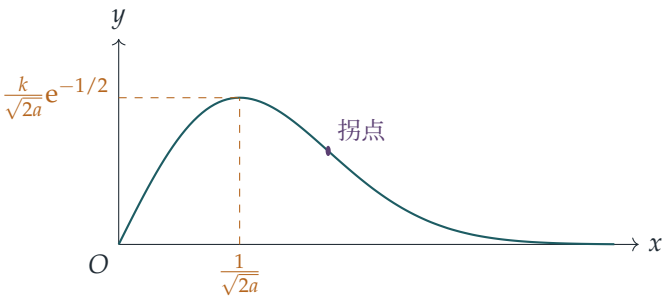
解答

草图

设 $g(x)=kx\mathrm{e}^{-ax^2}$，其中 $x\geq 0$。则 $g(0)=0$，当 $x>0$ 时 $g(x)>0$，而且由积法则（product rule）可得

$$g'(x)=k\mathrm{e}^{-ax^2}\left(1-2ax^2\right),$$

它在 $0\leq x<1/\sqrt{2a}$ 时为正，其后为负。所以 g 由原点开始上升（在该处的斜率为 k），在 $x=1/\sqrt{2a}$ 达到唯一的极大值，此时 $y=\dfrac{k}{\sqrt{2a}}\mathrm{e}^{-1/2}$，然后下降。当 $x\to\infty$ 时，指数因子压倒线性因子，所以 $y\to 0$：x 轴是渐近线（asymptote）。（再求导一次可知曲线在 $x=\sqrt{3/(2a)}$ 有一个拐点（point of inflection），曲线在该处由凹（concave）转为凸（convex）。）



常数 k

由于 $\dfrac{\mathrm{d}}{\mathrm{d}x}e^{-ax^2} = -2axe^{-ax^2}$，对于 $0 \le t \le 1$，

$$\int_0^t kxe^{-ax^2}\,\mathrm{d}x = \left[-\frac{k}{2a}e^{-ax^2}\right]_0^t = \frac{k}{2a}\left(1-e^{-at^2}\right).$$

总概率为 1，所以取 $t=1$ 可得 $\dfrac{k}{2a}(1-e^{-a}) = 1$，即

$$k = \frac{2a}{1-e^{-a}},$$

这正是所需的结果。（因为 $a > 0$，所以分母为正。）把这个值代回去，在 $[0,1]$ 上的累积分布函数为

$$F(t) = P(X \le t) = \frac{1-e^{-at^2}}{1-e^{-a}}.$$

众数

概率密度函数（density）就是我们刚才所画的曲线，在 $x=1$ 处截断。从草图可见，在 $[0,1]$ 上，如果转向点（turning point）位于区间之内，密度便在该点最大，否则在区间的右端点最大。

- 若 $a > \frac{1}{2}$，则 $2a > 1$，所以 $1/\sqrt{2a} < 1$：顶点位于区间之内，故 $m = \dfrac{1}{\sqrt{2a}}$。
- 若 $a < \frac{1}{2}$，则 $1/\sqrt{2a} > 1$，所以在整个 $[0,1)$ 上 $f'(x) > 0$：密度在整个区间上递增，故 $m = 1$。

（当 $a = \frac{1}{2}$ 时，两条公式都给出 $m = 1$。）

中位数

中位数 h 满足 $F(h) = \frac{1}{2}$：

$$1-e^{-ah^2} = \tfrac{1}{2}\left(1-e^{-a}\right) \quad\Longleftrightarrow\quad e^{-ah^2} = \frac{1+e^{-a}}{2}.$$

取对数，并利用 $h > 0$，

$$h = \sqrt{\frac{1}{a}\ln\left(\frac{2}{1+\mathrm{e}^{-a}}\right)}.$$

由于 $1+\mathrm{e}^{-a} < 2$，对数为正，所以这是一个实数。

何时 $h > m$?

若 $a \le \frac{1}{2}$，则 $m = 1$，而 $h < 1$（一半的概率在 h 以下，而 1 以上则没有概率），所以 $h < m$。因此只有当 $a > \frac{1}{2}$ 时才可能有 $h > m$，而此时 $m = 1/\sqrt{2a}$。比较两者的平方（两者皆为正），

$$\begin{aligned} h > m &\iff \frac{1}{a}\ln\left(\frac{2}{1+\mathrm{e}^{-a}}\right) > \frac{1}{2a} \iff \ln\left(\frac{2}{1+\mathrm{e}^{-a}}\right) > \frac{1}{2} \\ &\iff \frac{2}{1+\mathrm{e}^{-a}} > \mathrm{e}^{1/2} \iff 1+\mathrm{e}^{-a} < 2\mathrm{e}^{-1/2} \\ &\iff \mathrm{e}^{-a} < 2\mathrm{e}^{-1/2} - 1 \iff a > -\ln\left(2\mathrm{e}^{-1/2}-1\right). \end{aligned}$$

第一步我们乘以 $a > 0$，第三步用到 exp 是递增函数，最后一步用到 ln 是递增函数。最后一步还需要 $2\mathrm{e}^{-1/2} - 1 > 0$，这是成立的，因为 $\mathrm{e}^{1/2} < 2$（因为 $\mathrm{e} < 4$）。

> **想一想** 以上一连串的推导假设了 $a > \frac{1}{2}$。结论不是需要把它列为额外条件吗?

不需要，而这正是下一部分要求证明它的原因。我们在下面会证明 $c = -\ln\left(2\mathrm{e}^{-1/2}-1\right)$ 大于 $\frac{1}{2}$。所以若 $a > c$，则自动有 $a > \frac{1}{2}$，公式 $m = 1/\sqrt{2a}$ 适用，而上述推导说明 $h > m$。连同 $a \le \frac{1}{2}$ 的情况，这就证明了 $h > m$ 当且仅当 $a > c$。

该常数大于 $\frac{1}{2}$

由于 ln 是递增函数，

$$-\ln\left(2\mathrm{e}^{-1/2}-1\right) > \tfrac{1}{2} \iff 2\mathrm{e}^{-1/2} - 1 < \mathrm{e}^{-1/2} \iff \mathrm{e}^{-1/2} < 1,$$

而最后一个不等式成立。（数值上，$c \approx 1.546$。）

条件概率

假设 $a > c$。则 $m < h$，所以事件 $\{X > m\} \cap \{X < h\}$ 就是事件 $\{m < X < h\}$。由于 $P(X < h) = \frac{1}{2}$，

$$P(X > m \mid X < h) = \frac{P(m < X < h)}{P(X < h)} = \frac{F(h) - F(m)}{\frac{1}{2}} = 1 - 2F(m).$$

当 $m = 1/\sqrt{2a}$ 时，$am^2 = \frac{1}{2}$，所以 $F(m) = \dfrac{1 - \mathrm{e}^{-1/2}}{1 - \mathrm{e}^{-a}}$。因此

$$P(X > m \mid X < h) = \frac{(1 - \mathrm{e}^{-a}) - 2(1 - \mathrm{e}^{-1/2})}{1 - \mathrm{e}^{-a}} = \frac{2\mathrm{e}^{-1/2} - \mathrm{e}^{-a} - 1}{1 - \mathrm{e}^{-a}},$$

这正是所需的结果。

验算

- 当 $a = c$ 时，分子 $2\mathrm{e}^{-1/2} - \mathrm{e}^{-a} - 1$ 恰好为 0，与该处 $m = h$ 吻合；当 $a > c$ 时它为正，符合概率应有的性质。
- 当 $a \to \infty$ 时，答案趋向 $2\mathrm{e}^{-1/2} - 1 \approx 0.213$，小于 1。
- 取 $a = 2.5$：$k \approx 5.447$，$m \approx 0.447$，$h \approx 0.496$，而对密度作数值积分可得 $P(X > m \mid X < h) \approx 0.1427$，与公式一致。取 $a = 1$ 则得 $m \approx 0.707 > h \approx 0.616$，与 $1 < c$ 相符。

常见错误

- 在所有情况下都答 $m = 1/\sqrt{2a}$：当 $a < \frac{1}{2}$ 时，转向点落在 $x = 1$ 之外，而该处的密度为零。
- 忘记 F 里面的 k 必须是归一化后的值，才能使 $F(1) = 1$。
- 写成 $P(X > m \mid X < h) = P(X > m)/P(X < h)$；分子必须是两个事件同时发生的概率。

STEP 之外

这个密度是截断瑞利分布（truncated Rayleigh distribution）：若不在 $x = 1$ 处截断，取 $k = 2a$ 的 $kx\mathrm{e}^{-ax^2}$ 就是一个点与原点距离的分布，而该点的两个坐标是互相独立的正态变量（normal variables）。比较平均值（mean）、中位数和众数，以及截断如何使它们移动，是大学第一门统计课程中反复出现的主题。Grimmett and Welsh 的 *Probability: An Introduction*（概率入门教材）是进一步了解连续随机变量（continuous random variables）及其密度函数和分布函数的好去处。

详细解答

STEP II 2004，第 13 题

概率与统计 · 最优停止

STEP II 2004，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/04-s2-q13

关键思路

逐次抽球去想。假设她已经安全取得 n 个白球，再抽一次的话，要么多得 £1，要么把手上的 £n 全部输掉，所以这一次抽球的期望（expected）效果很容易写出来。以 N 个球为目标时，总期望奖金会上升，当且仅当多抽那一球的期望得益为正，因此最佳目标就是该得益仍然非负的最后一个目标。

解答

全题假设 $1 \le r < b$，即最少有一个红球和最少有一个白球。（若 $r = 0$，她只需把所有球取出。）

下一次抽球的期望得益

假设她首 n 次抽到的都是白球。袋中现在有 $b-n$ 个球：r 个红球和 $b-n-r$ 个白球，每个球在下一次被抽出的机会均等（要还有白球剩下，须有 $n \le b-r-1$；无论如何 $b-n \ge r \ge 1$）。

- 抽到白球的概率（probability）为 $\dfrac{b-n-r}{b-n}$，她的奖金由 £n 变为 £$(n+1)$：得益为 $+1$。
- 抽到红球的概率为 $\dfrac{r}{b-n}$，她会失去本来可得的 £n：得益为 $-n$。

所以她下一次抽球的期望得益是

$$G_n = \frac{(b-n-r)\cdot 1 + r\cdot(-n)}{b-n} = \frac{b-r-n(r+1)}{b-n}.$$

分母为正，所以 $G_n = 0$ 当且仅当 $b - r = n(r+1)$，亦即

$$n = \frac{b-r}{r+1},$$

这正是所需的结果。更有用的是：当 $n < \dfrac{b-r}{r+1}$ 时 $G_n > 0$，而当 $n > \dfrac{b-r}{r+1}$ 时 $G_n < 0$。

最佳的取球数目

设 P_N 为首 N 次抽球全为白球的概率，并设 E_N 为她以取出 N 个球为目标时的总期望奖金。若该 N 个球全为白球，她得到 £N，否则一无所得，所以

$$E_N = NP_N.$$

? **想一想** 单次抽球的期望得益与总期望奖金 E_N 有甚么关系?

比较以 N 为目标和以 $N+1$ 为目标这两个方案。只有当首 N 个球全为白球时，两个方案才有分别（否则在两个方案下她都已经输掉），这情况出现的概率为 P_N；而在这情况下，多抽的一球使她的奖金改变，改变量的期望值为 G_N。因此

$$E_{N+1} - E_N = P_N G_N = P_N \frac{b-r-N(r+1)}{b-N}.$$

（亦可由 $P_{N+1} = P_N \cdot \frac{b-r-N}{b-N}$ 直接验证。）

当 $N \le b-r$ 时，我们有 $P_N > 0$，所以 $E_{N+1} - E_N$ 的正负号与 $b-r-N(r+1)$ 相同。记 $k = N+1$，这就是下式的正负号：

$$(b+1) - k(r+1).$$

设 $m = \dfrac{b+1}{r+1}$。那么当 $k < m$ 时 $E_k > E_{k-1}$，当 $k = m$ 时 $E_k = E_{k-1}$，而当 $k > m$ 时 $E_k < E_{k-1}$。所以数列（sequence）$E_0, E_1, E_2, \ldots$ 严格递增至 $E_{\lfloor m \rfloor}$（若 m 为整数，则出现相等的情况 $E_{m-1} = E_m$），然后严格递减，至少在 $k-1 \le b-r$ 的范围内如此。超出这范围后 $P_N = 0$ 且 $E_N = 0$，不会更好。

我们应该验证 $\lfloor m \rfloor$ 是一个合理的目标，即 $\lfloor m \rfloor \le b-r$。若 $b = r+1$，则 $m = \frac{r+2}{r+1}$，所以 $\lfloor m \rfloor = 1 = b-r$。若 $b \ge r+2$，则

$$(b-r)(r+1) - (b+1) = r(b-r-1) - 1 \ge r - 1 \ge 0,$$

所以 $m \le b-r$。无论哪种情况，E_N 的最大值都在

$$n = \left\lfloor \frac{b+1}{r+1} \right\rfloor,$$

取得，即 $\frac{b+1}{r+1}$ 的整数部分，这正是所需的结果。

无论采用甚么停止规则（stopping rule），这个目标也是她所能做到的最佳选择：在决定是否继续之前，她唯一能得知的就是至今抽出的每个球都是白球（否则游戏已经结束），所以任何规则都等同于预先选定一个数 N。

$r = 1$ 的情况

只有一个红球时，首 N 次抽球避开红球，当且仅当红球位于最后 $b - N$ 个位置之一，而各个位置的机会均等。所以

$$P_N = \frac{b-N}{b}, \qquad E_N = \frac{N(b-N)}{b}.$$

最佳目标是 $n = \left\lfloor \frac{b+1}{2} \right\rfloor$。

*b 为偶数。*这时 $\frac{b+1}{2} = \frac{b}{2} + \frac{1}{2}$，所以 $n = \frac{b}{2}$，且

$$E_n = \frac{\frac{b}{2} \cdot \frac{b}{2}}{b} = \frac{b}{4}.$$

她的总期望奖金为 $£\frac{1}{4}b$，这正是所需的结果。

*b 为奇数。*这时 $n = \frac{b+1}{2}$，所以 $b - n = \frac{b-1}{2}$，且

$$E_n = \frac{\frac{b+1}{2} \cdot \frac{b-1}{2}}{b} = \frac{b^2-1}{4b}.$$

她的总期望奖金为 $£\frac{b^2-1}{4b}$。（这里 $m = \frac{b+1}{2}$ 是整数，所以以 $\frac{b-1}{2}$ 个球为目标会得到相同的值，由对称的公式 $N(b-N)/b$ 可以看出。）

验算

- 当 $r = 1$ 时，$E_N = N(b-N)/b$ 是关于 N 的开口向下的抛物线（parabola），顶点在 $N = b/2$，与两个答案都吻合：b 为偶数时，在顶点处恰好是 $b/4$；b 为奇数时则略小，为 $\frac{b}{4} - \frac{1}{4b}$。
- 取 $b = 7$,$r = 2$:公式给出 $n = \lfloor 8/3 \rfloor = 2$,而直接计算得 $E_0, \ldots, E_5 = 0, \frac{5}{7}, \frac{20}{21}, \frac{6}{7}, \frac{4}{7}, \frac{5}{21}$，在 $N = 2$ 处达到最高。我们列举了球的全部 7! 种排列次序来确认这一点，并就每个 $b \le 24$，以完整的最优停止计算核对了 n 的公式。
- 当 $b = 9$，$r = 1$ 时，答案是 $\frac{80}{36} = \frac{20}{9} \approx 2.22$，刚好略低于 $\frac{9}{4}$。

常见错误

- 忘记了抽到红球会令她失去已取得的 $£n$，而把期望得益只写成 $\frac{b-n-r}{b-n}$。

- 混淆“已抽出 n 个球”与“以 n 个球为目标”：得益 G_n 在 $n = \frac{b-r}{r+1}$ 时为零，所以最佳目标要再多一个，即 $\frac{b-r}{r+1} + 1 = \frac{b+1}{r+1}$。
- 遗漏了 $\frac{b+1}{r+1}$ 为整数时的相等情况：这时 n 和 $n-1$ 同为最优。

STEP 之外

这是一个小型的最优停止（optimal stopping）问题，而所用的论证就是“向前看一步”（one-step look-ahead）法则：一旦再走一步的期望得益为负，便立即停止。只要问题是单调（monotone）的，这个法则就是最优的；所谓单调，是指一旦继续下去显得不利，其后便一直不利，正如本题中 G_n 随 n 递减。同样的想法会引伸到秘书问题（secretary problem），以及大学一、二年级概率论中的鞅（martingale）方法。Grimmett and Welsh 的 *Probability: An Introduction*（概率论入门教材）是打好背后条件期望（conditional expectation）工具基础的好去处。

详细解答

STEP II 2004，第 14 题

概率与统计 · 容斥原理

STEP II 2004，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/04-s2-q14

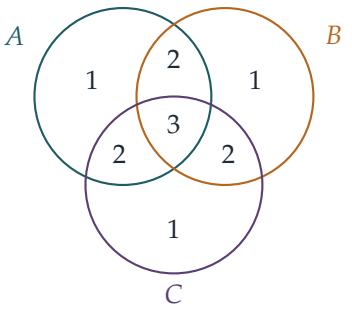
关键思路

容斥原理（inclusion–exclusion）是一种修正计数的方法：把 $\mathrm{P}(A)+\mathrm{P}(B)+\mathrm{P}(C)$ 加起来时，重叠的部分被多算了，而两两相交与三者相交的项正好把计数修正过来。对于布甸这一题，考虑补事件（complement）会容易得多：设 E_i 为布甸 i 是空的这一事件（event）。某个布甸是空的概率是一个并集（union）的概率，这正是容斥原理所处理的情况，而且各个 E_i 的每一个交集（intersection）的概率都十分简单。

解答

容斥公式

把 $A \cup B \cup C$ 分成温氏图（Venn diagram）中七个互不相交（disjoint）的区域。任何由若干个完整区域组成的事件，其概率等于这些区域的概率之和，所以只需验证每个区域在等式右边都恰好被计算一次。



图中的数字表示每个区域落在 A、B、C 当中的多少个之内。取一个恰好落在三个事件中

k 个之内的区域，其中 $k = 1$、2 或 3。

- 它出现在 $\mathrm{P}(A), \mathrm{P}(B), \mathrm{P}(C)$ 这些项当中的 k 项。
- 它出现在 $\binom{k}{2}$ 个两两交集之中，因为（举例说）它落在 $A \cap B$ 之内，当且仅当它同时落在 A 和 B 之内。
- 只有当 $k = 3$ 时，它才出现在三者的交集之中，即出现 $\binom{k}{3}$ 次。

因此它被计算的净次数为

$$k - \binom{k}{2} + \binom{k}{3} = \begin{cases} 1 - 0 + 0 = 1 & (k = 1), \\ 2 - 1 + 0 = 1 & (k = 2), \\ 3 - 3 + 1 = 1 & (k = 3). \end{cases}$$

$A \cup B \cup C$ 的每个区域都恰好被计算一次，而并集以外的区域则完全没有被计算。所以等式右边等于 $\mathrm{P}(A \cup B \cup C)$。

（另一个等价的做法：把两个事件的法则 $\mathrm{P}(X \cup Y) = \mathrm{P}(X) + \mathrm{P}(Y) - \mathrm{P}(X \cap Y)$ 应用于 $X = A \cup B$ 和 $Y = C$，然后再应用于 $(A \cap C) \cup (B \cap C)$。）

每个布匐最少有一枚六便士硬币

模型。彻底搅拌后再平均分成几份，意思是每枚六便士硬币（sixpence）落入布匐 1、2 或 3 的概率各为 $\frac{1}{3}$，并且与其他硬币独立（independent）。取 $r \geq 1$。

设 E_i 为布匐 i 内没有六便士硬币这一事件。

- 当全部 r 枚硬币都没有落入布匐 1 时，E_1 发生。每枚硬币没有落入布匐 1 的概率是 $\frac{2}{3}$，所以由独立性可得 $\mathrm{P}(E_1) = \left(\frac{2}{3}\right)^r$，$E_2$、$E_3$ 亦同理。
- 当每枚硬币都落入布匐 3 时，$E_1 \cap E_2$ 发生，所以 $\mathrm{P}(E_1 \cap E_2) = \left(\frac{1}{3}\right)^r$，其余各对亦同理。
- $E_1 \cap E_2 \cap E_3$ 要求三个布匐全是空的，但由于 $r \geq 1$，这是不可能的。所以它的概率是 0。

由容斥原理，最少有一个布匐是空的概率为

$$p_r = \mathrm{P}(E_1 \cup E_2 \cup E_3) = 3\left(\tfrac{2}{3}\right)^r - 3\left(\tfrac{1}{3}\right)^r.$$

每个布匐最少有一枚六便士硬币的概率就是其补事件的概率：

$$1 - 3\left(\tfrac{2}{3}\right)^r + 3\left(\tfrac{1}{3}\right)^r = \frac{3^r - 3 \cdot 2^r + 3}{3^r}.$$

最少六枚六便士硬币

我们要找出使 $p_r < \frac{1}{3}$ 的 r 值。直接计算可得

$$p_5 = \frac{3\cdot 32 - 3}{243} = \frac{93}{243} = \frac{31}{81} > \frac{27}{81} = \frac{1}{3}, \qquad p_6 = \frac{3\cdot 64 - 3}{729} = \frac{189}{729} = \frac{7}{27} < \frac{9}{27} = \frac{1}{3}.$$

为了确定没有更小的 r 符合要求，我们证明 p_r 是递减的。对于 $r \geq 1$，

$$p_{r+1} - p_r = 3\left(\tfrac{2}{3}\right)^r\left(\tfrac{2}{3} - 1\right) - 3\left(\tfrac{1}{3}\right)^r\left(\tfrac{1}{3} - 1\right) = -\left(\tfrac{2}{3}\right)^r + 2\left(\tfrac{1}{3}\right)^r = \left(\tfrac{1}{3}\right)^r\left(2 - 2^r\right),$$

当 $r = 1$ 时此式为 0，当 $r \geq 2$ 时为负数。所以 $p_1 = p_2 \geq p_3 \geq p_4 \geq p_5 = \frac{31}{81}$，而每个 $r \leq 5$ 都有 $p_r > \frac{1}{3}$。（当 $r = 0$ 时，每个布甸必然是空的。）因此，只有当 $r \geq 6$ 时，某个布甸没有六便士硬币的机会才会低于 $\frac{1}{3}$；又由于 p_r 持续递减，每个 $r \geq 6$ 确实都有 $p_r \leq p_6 < \frac{1}{3}$。所以厨师必须使用 6 枚或以上的六便士硬币，这正是所要证明的。

每个布甸有两枚六便士硬币

现在 $r = 6$。设 T 为每个布甸恰好有两枚六便士硬币这一事件，S 为每个布甸最少有一枚这一事件。由上一部分，

$$\mathrm{P}(S) = 1 - p_6 = 1 - \frac{7}{27} = \frac{20}{27}.$$

把六枚（可区分的）六便士硬币分配到各个布甸，共有 $3^6 = 729$ 种等可能的方法。其中每个布甸得到两枚的方法数目，就是把六件物件分成三个有标号的对的方法数目，即 $\dfrac{6!}{2!\,2!\,2!} = 90$。所以 $\mathrm{P}(T) = \frac{90}{729} = \frac{10}{81}$。

> **想一想** 为甚么 $\mathrm{P}(T \cap S)$ 就是 $\mathrm{P}(T)$?

如果每个布甸都有两枚六便士硬币，那么每个布甸当然最少有一枚，所以 $T \subseteq S$，而 $T \cap S = T$。因此

$$\mathrm{P}(T \mid S) = \frac{\mathrm{P}(T)}{\mathrm{P}(S)} = \frac{10/81}{20/27} = \frac{10}{81}\cdot\frac{27}{20} = \frac{1}{6}.$$

验算

对于 $r = 1$ 和 $r = 2$，公式给出每个布甸都有六便士硬币的概率为 0，这是对的：少于三枚硬币不可能令三个布甸都有硬币。对于 $r = 3$，公式给出 $\frac{27-24+3}{27} = \frac{2}{9}$，与直接计数所得的 $3!/3^3 = 6/27$ 吻合。对 $r \leq 9$ 逐一列举全部 3^r 种分配方法，结果与公式一致，并确认 $p_5 = \frac{31}{81} \approx 0.383$ 和 $p_6 = \frac{7}{27} \approx 0.259$，同时得出最终答案 $\frac{1}{6} \approx 0.167$（模拟 200 000 次布甸制作所得的结果是 0.166）。

常见错误

- 把六便士硬币当作不可区分，并把 $6 = a + b + c$ 的 28 种写法视为等可能。它们并非等可能：$(2,2,2)$ 可以有 90 种出现方式，而 $(6,0,0)$ 只有一种。
- 只检查 $r = 6$ 和 $r = 5$。“必须使用 6 枚或以上”还要求每个更小的 r 都不符合要求，这就是单调性（monotonicity）那一步之所以重要的原因。
- 忘记三者的交集是空集，或者不加说明就把它的概率写成 $(0)^r$。

STEP 之外

对于 n 个事件的一般容斥原理，涉及所有 k 重交集的交错和，其证明用的正是上面的计数论证，当中用到 $\sum_{j\geq 1}(-1)^{j-1}\binom{k}{j} = 1$。把它应用于 n 个布匈，所数算的就是满射（surjection）的数目，并与第二类斯特灵数（Stirling numbers of the second kind）以及经典的优惠券收集问题（coupon-collector problem）相关。Grimmett 与 Welsh 的 *Probability: An Introduction*（概率论入门教材）对容斥原理和这类占位问题（occupancy problem）有详细阐述。

详细解答

STEP III 2004，第 1 题

纯数学 · 双曲函数积分

STEP III 2004，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/04-s3-q1

关键思路

两个被积函数（integrand）都是一个双曲函数（hyperbolic function）乘以由另一个双曲函数构成的式子，而 $\sinh x\,\mathrm{d}x$ 和 $\cosh x\,\mathrm{d}x$ 正好分别是 $\cosh x$ 和 $\sinh x$ 的微分（differential）。因此，代换（substitution）$c=\cosh x$ 和 $s=\sinh x$ 可把它们化为有理函数的积分：一个要用部分分式（partial fractions），另一个则得出反正切（arctangent）。至于题目中“由此”（hence）的一步，关键在于留意到两个分母其实是同一个函数，因为 $2\cosh^2 x-1=1+2\sinh^2 x$。

解答

第一个积分

设 $c=\cosh x$，则 $\mathrm{d}c=\sinh x\,\mathrm{d}x$；积分限 $x=0$ 和 $x=a$ 变为 $c=1$ 和 $c=\cosh a$。于是

$$\int_0^a \frac{\sinh x}{2\cosh^2 x-1}\,\mathrm{d}x=\int_1^{\cosh a}\frac{\mathrm{d}c}{2c^2-1}.$$

分母可因式分解为 $(\sqrt{2}\,c-1)(\sqrt{2}\,c+1)$，而

$$\frac{1}{(\sqrt{2}\,c-1)(\sqrt{2}\,c+1)}=\frac{1}{2}\left(\frac{1}{\sqrt{2}\,c-1}-\frac{1}{\sqrt{2}\,c+1}\right),$$

把右边通分合并即可验证：分子是 $\frac{1}{2}\big((\sqrt{2}c+1)-(\sqrt{2}c-1)\big)=1$。每一项积分后都得出一个对数，并因链式法则（chain rule）而多出一个因子 $1/\sqrt{2}$。当 $c\geq 1$ 时，$\sqrt{2}c\pm 1$ 两

者皆为正数，所以无须加上绝对值符号：

$$\int_1^{\cosh a} \frac{\mathrm{d}c}{2c^2-1} = \frac{1}{2\sqrt{2}}\Big[\ln(\sqrt{2}\,c-1)-\ln(\sqrt{2}\,c+1)\Big]_1^{\cosh a}$$
$$= \frac{1}{2\sqrt{2}}\ln\left(\frac{\sqrt{2}\cosh a-1}{\sqrt{2}\cosh a+1}\right)-\frac{1}{2\sqrt{2}}\ln\left(\frac{\sqrt{2}-1}{\sqrt{2}+1}\right).$$

由于 $-\ln\frac{\sqrt{2}-1}{\sqrt{2}+1}=\ln\frac{\sqrt{2}+1}{\sqrt{2}-1}$，这正是所求的结果。

第二个积分

现在设 $s=\sinh x$，则 $\mathrm{d}s=\cosh x\,\mathrm{d}x$，积分限为 $s=0$ 和 $s=\sinh a$：

$$\int_0^a \frac{\cosh x}{1+2\sinh^2 x}\,\mathrm{d}x = \int_0^{\sinh a}\frac{\mathrm{d}s}{1+(\sqrt{2}\,s)^2} = \frac{1}{\sqrt{2}}\Big[\arctan(\sqrt{2}\,s)\Big]_0^{\sinh a}$$
$$= \frac{1}{\sqrt{2}}\arctan\big(\sqrt{2}\sinh a\big).$$

无穷积分

? **想一想** 两个被积函数的分母看来并不相同，我们怎样才可以把它们相减？

利用 $\cosh^2 x=1+\sinh^2 x$，

$$2\cosh^2 x-1=2+2\sinh^2 x-1=1+2\sinh^2 x,$$

所以两个分母完全相同，而

$$\int_0^a \frac{\cosh x-\sinh x}{1+2\sinh^2 x}\,\mathrm{d}x = \frac{1}{\sqrt{2}}\arctan(\sqrt{2}\sinh a)-\frac{1}{2\sqrt{2}}\ln\left(\frac{\sqrt{2}\cosh a-1}{\sqrt{2}\cosh a+1}\right)$$
$$-\frac{1}{2\sqrt{2}}\ln\left(\frac{\sqrt{2}+1}{\sqrt{2}-1}\right).$$

现在令 $a\to\infty$。由于 $\sinh a\to\infty$，故 $\arctan(\sqrt{2}\sinh a)\to\frac{\pi}{2}$。由于 $\cosh a\to\infty$，

$$\frac{\sqrt{2}\cosh a-1}{\sqrt{2}\cosh a+1}=\frac{1-1/(\sqrt{2}\cosh a)}{1+1/(\sqrt{2}\cosh a)}\to 1,$$

而 ln 在 1 处连续（continuous），所以它的对数趋向 0。三项都有有限的极限，因此这个无穷积分收敛（converge），并且

$$\int_0^\infty \frac{\cosh x-\sinh x}{1+2\sinh^2 x}\,\mathrm{d}x=\frac{\pi}{2\sqrt{2}}-\frac{1}{2\sqrt{2}}\ln\left(\frac{\sqrt{2}+1}{\sqrt{2}-1}\right),$$

正如所求。

最后一个积分

? **想一想** 以 e^x 表示的话，$\cosh x - \sinh x$ 和 $1 + 2\sinh^2 x$ 会是甚么样子?

设 $u = e^x$，则 $\cosh x - \sinh x = e^{-x} = 1/u$；再利用二倍角公式（double-angle formula），

$$1 + 2\sinh^2 x = \cosh 2x = \frac{u^2 + u^{-2}}{2} = \frac{u^4 + 1}{2u^2}.$$

此外 $\mathrm{d}u = e^x\,\mathrm{d}x$，所以 $\mathrm{d}x = \mathrm{d}u/u$，而积分限 $x = 0$、$x \to \infty$ 变为 $u = 1$、$u \to \infty$。因此

$$\begin{aligned}\int_0^\infty \frac{\cosh x - \sinh x}{1 + 2\sinh^2 x}\,\mathrm{d}x &= \int_1^\infty \frac{1}{u} \cdot \frac{2u^2}{1 + u^4} \cdot \frac{\mathrm{d}u}{u} \\ &= 2\int_1^\infty \frac{\mathrm{d}u}{1 + u^4}.\end{aligned}$$

把上一个结果减半，

$$\int_1^\infty \frac{\mathrm{d}u}{1 + u^4} = \frac{\pi}{4\sqrt{2}} - \frac{1}{4\sqrt{2}} \ln\left(\frac{\sqrt{2} + 1}{\sqrt{2} - 1}\right).$$

这个式子还可以化简：$\dfrac{\sqrt{2} + 1}{\sqrt{2} - 1} = \dfrac{(\sqrt{2} + 1)^2}{(\sqrt{2} - 1)(\sqrt{2} + 1)} = (\sqrt{2} + 1)^2$，所以该对数等于 $2\ln(1 + \sqrt{2})$，而

$$\int_1^\infty \frac{\mathrm{d}u}{1 + u^4} = \frac{\pi}{4\sqrt{2}} - \frac{\ln(1 + \sqrt{2})}{2\sqrt{2}} \approx 0.2437.$$

验算

- 当 $a = 0$ 时，第一条公式给出 $\frac{1}{2\sqrt{2}}\left(\ln\frac{\sqrt{2}-1}{\sqrt{2}+1} + \ln\frac{\sqrt{2}+1}{\sqrt{2}-1}\right) = 0$，在空区间上的积分理应如此。
- 最后的答案是正数，并且小于 $\int_1^\infty u^{-4}\,\mathrm{d}u = \frac{1}{3}$；由于 $1 + u^4 > u^4$，这是合理的。
- 对 $\int_1^\infty \frac{\mathrm{d}u}{1+u^4}$ 使用辛普森法则（Simpson's rule）（经代换 $u = 1/t$ 后，它变为 $\int_0^1 \frac{t^2}{1+t^4}\,\mathrm{d}t$），得出 0.243748，与公式吻合至小数点后十二位。以 x 表示的无穷积分，数值验算结果为 0.487495，正是它的两倍。

常见错误

- 对 $\frac{1}{\sqrt{2}c \pm 1}$ 或 $\frac{1}{1+2s^2}$ 积分时，漏掉因子 $\frac{1}{\sqrt{2}}$。
- 代换时没有更改积分限，或者答案仍以 c 或 s 表示。
- 在最后一部分，忘记 $\mathrm{d}x = \mathrm{d}u/u$，结果得出 $2u/(1 + u^4)$ 和一个发散（divergent）的积分。

STEP 之外

最后的结果是一个经典积分的一部分：$\int_0^\infty \frac{\mathrm{d}u}{1+u^4} = \frac{\pi}{2\sqrt{2}}$，而代换 $u \mapsto 1/u$ 显示，在 $[0,1]$ 和 $[1,\infty)$ 上的两部分分别是 $\int_0^1 \frac{1}{1+t^4}\,\mathrm{d}t$ 和 $\int_0^1 \frac{t^2}{1+t^4}\,\mathrm{d}t$，它们的和与差都可以用初等方法求出。在大学一年级的课程中，你还会遇到一条更巧妙的途径，就是围道积分（contour integration）和留数（residue），它可以一举求出 $\int_0^\infty \frac{\mathrm{d}u}{1+u^n} = \frac{\pi/n}{\sin(\pi/n)}$。Needham 的 *Visual Complex Analysis*（《复分析：可视化方法》）是学习这个想法的绝佳读物。

详细解答

STEP III 2004，第 2 题

纯数学 · 渐近线与曲线描绘

STEP III 2004，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/04-s3-q2

关键思路

把 $f(x)$ 写成直线 $y = x - 4$ 加上一个修正项。修正项的分子是完全平方（perfect square），所以它在某一点处有二重零点，而这个二重零点正正就是相切（tangency）的意思。至于 $f(x) = 0$ 的根，只要消去分母，便会得到一个平方差（difference of two squares），可以立即因式分解。

解答

在整个解答中，记

$$f(x) = (x-4) - \frac{16(2x+1)^2}{x^2(x-4)}.$$

第 (i) 部分

分母在 $x = 0$ 和 $x = 4$ 处为零。分子 $16(2x+1)^2$ 在这两处都不为零，所以垂直渐近线（vertical asymptote）是

$$x = 0 \qquad 及 \qquad x = 4.$$

当 $|x|$ 很大时，分式的分子是二次式，分母是三次式：

$$\frac{16(2x+1)^2}{x^2(x-4)} = \frac{64x^2+64x+16}{x^3-4x^2} \approx \frac{64}{x} \to 0 \quad 当 x \to \pm\infty,$$

所以 $f(x) - (x-4) \to 0$，而斜渐近线（oblique asymptote）是

$$y = x - 4.$$

想一想 渐近线怎可能同时又是切线（tangent）?

渐近线只描述曲线在远处的情况，所以没有甚么会阻止这条直线同时在某个有限的点与曲线相切。要找出直线与曲线的交点，设 $\mathrm{f}(x) = x - 4$:

$$\mathrm{f}(x) - (x-4) = -\frac{16(2x+1)^2}{x^2(x-4)} = 0 \iff x = -\tfrac{1}{2}.$$

由于因式 $(2x+1)$ 带有平方，$x = -\frac{1}{2}$ 是 $\mathrm{f}(x) - (x-4) = 0$ 的二重根（double root）。这表示直线在该处与曲线相切。要直接验证，设 $D(x) = \mathrm{f}(x) - (x-4) = (2x+1)^2\, h(x)$，其中 $h(x) = -16/\big(x^2(x-4)\big)$。函数 h 在 $x = -\frac{1}{2}$ 附近可导（differentiable），所以

$$D'(x) = 4(2x+1)h(x) + (2x+1)^2 h'(x),$$

它在 $x = -\frac{1}{2}$ 处等于 0。因此 $\mathrm{f}'(-\frac{1}{2}) = 1$，即渐近线的斜率（gradient）。此外 $\mathrm{f}(-\frac{1}{2}) = -\frac{9}{2}$，这一点位于 $y = x - 4$ 上。所以斜渐近线就是曲线在 $\left(-\frac{1}{2}, -\frac{9}{2}\right)$ 处的切线。

第 (ii) 部分

把所有项通分，公分母为 $x^2(x-4)$:

$$\mathrm{f}(x) = \frac{x^2(x-4)^2 - 16(2x+1)^2}{x^2(x-4)} = \frac{\big[x(x-4)\big]^2 - \big[4(2x+1)\big]^2}{x^2(x-4)}.$$

分子是一个平方差:

$$\begin{aligned}\big[x^2 - 4x - 4(2x+1)\big]\big[x^2 - 4x + 4(2x+1)\big] &= (x^2 - 12x - 4)(x^2 + 4x + 4)\\ &= (x^2 - 12x - 4)(x+2)^2.\end{aligned}$$

所以

$$\mathrm{f}(x) = \frac{(x+2)^2(x^2 - 12x - 4)}{x^2(x-4)}.$$

因式 $(x+2)^2$ 显示 $x = -2$ 是 $\mathrm{f}(x) = 0$ 的二重根，而且 $x = -2$ 并非分母的零点，正如题目所求。另外两个根来自 $x^2 - 12x - 4 = 0$:

$$x = 6 \pm 2\sqrt{10} \approx 12.32,\ -0.32.$$

第 (iii) 部分

我们先整理绘画草图所需的各项特征。

在 $x = 0$ 附近。当 $x \to 0$ 时，$\mathrm{f}(x) \approx -\dfrac{16}{x^2 \cdot (-4)} = \dfrac{4}{x^2}$，所以从两侧趋近时都有 $\mathrm{f}(x) \to +\infty$。

在 $x = 4$ 附近。当 $x \to 4$ 时，$\mathrm{f}(x) \approx -\dfrac{16 \cdot 81}{16(x-4)} = -\dfrac{81}{x-4}$。所以当 $x \to 4^-$ 时 $\mathrm{f}(x) \to +\infty$，而当 $x \to 4^+$ 时 $\mathrm{f}(x) \to -\infty$。

曲线在斜渐近线的哪一侧？$\mathrm{f}(x) - (x-4) = -16(2x+1)^2/\big(x^2(x-4)\big)$。当 $x < 4$ 时（$x = 0$ 和 $x = -\frac{1}{2}$ 除外）这个量为正，当 $x > 4$ 时为负。所以在 $x = 4$ 的左方，曲线位于 $y = x - 4$ 之上，只在 $x = -\frac{1}{2}$ 处与它相切；在右方，曲线则位于它之下。

f 的正负号。从因式分解后的形式可见，因式 $(x+2)^2$ 和 x^2 永不为负。所以 f 的正负号与 $(x^2 - 12x - 4)/(x-4)$ 的正负号相同：

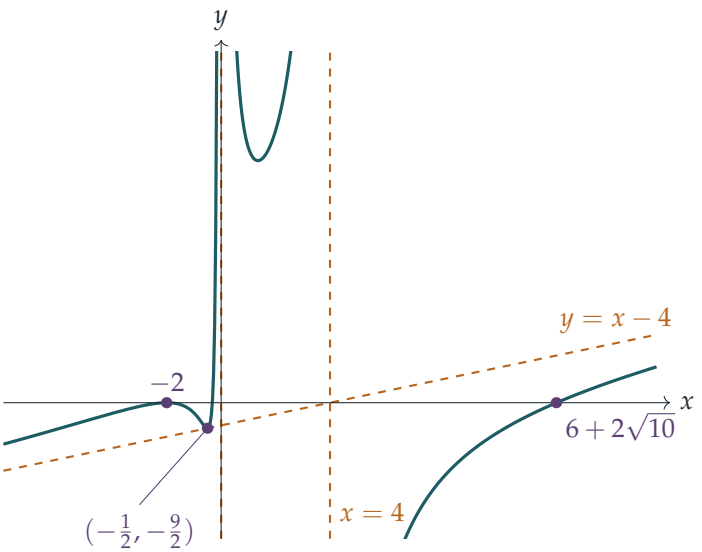
- 当 $x < 6 - 2\sqrt{10}$ 时（$\mathrm{f}(-2) = 0$ 除外），以及当 $4 < x < 6 + 2\sqrt{10}$ 时，$\mathrm{f} < 0$；
- 当 $6 - 2\sqrt{10} < x < 4$ 时（$x \neq 0$），以及当 $x > 6 + 2\sqrt{10}$ 时，$\mathrm{f} > 0$。

曲线在 $x = -2$ 的两侧都保持为负，并在该处与 x 轴相切，所以 $(-2, 0)$ 是局部极大点（local maximum）。

转向点（*turning point*）。STEP 的草图并不需要这些转向点，但有了它们，草图会更准确。求导可得

$$\mathrm{f}'(x) = 1 + \frac{16(2x+1)(2x^2+3x-8)}{x^3(x-4)^2}.$$

在 $(-2,\ 6 - 2\sqrt{10})$ 上，曲线的起点和终点高度都是 0，中间为负，所以在这区间内有一个极小点。以数值计算，这个极小点约在 $(-0.51,\ -4.50)$，刚好在切点的左方。在 $(0, 4)$ 上，曲线在两端都趋向 $+\infty$，所以这里也有一个极小点，约在 $(1.35,\ 42.7)$。在 $\mathrm{f}'(x) = 0$ 中消去分母，得 $(x+2)(x^4 - 10x^3 + 100x^2 - 72x - 64) = 0$。这个四次式恰好只有上述两个实根，所以再没有其他转向点。特别地，f 在整个 $x > 4$ 的范围内都是递增的。



这幅草图是按比例绘画的，因此 $x = -2$ 与 $x \approx -0.32$ 之间的凹陷看起来很浅。那里的最低点只在轴下约 4.5 处，而曲线在 $\left(-\frac{1}{2}, -\frac{9}{2}\right)$ 与渐近线相切。

验算

- $f(-\frac{1}{2}) = -\frac{9}{2} - 0 = -\frac{9}{2}$，这一点在 $y = x - 4$ 上。在 $x = -\frac{1}{2}$ 处以数值方法求导数，得 1.000。
- 展开 $(x+2)^2(x^2 - 12x - 4)$，并与 $x^2(x-4)^2 - 16(2x+1)^2$ 比较：两者都是 $x^4 - 8x^3 - 48x^2 - 64x - 16$。
- $f(1000) - (1000 - 4) \approx -0.064 \approx -64/1000$，与 $64/x$ 这个估算所预测的一致。

常见错误

- 忘记了在 $x = 0$ 附近，曲线在两侧都趋向 $+\infty$，原因是有 x^2 这个因式。两条垂直渐近线的表现并不相同。
- 证明相切时，只证明了直线与曲线相交。你还需要证明斜率相同，或者等价地，证明有重根。
- 在第 (ii) 部分把四次的分子展开，然后尝试凭观察找出它的根，而没有看出那是一个平方差。

STEP 之外

一条直线与一条曲线相交而出现“二重根”，恰好就是两者相切的时候。这就是**相交重数**（multiplicity of intersection）的概念，它是代数几何（algebraic geometry）的基础，并引出一些结果，例如关于两条曲线相交多少次的贝祖定理（Bézout's theorem）。有理函数（rational function）在无穷远处可以用它的多项式部分来近似，而余项的大小（这里是 $64/x$）说明了曲线趋近渐近线的速度有多快。这是对渐近展开（asymptotic expansion）的初步体验。Siklos 的 *Advanced Problems in Mathematics* 有更多这种风格的曲线描绘题目。

详细解答

STEP III 2004，第 3 题

纯数学 · 凸性与级数的界限

STEP III 2004，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/04-s3-q3

关键思路

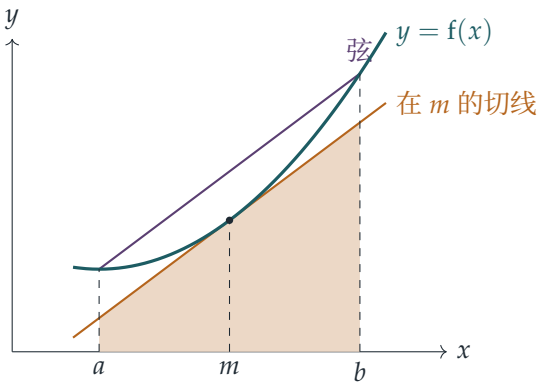
若 $\mathrm{f}'' > 0$，曲线向上弯。因此曲线位于每一条切线（tangent）之上，亦位于每一条弦（chord）之下。在中点的切线，以及连接两端的弦，各自截出一个梯形（trapezium），其面积很容易写出，而这两个面积把积分夹在中间。取 $\mathrm{f}(x) = 1/x^2$ 及区间 $[n-1, n]$，积分便会裂项相消（telescope）。再把各不等式对 n 相加，便得出 $\sum 1/n^2$ 的上下界。

解答

积分的两个界

设 $m = \frac{1}{2}(a+b)$。由于在 $[a,b]$ 上 $\mathrm{f}'' > 0$，斜率 f' 严格递增，所以曲线是凸的（convex）。由此得到两个事实。

- 曲线位于它在 m 的切线之上。设 $t(x) = \mathrm{f}(m) + \mathrm{f}'(m)(x-m)$ 及 $g = \mathrm{f} - t$。则 $g(m) = g'(m) = 0$，且 $g'' = \mathrm{f}'' > 0$，所以 g' 递增。因此在 m 的左边 $g' < 0$，在右边 $g' > 0$，故 g 在 m 取得严格极小值 0，而当 $x \neq m$ 时 $\mathrm{f}(x) > t(x)$。
- 曲线位于它的弦之下。设 $c(x)$ 为通过 $(a, \mathrm{f}(a))$ 和 $(b, \mathrm{f}(b))$ 的直线，并设 $h = \mathrm{f} - c$。则 $h(a) = h(b) = 0$ 且 $h'' > 0$。假如 h 在 (a,b) 内取得某个 ≥ 0 的值，它便会在该区间内有一个极大值，而在该处 $h'' \leq 0$，这是不可能的。所以当 $a < x < b$ 时 $\mathrm{f}(x) < c(x)$。



现在比较 $[a,b]$ 上的面积。切线下的面积是一个阔度为 $b-a$ 的梯形。它的平均高度就是 t 在中点的高度 $t(m)=\mathrm{f}(m)$，所以面积是 $(b-a)\,\mathrm{f}(m)$（阴影部分）。弦下的面积是一个梯形，两条平行边的长度为 $\mathrm{f}(a)$ 和 $\mathrm{f}(b)$，所以面积是 $(b-a)\frac{1}{2}\big(\mathrm{f}(a)+\mathrm{f}(b)\big)$。除了个别孤立点外，曲线严格位于两条直线之间。所以

$$(b-a)\,\mathrm{f}\Big(\frac{a+b}{2}\Big) < \int_a^b \mathrm{f}(x)\,\mathrm{d}x < (b-a)\,\frac{\mathrm{f}(a)+\mathrm{f}(b)}{2}.$$

（草图所示的 f 是正的。但论证其实只用到 $\int(\mathrm{f}-t)>0$ 和 $\int(c-\mathrm{f})>0$，所以 f 的正负并无影响。）

选取 a、b 和 f

中间的一项 $\frac{1}{n-1}-\frac{1}{n}$ 看来像 $\big[-\frac{1}{x}\big]_{n-1}^{n}$，所以取 $\mathrm{f}(x)=1/x^2$、$a=n-1$、$b=n$。于是 $b-a=1$，而且因为 $n-1\geq 1>0$，在 $[n-1,n]$ 上有 $\mathrm{f}''(x)=6/x^4>0$。此外

$$\int_{n-1}^{n}\frac{\mathrm{d}x}{x^2}=\frac{1}{n-1}-\frac{1}{n},\qquad \mathrm{f}\Big(\frac{2n-1}{2}\Big)=\frac{4}{(2n-1)^2},$$

所以由第一部分得

$$\frac{4}{(2n-1)^2} < \frac{1}{n-1}-\frac{1}{n} < \frac{1}{2}\Big(\frac{1}{n^2}+\frac{1}{(n-1)^2}\Big)\qquad (n\geq 2),$$

这正是所要证明的。

推导关于 1 的两个界

设 $S=\sum_{n=1}^{\infty}\frac{1}{n^2}$。这个级数收敛（converge），因为当 $n\geq 2$ 时有 $\frac{1}{n^2}<\frac{1}{n(n-1)}=\frac{1}{n-1}-\frac{1}{n}$，而这些项会裂项相消。把 $n=2,3,\ldots,N$ 的不等式相加。中间各项裂项相消后得 $1-\frac{1}{N}$，它趋向 1。

想一想 当我们令 $N\to\infty$ 时，严格不等式通常会变成 $\leq$。怎样才能保持严格？

看看当中的差距。对每个 $n \geq 2$，差距 $d_n = \left(\frac{1}{n-1} - \frac{1}{n}\right) - \frac{4}{(2n-1)^2}$ 是正数。所以这些差距的无穷和至少等于它的第一项，即 $d_2 = \frac{1}{2} - \frac{4}{9} = \frac{1}{18} > 0$。这里每个级数都收敛，所以可以逐项求和，而两个和之差严格为正。同样的论证亦适用于右边的不等式。

在左边，$n = 2, 3, 4, \ldots$ 给出 $2n - 1 = 3, 5, 7, \ldots$，所以

$$4\left(\frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \cdots\right) < 1.$$

在右边，

$$\sum_{n=2}^{\infty} \frac{1}{2}\left(\frac{1}{n^2} + \frac{1}{(n-1)^2}\right) = \frac{1}{2}(S-1) + \frac{1}{2}S = \frac{1}{2} + (S-1),$$

由此得

$$1 < \frac{1}{2} + \left(\frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots\right).$$

与奇数平方作比较

把左边的级数每两个相邻的项分成一组，每组先是一个奇数项，然后是一个偶数项。对 $k \geq 2$，

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

因为 $2k > 2k - 1 > 0$。取 $k = 2, 3, 4, \ldots$，左边的各组是 $(3^2, 4^2)$、$(5^2, 6^2), \ldots$，合起来刚好把 $\frac{1}{3^2}, \frac{1}{4^2}, \frac{1}{5^2}, \ldots$ 每一项各用一次。（可以这样分组，是因为所有项都是正数，而且级数收敛。）右边则是 $\frac{1}{3^2}, \frac{1}{5^2}, \ldots$。对 k 求和，并如前所述保持严格不等，便得

$$\frac{1}{2}\left(\frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \frac{1}{6^2} + \cdots\right) < \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \cdots.$$

S 的上下界

下界（*lower bound*）。由右边的不等式，$1 < \frac{1}{2} + (S-1)$，所以 $S > \frac{3}{2}$。

上界（*upper bound*）。设 $T = \frac{1}{3^2} + \frac{1}{4^2} + \cdots = S - 1 - \frac{1}{4}$。把刚才的比较与左边的不等式结合，得

$$\frac{1}{2}T < \frac{1}{3^2} + \frac{1}{5^2} + \cdots < \frac{1}{4},$$

所以 $T < \frac{1}{2}$。因此 $S = 1 + \frac{1}{4} + T < \frac{7}{4}$。总括而言，

$$\frac{3}{2} < \sum_{n=1}^{\infty} \frac{1}{n^2} < \frac{7}{4}.$$

验算

▶ 真确值是 $S = \pi^2/6 \approx 1.6449$，位于 $(1.5, 1.75)$ 之内。

- 数值上，$4\sum_{k\ge 2}(2k-1)^{-2} = 4(\pi^2/8 - 1) \approx 0.935 < 1$，而 $\frac{1}{2} + (S-1) \approx 1.145 > 1$。
- 至于那个比较，$\frac{1}{2}T \approx 0.197 < 0.234 \approx \sum_{k\ge 2}(2k-1)^{-2}$。
- 试代 $n = 2$：$\frac{4}{9} \approx 0.444 < \frac{1}{2} < \frac{5}{8} = 0.625$。以精确分数对 $2 \le n \le 2000$ 逐一检查，三个不等式全部成立。

常见错误

- 忘记了缺少的 $\frac{1}{4}$：括号由 $\frac{1}{3^2}$ 开始，所以 $T = S - \frac{5}{4}$，而不是 $S - 1$。若用 $S - 1$，便会得出 $S < \frac{3}{2}$ 这个错误结论。
- 把不等号的方向弄反。对于向上弯的曲线，切线（中点）估算偏小，而弦（梯形）估算偏大。
- 过渡到无穷和时，不加说明便保留“$<$”，没有解释为何它取极限后仍然成立。

STEP 之外

这两个界就是中点法则（midpoint rule）和梯形法则（trapezium rule）。凸性告诉你每个法则的误差偏向哪一边，而泰勒定理（Taylor's theorem）则显示两者的误差都以 $(b-a)^3$ 的速度缩小。精确值 $\sum 1/n^2 = \pi^2/6$ 是欧拉（Euler）对巴塞尔问题（Basel problem）的解答。大学的第一门分析课程会用傅里叶级数（Fourier series）证明它，也会把“严格不等式取极限”这一步严谨化。Abbott 的 *Understanding Analysis* 涵盖了级数的收敛和凸函数这两个课题。

详细解答

STEP III 2004，第 4 题

纯数学 · 相切圆与级数

STEP III 2004，第 4 题

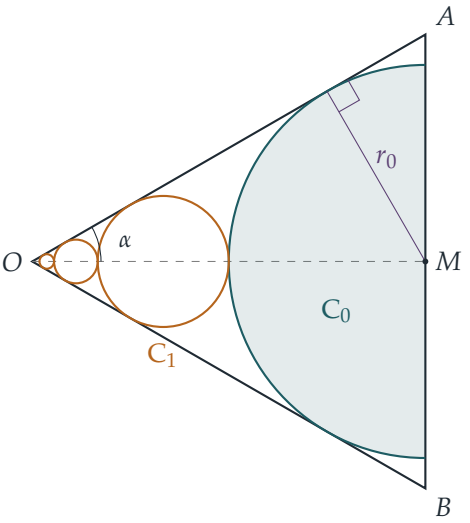
继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

step.maths.org/questions/04-s3-q4

关键思路

凡是同时与 OA 和 OB 相切的圆，圆心都在角 AOB 的角平分线（angle bisector）上，而圆心与 O 的距离由半径决定：在 O 处夹角为 α 的直角三角形给出距离 $= r/\sin\alpha$。把“两圆相切”写成关于这些距离的方程，便把相邻的半径联系起来，从而得知各半径构成等比数列（geometric progression）。于是各面积构成等比级数（geometric series），而最后一部分是单变量的求极大值问题（或者选一个合适的 α 值即可）。

解答



半径之比

设 M 为 AB 的中点。由于 $OA = OB$，直线 OM 平分角 AOB，并且垂直于 AB；角 AOM 和角 BOM 各为 α。

任取一个半径为 r、同时与直线 OA 和 OB 相切的圆。它的圆心与这两条直线等距，所以位于角平分线 OM 上。设圆心为 P，圆与 OA 相切于 T，则 $PT \perp OA$，$PT = r$，且角 $POT = \alpha$，所以在直角三角形 OTP 中

$$OP = \frac{r}{\sin\alpha}.$$

这个结果适用于 C_0（圆心为 M），也适用于每一个 C_n。记 $d_n = r_n/\sin\alpha$ 为 O 到 C_n 圆心的距离。

圆 C_{n+1} 位于 C_n 与 O 之间的角落里（它必定在角的内部，并在 C_n 靠近 O 的一侧，因为它与边 OA 和 OB 相切，而不是与它们的延长线相切）。两圆外切（touch externally）时，圆心相距 $r_n + r_{n+1}$，而这里两个圆心都在 OM 上，所以

$$d_n - d_{n+1} = r_n + r_{n+1}.$$

代入 $d_n = r_n/\sin\alpha$，再乘以 $\sin\alpha$：

$$\begin{aligned} r_n - r_{n+1} &= (r_n + r_{n+1})\sin\alpha \\ r_n(1-\sin\alpha) &= r_{n+1}(1+\sin\alpha), \end{aligned}$$

因此

$$\frac{r_{n+1}}{r_n} = \frac{1-\sin\alpha}{1+\sin\alpha},$$

这正是所要证明的结果。$n = 0$ 时这个论证同样成立：C_1 与半圆的弧形部分相切，而这部分正是 C_0 面向 O 的那一半。

总面积

记 $k = \dfrac{1-\sin\alpha}{1+\sin\alpha}$。由于 $0 < \sin\alpha < 1$，我们有 $0 < k < 1$，且 $r_n = k^n r_0$。半圆的面积为 $\frac{1}{2}\pi r_0^2$，而 C_n 的面积为 $\pi k^{2n} r_0^2$。级数 $\sum_{n\geq 1} k^{2n}$ 是公比为 $k^2 < 1$ 的等比级数，所以它收敛（converge）于 $k^2/(1-k^2)$。于是

$$S = \pi r_0^2 \left(\frac{1}{2} + \frac{k^2}{1-k^2}\right).$$

现在

$$1 - k^2 = \frac{(1+\sin\alpha)^2 - (1-\sin\alpha)^2}{(1+\sin\alpha)^2} = \frac{4\sin\alpha}{(1+\sin\alpha)^2}, \qquad \text{所以} \qquad \frac{k^2}{1-k^2} = \frac{(1-\sin\alpha)^2}{4\sin\alpha}.$$

因此

$$S = \pi r_0^2 \frac{2\sin\alpha + 1 - 2\sin\alpha + \sin^2\alpha}{4\sin\alpha} = \frac{1+\sin^2\alpha}{4\sin\alpha}\pi r_0^2,$$

这就是我们要的结果。

与三角形比较

由第一部分，$OM = d_0 = r_0/\sin\alpha$。在直角三角形 OMA 中，$AM = OM\tan\alpha$，所以三角形 OAB 的面积为

$$T = OM\cdot AM = OM^2\tan\alpha = \frac{r_0^2}{\sin^2\alpha}\cdot\frac{\sin\alpha}{\cos\alpha} = \frac{r_0^2}{\sin\alpha\cos\alpha}.$$

（半圆确实放得进底边上：$r_0 = OM\sin\alpha < OM\tan\alpha = AM$。）两式相除，得

$$\frac{S}{T} = \frac{\pi}{4}(1+\sin^2\alpha)\cos\alpha.$$

想一想 我们应该试哪一个 α 值?

当 $\alpha\to 0$ 时，这个比趋向 $\pi/4\approx 0.785$，仅仅略小于 0.8，所以非常瘦长的三角形并不可行。很自然的第一个尝试，是令 O 处为直角，即 $\alpha = \pi/4$。这时 $\sin^2\alpha = \frac12$，$\cos\alpha = 1/\sqrt{2}$，所以

$$\frac{S}{T} = \frac{\pi}{4}\cdot\frac{3}{2}\cdot\frac{1}{\sqrt{2}} = \frac{3\pi}{8\sqrt{2}}.$$

这个值大于 $\frac45$，当且仅当 $15\pi > 32\sqrt{2}$。两边均为正数，所以可以平方：条件变成 $225\pi^2 > 2048$，即 $\pi^2 > 9.102\ldots$。由于 $\pi > 3.1$，我们有 $\pi^2 > 9.61$，所以不等式成立。因此当 $\alpha = \pi/4$ 时（并且根据连续性（continuity），对所有接近 $\pi/4$ 的 α 亦然），这些圆覆盖了三角形超过五分之四的面积。

顺带一提，最佳的 α 可由求 $f(c) = (2-c^2)c$ 的极大值得出，其中 $c = \cos\alpha$：$f'(c) = 2-3c^2 = 0$ 给出 $\cos\alpha = \sqrt{2/3}$，即 $\sin\alpha = 1/\sqrt{3}$，而这个比的极大值为 $\pi\sqrt{6}/9\approx 0.855$。

验算

- 当 $\alpha\to\pi/2$ 时，三角形变得扁平，且 $k\to 0$：各圆 C_n 消失，$S\to\frac12\pi r_0^2$，与 $\frac{1+1}{4}\pi r_0^2$ 吻合。
- 取 $\alpha = \pi/6$，则 $k = \frac13$，$S = \pi r_0^2\left(\frac12 + \frac{1/9}{8/9}\right) = \frac58\pi r_0^2$；公式给出 $\frac{1+1/4}{2}\pi r_0^2 = \frac58\pi r_0^2$。
- 以数值方法验证：对随机选取的 α，直接按相切条件构造 3000 个圆，所得结果与两条公式吻合，误差约为 10^{-15}；而 $\alpha = \pi/4$ 时的比为 0.8330。

常见错误

- 把 C_0 当作整个圆来计算：它是半圆，这正是 $\frac{1}{2}$ 出现的原因。
- 使用 $d_n + d_{n+1} = r_n + r_{n+1}$ 或 $d_n - d_{n+1} = r_n - r_{n+1}$：这些圆是外切的，而且两个圆心都在 O 的同一侧。
- 在最后一部分尝试很小的 α：这时比趋向 $\pi/4 < \frac{4}{5}$。

STEP 之外

这一串圆是由单一个以 O 为中心、比例因子为 k 的位似变换（dilation，即放大 enlargement）生成的：它把角落区域映射到自身，并把每个圆映射到下一个圆。这类自相似（self-similar）的构形通向分形（fractal）和阿波罗尼奥斯圆填充（Apollonian circle packing），其中互相相切的圆的曲率（curvature）满足笛卡儿圆定理（Descartes' circle theorem）。反演（inversion）是一种把圆变成直线的变换，能令许多这类相切问题变得容易；Needham 的 *Visual Complex Analysis* 对此有精彩的论述。

详细解答

STEP III 2004，第 5 题

纯数学 · 三角方程

> **STEP III 2004，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/04-s3-q5

关键思路

两个余弦值相等，当且仅当两个角在相差一个正负号及 2π 的整数倍之下相同。改用正切（tangent）之后，用来区分 θ 与 $\theta+\pi$ 的信息便会丢失，所以这个推论只在一个方向成立，而这正是反例（counterexample）的用意所在。在第 (i) 及 (ii) 部分，我们把 $a\cos x + b\sin x$ 写成 $R\cos(x-\alpha)$ 的形式，利用第一个结果以 $\tan\omega$ 表示 $\tan x$，然后根据各角的大小选出正确的象限（quadrant）。

解答

开首的结果

若 $\cos(x-\alpha)=\cos\beta$，则对某个整数 n 有 $x-\alpha=\pm\beta+2n\pi$，因为余弦是偶函数（even function），且周期（period）为 2π。所以

$$x=\alpha+\beta+2n\pi \quad 或 \quad x=\alpha-\beta+2n\pi.$$

由于 tan 的周期为 π，2π 当然也是它的周期，由此可得 $\tan x=\tan(\alpha+\beta)$ 或 $\tan x=\tan(\alpha-\beta)$（只要这些正切值有定义）。

逆命题（converse）不成立，因为 tan 的周期为 π，而 cos 却不是。取 $\alpha=\beta=0$ 及 $x=\pi$。这时 $\tan x=0=\tan(\alpha+\beta)$，但

$$\cos(x-\alpha)=\cos\pi=-1\neq 1=\cos\beta.$$

一般而言，$x=\alpha+\beta+\pi$ 会给出 $\cos(x-\alpha)=-\cos\beta$。

关于 ω 有两项事实稍后会用得着。由于 $\tan\omega=\frac{4}{3}$，我们有 $\cos\omega=\frac{3}{5}$ 及 $\sin\omega=\frac{4}{5}$。又

由于 $\frac{4}{3} > 1$，我们亦有 $\frac{\pi}{4} < \omega < \frac{\pi}{2}$。

第 (i) 部分

这里 $\sqrt{1^2+7^2} = 5\sqrt{2}$，所以我们把等式两边除以 $5\sqrt{2}$：

$$\tfrac{1}{5\sqrt{2}}\cos x - \tfrac{7}{5\sqrt{2}}\sin x = \tfrac{1}{\sqrt{2}}.$$

设 φ 为满足 $\cos\varphi = \frac{1}{5\sqrt{2}}$ 及 $\sin\varphi = \frac{7}{5\sqrt{2}}$ 的锐角（acute angle），于是 $\tan\varphi = 7$。方程变成

$$\cos(x+\varphi) = \cos\tfrac{\pi}{4},$$

这正是开首的结果中取 $\alpha = -\varphi$ 及 $\beta = \frac{\pi}{4}$ 的情况。利用 $\tan(A \pm B) = \frac{\tan A \pm \tan B}{1 \mp \tan A \tan B}$，

$$\tan(\alpha+\beta) = \tan\left(\tfrac{\pi}{4} - \varphi\right) = \frac{1-7}{1+7} = -\frac{3}{4},$$
$$\tan(\alpha-\beta) = -\tan\left(\varphi + \tfrac{\pi}{4}\right) = -\frac{7+1}{1-7} = \frac{4}{3}.$$

想一想 这两个正切值各自对应 $[0, 2\pi]$ 内的两个角，但方程在该范围内只有两个解。是哪两个呢?

正如反例所示，单凭正切值无法判断，所以我们回到 $x = \alpha \pm \beta + 2n\pi$，找出 $\alpha \pm \beta$ 实际上位于何处。由于 $\tan\varphi = 7 > 1$，我们有 $\frac{\pi}{4} < \varphi < \frac{\pi}{2}$。

- $\alpha + \beta = \frac{\pi}{4} - \varphi$ 位于 $(-\frac{\pi}{4}, 0)$ 内，其正切值为 $-\frac{3}{4}$。而 $\tan(\frac{\pi}{2} - \omega) = \cot\omega = \frac{3}{4}$，且 $\frac{\pi}{2} - \omega \in (0, \frac{\pi}{4})$。因此 $\alpha + \beta = \omega - \frac{\pi}{2}$，加上 2π 便得 $x = \frac{3\pi}{2} + \omega$。
- $\alpha - \beta = -\varphi - \frac{\pi}{4}$ 位于 $(-\frac{3\pi}{4}, -\frac{\pi}{2})$ 内，其正切值为 $\frac{4}{3}$。角 $\omega - \pi$ 同样位于这个区间内，正切值亦为 $\frac{4}{3}$。在长度小于 π 且不包含渐近线（asymptote）的区间上，正切函数是一对一（one-to-one）的，所以 $\alpha - \beta = \omega - \pi$。加上 2π 便得 $x = \pi + \omega$。

由于 $0 < \omega < \frac{\pi}{2}$，这两个值都在 $[0, 2\pi]$ 内，而 n 取其他值时都不会得出此范围内的值。因此

$$x = \pi + \omega \quad \text{或} \quad x = \tfrac{3\pi}{2} + \omega.$$

验证一下：当 $x = \pi + \omega$ 时，$\cos x - 7\sin x = -\frac{3}{5} + \frac{28}{5} = 5$；当 $x = \frac{3\pi}{2} + \omega$ 时，$\cos x = \sin\omega$ 及 $\sin x = -\cos\omega$，由此得 $\frac{4}{5} + \frac{21}{5} = 5$。

第 (ii) 部分

这里 $\sqrt{2^2+11^2} = \sqrt{125} = 5\sqrt{5}$，所以

$$\tfrac{2}{5\sqrt{5}}\cos x + \tfrac{11}{5\sqrt{5}}\sin x = \tfrac{2}{\sqrt{5}}.$$

设 α 为锐角，满足 $\cos\alpha = \frac{2}{5\sqrt{5}}$ 及 $\sin\alpha = \frac{11}{5\sqrt{5}}$，于是 $\tan\alpha = \frac{11}{2}$。设 β 为锐角，满足 $\cos\beta = \frac{2}{\sqrt{5}}$，于是 $\sin\beta = \frac{1}{\sqrt{5}}$ 及 $\tan\beta = \frac{1}{2}$。方程就是 $\cos(x-\alpha) = \cos\beta$，而

$$\tan(\alpha-\beta) = \frac{\frac{11}{2}-\frac{1}{2}}{1+\frac{11}{4}} = \frac{5}{15/4} = \frac{4}{3},$$
$$\tan(\alpha+\beta) = \frac{\frac{11}{2}+\frac{1}{2}}{1-\frac{11}{4}} = \frac{6}{-7/4} = -\frac{24}{7}.$$

第二个值就是 $\tan 2\omega$，因为

$$\tan 2\omega = \frac{2\cdot\frac{4}{3}}{1-\frac{16}{9}} = \frac{8/3}{-7/9} = -\frac{24}{7}.$$

我们再次确定各角的位置。α 和 β 都是锐角，所以

- $\alpha-\beta \in (-\frac{\pi}{2}, \frac{\pi}{2})$，正切值为 $\frac{4}{3}$，所以 $\alpha-\beta = \omega$；
- $\alpha+\beta \in (0,\pi)$，正切值为负，所以它位于 $(\frac{\pi}{2},\pi)$ 内。由于 $\frac{\pi}{4} < \omega < \frac{\pi}{2}$，角 2ω 同样位于 $(\frac{\pi}{2},\pi)$ 内，且正切值相同，所以 $\alpha+\beta = 2\omega$。

ω 和 2ω 都在 $[0,2\pi]$ 内，而加上 $\pm 2\pi$ 便会超出这个范围。因此方程的解为

$$x = \omega \quad 及 \quad x = 2\omega,$$

其中一个解是另一个的两倍，正如题目所求。直接验算：$x=\omega$ 给出 $\frac{6}{5}+\frac{44}{5} = 10$。当 $x = 2\omega$ 时，$\cos 2\omega = \frac{9-16}{25} = -\frac{7}{25}$ 及 $\sin 2\omega = \frac{24}{25}$，由此得 $-\frac{14}{25}+\frac{264}{25} = 10$。

验算

- 每条方程在 $[0,2\pi]$ 内应该恰好有两个根，因为 $5 < 5\sqrt{2}$ 及 $10 < 5\sqrt{5}$，所以右边的值严格介乎 $-R$ 与 R 之间。
- 以数值计算，$\omega \approx 0.9273$。用变号搜索法可找到 (i) 的根为 4.0689 及 5.6397，即 $\pi+\omega$ 及 $\frac{3\pi}{2}+\omega$。对于 (ii)，则找到 0.9273 及 1.8546，即 ω 及 2ω。

常见错误

- 解 $\tan x = \frac{4}{3}$ 或 $\tan x = -\frac{3}{4}$，然后把 $[0,2\pi]$ 内的四个角全部列出。其中两个是增根，原因正是反例所揭示的那一点。
- 在 (i) 中把 φ 的正负号弄错，例如把 $\cos x - 7\sin x$ 写成 $\cos(x-\varphi)$ 而不是 $\cos(x+\varphi)$。

STEP 之外

这里的困难在于 tan 不是单射（injective），所以“$\tan x = \tan y$”只能告诉你 x 模 π 的值。在复分析（complex analysis）中，同样的问题出现在辐角（argument）以及 $\log z$ 的分支（branch）的定义之中：在那里，角只能确定到相差 2π 的整数倍。在这个框架下，$a\cos x + b\sin x = \mathrm{Re}\left((a - \mathrm{i}b)\mathrm{e}^{\mathrm{i}x}\right)$，而 R 和 α 不过是 $a + \mathrm{i}b$ 的模（modulus）和辐角。Needham 的 *Visual Complex Analysis*（《复分析：可视化方法》）把这种几何观点阐述得十分精彩。

详细解答

STEP III 2004，第 6 题

纯数学 · 递推关系

> **STEP III 2004，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/04-s3-q6

关键思路

F_n 这个式子是附在数列每一对相邻两项上的量，而开首的恒等式（identity）正好告诉我们，由一对相邻项到下一对时，它在甚么情况下保持不变。对于 (i) 中的递推关系（recurrence），第二个因式为零，所以 F_n 是一个不变量（invariant），我们只需求出它在起点的值。第 (ii) 部分把论证倒过来做：若 F_n 是常数，乘积必定为零，所以每一步两个因式之中总有一个是零。于是每一步都有一个选择，而周期（period）为 4 的数列（sequence）就是靠选得巧妙而得出的。

解答

恒等式

把两项都写出来，w_{n-1}^2 项互相抵消：

$$\begin{aligned} F_n - F_{n-1} &= \left(w_n^2 + w_{n-1}^2 - 4w_n w_{n-1}\right) - \left(w_{n-1}^2 + w_{n-2}^2 - 4w_{n-1}w_{n-2}\right) \\ &= \left(w_n^2 - w_{n-2}^2\right) - 4w_{n-1}\left(w_n - w_{n-2}\right) \\ &= (w_n - w_{n-2})(w_n + w_{n-2}) - 4w_{n-1}(w_n - w_{n-2}) \\ &= (w_n - w_{n-2})(w_n + w_{n-2} - 4w_{n-1}), \end{aligned}$$

此式对 $n \geq 2$ 成立（因为需要 w_{n-2} 存在），正是所求。

还有一点观察，往后会一直用到：F_n 对 w_n 和 w_{n-1} 是对称（symmetric）的。把两者对调，$w_n^2 + w_{n-1}^2 - 4w_n w_{n-1}$ 保持不变。

第 (i) 部分

取 $w_n = u_n$，并设 F_n 为相应的量。当 $n \geq 2$ 时，递推关系给出 $u_n + u_{n-2} - 4u_{n-1} = 0$，所以恒等式中的第二个因式为零，于是

$$F_n = F_{n-1} \quad (n \geq 2).$$

因此对所有 $n \geq 1$ 都有 $F_n = F_1$，而

$$F_1 = u_1^2 + u_0^2 - 4u_1u_0 = 4 + 1 - 8 = -3.$$

所以 $u_n^2 + u_{n-1}^2 - 4u_nu_{n-1} = -3$，即对所有 $n \geq 1$ 都有 $u_n^2 + u_{n-1}^2 = 4u_nu_{n-1} - 3$，正是所求。（这个数列是 $1, 2, 7, 26, 97, \ldots$。）

第 (ii)(a) 部分

把 $n = 1$ 和 $v_0 = 1$ 代入 $(*)$：

$$v_1^2 + 1 = 4v_1 - 3 \iff v_1^2 - 4v_1 + 4 = 0 \iff (v_1 - 2)^2 = 0,$$

所以 $v_1 = 2$。这里没有选择的余地，因为它是重根（repeated root）。

现在取 $w_n = v_n$。条件 $(*)$ 说的正是：对所有 $n \geq 1$，$F_n = -3$。所以当 $n \geq 2$ 时，F_n 和 F_{n-1} 都等于 -3，由恒等式得

$$0 = F_n - F_{n-1} = (v_n - v_{n-2})(v_n + v_{n-2} - 4v_{n-1}).$$

两个实数的乘积为零，仅当其中一个为零，所以 $v_n = v_{n-2}$ 或 $v_n = 4v_{n-1} - v_{n-2}$，正是所求。

> ? **想一想** 这样就完了吗？两个选项是否真的都给出合乎条件的下一项？

是的。固定 v_{n-1}，条件 $(*)$ 就是以 $x = v_n$ 为未知数的二次方程（quadratic）$x^2 - 4v_{n-1}x + (v_{n-1}^2 + 3) = 0$。我们知道 $x = v_{n-2}$ 是一个根，因为 F 是对称的，而且 $F_{n-1} = -3$。两根之和为 $4v_{n-1}$，所以另一个根是 $4v_{n-1} - v_{n-2}$。因此 (a) 中的两个选项恰好就是这两个根。每一步我们都可以任选其一，所得结果总是满足 $(*)$。（当两根重合时，例如 $n = 1$ 的情况，就只有一个选项。）

第 (ii)(b) 部分

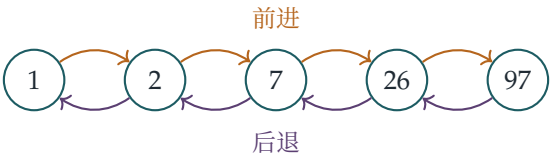
这里 $v_0 = 1$，符合要求，而每一对相邻项 (v_{n-1}, v_n) 不是 $(1, 2)$ 就是 $(2, 1)$。由于 F 是对称的，两者都给出

$$v_n^2 + v_{n-1}^2 - 4v_nv_{n-1} = 1 + 4 - 8 = -3,$$

所以 $(*)$ 对所有 $n \geq 1$ 成立。用 (a) 的说法，这个数列每一步都取 $v_n = v_{n-2}$ 这个选项。

第 (ii)(c) 部分

我们有 $v_0 = 1$ 和 $v_1 = 2$。由 (a)，其后每一项要么等于前两位的那一项（“后退”），要么等于 $4\times$（前一项）$-$（再前一项）（“前进”）。由 $(1,2)$ 出发，前进得出 7，然后是 26、97，如此类推，这就是数列 u_n。后退则把最近一次前进抵消。



所以这个数列就是沿着这条链的一次游走，由 1 出发，每次移到相邻的一个数。由于 1 只有一个邻居，游走到达 1 之后必定立刻走到 2。

要在 $v_0 = 1$ 的情况下得到周期 4，我们需要 $v_4 = 1$。(b) 中周期为 2 的数列只是在 1 和 2 之间来回往返。要做出新的东西，我们多走一步：

$$v_0 = 1, \quad v_1 = 2, \quad v_2 = 4\cdot 2 - 1 = 7, \quad v_3 = v_1 = 2, \quad v_4 = v_2 \text{ 或 } 4\cdot 2 - 7 = 1.$$

选取 $v_4 = 1$，然后不断重复，得出

$$v_n = \begin{cases} 1 & n \equiv 0 \pmod 4, \\ 2 & n \text{ 为奇数}, \\ 7 & n \equiv 2 \pmod 4, \end{cases} \qquad \text{即} \quad 1,2,7,2,1,2,7,2,\ldots$$

直接验证 $(*)$：每一对相邻项都是 $(1,2)$、$(2,7)$、$(7,2)$ 或 $(2,1)$。由对称性，我们只需验证

$$1 + 4 - 8 = -3 \quad \text{和} \quad 4 + 49 - 56 = -3,$$

两式都成立。所以 $(*)$ 对所有 $n \geq 1$ 成立，而这个数列的周期为 4（而不是 2，因为 $v_2 \neq v_0$）。

事实上，周期恰好为 4 的答案只有这一个。在 1, 2 之后，游走必须有 $v_4 = 1$，所以它只能在 v_2 到达 7，然后必须立刻折返。

验算

- 计算 u_n 直至 $n = 40$，确认每一次都有 $u_n^2 + u_{n-1}^2 - 4u_nu_{n-1} = -3$。例如 $(26,97)$：$676 + 9409 - 10088 = -3$。
- 用数以千计的随机整数三元组检验开首的恒等式。
- 由 1, 2 开始，列出 (a) 所容许的每一个数列，直至 v_8。当中自始至终满足 $v_{k+4} = v_k$ 的只有 $1,2,1,2,\ldots$ 和 $1,2,7,2,\ldots$。

常见错误

- 在 (ii) 中断定 $F_n - F_{n-1} = 0$ 却没有说明原因。原因是 $(*)$ 令每一个 F_n 都等于 -3。
- 忘记 v_1 是唯一确定的。该二次方程有重根 2，所以无需分情况讨论。
- 在 (c) 中只检查了一个周期之内的各对相邻项，而忘记了“首尾相接”的那一对 $(v_3, v_4) = (2, 1)$。

STEP 之外

F_n 这个量是递推关系的一个不变量：一个守恒量，就像力学中的能量。方程 $x^2 + y^2 - 4xy = -3$ 是一个马尔可夫（Markov）型（或佩尔（Pell）型）的丢番图方程（Diophantine equation）。把二次方程的一个根换成另一个根，即 $x \mapsto 4y - x$，正是经典的“韦达跳跃”（Vieta jumping）手法，它能由最小的解生成方程的全部整数解。Davenport 的 *The Higher Arithmetic*（高等算术）讲述了佩尔方程，以及如何用递推关系求出它的解。

详细解答

STEP III 2004，第 7 题

纯数学 · 归约公式

STEP III 2004，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/04-s3-q7

关键思路

I_{n+1} 的被积函数（integrand）等于 I_n 的被积函数乘以 $\frac{t}{t+1}$，所以只要估计这一个因子的上界（upper bound），就能得到联系相邻两个积分的不等式。分部积分法（integration by parts）也给出一个精确的关系式；把不等式与等式并排放在一起，就能挤出一个只涉及 I_n 的上界。这个精确的关系式可以逐项相消（telescope），一直化到 I_1，而它是我们熟悉的对数；于是整件事就变成 $\ln 2$ 的一个级数（series），而且误差项在我们掌握之中。

解答

不等式 $I_{n+1} < \frac{1}{2} I_n$

在 $0 \le t \le 1$ 上，

$$\frac{t}{t+1} = 1 - \frac{1}{t+1},$$

它随 t 递增，所以最大值是 $\frac{1}{2}$，而且只在 $t = 1$ 时取得。现在把 I_{n+1} 的被积函数写成

$$\frac{t^n}{(t+1)^{n+1}} = \frac{t}{t+1} \cdot \frac{t^{n-1}}{(t+1)^n}.$$

因此

$$\frac{1}{2} I_n - I_{n+1} = \int_0^1 \left(\frac{1}{2} - \frac{t}{t+1} \right) \frac{t^{n-1}}{(t+1)^n} \, \mathrm{d}t.$$

这里的被积函数是连续（continuous）的，永不为负，而且在 $0 < t < 1$ 时严格为正（两个因子在该处均为正）。所以这个积分严格为正，即 $I_{n+1} < \frac{1}{2} I_n$。

? **想一想** 既然 $\frac{t}{t+1}$ 确实会达到 $\frac{1}{2}$，为甚么不等式仍然是严格的?

它只在 $t=1$ 这一点达到 $\frac{1}{2}$。一个连续的非负函数，只要在整个区间上为正，其积分就是正数，所以单单一点取等号并无影响。

递推关系

对 I_{n+1} 作分部积分，取 $u=t^n$ 及 $\dfrac{\mathrm{d}v}{\mathrm{d}t}=(t+1)^{-(n+1)}$，于是 $v=-\dfrac{1}{n(t+1)^n}$（因为 $n\geq 1$，此式成立）：

$$\begin{aligned} I_{n+1} &= \Big[-\frac{t^n}{n(t+1)^n}\Big]_0^1 + \int_0^1 \frac{nt^{n-1}}{n(t+1)^n}\,\mathrm{d}t \\ &= -\frac{1}{n\,2^n} + I_n, \end{aligned}$$

这是因为方括号内的式子在 $t=1$ 时等于 $-\frac{1}{n2^n}$，在 $t=0$ 时等于 0。这就是所要求的结果，即题目中的递推关系（recurrence）。

I_n 的上界

把递推关系代入不等式：

$$I_n - \frac{1}{n\,2^n} = I_{n+1} < \frac{1}{2}I_n \quad\Longrightarrow\quad \frac{1}{2}I_n < \frac{1}{n\,2^n} \quad\Longrightarrow\quad I_n < \frac{1}{n\,2^{n-1}}.$$

$\ln 2$ 的级数

首先，

$$I_1 = \int_0^1 \frac{1}{t+1}\,\mathrm{d}t = \big[\ln(t+1)\big]_0^1 = \ln 2.$$

递推关系指出，对每个 $r\geq 1$ 都有 $I_r - I_{r+1} = \dfrac{1}{r\,2^r}$。把 $r=1,\ldots,n$ 的各式相加，左边逐项相消：

$$I_1 - I_{n+1} = \sum_{r=1}^{n}\frac{1}{r\,2^r}, \qquad \text{所以} \qquad \ln 2 = \sum_{r=1}^{n}\frac{1}{r\,2^r} + I_{n+1}.$$

上下界 $\frac{2}{3} < \ln 2 < \frac{17}{24}$

关于余项（remainder），我们知道两件事：$I_{n+1}>0$，因为它的被积函数在 $(0,1]$ 上为正；以及 $I_{n+1} < \dfrac{1}{(n+1)2^n}$，这是把上文的上界中的 n 换成 $n+1$ 而得。所以对每个 n，

$$\sum_{r=1}^{n}\frac{1}{r\,2^r} < \ln 2 < \sum_{r=1}^{n}\frac{1}{r\,2^r} + \frac{1}{(n+1)2^n}.$$

部分和（partial sums）依次为

$$\frac{1}{2}, \qquad \frac{1}{2}+\frac{1}{8}=\frac{5}{8}, \qquad \frac{5}{8}+\frac{1}{24}=\frac{16}{24}=\frac{2}{3}.$$

想一想 求上界和下界时，各应取哪个 n 值?

两者并不相同。求下界（lower bound）时取 $n=3$：$\ln 2 > \frac{2}{3}$。求上界时取 $n=2$，此时误差的上界是 $\frac{1}{3\cdot 4}=\frac{1}{12}$：

$$\ln 2 < \frac{5}{8}+\frac{1}{12}=\frac{15}{24}+\frac{2}{24}=\frac{17}{24}.$$

合起来便得 $\frac{2}{3}<\ln 2<\frac{17}{24}$，正是所求。

验算

以数值计算，$\ln 2\approx 0.6931$，确实介乎 $\frac{2}{3}\approx 0.6667$ 与 $\frac{17}{24}\approx 0.7083$ 之间。用辛普森法则（Simpson's rule）可得 $I_3\approx 0.0681$，明显小于上界 $\frac{1}{12}\approx 0.0833$，而且 $\frac{5}{8}+0.0681=0.6931$ 亦确实成立。我们也用数值方法验算了递推关系和两个不等式，n 的取值直至 24。

常见错误

- 求上下界时用同一个 n：取 $n=3$ 会得到上界 $\frac{2}{3}+\frac{1}{32}=\frac{67}{96}$，这个上界是对的，而且更精确，但并不是题目要求的 $\frac{17}{24}$。
- 分部积分时弄错边界项的正负号；v 是负的。
- 忘记说明为甚么 $I_{n+1}>0$，而严格的下界正需要这一点。

STEP 之外

这就是级数 $\ln 2=\sum_{r\geq 1}\frac{1}{r2^r}$，亦即 $-\ln(1-x)=\sum x^r/r$ 在 $x=\frac{1}{2}$ 时的情况，并且附有取 n 项之后余项的明确上界。用积分来估计余项，正是带积分余项的泰勒定理（Taylor's theorem）的做法，那是大学一年级分析课的基本内容。关于幂级数（power series）及其余项的理论，可参阅 Abbott 的 *Understanding Analysis*。

详细解答

STEP III 2004，第 8 题

纯数学 · 微分方程

> **STEP III 2004，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/04-s3-q8

关键思路

由于有 $1/y$ 这一项，方程对 y 而言并非线性（linear）。但把方程两边同乘 y 后，$y\frac{\mathrm{d}y}{\mathrm{d}x}$ 就变成 y^2 的导数的一半，而 $1/y$ 项则变成纯粹是 x 的函数。因此以 $u = y^2$ 表示时，方程是线性的（这是一条伯努利型（Bernoulli type）方程），再用积分因子（integrating factor）便可完成求解。三条曲线都是圆锥曲线（conics），辨认出它们之后，作图就很容易。

解答

代换

设 $u = y^2$。由链式法则（chain rule），

$$\frac{\mathrm{d}u}{\mathrm{d}x} = 2y\frac{\mathrm{d}y}{\mathrm{d}x} = 2y\left(f(x)\,y + \frac{g(x)}{y}\right) = 2f(x)\,y^2 + 2g(x).$$

以 u 取代 y^2，

$$\frac{\mathrm{d}u}{\mathrm{d}x} - 2f(x)\,u = 2g(x),$$

这对 u 而言是线性的：u 和 $\frac{\mathrm{d}u}{\mathrm{d}x}$ 都只以一次幂出现，而系数只取决于 x。这正是我们要证明的。

解方程

这里 $f(x) = \frac{1}{x}$ 而 $g(x) = -1$，所以线性方程是

$$\frac{\mathrm{d}u}{\mathrm{d}x} - \frac{2}{x}\,u = -2.$$

积分因子是 $\exp\left(-\int \frac{2}{x}\,\mathrm{d}x\right) = \mathrm{e}^{-2\ln|x|} = \dfrac{1}{x^2}$。把方程两边同乘这个因子，

$$\frac{1}{x^2}\frac{\mathrm{d}u}{\mathrm{d}x} - \frac{2}{x^3}u = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{u}{x^2}\right) = -\frac{2}{x^2}.$$

积分后得 $\dfrac{u}{x^2} = \dfrac{2}{x} + C$，其中 C 为常数，所以通解（general solution）是

$$y^2 = 2x + Cx^2.$$

（原方程要求 $x \neq 0$ 及 $y \neq 0$，所以每个解都定义在一个不含 0 的 x 区间上，在该区间上 $2x + Cx^2 > 0$，而且 y 始终保持同一符号。）

三条曲线

三个点都满足 $x > 0$ 及 $y > 0$，所以每条解曲线（solution curve）都是 $y^2 = 2x + Cx^2$ 在包含该点的正 x 区间上 $y > 0$ 的那一部分。

通过 $(1,1)$：$1 = 2 + C$，所以 $C = -1$，$y^2 = 2x - x^2$。配方后得

$$(x-1)^2 + y^2 = 1,$$

即圆心为 $(1,0)$、半径为 1 的圆。解曲线是上半圆 $y = \sqrt{2x - x^2}$，其中 $0 < x < 2$。

通过 $(2,2)$：$4 = 4 + 4C$，所以 $C = 0$，$y^2 = 2x$：这是顶点在原点、对称轴沿 x 轴的抛物线。解曲线是 $y = \sqrt{2x}$，其中 $x > 0$。

通过 $(4,4)$：$16 = 8 + 16C$，所以 $C = \frac{1}{2}$，$y^2 = 2x + \frac{1}{2}x^2$。配方后得 $\frac{1}{2}(x+2)^2 - y^2 = 2$，即

$$\frac{(x+2)^2}{4} - \frac{y^2}{2} = 1,$$

这是一条双曲线，中心为 $(-2,0)$，顶点为 $(0,0)$ 和 $(-4,0)$，渐近线（asymptotes）为 $y = \pm\frac{1}{\sqrt{2}}(x+2)$。解曲线是右支的上半部分，$y = \sqrt{2x + \frac{1}{2}x^2}$，其中 $x > 0$。

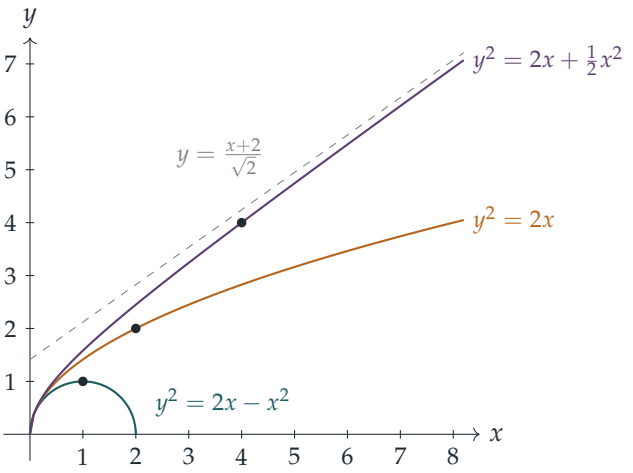
> ? **想一想** 为甚么不包括双曲线的左支，或者各曲线的下半部分?

微分方程的解必须在某个区间上可导，并在该区间上满足方程，而方程在 $x = 0$ 或 $y = 0$ 时没有定义。从 $(4,4)$ 出发向左走，曲线会到达 $(0,0)$，在那里两个条件都不成立，所以解不能延拓过这一点。左支（$x \leq -4$）和下半部分虽然满足 $y^2 = 2x + \frac{1}{2}x^2$，但它们是另外的解，并不是通过 $(4,4)$ 的那一个。

略图

三条曲线都从原点出发，在该处的切线是铅垂的（在 $x = 0$ 附近，每条曲线的表现都近似 $y \approx \sqrt{2x}$），而在所给各点的斜率分别是 $\frac{1}{1} - \frac{1}{1} = 0$、$\frac{2}{2} - \frac{1}{2} = \frac{1}{2}$ 和 $\frac{4}{4} - \frac{1}{4} = \frac{3}{4}$。当 x 较大时，

圆会回落到 $(2,0)$，抛物线按 $\sqrt{x}$ 的速度增长，而双曲线则趋近其渐近线 $y=\frac{1}{\sqrt{2}}(x+2)$。由于当 $x>0$ 时 $2x-x^2<2x<2x+\frac{1}{2}x^2$，圆位于抛物线之下，而抛物线又位于双曲线之下。



验算

- 代回原方程：若 $y^2=2x+Cx^2$，则 $2yy'=2+2Cx$，所以 $y'=\frac{1+Cx}{y}$，而 $\frac{y}{x}-\frac{1}{y}=\frac{y^2-x}{xy}=\frac{x+Cx^2}{xy}=\frac{1+Cx}{y}$。两者一致。
- 用数值方法（龙格—库塔法（Runge–Kutta））对原方程由 $(4,4)$ 逐步计算至 $x=8$，得 $y\approx 6.92820$，与 $\sqrt{16+32}=\sqrt{48}$ 吻合；由 $(1,1)$ 和 $(2,2)$ 出发作类似计算，结果与半圆和抛物线吻合，误差约为 10^{-12}。

常见错误

- 把积分因子写成 x^2 而不是 x^{-2}：u 的系数是 $-\frac{2}{x}$，所以指数是 $-2\ln x$。
- 漏掉 $\frac{\mathrm{d}u}{\mathrm{d}x}=2y\frac{\mathrm{d}y}{\mathrm{d}x}$ 中的因子 2，结果得出 $y^2=x+Cx^2$；这通不过代回原方程的验算。
- 把整个圆、抛物线的上下两半或双曲线的两支都画出来，当作“解曲线”。

STEP 之外

这里的技巧适用于任何*伯努利方程*（Bernoulli equation）$y'=f(x)y+g(x)y^n$：代换 $u=y^{1-n}$ 总能把它化为线性方程（这里 $n=-1$，所以 $u=y^2$）。在 $x=0$ 和 $y=0$ 处所需的谨慎，让我们初次窥见存在与唯一性定理（existence and uniqueness theorem）：凡在

右边表现良好的点，都有且仅有一个解通过，而曲线族只能在右边表现不良好的点相交，正如我们的所有曲线都在原点相交一样。Boyce 与 DiPrima 所著的 *Elementary Differential Equations*（初等微分方程教科书）涵盖了伯努利方程和这个定理。

详细解答

STEP III 2004，第 9 题

力学 · 力矩与摩擦力

STEP III 2004，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/04-s3-q9

关键思路

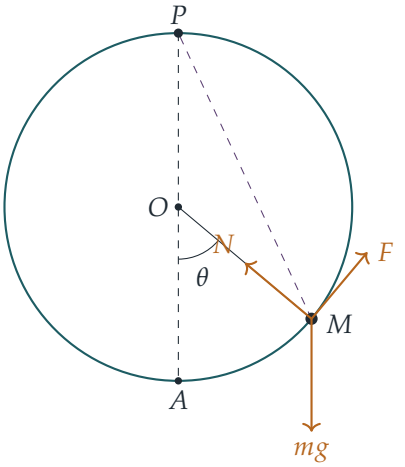
圆环的重量作用于转轴正下方的 O，因此对 P 没有力矩（moment）。除了转轴处的反作用力之外，圆环所受的另一个力就只有老鼠施加的接触力（contact force），所以圆环保持不动，当且仅当这个接触力对 P 的力矩为零，也就是它的作用线（line of action）通过 P。整道题其实只在问：老鼠所受的接触力何时可以沿 MP 的方向？这个问题要问两次，一次求速率，一次求摩擦力的极限。

解答

模型

把老鼠视为质量为 m 的质点（particle），把圆环视为刚性（rigid）圆环（稍后可见它的质量并不影响结果），圆环可绕通过 P 的光滑水平轴自由转动。设 $\theta = \angle AOM$，由向下的铅垂线起量度，老鼠朝 θ 递增的方向奔跑。圆环静止时，老鼠在一个半径为 a、圆心为 O 的固定圆上以恒定速率 u 运动。

老鼠所受的力有：重量 mg、圆环施加并指向 O 的法向反作用力（normal reaction）N，以及沿切线方向的摩擦力（friction）F，我们取 θ 递增的方向（沿圆环向上）为正。



对老鼠沿切线方向（速率恒定，故没有切向加速度）和沿半径方向（加速度为 u^2/a，指向 O）分解：

$$F = mg\sin\theta,$$
$$N - mg\cos\theta = \frac{mu^2}{a}.$$

圆环何时处于平衡?

根据牛顿（Newton）第三定律，老鼠推圆环的力，与老鼠所受的接触力（N 和 F）大小相等、方向相反。圆环的重量作用于 O，位于 P 的正下方，而转轴的反作用力作用于 P，所以两者对 P 都没有力矩。因此，圆环保持平衡（equilibrium）且 POA 铅垂，当且仅当接触力对 P 的力矩为零，也就是它的作用线通过 P。

三角形 OPM 是等腰三角形（$OP = OM = a$），且 $\angle POM = \pi - \theta$，所以 $\angle OMP = \frac{1}{2}\theta$。于是直线 MP 与半径 MO 的夹角为 $\frac{1}{2}\theta$，并偏向 θ 递增的一方。因此，老鼠所受的接触力沿 MP 的方向，当且仅当

$$\frac{F}{N} = \tan\tfrac{1}{2}\theta. \qquad (*)$$

（等价地，直接对 P 取力矩，老鼠施于圆环的力的力矩为 $a\left[N\sin\theta - F(1+\cos\theta)\right]$，而 $\sin\theta/(1+\cos\theta) = \tan\frac{1}{2}\theta$。）

想一想 怎样把这个条件化为关于 u 的条件?

利用运动方程。老鼠所受的接触力等于质量乘以加速度，再减去重量：

$$\mathbf{R} = m\cdot\frac{u^2}{a}\,\widehat{MO} + mg\,\widehat{\text{上}}.$$

但 $\overrightarrow{MP} = \overrightarrow{MO} + \overrightarrow{OP} = a\,\widehat{MO} + a\,\widehat{上}$，两个系数相等。所以 $\mathbf{R}$ 平行于 MP 对每个 θ 都成立，当且仅当 $u^2/a = g$，即

$$u = \sqrt{ag}.$$

用代数方法：当 $u^2 = ag$ 时，运动方程给出 $F = mg\sin\theta$ 和 $N = mg(1+\cos\theta)$，所以

$$\frac{F}{N} = \frac{\sin\theta}{1+\cos\theta} = \frac{2\sin\frac{1}{2}\theta\cos\frac{1}{2}\theta}{2\cos^2\frac{1}{2}\theta} = \tan\tfrac{1}{2}\theta,$$

这正是 $(*)$。对于一般的速率，对 P 的力矩算出来是 $m\sin\theta\,(u^2 - ag)$，它在每个位置都等于零，当且仅当 $u^2 = ag$。所以老鼠以速率 $\sqrt{ag}$ 奔跑，就能令圆环保持平衡，正如题目所要求，前提是圆环确实能够提供所需的摩擦力。接下来我们就考察这一点。

摩擦力的极限

只有当 $|F| \le \mu N$ 时，老鼠的脚才能抓紧而不打滑。取 $u = \sqrt{ag}$，并考虑圆环的下半部分，该处 $|\theta| < \frac{1}{2}\pi$，故 $\cos\frac{1}{2}\theta > 0$，这个条件就是

$$\tan\tfrac{1}{2}|\theta| \le \mu \quad\Longleftrightarrow\quad |\theta| \le 2\arctan\mu.$$

所以在 $\angle AOM = 2\arctan\mu$ 之前，一切都行得通。（只有当 $\mu < 1$ 时，这一点才位于圆环的下半部分；若 $\mu \ge 1$，老鼠永远到不了这一点。）

过了这一点，无论选甚么速率都无补于事。平衡需要 $F/N = \tan\frac{1}{2}\theta > \mu$，而摩擦力无法做到。不论速率是多少，只要摩擦力满足 $|F| \le \mu N$，老鼠施于圆环的力对 P 的力矩都是

$$a\big[N\sin\theta - F(1+\cos\theta)\big] \;\ge\; aN(1+\cos\theta)\big(\tan\tfrac{1}{2}\theta - \mu\big) > 0.$$

会发生甚么事

老鼠一旦越过 $\angle AOM = 2\arctan\mu$，就再也抓不够紧。牠的脚开始打滑，摩擦力降至极限值（limiting value）μN，老鼠无法维持原有速率：牠开始向后滑。这时牠施于圆环的接触力从 P 的旁边经过，因而力矩不为零，圆环不再处于平衡。圆环开始绕 P 转动。上面那个力矩的方向表明，老鼠的法向推力胜过已减弱的摩擦力，所以圆环摆动时，A 会移向老鼠所在的一侧，也就是老鼠原本奔跑的方向。从这一刻起，运动不再是简单的平衡情况：圆环绕 P 摆动，而老鼠则在圆环上滑动。

验算

- 当 $\theta = 0$ 时，老鼠位于 A，接触力 $N = 2mg$ 是铅垂的，并通过 P，而且不需要摩擦力，与预期相符。
- 速率 $\sqrt{ag}$ 具有速率的单位，既不取决于 μ，也不取决于圆环的质量，而且它令法向反作用力等于 $mg(1+\cos\theta)$，在整个下半部分都是正数。

- 当 $\mu = 0.5$ 时，临界角为 $2\arctan 0.5 \approx 0.927$ 弧度 $\approx 53.1^\circ$。在该处 $\sin\theta = 0.8$，$1 + \cos\theta = 1.6$，所以 $F/N = 0.5 = \mu$，恰好相等。以 $u^2 = ag$ 在随机位置对 P 的力矩作数值检验，结果为零，误差在 $10^{-15}mga$ 以内。

常见错误

- 对 O 而不是对 P 取力矩。圆环是绕 P 转动的，而法向反作用力（它通过 O）对 P 确实有力矩。
- 漏掉 N 中的向心（centripetal）项，结果得出 $N = mg\cos\theta$ 和错误的摩擦力条件。
- 声称老鼠只是“掉下来”。在下半部分，$N > 0$ 始终成立。失效的是抓地力，而不是接触。

STEP 之外

这里的巧妙之处在于：一个物体绕枢轴（pivot）处于平衡，当且仅当其余各力的合力（resultant）通过枢轴。“作用线通过 P”这个论证，省去了一整页的分解计算。题目的后半部分则让人初尝耦合系统（coupled systems）的滋味：打滑的接触把一个物体的运动传到另一个物体上。要分析打滑之后的运动，就需要圆环的转动惯量（moment of inertia）和完整的刚体转动方程。Kleppner 与 Kolenkow 合著的 *An Introduction to Mechanics*（力学入门教科书）正是以这种风格讲解绕固定轴的力矩（torque）和受摩擦力限制的运动。

详细解答

STEP III 2004，第 10 题

力学 · 弹性绳与简谐运动

STEP III 2004，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：

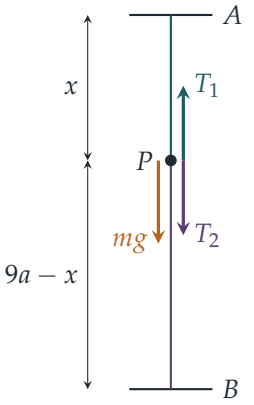
step.maths.org/questions/04-s3-q10

关键思路

每条弦线只有在拉紧时才遵从虎克定律（Hooke's law），所以要按哪些弦线处于拉紧状态，把运动分成几个阶段。在每个阶段中，加速度都是 $7a - x$ 的常数倍，因此质点以 $x = 7a$ 为中心作简谐运动（simple harmonic motion），只是各阶段的 ω 不同。我们先求出第一阶段结束时的时间和速率，再以此作为第二阶段的起始条件。

解答

我们把 P 视为质点（particle），并把弦线视为轻弦：伸长时遵从虎克定律，松弛时不施力。取 $x = AP$，由 A 向下量度，于是 $PB = 9a - x$，并以向下为正方向。



运动方程

当 $x > 6a$ 时，弦线 AP 的伸长量为 $x - 6a$，所以它把 P 向上拉的张力（tension）为

$$T_1 = \frac{6mg(x-6a)}{6a} = \frac{mg(x-6a)}{a}.$$

当 $x < 7a$ 时，弦线 BP 的伸长量为 $(9a - x) - 2a = 7a - x$，所以它把 P 向下拉向 B 的张力为

$$T_2 = \frac{6mg(7a-x)}{2a} = \frac{3mg(7a-x)}{a},$$

而当 $x \geq 7a$ 时，BP 松弛，$T_2 = 0$。

当 $6a < x < 7a$ 时，两条弦线都拉紧，沿向下方向应用牛顿第二定律（Newton's second law），得

$$\begin{aligned} m\ddot{x} &= mg - \frac{mg(x-6a)}{a} + \frac{3mg(7a-x)}{a} \\ &= \frac{mg}{a}(a - x + 6a + 21a - 3x) = \frac{mg}{a}(28a - 4x), \end{aligned}$$

所以 $\ddot{x} = \dfrac{4g}{a}(7a - x)$，正是所需结果。

当 $7a < x < 9a$ 时，只有 AP 拉紧，所以

$$m\ddot{x} = mg - \frac{mg(x-6a)}{a} = \frac{mg}{a}(7a - x),$$

由此得 $\ddot{x} = \dfrac{g}{a}(7a - x)$，正是所需结果。

到达 B 所需的时间

第一阶段：$6a \leq x \leq 7a$。设 $y = x - 7a$，则 $\ddot{y} = -\omega_1^2 y$，其中 $\omega_1^2 = 4g/a$，这是以 $x = 7a$ 为中心的简谐运动。质点由 $y = -a$ 从静止开始运动，故

$$x = 7a - a\cos\omega_1 t, \qquad \dot{x} = a\omega_1 \sin\omega_1 t.$$

质点在 $x = 6a$ 从静止开始，向下加速（因为该处 $\ddot{x} > 0$），所以 x 由 $6a$ 起递增，并一直处于这个阶段，直至 $x = 7a$。这在 $\cos\omega_1 t = 0$ 时发生，首次出现于 $\omega_1 t = \pi/2$，即经过时间

$$t_1 = \frac{\pi}{2\omega_1} = \frac{\pi}{4}\sqrt{\frac{a}{g}},$$

此时速率为 $a\omega_1 = 2\sqrt{ag}$，方向向下。

> ? **想一想** 方程在 $x = 7a$ 处改变了。我们该怎样继续?

位置和速度在 $x = 7a$ 处都是连续的，所以第一阶段的终结状态就是第二阶段的初始条件（starting conditions）。我们在该处重新计时。

第二阶段：$7a \le x \le 9a$。这时 $\ddot{y} = -\omega_2^2 y$，其中 $\omega_2^2 = g/a$，质点由中心 $y = 0$ 以速率 $2\sqrt{ag}$ 出发。因此，设 τ 为自 $x = 7a$ 起计的时间，则

$$x = 7a + \frac{2\sqrt{ag}}{\omega_2} \sin \omega_2 \tau = 7a + 2a \sin \omega_2 \tau.$$

振幅（amplitude）刚好是 $2a$，所以当 $\sin \omega_2 \tau = 1$ 时，质点到达 $x = 9a$，亦即 B 点，首次出现于 $\omega_2 \tau = \pi/2$。在该刻 $\dot{x} = 2\sqrt{ag} \cos \omega_2 \tau = 0$，所以质点到达 B 时速率为零。这个阶段需时

$$t_2 = \frac{\pi}{2\omega_2} = \frac{\pi}{2}\sqrt{\frac{a}{g}}.$$

因此，质点到达 B 所需的时间为

$$t_1 + t_2 = \frac{\pi}{4}\sqrt{\frac{a}{g}} + \frac{\pi}{2}\sqrt{\frac{a}{g}} = \frac{3\pi}{4}\sqrt{\frac{a}{g}}.$$

验算

能量的计算可以确认 P 刚好到达 B 并在该处静止。比较起点（BP 伸长了 a）和到达 B 的一刻（AP 伸长了 $3a$）。弹性势能（elastic energy）由 $\frac{6mg\,a^2}{2\cdot 2a} = \frac{3}{2}mga$ 变为 $\frac{6mg\,(3a)^2}{2\cdot 6a} = \frac{9}{2}mga$。增加了 $3mga$，刚好等于下降 $3a$ 所失去的重力势能（gravitational energy），所以在 B 处的动能（kinetic energy）为零。数值上，取 $a = 0.7$ 和 $g = 9.8$，由 $x = 6a$ 从静止开始，对完整的运动方程逐步作数值计算，得出在 $t \approx 0.2099$ 时 $x = 7a$，在 $t \approx 0.6297$ 时到达最低点 $x = 9a$，而 $\frac{3\pi}{4}\sqrt{a/g} = 0.6297$。

常见错误

- 在 $x > 7a$ 时仍使用 T_2：弦线不能推物，所以一旦 BP 短于 $2a$，它便不施力。
- 整个运动都使用 ω_1，或者忘记以第一段简谐运动所留下的速率 $2\sqrt{ag}$ 重新开始第二段简谐运动。
- 没有检查振幅 $2a$ 是否真的能把 P 带到 B，便把第二阶段说成是四分之一个周期（period）。

STEP 之外

这是一个分段线性振子（piecewise-linear oscillator）：回复力（restoring force）是线性的，但劲度（stiffness）在 $x = 7a$ 处改变。两个部分各有自己的相平面（phase plane）椭圆，而轨迹则由两个不同椭圆的弧组成，在弦线变松之处相接。由弹跳的球到碰撞振子（impact oscillator），这类系统是非线性动力学（nonlinear dynamics）的标准入门例子。

Kleppner 与 Kolenkow 所著的 *An Introduction to Mechanics*（力学导论）深入讨论了振动和能量方法。

详细解答

STEP III 2004，第 11 题

力学 · 二体排斥

STEP III 2004，第 11 题

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

step.maths.org/questions/04-s3-q11

本题在 *Beyond STEP* 第 10.6 节二体化为一体中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch10/two-body.html

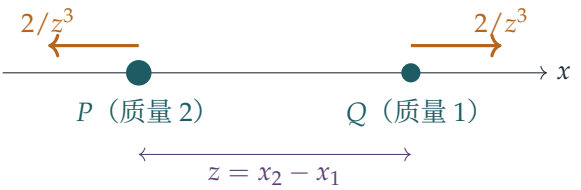
关键思路

两个质点互相推开，会得出两条运动方程（equation of motion），而只要把它们巧妙地组合两次，问题便迎刃而解。*差* $x_2 - x_1$ 给出一条只涉及间距 z 的方程，利用 $v\,\mathrm{d}v/\mathrm{d}z$ 这个技巧便可积分一次（这其实就是能量守恒）。*加权和* $2x_1 + x_2$ 是质心（centre of mass）位置的三倍，而由于这些力都是内力，它完全不受任何净力作用。

解答

z 的方程

把 P 和 Q 视为直线上的质点（particle），x 沿由 P 指向 Q 的方向递增（因为 $x_2 > x_1$）。斥力（repulsion）把 P 推向负方向，把 Q 推向正方向，两者的大小均为 $2/z^3$。



对每个质点应用牛顿第二定律（Newton's second law），得

$$2\ddot{x}_1 = -\frac{2}{z^3}, \qquad \ddot{x}_2 = \frac{2}{z^3}.$$

所以 $\ddot{x}_1 = -1/z^3$，两式相减，得

$$\frac{\mathrm{d}^2 z}{\mathrm{d}t^2} = \ddot{x}_2 - \ddot{x}_1 = \frac{2}{z^3} + \frac{1}{z^3} = \frac{3}{z^3},$$

正是所求。

求解 z

先看初始条件（initial conditions）。在 $t = 0$ 时，$z = 1$。P 朝着 Q 运动，即沿正方向运动，所以 $\dot{x}_1 = 1$，而 $\dot{x}_2 = 0$。因此在 $t = 0$ 时，$v = \dot{z} = 0 - 1 = -1$：两者的间距正在缩小。

利用 $\ddot{z} = v\,\mathrm{d}v/\mathrm{d}z$，方程变成

$$v\frac{\mathrm{d}v}{\mathrm{d}z} = \frac{3}{z^3} \quad \Longrightarrow \quad \tfrac{1}{2}v^2 = -\frac{3}{2z^2} + C.$$

当 $z = 1$ 时，$v^2 = 1$，所以 $\frac{1}{2} = -\frac{3}{2} + C$，即 $C = 2$。因此

$$v^2 = 4 - \frac{3}{z^2}, \qquad \text{或} \qquad z^2\left(\frac{\mathrm{d}z}{\mathrm{d}t}\right)^2 = 4z^2 - 3.$$

左边是 $\left(\frac{1}{2}\,\mathrm{d}(z^2)/\mathrm{d}t\right)^2$，这提示我们设 $u = \sqrt{4z^2 - 3} \geq 0$。于是 $u^2 = 4z^2 - 3$，所以 $u\,\dot{u} = 4z\dot{z}$，而上面的方程表示 $(z\dot{z})^2 = u^2$，即 $z\dot{z} = \pm u$。因此

$$\dot{u} = \frac{4z\dot{z}}{u} = \pm 4.$$

起初 $u = 1$ 而 $z\dot{z} = -1 < 0$，所以取负号：$u = 1 - 4t$。

想一想 当 u 降至 0 时会发生甚么事？$\dot{z}$ 的符号会一直保持为负吗？

不会。在 $t = \frac{1}{4}$ 时 $u = 0$，所以 $z = \sqrt{3}/2$ 而 $v = 0$：两个质点在这一刻最为接近。此时 $\ddot{z} = 3/z^3 > 0$，所以 z 开始增加，$\dot{z}$ 变为正数，改为取正号，于是当 $t \geq \frac{1}{4}$ 时 $u = 4t - 1$（在 $t = \frac{1}{4}$ 处连续）。在两种情况下均有 $u^2 = (4t - 1)^2$，所以对所有 $t \geq 0$，

$$4z^2 - 3 = 16t^2 - 8t + 1 \quad \Longrightarrow \quad z^2 = 4t^2 - 2t + 1,$$

又因为 $z > 0$，

$$z = \sqrt{4t^2 - 2t + 1}.$$

这就是所求的结果。（注意 $4t^2 - 2t + 1 = 4(t - \frac{1}{4})^2 + \frac{3}{4} > 0$，所以平方根恒有定义。）

求 x_1 和 x_2

由两条运动方程，

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}(2x_1+x_2)=2\ddot{x}_1+\ddot{x}_2=-\frac{2}{z^3}+\frac{2}{z^3}=0.$$

这就是动量守恒（conservation of momentum）：两个力大小相等、方向相反。所以 $2x_1+x_2$ 是 t 的线性函数。在 $t=0$ 时它等于 $2\cdot 0+1=1$，而它的变化率是 $2\dot{x}_1+\dot{x}_2=2\cdot 1+0=2$。因此

$$2x_1+x_2=2t+1,\qquad x_2-x_1=\sqrt{4t^2-2t+1}.$$

第一式减去第二式得 $3x_1$，第一式加上第二式的两倍得 $3x_2$：

$$x_1=\frac{2t+1-\sqrt{4t^2-2t+1}}{3},\qquad x_2=\frac{2t+1+2\sqrt{4t^2-2t+1}}{3}.$$

验算

- 在 $t=0$ 时，以上两式给出 $x_1=0$ 和 $x_2=1$。求导后，$\dot{x}_1(0)=\frac{1}{3}\big(2-\frac{1}{2}\cdot(-2)\big)=1$ 及 $\dot{x}_2(0)=\frac{1}{3}(2-2)=0$，符合要求。
- 求 z 的一条捷径：设 $w=z^2$，则 $\ddot{w}=2\dot{z}^2+2z\ddot{z}=2\big(4-3/z^2\big)+6/z^2=8$，所以由 $w(0)=1$ 和 $\dot{w}(0)=2z\dot{z}=-2$ 可得 $w=4t^2-2t+1$。这样便完全不必为符号操心。
- 当 $t\to\infty$ 时，$z\approx 2t-\frac{1}{2}$，所以 $x_1\to\frac{1}{2}$ 而 $\dot{x}_2\to 2$：P 最终静止，而 Q 带走全部动量（$1\times 2=2\times 1$）和全部能量。势能（potential energy）是 $1/z^2$，所以开始时的能量是 $1+1=2$，最终的能量是 $\frac{1}{2}\cdot 1\cdot 2^2=2$。
- 以数值方法逐步计算原来的方程（龙格－库塔法（Runge–Kutta），步长 10^{-4}）直至 $t=5$，所得结果与 x_1、x_2 和 z 的公式相差不超过 10^{-11}，最接近时的间距为 $z=0.8660\ldots=\sqrt{3}/2$。

常见错误

- 取 $\dot{z}(0)=+1$。P 正朝着 Q 运动，所以间距正在缩小。
- 写出 $\mathrm{d}z/\mathrm{d}t=-\sqrt{4-3/z^2}$ 并以为它在所有时刻都成立，忽略了 $t=\frac{1}{4}$ 处的转向点（turning point）。
- 在 P 的方程中漏掉质量 2，结果得出 $\ddot{z}=4/z^3$。

STEP 之外

这是二体问题（two-body problem）的缩影：质心作匀速运动，而相对坐标的表现就像一个约化质量（reduced mass）为 $m_1m_2/(m_1+m_2)=\frac{2}{3}$ 的单一质点受着同一个力，因为 $\frac{2}{3}\ddot{z}=2/z^3$。同样的分拆方法可把行星运动化为单一轨道的问题。平方反比以外，立方反比力（inverse-cube force）也很特别：z^2 原来恰好是 t 的二次函数，这是 $1/z^2$ 势所隐藏的一种对称性，属于共形力学（conformal mechanics）的研究课题。Kleppner 与 Kolenkow 所著的 *An Introduction to Mechanics*（力学入门）对约化质量和二体运动有细致的论述。

详细解答

STEP III 2004，第 12 题

概率与统计 · 指示变量与错排

STEP III 2004，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/04-s3-q12

关键思路

把一个计数写成指示变量（indicator variable）之和，是概率论中最有用的技巧之一。指示变量的期望值（expectation）其实就是概率，所以 $\mathrm{E}(N)$ 唾手可得；方差（variance）则需要用到每一对变量之间的协方差（covariance），这些协方差很小但不是零，因为球衣并不是独立地分配的。一旦知道 N 的均值和方差都等于 1，参数为 1 的泊松分布（Poisson distribution）自然就是值得一试的模型。

解答

单一指示变量 C_1

球衣的 $m!$ 种分配方式每一种都等可能，所以球衣 1 分给 m 名球员中任何一人的可能性都相同。因此 $\mathrm{P}(C_1 = 1) = \frac{1}{m}$，而由于 C_1 只取 0 和 1 两个值，

$$\mathrm{E}(C_1) = 1 \cdot \mathrm{P}(C_1 = 1) = \frac{1}{m},$$

正如所求。因为 $C_1^2 = C_1$（因 $0^2 = 0$ 及 $1^2 = 1$），我们亦有 $\mathrm{E}(C_1^2) = \frac{1}{m}$，所以

$$\mathrm{Var}(C_1) = \frac{1}{m} - \frac{1}{m^2} = \frac{m-1}{m^2}.$$

C_1 与 C_2 的协方差

这里我们需要 $m \geq 2$。乘积 C_1C_2 恰好在球员 1 和 2 都拿到自己的球衣时等于 1，否则等于 0。已知球员 1 拿到球衣 1，余下的 $m-1$ 件球衣便随机分给其余 $m-1$ 名球员，所以

球员 2 拿到球衣 2 的概率是 $\frac{1}{m-1}$。因此

$$\mathrm{E}(C_1C_2) = \mathrm{P}(C_1 = 1,\, C_2 = 1) = \frac{1}{m} \cdot \frac{1}{m-1},$$

而

$$\begin{aligned}\mathrm{Cov}(C_1, C_2) &= \mathrm{E}(C_1C_2) - \mathrm{E}(C_1)\,\mathrm{E}(C_2) = \frac{1}{m(m-1)} - \frac{1}{m^2} \\ &= \frac{m-(m-1)}{m^2(m-1)} = \frac{1}{m^2(m-1)}.\end{aligned}$$

这个值是正数：知道球员 1 拿对了球衣，就等于从球员 2 可能拿到的球衣中剔除了一件错的，令其拿对的机会略为提高。

> **想一想** 为甚么协方差不是零？各球员的球衣不是分开选的吗？

并不是。球衣是每人分配一件的，所以球员 1 拿到甚么，会影响其他人剩下甚么可拿。各个 C_i 是相依的，这正是计算 $\mathrm{Var}(N)$ 时必须计入协方差项的原因。

N 的均值与方差

根据对称性，每个 C_i 的分布都与 C_1 相同，而每一对 C_i, C_j（其中 $i \neq j$）的联合分布（joint distribution）都与 C_1, C_2 相同（把球员和球衣一并重新编号，并不会改变这个随机分配）。利用期望值的线性性质（linearity of expectation），而这性质并不需要独立性，

$$\mathrm{E}(N) = \sum_{i=1}^{m} \mathrm{E}(C_i) = m \cdot \frac{1}{m} = 1.$$

至于和的方差，我们把所有方差加起来，再加上所有有序对 $i \neq j$ 的协方差，这样的有序对共有 $m(m-1)$ 个：

$$\begin{aligned}\mathrm{Var}(N) &= \sum_{i=1}^{m} \mathrm{Var}(C_i) + \sum_{i \neq j} \mathrm{Cov}(C_i, C_j) \\ &= m \cdot \frac{m-1}{m^2} + m(m-1) \cdot \frac{1}{m^2(m-1)} \\ &= \frac{m-1}{m} + \frac{1}{m} = 1.\end{aligned}$$

所以 $\mathrm{E}(N) = \mathrm{Var}(N) = 1$，正如所求。（若 $m = 1$，则必有 $N = 1$，于是 $\mathrm{Var}(N) = 0$；题目其实隐含假设了 $m \geq 2$，一如计算协方差时已经需要的那样。）

用哪一种近似？

无论 m 取甚么值，正态近似（Normal approximation）的效果都很差。N 只取整数值 $0, 1, 2, \ldots$，而由于均值为 1、标准差为 1，它的概率几乎全部集中在 $0, 1, 2, 3$ 这三四个值上。

这个分布在 0 处被截断，那里只比均值低一个标准差，所以它明显向右偏斜；而 $\mathrm{N}(1,1)$ 分布是对称的，并且在 $-\frac{1}{2}$ 以下有不可忽视的概率（即使作了连续性校正（continuity correction），也约有 7%），但 N 根本不可能取到那里的值。最关键的是，增大 m 也无济于事：对每一个 m，均值和方差都维持在 1，所以分布的形状永远不会展开成钟形。

相反，泊松近似则合情合理。N 所数的，是在为数甚多的 m 个事件之中“成功”的个数，每个事件的概率 $\frac{1}{m}$ 都很小，而这些事件之间只有很弱的相依性（协方差 $\frac{1}{m^2(m-1)}$ 极小）。计数近似服从泊松分布，正是在这种情况下发生的。此外，泊松变量的均值等于方差，而 N 恰好具有这个性质，所以 $\mathrm{Po}(1)$ 是自然的选择。

$m = 4$ 的情况

共有 $4! = 24$ 种等可能的分配方式。把一种分配写成球员 $1,2,3,4$ 所穿球衣的号码序列，则没有人穿上自己球衣的分配是

$$2143,\quad 2341,\quad 2413,\quad 3142,\quad 3412,$$
$$3421,\quad 4123,\quad 4312,\quad 4321.$$

（有系统地列举的方法：球员 1 穿的是球衣 2、3 或 4，而每种情况下都恰好有三种方法完成余下的分配，所以共有 $3 \times 3 = 9$ 种。）因此

$$\mathrm{P}(N = 0) = \frac{9}{24} = \frac{3}{8} = 0.375.$$

参数为 1 的泊松近似给出 $\mathrm{P}(N = 0) \approx \mathrm{e}^{-1}$。取 $\mathrm{e} \approx 2.72$，

$$\mathrm{e}^{-1} \approx \frac{100}{272} = \frac{25}{68}.$$

以真确值为基准的相对误差（relative error）是

$$\frac{\frac{3}{8} - \frac{25}{68}}{\frac{3}{8}} = 1 - \frac{25}{68} \cdot \frac{8}{3} = 1 - \frac{200}{204} = \frac{1}{51} \approx 0.0196,$$

即大约 2%，与题目所述相符。（如果改为以近似值为基准来量度误差，则得 $\frac{1}{50}$，恰好是 2%；无论用哪一种量法，对这么小的 m 来说，吻合程度都好得惊人。）

验算

- 当 $m = 2$ 时，只有两种可能：两人都拿对（$N = 2$）或两人都拿错（$N = 0$），概率各为 $\frac{1}{2}$：均值为 1，方差为 1。吻合。
- 用电脑枚举 $m = 2, \ldots, 7$ 的所有排列（permutation），确认了 $\mathrm{E}(C_1) = \frac{1}{m}$、$\mathrm{Var}(C_1) = \frac{m-1}{m^2}$、$\mathrm{Cov}(C_1, C_2) = \frac{1}{m^2(m-1)}$ 及 $\mathrm{E}(N) = \mathrm{Var}(N) = 1$ 全部准确成立，并找到 4 件物件的错排（derangement）共 9 个，所以 $\mathrm{P}(N = 0) = 0.375$，对比 $1/2.72 \approx 0.3676$。

- 留意 $N = m - 1$ 是不可能的：如果除一人外其他球员都拿对了球衣，最后那一人也必定拿对。这再次显示 N 与正态分布相去甚远。

常见错误

- 把各个 C_i 当作独立，结果得出 $\mathrm{Var}(N) = \frac{m-1}{m}$ 而不是 1。
- 只数了 $\binom{m}{2}$ 个协方差项而不是 $m(m-1)$ 个，即忘记了 $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$ 中的因子 2。
- 列举错排时粗心大意：有系统地数（3×3）可避免遗漏。

STEP 之外

这是经典的“配对”问题，又称错排问题。利用容斥原理（inclusion–exclusion）可以证明 $\mathrm{P}(N=0) = \sum_{k=0}^{m} \frac{(-1)^k}{k!}$，这个值极快地趋向 e^{-1}，而事实上当 $m \to \infty$ 时，N 依分布收敛（converges in distribution）于 $\mathrm{Po}(1)$。更广泛的原理是：由许多罕见而弱相依的事件所构成的计数，近似服从泊松分布；Chen–Stein 方法把这个原理严格地表述出来。Blitzstein 与 Hwang 的 *Introduction to Probability*（概率论入门教材）一书，反复以指示变量和配对问题作为例子。

详细解答

STEP III 2004，第 13 题

概率与统计 · 几何分布

STEP III 2004，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/04-s3-q13

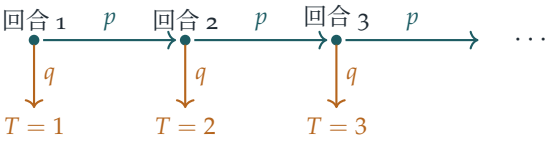
关键思路

暂且放下整场比赛，只跟着一位参赛者看。他出局的回合是一个几何随机变量（geometric random variable），而三位参赛者的出局回合互相独立（independent）。于是每一部分都是关于三个独立几何变量的问题。原来大奖是否颁发，只取决于最迟的出局回合是否有人并列，所以最后一部分就是前两部分合起来的补事件（complement）。

解答

设 $q = 1 - p$，并假设 $0 < p < 1$。对参赛者 j（$j = 1, 2, 3$），设 T_j 为他失败的回合。他在第 k 回合失败，当且仅当他通过了第 $1, \ldots, k-1$ 回合，然后在第 k 回合失败，所以

$$\mathrm{P}(T_j = k) = p^{k-1}q, \qquad k = 1, 2, 3, \ldots$$



由于每一回合的通过与否互相独立，而各参赛者的表现亦互相独立，所以 T_1、T_2、T_3 互相独立。另外留意

$$\mathrm{P}(T_j \leq r - 1) = 1 - \mathrm{P}(\text{通过回合} 1, \ldots, r-1) = 1 - p^{r-1}.$$

第 (i) 部分

根据独立性，三人同在第 k 回合出局的概率是 $(p^{k-1}q)^3$。对 k 求这个等比级数（geometric series）之和（公比（common ratio）$p^3 < 1$），得

$$\mathrm{P}(T_1 = T_2 = T_3) = \sum_{k=1}^{\infty} q^3 p^{3(k-1)} = \frac{q^3}{1-p^3} = \frac{(1-p)^2}{1+p+p^2}.$$

第 (ii) 部分

较早出局的参赛者有 3 种选法。选定之后，其余两人同在第 r 回合失败的概率是 $(p^{r-1}q)^2$，而较早出局的那一位在第 r 回合之前某一回合失败的概率是 $1-p^{r-1}$。不同选法所对应的事件互斥（disjoint），所以

$$\mathrm{P}(\text{两人在回合}r\text{，第三人较早}) = 3p^{2r-2}q^2\left(1-p^{r-1}\right).$$

第 (iii) 部分

想一想 用 T_1, T_2, T_3 来表达，大奖究竟在甚么情况下才会颁发?

假设 T_j 之中的最大值只由一位参赛者取得，例如 $T_3 > T_1, T_2$，并设 $m = \max(T_1, T_2)$。在第 m 回合，参赛者 3 仍然在场，并且通过该回合。当时仍在场的其他参赛者都有 $T_j = m$，所以都在该回合失败。因此参赛者 3 在第 m 回合赢得大奖。

反过来，假设最大值有人并列，例如 $T_2 = T_3 = M \geq T_1$。在第 M 回合之前，参赛者 2 和 3 都仍然在场，而且每一回合都通过，所以没有任何一个回合只剩一人过关。到了第 M 回合，余下的人全部失败，所以根本无人过关。因此大奖不会颁发。

由此可见，大奖不颁发，当且仅当最迟的两个出局回合相等：要么三者全部相等（第 (i) 部分），要么其中两者在某个第 r 回合相等，而第三者较早。第二种情况需要 $r \geq 2$，而它正是第 (ii) 部分的事件。把第 (ii) 部分的结果对 $r \geq 2$ 求和，并设 $s = r-1$，

$$\begin{aligned}\sum_{r=2}^{\infty} 3q^2 p^{2r-2}\left(1-p^{r-1}\right) &= 3q^2\sum_{s=1}^{\infty}\left(p^{2s}-p^{3s}\right) = 3q^2\left(\frac{p^2}{1-p^2}-\frac{p^3}{1-p^3}\right)\\ &= 3q^2\,\frac{p^2(1-p^3)-p^3(1-p^2)}{(1-p^2)(1-p^3)} = 3q^2\,\frac{p^2(1-p)}{(1-p^2)(1-p^3)}\\ &= \frac{3p^2q^2}{(1+p)(1-p^3)}.\end{aligned}$$

加上第 (i) 部分的结果，并利用 $1-p^3 = q(1+p+p^2)$，

$$\mathrm{P}(\text{没有大奖}) = \frac{q^3(1+p)+3p^2q^2}{(1+p)(1-p^3)} = \frac{q^2(1+2p^2)}{(1+p)(1-p^3)} = \frac{(1-p)(1+2p^2)}{(1+p)(1+p+p^2)}.$$

由于 $(1+p)(1+p+p^2) = 1+2p+2p^2+p^3$ 及 $(1-p)(1+2p^2) = 1-p+2p^2-2p^3$，两者之差是 $3p+3p^3$。因此

$$\mathrm{P}(\text{颁发大奖}) = \frac{3p(1+p^2)}{(1+p)(1+p+p^2)}.$$

验算

- 当 $p \to 0$ 时，人人都在第 1 回合失败，而 (iii) 的答案趋向 0。当 $p \to 1$ 时，并列的机会变得很小，而答案趋向 $\frac{3\cdot 2}{2\cdot 3} = 1$。
- 直接计算：当其中一个 T_j 大于另外两个时，大奖便会颁发，所以其概率是 $3\sum_k p^{k-1}q\,(1-p^{k-1})^2$。取 $p = \frac{1}{2}$ 作精确求和，得 $\frac{5}{7} \approx 0.714$，与 $\frac{3\cdot\frac{1}{2}\cdot\frac{5}{4}}{\frac{3}{2}\cdot\frac{7}{4}} = \frac{5}{7}$ 吻合。
- 直接模拟（simulation）比赛本身（取 $p = 0.6$，进行 2×10^5 次），结果有 77.9% 的次数颁发了大奖，而公式给出的是 0.7806。

常见错误

- 在第 (ii) 部分忘记了因数 3，即选择哪一位参赛者较早出局的选法数目。
- 以为最后两人必须在不同回合出局，大奖才会颁发。如果有两人同在第 m 回合出局，而第三人通过了该回合，他仍然获奖。只有在最高处出现并列，大奖才不会颁发。

STEP 之外

这道题目所问的，是独立几何随机变量的最大值，以及这个最大值是否唯一。对 n 位参赛者而言，当 $n \to \infty$ 时，最高处出现并列的概率并不会稳定下来趋向一个极限。它会随 $\log n$ 作极轻微的振荡，这是领袖选举（leader election）等算法分析中一个著名的意外现象。这里用到的工具，即对一个变量取条件（conditioning）以及等比级数求和，在 Grimmett 与 Welsh 的 *Probability: An Introduction* 中有详细阐述。

详细解答

STEP III 2004，第 14 题

概率与统计 · 假设检验的错误

STEP III 2004，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/04-s3-q14

关键思路

在两个假设之下，$\overline{X}$ 的中心同样是 μ；两者的分别只在于 $\overline{X}$ 有多分散。因此，否定 H_0 的证据就是 $\overline{X}$ 落在*远离* μ 的位置，不论在哪一边，而对称的双侧拒绝域（rejection region）正好能侦测这种情况。之后的一切，都归结为把 $\overline{X}$ 标准化（standardise），在每个假设之下各做一次。

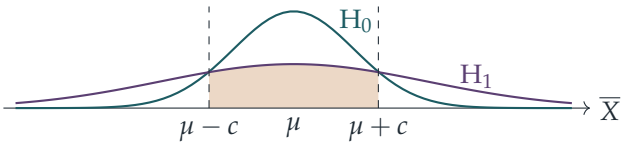
解答

为何采用双侧检验

由于 $X \sim \mathrm{N}(\mu, \sigma^2)$，而且各样本值互相独立，样本平均值（sample mean）满足

$$\overline{X} \sim \mathrm{N}\left(\mu,\ \frac{\sigma^2}{n}\right).$$

无论在哪一个假设之下，$\overline{X}$ 都以已知的值 μ 为中心；唯一的分别是在 H_1 之下，它的标准差（standard deviation）$\sigma_1/\sqrt{n}$ 比 $\sigma_0/\sqrt{n}$ 大。$\overline{X}$ 接近 μ 的值在 H_0 之下相对较有可能出现，而远离 μ 的值则在 H_1 之下相对较有可能出现（较阔的曲线尾部较厚；见附图）。由于两个分布都关于 μ 对称，一个很大的 $\overline{X}$ 值与一个同样程度地小的值相比，并不是支持 H_1 的更强证据。所以，自然的检验方法是当 $|\overline{X} - \mu|$ 很大时拒绝 H_0，不论方向；而根据对称性，接受区间应以 μ 为中心：当 $\mu - c < \overline{X} < \mu + c$ 时接受 H_0。在这里采用单侧检验（one-sided test）是错的，因为这样会忽略一半支持分散程度较大的证据。



阴影部分的面积（H_1 为真却接受 H_0）是 β；两条虚线以外的蓝绿色曲线下面积是 α。

求 c

在 H_0 之下，$Z = \dfrac{\sqrt{n}\,(\overline{X} - \mu)}{\sigma_0}$ 服从标准正态分布（standard normal distribution）。当 $|\overline{X} - \mu| \geq c$ 时，即当 $|Z| \geq c\sqrt{n}/\sigma_0$ 时，我们接受 H_1。根据标准正态分布的对称性，

$$\alpha = \mathrm{P}\left(|Z| \geq \frac{c\sqrt{n}}{\sigma_0}\right) = 2\left(1 - \Phi\left(\frac{c\sqrt{n}}{\sigma_0}\right)\right), \qquad \text{所以} \qquad \Phi\left(\frac{c\sqrt{n}}{\sigma_0}\right) = 1 - \tfrac{1}{2}\alpha.$$

由于 Φ 严格递增，这表示 $c\sqrt{n}/\sigma_0 = z_\alpha$，因此

$$c = \frac{z_\alpha\,\sigma_0}{\sqrt{n}}.$$

β 与 n 无关

在 H_1 之下，$W = \dfrac{\sqrt{n}\,(\overline{X} - \mu)}{\sigma_1}$ 服从标准正态分布，而

$$\begin{aligned}\beta = \mathrm{P}(|\overline{X} - \mu| < c) &= \mathrm{P}\left(|W| < \frac{c\sqrt{n}}{\sigma_1}\right) = \mathrm{P}\left(|W| < \frac{z_\alpha\sigma_0}{\sigma_1}\right) \\ &= 2\,\Phi\left(\frac{z_\alpha\sigma_0}{\sigma_1}\right) - 1,\end{aligned}$$

其中我们代入了上一部分得到的 $c\sqrt{n} = z_\alpha\sigma_0$。$c$ 之中的 $\sqrt{n}$ 与标准化时出现的 $\sqrt{n}$ 互相抵消，所以 β 只取决于 α 和 σ_0/σ_1，与 n 无关。

想一想 样本愈大，检验理应愈好吧?

通常确是如此，因为随着 n 增大，两个抽样分布（sampling distribution）会互相分开。但在这里它们永远不会分开：两者都以 μ 为中心，而增大 n 会令两个标准差按同一个因子 $\sqrt{n}$ 缩小。整幅图只是按比例缩放，所以面积 α 和 β 都不变。用样本平均值来检验方差（variance），是一个差劣的统计量；它把各个数值所包含的关于分散程度的信息都丢掉了。

比值 σ_1/σ_0 的最小值

设 $r = \sigma_1/\sigma_0 > 1$。

条件 $\alpha < 0.05$。这表示 $\Phi(z_\alpha) = 1 - \frac{1}{2}\alpha > 0.975 \approx \Phi(1.960)$，所以，由于 Φ 递增，$z_\alpha > 1.960$。

条件 $\beta < 0.05$。这表示 $2\Phi(z_\alpha/r) - 1 < 0.05$，即 $\Phi(z_\alpha/r) < 0.525 \approx \Phi(0.063)$，所以 $z_\alpha/r < 0.063$，亦即

$$r > \frac{z_\alpha}{0.063}.$$

综合两者，我们需要 $r > z_\alpha/0.063 > 1.960/0.063$。反过来说，若 $r > 1.960/0.063$，我们可以选取刚好大于 1.960 且满足 $z_\alpha < 0.063\,r$ 的 z_α，这样两个条件都成立。因此这个比值必须大于

$$\frac{1.960}{0.063} \approx 31.1,$$

所以 σ_1/σ_0 的最小值约为 31。两个标准差必须相差约三十倍，样本平均值才能如此可靠地把它们区分开来，而且不论样本大小如何都是这样。

验算

- 以 $\mu = 3$、$\sigma_0 = 1$、$\sigma_1 = 2.5$、$\alpha = 0.1$ 及 $n = 1, 4, 9, 25$ 进行模拟，在 H_0 之下的拒绝率每次都约为 0.10，而在 H_1 之下的接受率对每个 n 都约为 0.489，与 $2\Phi(1.645/2.5) - 1 \approx 0.489$ 吻合。
- 当 $r = 1.960/0.063$ 而 $\alpha = 0.05$ 恰好成立时，$\beta = 2\Phi(0.063) - 1 \approx 0.050$，一如所料正好在边界上。若使用 Φ 的精确反函数值，会得到 $r \approx 31.3$，所以答案 31 不受题目所给表值的舍入误差影响。
- 合理性检查：若 $\sigma_1 = \sigma_0$，则 $\beta = 2\Phi(z_\alpha) - 1 = 1 - \alpha$，当两个假设重合时必然如此。

常见错误

- 使用单侧的拒绝域，例如 $\overline{X} > \mu + c$：在 H_1 之下平均值并没有移动，所以没有理由偏重其中一边的尾部。
- 写成 $\alpha = 1 - \Phi(c\sqrt{n}/\sigma_0)$，忘记了拒绝域有两个尾部。
- 把关于 r 的不等式方向弄反：β 小需要 z_α/r 小，所以 r 要大。

STEP 之外

这道题展示了两类错误之间的取舍，以及检验统计量（test statistic）的选择为何重要。奈曼–皮尔逊引理（Neyman–Pearson lemma）指出，最强大的检验是比较两个假设之下的似然（likelihood）；在这里，这会引出统计量 $\sum(X_i - \mu)^2$，它服从的 χ^2_n 分布确实会随着

n 增大而把两个假设区分开来，这与 $\bar{X}$ 不同。想找一本易读的入门书了解这些概念，可参阅 Blitzstein 与 Hwang 的 *Introduction to Probability*（概率论入门），或 Grimmett 与 Welsh 的 *Probability: An Introduction*（概率导论）。

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

2004 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。

