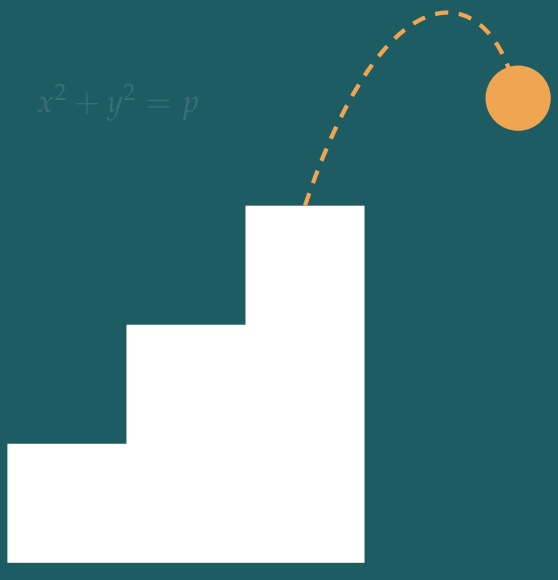
STEP 备试 · 大学数学

Beyond STEP

STEP I, II 及 III

1998 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

1998 年详细解答

Keith McNulty

Beyond STEP：以历届 STEP 试题为门，走进大学数学

1998 年详细解答

1998 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 数字模式的计数 纯数学 3

第 2 题 代换积分法 纯数学 6

第 3 题 真假命题的证明 纯数学 9

第 4 题 约束最优化 纯数学 13

第 5 题 复数与旋转 纯数学 17

第 6 题 三角函数递推关系 纯数学 20

第 7 题 等比级数 纯数学 23

第 8 题 二阶微分方程 纯数学 26

第 9 题 摆的碰撞 力学 29

第 10 题 抛体运动 力学 32

第 11 题 跨过滑轮的连接质点 力学 35

第 12 题 骰数与独立性 概率与统计 38

第 13 题 放回抽样 概率与统计 42

第 14 题 几何概率 概率与统计 45

STEP II

第 1 题 丢番图方程与奇偶性 纯数学 48

第 2 题 二项式级数近似 纯数学 51

第 3 题 裂项相消级数 纯数学 54

第 4 题 分部积分法 纯数学 57

第 5 题 复数的三角不等式 纯数学 60

第 6 题 摆线与法线 纯数学 63

第 7 题 以导数证明不等式 纯数学 67

第 8 题 向量几何 纯数学 70

第 9 题 桁架的静力学 力学 . . . 74
第 10 题 移动楔上的质点 力学 . . . 78
第 11 题 抛体运动 力学 . . . 82
第 12 题 贝叶斯定理 概率与统计 . . . 86
第 13 题 整数分布 概率与统计 . . . 89
第 14 题 二项分布与泊松分布的期望值 概率与统计 . . . 92

STEP III

第 1 题 曲线描绘 纯数学 . . . 95
第 2 题 Beta 积分的递推关系 纯数学 . . . 99
第 3 题 线性递推关系 纯数学 . . . 102
第 4 题 极坐标玫瑰线 纯数学 . . . 106
第 5 题 矩阵指数 纯数学 . . . 110
第 6 题 多面体的体积 纯数学 . . . 114
第 7 题 曲线描绘与积分 纯数学 . . . 118
第 8 题 球面的向量几何 纯数学 . . . 122
第 9 题 转动惯量 力学 . . . 125
第 10 题 斜碰撞 力学 . . . 129
第 11 题 振幅摆 力学 . . . 133
第 12 题 网络可靠度 概率与统计 . . . 137
第 13 题 最大似然估计 概率与统计 . . . 140
第 14 题 无偏估计量 概率与统计 . . . 144

详细解答

STEP I 1998，第 1 题

纯数学 · 数字模式的计数

> **STEP I 1998，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/98-s1-q1

关键思路

不要尝试把数逐个列出。应该先选定出现的是哪两个数字，再计数同时用上这两个数字的五位数字串（string）。唯一的麻烦是一个数不能以 0 开头，所以我们按 0 是否两个数字之一来分开计数。另外，别忘了检查范围的两端。

解答

由 10 000 到 99 999 的每个整数都恰好有五位，而且首位数字不是零。范围内余下的唯一整数是 100 000，我们留到最后处理。

一个五位数恰好含有两个不同的数字，即是说它的所有数字都来自某一对 $\{a,b\}$，而且 a 和 b 都确实出现。每个这样的数都决定了它所属的那一对，所以我们可以逐对计数，然后相加。

情况 1：两个数字都不是 0

从 $1,2,\ldots,9$ 中选出两个不同的数字，共有 $\binom{9}{2}=36$ 种选法。对于固定的一对 $\{a,b\}$，五个位置各自可以是 a 或 b，得出 $2^5=32$ 个数字串。其中两个只用了一个数字（$aaaaa$ 和 $bbbbb$），所以有 $32-2=30$ 个同时用上两个数字。它们全都不以 0 开头，所以这 30 个都是真正的五位数。这个情况得出

$$36\times 30=1080.$$

? **想一想** 为甚么不能用完全相同的方法处理 $\{0,a\}$ 这一对？

因为在 32 个由 0 和 a 组成的数字串之中，有一半以 0 开头，而这些并不是五位数（例如

$0a0aa$ 其实是一个四位数）。所以我们必须先固定首位数字，然后才计数。

情况 2：其中一个数字是 0

另一个数字 a 可以是 $1,\ldots,9$ 中的任何一个，共 9 种选择。首位数字不能是 0，所以必定是 a。余下四个位置各自可以是 0 或 a，得出 $2^4 = 16$ 种可能。当中恰好有一个，即 $aaaa$，不含 0，这样整个数便只有一个不同的数字。所以对每个 a 都有 $16 - 1 = 15$ 个数，这个情况得出

$$9 \times 15 = 135.$$

端点 100 000

这个范围包括两个端点，而 100 000 只用了 1 和 0 这两个数字，所以它也要计算在内。

总数

$$1080 + 135 + 1 = 1216.$$

所以这样的整数共有 **1216** 个。

验算

- 用电脑检查由 10 000 到 100 000 的每一个整数，计数其中数字所组成的集合大小为 2 的数，结果是 1216。这也印证了上面的分拆：1080 个不含 0，135 个含有它，再加上 100 000。
- 把同一方法用于一个可以用手逐一列出的较小情况。对于两位数，这方法得出 $\binom{9}{2}(2^2 - 2) + 9(2^1 - 1) = 72 + 9 = 81$。这与直接计数的结果吻合：共有 90 个两位数，减去 9 个各位数字相同的数（repdigit）$11, 22, \ldots, 99$。

常见错误

- 忘记了 100 000，得出 1215。遇到“包括首尾”的范围，务必留意两个端点。
- 计算成 $\binom{10}{2} \times 30 = 1350$，这把以 0 开头的数也算了进去。
- 在情况 2 中减去 2 而不是 1：全是零的数字串在我们把首位数字固定为 a 时已经排除了。

STEP 之外

从所有由 $\{a,b\}$ 组成的数字串中减去只用一个数字的数字串，这个技巧是容斥原理（inclusion–exclusion principle）最简单的情况。一般而言，长度为 n 而且用上全部 k 个给定符号的字串，数目是 $\sum_{j=0}^{k}(-1)^j\binom{k}{j}(k-j)^n$，也就是 $k!$ 乘以一个第二类斯特灵数（Stirling number of the second kind）。这类计数结果在组合数学（combinatorics）和概率（probability）中随处可见，例如优惠券收集问题（coupon collector problem）。想多做这类计数题，可以试试 Grimmett and Welsh 的 *Probability: An Introduction*，或者 Siklos 的 *Advanced Problems in Mathematics* 中的组合数学题目。

详细解答

STEP I 1998，第 2 题

纯数学 · 代换积分法

STEP I 1998，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/98-s1-q2

关键思路

那个不好处理的式子 $(x^2+1)^{1/2}+x$ 位于 f 之内，所以最自然的做法就是把它整个设为 t。出人意表的是，这个关系很容易反过来解：先把根式单独移到一边再平方，x^2 项便会互相抵消，剩下的 x 只是 t 的一个简单函数。第一条恒等式（identity）一旦到手，第二个积分不过是 $\mathrm{f}(t)=t^{-3}$ 的特殊情况。

解答

变量代换

设

$$t=(x^2+1)^{1/2}+x.$$

当 x 由 0 走到 ∞ 时，$(x^2+1)^{1/2}$ 和 x 都在递增，所以 t 严格递增（strictly increasing），由 $x=0$ 时的 $t=1$ 出发，并随 $x\to\infty$ 趋向 ∞。因此这个代换（substitution）是由 $[0,\infty)$ 到 $[1,\infty)$ 的一对一（one-to-one）满射，新的积分上下限是 1 和 ∞。

想一想 怎样把 x 用 t 表示出来?

把平方根单独移到一边，然后平方：由 $t-x=(x^2+1)^{1/2}$ 得

$$t^2-2tx+x^2=x^2+1 \quad\Longrightarrow\quad x=\frac{t^2-1}{2t}=\frac{1}{2}\left(t-t^{-1}\right),$$

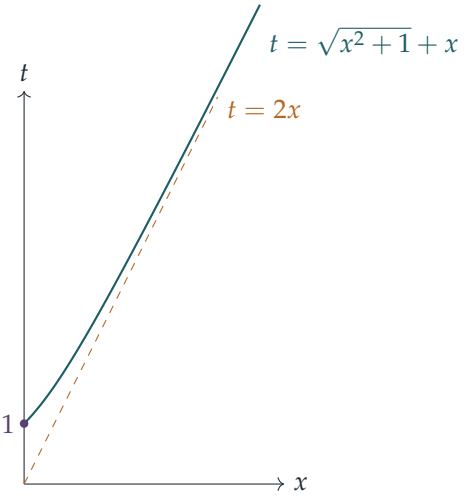
其中因为 $t\geq 1$, 除以 t 并无问题。(平方可能会引入增根, 但这里 $t-x=(t+t^{-1})/2>0$, 所以我们平方的那个根确实是正数。) 求导得

$$\frac{\mathrm{d}x}{\mathrm{d}t}=\frac{1}{2}\left(1+t^{-2}\right).$$

因此

$$\int_0^\infty \mathrm{f}\big((x^2+1)^{1/2}+x\big)\,\mathrm{d}x = \int_1^\infty \mathrm{f}(t)\cdot\frac{1}{2}\big(1+t^{-2}\big)\,\mathrm{d}t = \frac{1}{2}\int_1^\infty \big(1+t^{-2}\big)\mathrm{f}(t)\,\mathrm{d}t,$$

这正是所求的恒等式。



从草图可见，t 由 1 开始稳定上升，所以值域中没有任何部分被重复覆盖。

最后部分

取 $\mathrm{f}(t) = t^{-3}$。恒等式的左边正好就是我们要求的积分，所以

$$\begin{aligned}\int_0^\infty \big((x^2+1)^{1/2}+x\big)^{-3}\,\mathrm{d}x &= \frac{1}{2}\int_1^\infty \big(t^{-3}+t^{-5}\big)\,\mathrm{d}t \\ &= \frac{1}{2}\Big[-\tfrac{1}{2}t^{-2}-\tfrac{1}{4}t^{-4}\Big]_1^\infty \\ &= \frac{1}{2}\Big(\frac{1}{2}+\frac{1}{4}\Big) = \frac{3}{8},\end{aligned}$$

正如所求。两个广义积分（improper integral）都收敛（converge），因为 t^{-3} 和 t^{-5} 在 $[1,\infty)$ 上可积。

> **想一想** 有没有更简洁的方法去理解这个代换?

有：设 $x = \sinh\theta$。这样 $(x^2+1)^{1/2} = \cosh\theta$，所以 $t = \cosh\theta + \sinh\theta = \mathrm{e}^\theta$。我们的公式 $x = \frac{1}{2}(t - t^{-1})$ 其实就是 $\sinh\theta = \frac{1}{2}(\mathrm{e}^\theta - \mathrm{e}^{-\theta})$ 换了个样子而已。

验算

- 当 $x = 0$ 时，$t = 1$，而 $x = \frac{1}{2}(1-1) = 0$：反函数公式吻合。当 $x = \frac{3}{4}$ 时，$t = \frac{5}{4} + \frac{3}{4} = 2$，而 $\frac{1}{2}(2 - \frac{1}{2}) = \frac{3}{4}$。
- 当 x 很大时，$t \approx 2x$，所以最后部分的被积函数（integrand）表现得像 $(2x)^{-3}$，积分必定收敛。
- 数值上，对原来的积分使用辛普森法则（Simpson's rule）得 0.375000，与 $\frac{3}{8}$ 相符。第一部分的恒等式亦可用数值方法验证：取 $\mathrm{f}(t) = \mathrm{e}^{-t}$（两边均为 0.258187）和 $\mathrm{f}(t) = 1/(1+t^4)$（两边均为 0.188387）时都成立。

常见错误

- 忘记更改积分上下限：下限变成 $t = 1$，而不是 $t = 0$。
- 把 $\mathrm{d}t$ 用 x 表示，然后苦苦设法消去 x；先把 x 用 t 表示，再对它求导，会容易得多。

STEP 之外

这条恒等式说的是在单调（monotonic）映射下的变量代换；在分析学中，对于广义积分，这是通过取 $[0, X]$ 上积分的极限来证明其合理性的。这个代换本身就是 sinh 的反函数：$\ln\left(x + \sqrt{x^2+1}\right) = \operatorname{arsinh} x$，这正是 t 原来等于 $\mathrm{e}^{\operatorname{arsinh} x}$ 的原因。如欲参考有关广义积分和代换的严谨论述，可参阅 Bartle 与 Sherbert 的 *Introduction to Real Analysis*（实分析入门）。

详细解答

STEP I 1998，第 3 题

纯数学 · 真假命题的证明

> **STEP I 1998，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/98-s1-q3

关键思路

每一部分所需的工具都不同，而首要工作是判断该命题是真是假。答“真”需要一个涵盖所有情况的证明；答“假”则只需一个反例（counterexample），或者一个说明符合要求的对象根本不可能存在的论证。取对数（logarithm）可解决 (i)，选一个合适的 θ 值可解决 (ii)，多项式（polynomial）在 θ 很大时的表现可解决 (iii)，而一个完全平方（perfect square）可解决 (iv)。

解答

第 (i) 部分：真

由于 $a > 0$，可以写成 $a = \mathrm{e}^{\ln a}$。于是

$$a^{\ln b} = \left(\mathrm{e}^{\ln a}\right)^{\ln b} = \mathrm{e}^{\ln a \,\ln b},$$
$$b^{\ln a} = \left(\mathrm{e}^{\ln b}\right)^{\ln a} = \mathrm{e}^{\ln b \,\ln a}.$$

因为乘法符合交换律，两个指数相等，所以对所有 $a, b > 0$ 都有 $a^{\ln b} = b^{\ln a}$。（等价的说法是：两边都是正数，而且对数相同，都是 $\ln a \ln b$。）

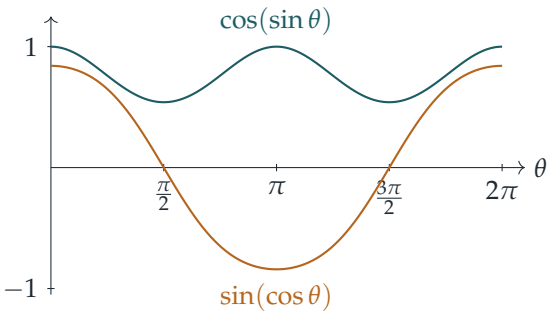
第 (ii) 部分：假

一个反例已经足够。取 $\theta = 0$：

$$\cos(\sin 0) = \cos 0 = 1, \qquad \sin(\cos 0) = \sin 1.$$

由于 $0 < 1 < \frac{\pi}{2}$，可知 $0 < \sin 1 < 1$（事实上 $\sin 1 \approx 0.841$），所以两边并不相等，命题为假。

事实上两边永远不会相等；从图像可见，$\cos(\sin\theta)$ 总是在 $\sin(\cos\theta)$ 之上。



要证明这一点（本题并不要求），留意 $\sin(\cos\theta) \le \sin|\cos\theta| = \cos\left(\frac{\pi}{2} - |\cos\theta|\right)$，以及 $\cos(\sin\theta) = \cos|\sin\theta|$。$|\sin\theta|$ 和 $\frac{\pi}{2} - |\cos\theta|$ 都落在 $[0, \frac{\pi}{2}]$ 之内，而余弦在这个区间上递减；又因为 $|\sin\theta| + |\cos\theta| \le \sqrt{2} < \frac{\pi}{2}$，所以 $|\sin\theta| < \frac{\pi}{2} - |\cos\theta|$。因此对每一个 θ 都有 $\cos(\sin\theta) > \sin(\cos\theta)$。

第 (iii) 部分：假

设 P 为任意多项式。我们要证明存在某个实数 θ 使得 $|\mathrm{P}(\theta) - \cos\theta| > 10^{-6}$，以下分两种情况讨论。

P 为常数，设 $\mathrm{P}(\theta) = c$。取 $\theta = 0$ 和 $\theta = \pi$，此时 $\cos\theta$ 分别是 1 和 -1，由三角不等式（triangle inequality）得

$$|c-1| + |c+1| \ge |(c+1) - (c-1)| = 2,$$

所以 $|c-1|$、$|c+1|$ 之中至少有一个不小于 1，远远大于 10^{-6}。

P 的次数（*degree*）为 $n \ge 1$，设 $\mathrm{P}(\theta) = a_n\theta^n + \cdots + a_0$，其中 $a_n \ne 0$。当 $\theta \ne 0$ 时，

$$\mathrm{P}(\theta) = a_n\theta^n\left(1 + \frac{a_{n-1}}{a_n\theta} + \cdots + \frac{a_0}{a_n\theta^n}\right),$$

当 $\theta \to \infty$ 时，括号内的式子趋向 1，而 $|a_n\theta^n| \to \infty$。所以 $|\mathrm{P}(\theta)| \to \infty$，我们可以选一个 θ 使得 $|\mathrm{P}(\theta)| > 2$。对这个 θ，由于 $|\cos\theta| \le 1$，

$$|\mathrm{P}(\theta) - \cos\theta| \ge |\mathrm{P}(\theta)| - |\cos\theta| > 2 - 1 = 1 > 10^{-6}.$$

在两种情况下，该不等式都会在某处不成立，所以这样的多项式并不存在。

想一想 可是 $\cos\theta$ 的麦克劳林级数（Maclaurin series）是一个类似多项式的和，而且可以要多接近就有多接近。为甚么把它截断不行呢?

在任何固定的有界区间（bounded interval）上，只要取足够多项，截断后的麦克劳林级数与 $\cos\theta$ 的差距*确实*在 10^{-6} 之内。但命题要求的是同一个多项式对*所有实数* θ 都适用，而每一个非常数多项式最终都会奔向 $\pm\infty$，$\cos\theta$ 却始终困在 $[-1,1]$ 之内。正是“对所有实数 θ”这几个字令命题为假。

第 (iv) 部分：真

对 $x>0$,

$$x^4+3+x^{-4}-5=x^4-2+x^{-4}=\left(x^2-x^{-2}\right)^2\geq 0,$$

因为实数的平方必为非负数。因此对所有 $x>0$ 都有 $x^4+3+x^{-4}\geq 5$，而等号恰好在 $x^2=x^{-2}$ 时成立，即 $x=1$ 时。（另一个做法：由算术平均—几何平均不等式（AM–GM inequality）得 $x^4+x^{-4}\geq 2\sqrt{x^4\cdot x^{-4}}=2$。）

总结：(i) 真，(ii) 假，(iii) 假，(iv) 真。

验算

- (i)：取 $a=2$、$b=3$，$2^{\ln 3}$ 和 $3^{\ln 2}$ 都约等于 2.141；随机选取正数 a,b，两边相符，相对误差约为 10^{-15}。
- (ii)：在一整个周期内对 θ 取样，$\cos(\sin\theta)-\sin(\cos\theta)$ 从未低于约 0.107。
- (iii)：cos 的 20 次麦克劳林多项式在 $\theta=1$ 时很准确，但在 $\theta=10$ 时误差约为 7.5，在 $\theta=30$ 时误差约为 10^{11}。
- (iv)：在细密的格点上，x^4+3+x^{-4} 的最小值是 5，在 $x=1$ 时取得。

常见错误

- 只检查几个数值就想“证明”一个假命题，或者因为代数运算出错而“否证”了一个真命题：动笔之前，先判断每一部分是真是假。
- 在 (iii) 中忘记了常数多项式：“趋向无穷”的论证需要次数至少为 1。
- 在 (iii) 中引用泰勒级数（Taylor series），把它当作处处适用的多项式；它并不是。

STEP 之外

第 (iii) 部分与魏尔斯特拉斯逼近定理（Weierstrass approximation theorem）形成对比：在任何有界闭区间上，每一个连续函数（continuous function）都可以用多项式一致逼近（uniform approximation），精确度要多高就有多高。但在整条实数线上一致逼近是不可能的，原因正是上面的增长论证。Abbott 的 *Understanding Analysis*（一本实分析入门书）讨论了魏尔斯特拉斯定理和一致收敛（uniform convergence）。

详细解答

STEP I 1998，第 4 题

纯数学 · 约束最优化

STEP I 1998，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/98-s1-q4

关键思路

内接（inscribed）于圆的长方形，其对角线就是直径（diameter），所以它的边长 a 和 b 由毕氏定理（Pythagoras' theorem）连系起来：a^2+b^2 是固定的。于是整个问题就变成：当 a^2+b^2 保持不变时，$a+b$、a^2+b^2 和 a^3+b^3 会怎样变化。$n=2$ 这个情况是转捩点：低于此值时正方形胜出，恰在此值时全部打成平手，高于此值时正方形反而敬陪末座。

解答

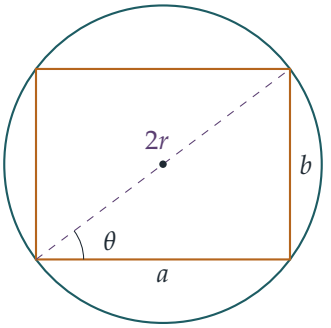
设圆的半径为 r，长方形的边长为 $a,b>0$。长方形的每个角都是直角，所以它所对的弦是直径（半圆上的圆周角）；因此每条对角线都是直径，并且

$$a^2+b^2=(2r)^2=4r^2.$$

反过来，任何满足 $a^2+b^2=4r^2$ 的 $a,b>0$ 都给出一个可放进圆内、且顶点都在圆上的长方形。所以不妨写成

$$a=2r\cos\theta,\qquad b=2r\sin\theta,\qquad 0<\theta<\tfrac{\pi}{2},$$

其中 θ 是对角线与长度为 a 的边之间的夹角。



周界

周界（perimeter）为 $P = 2(a+b)$。由于 $(a-b)^2 \geq 0$，我们有 $2ab \leq a^2 + b^2$，所以

$$(a+b)^2 = a^2 + b^2 + 2ab \leq 2(a^2+b^2) = 8r^2,$$

等号成立当且仅当 $a = b$。因为 $a + b > 0$，由此得 $a + b \leq 2\sqrt{2}\,r$，即

$$P \leq 4\sqrt{2}\,r,$$

而 $4\sqrt{2}\,r$ 这个值恰好在 $a = b = \sqrt{2}\,r$ 时取得。所以周界最大的长方形就是正方形，正如题目所要求证明的。

$n = 2$ 的情况

四条边的平方和是

$$2(a^2+b^2) = 8r^2,$$

不论是哪一个长方形都一样。所以每个内接长方形都给出相同的值：它们之间无分高下，特别是正方形并不比其他长方形优胜。

$n = 3$ 的情况

现在考虑 $S = 2(a^3+b^3)$。由于每条边都比直径短，$0 < a < 2r$ 且 $0 < b < 2r$，所以 $a^3 < 2r\,a^2$ 及 $b^3 < 2r\,b^2$。两式相加，

$$S = 2(a^3+b^3) < 2 \cdot 2r(a^2+b^2) = 16r^3.$$

另一方面，当 $\theta \to 0$ 时（一个又长又窄的长方形，其中 $a \to 2r$ 而 $b \to 0$），$S \to 2(2r)^3 = 16r^3$。所以 S 可以任意接近 $16r^3$，却永远达不到它：*并不存在使 S 最大的长方形*。唯一能达到 $16r^3$ 的“形状”是退化（degenerate）的情形，即把一条直径来回描两次，而那并不是长方形。

想一想 那么正方形又怎样了?

它由最好变成了最差。设 $f(\theta)=\cos^3\theta+\sin^3\theta$，则 $S=16r^3f(\theta)$，而

$$f'(\theta)=-3\cos^2\theta\sin\theta+3\sin^2\theta\cos\theta=3\sin\theta\cos\theta\,(\sin\theta-\cos\theta).$$

在 $0<\theta<\frac{\pi}{2}$ 上，因子 $\sin\theta\cos\theta$ 是正的，而 $\sin\theta-\cos\theta$ 在 $\theta<\frac{\pi}{4}$ 时为负，在 $\theta>\frac{\pi}{4}$ 时为正。所以 f 先递减后递增，而正方形（$\theta=\frac{\pi}{4}$）给出的是最小值，

$$S_{\min}=16r^3\left(2\cdot\frac{1}{2\sqrt{2}}\right)=8\sqrt{2}\,r^3.$$

总括而言，对于 $n=3$，S 的取值填满区间 $[8\sqrt{2}\,r^3,\,16r^3)$：正方形给出最小值，而没有任何长方形给出最大值。

对每个 $n>2$ 情况都一样：在 $0<\theta<\frac{\pi}{2}$ 上有 $\cos^n\theta+\sin^n\theta<\cos^2\theta+\sin^2\theta=1$，因为 $\cos\theta$ 和 $\sin\theta$ 都严格介乎 0 与 1 之间，所以只有把长方形压扁，才能趋近上确界 $2(2r)^n$。

验算

- 取 $r=1.7$，我们以数值方法让 θ 扫过 $(0,\frac{\pi}{2})$：周界在 $\theta=\frac{\pi}{4}$ 处达到峰值，其值为 $4\sqrt{2}\,r\approx 9.617$；平方和恒等于 $8r^2=23.12$；而立方和则由接近 $16r^3\approx 78.61$ 下降至正方形处的 $8\sqrt{2}\,r^3\approx 55.58$，然后再度上升，始终达不到 78.61。
- 一个简单的合理性检查：单位圆内的 $\sqrt{2}\times\sqrt{2}$ 正方形周界为 $4\sqrt{2}\approx 5.66$，而一个 1.9×0.62 的长方形（几乎就是一条直径）周界约为 5.04。

常见错误

- 说 $n=3$ 时的最大值是 $16r^3$，“在 $\theta=0$ 处取得”。那是一条线段，不是长方形，所以最大值并不存在；你必须指出这一点，并说明为何不等式是严格的。
- 求出 $n=3$ 时的驻点（stationary point）$\theta=\frac{\pi}{4}$ 后，没有检查其性质就把它当作极大点。

STEP 之外

$n=3$ 的情况让我们初次体会上确界（supremum）与最大值（maximum）的分别：开区间上的连续函数（continuous function）不一定取得它的界，这正是极值定理（extreme value theorem）坚持要求有界闭区间的原因。在 $n=2$ 处出现的逆转，是幂平均不等式（power-mean inequalities）的一个例子，这类不等式比较不同 p 值下的 $\left(\frac{a^p+b^p}{2}\right)^{1/p}$。

Abbott 的 *Understanding Analysis*（一本数学分析入门书）对上确界和极值定理有严谨细致的论述。

详细解答

STEP I 1998，第 5 题

纯数学 · 复数与旋转

> **STEP I 1998，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/98-s1-q5

关键思路

两部分都是让复数（complex number）的代数运算替我们完成工作。在第 (i) 部分，由正六边形（regular hexagon）的一个顶点走到下一个顶点，就是绕中心旋转（rotation）$60°$，而绕原点的旋转不过是乘以一个模（modulus）为 1 的复数，所以我们相对于 Q 来计算。在第 (ii) 部分，一个数是实数，当且仅当它的虚部（imaginary part）为零，所以我们把每个分母实数化（realise），再令虚部等于零。

解答

第 (i) 部分

相对于中心来计算。由 Q 到 A 的向量为

$$A - Q = (10 + 2i) - (4 + 6i) = 6 - 4i.$$

正六边形的六个顶点均匀地分布在以 Q 为圆心的圆上，所以角 AQB 是 $360°/6 = 60°$，而 QB 与 QA 等长。顺时针移动即是转过 $-60°$，也就是乘以

$$\mathrm{e}^{-i\pi/3} = \cos\tfrac{\pi}{3} - i\sin\tfrac{\pi}{3} = \tfrac{1}{2} - \tfrac{\sqrt{3}}{2}\,i.$$

> ? **想一想** 为甚么要乘 $A - Q$，而不是 A 本身？

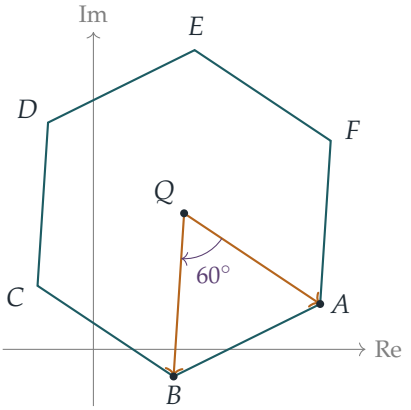
乘以 $\mathrm{e}^{-i\pi/3}$ 是绕原点旋转。要绕 Q 旋转，我们先把 Q 移到原点（减去 Q），旋转，然后

移回去（加上 Q）。所以

$$\begin{aligned} B - Q &= (6-4i)\left(\tfrac{1}{2} - \tfrac{\sqrt{3}}{2}\,i\right) \\ &= 3 - 3\sqrt{3}\,i - 2i + 2\sqrt{3}\,i^2 \\ &= (3-2\sqrt{3}) - (2+3\sqrt{3})\,i, \end{aligned}$$

再加上 $Q = 4+6i$，得

$$\boxed{B = (7-2\sqrt{3}) + (4-3\sqrt{3})\,i} \approx 3.536 - 1.196\,i.$$



图形印证了方向：A 位于 Q 的右下方，由它出发顺时针走，下一个顶点几乎在 Q 的正下方，正如我们求得的 B 那样。

第 (ii) 部分

把分子和分母同乘以共轭复数（conjugate），将每个分母实数化，当中用到 $(1+ki)(1-ki) = 1+k^2$：

$$\frac{1}{1+i} = \frac{1-i}{2}, \qquad \frac{1}{1+2i} = \frac{1-2i}{5}, \qquad \frac{1}{1+3i} = \frac{1-3i}{10}.$$

由于 a、b、c 都是实数，整个算式的虚部为

$$-\frac{a}{2} - \frac{2b}{5} - \frac{3c}{10} = -\frac{5a+4b+3c}{10}.$$

这个算式是实数，当且仅当这个虚部为零，所以所求的条件是

$$\boxed{5a+4b+3c=0}.$$

（任何非零整数倍，例如 $-5a-4b-3c=0$，都是同一个条件。）事实上，由于我们求出的是虚部的准确值，这个条件既是充分条件（sufficient），也是必要条件（necessary）。

验算

- 在第 (i) 部分，$|A-Q| = |6-4i| = \sqrt{52} = 2\sqrt{13}$。我们求得的 B 给出 $|B-Q|^2 = (3-2\sqrt{3})^2 + (2+3\sqrt{3})^2 = 21 - 12\sqrt{3} + 31 + 12\sqrt{3} = 52$，所以 B 确实在正确的圆上；而数值上亦有 $|B-A| \approx 7.211 = |A-Q|$，这正是等边三角形 QAB 所必须满足的。
- 在第 (ii) 部分，试取 $a=3$、$b=0$、$c=-5$（它们满足 $5a+4b+3c=0$）：$\frac{3(1-i)}{2} - \frac{5(1-3i)}{10} = \frac{3}{2} - \frac{1}{2} + \left(-\frac{3}{2} + \frac{3}{2}\right) i = 1$，这是实数。
- 以 10 000 组满足 $c = -(5a+4b)/3$ 的随机三元组作数值检验，每一次虚部都是零。

常见错误

- 逆时针旋转（乘以 $e^{i\pi/3}$）会得到 $(7+2\sqrt{3}) + (4+3\sqrt{3})i$，那是顶点 F，而不是 B。
- 把 A 绕原点旋转，而不是绕 Q 旋转。
- 在第 (ii) 部分，令实部等于零；或者忘记了 a、b、c 必须是实数，才可以这样直接读出虚部。

STEP 之外

乘以 $e^{i\theta}$ 是一个旋转，而映射 $z \mapsto \alpha z + \beta$ 则是先旋转和放大（enlargement），再作平移（translation）；这些都是莫比乌斯变换（Möbius transformations）最简单的例子，而莫比乌斯变换是复分析（complex analysis）入门课程的研习内容。以原点为中心的正多边形，其顶点就是 z_0 乘以各个单位根（roots of unity），这个事实是大量代数和几何的起点。Needham 的 *Visual Complex Analysis*（《复分析：可视化方法》）把这种从几何角度看复数乘法的观点发挥得十分精彩。

详细解答

STEP I 1998，第 6 题

纯数学 · 三角函数递推关系

STEP I 1998，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/98-s1-q6

关键思路

递推关系（recurrence）的每一步都是把某个量与一个余弦取平均，而半角恒等式（half-angle identity）$\frac{1}{2}(1+\cos\theta) = \cos^2\frac{\theta}{2}$ 会把这个平均变成一个完全平方。b_{n+1} 的公式中的平方根随即又把这个平方消去。一旦留意到每一项都以前一项为因子，一般规律就不难猜出，而验证它只需一步归纳法（induction）。

解答

全题中 $0 < x < \frac{\pi}{2}$，所以每个角 $x/2^k$（$k \geq 1$）都位于 $(0, \frac{\pi}{4})$ 内，其余弦为正。在 b_{n+1} 的定义中，我们取正的平方根。

求 a_2 和 b_2

利用 $1+\cos\theta = 2\cos^2\frac{\theta}{2}$，

$$a_2 = \tfrac{1}{2}(\cos x + 1) = \cos^2\frac{x}{2}, \qquad b_2 = (a_2 b_1)^{1/2} = \left(\cos^2\frac{x}{2}\right)^{1/2} = \cos\frac{x}{2},$$

其中最后一步用到 $\cos\frac{x}{2} > 0$。

求 a_3 和 b_3

抽出公因子 $\cos\frac{x}{2}$，然后再用一次半角恒等式，这次取 $\theta = \frac{x}{2}$：

$$a_3 = \tfrac{1}{2}\left(\cos^2\frac{x}{2} + \cos\frac{x}{2}\right) = \cos\frac{x}{2}\cdot\tfrac{1}{2}\left(1+\cos\frac{x}{2}\right) = \cos\frac{x}{2}\cos^2\frac{x}{4}.$$

于是

$$b_3 = (a_3 b_2)^{1/2} = \left(\cos^2\frac{x}{2}\cos^2\frac{x}{4}\right)^{1/2} = \cos\frac{x}{2}\cos\frac{x}{4},$$

因为两个余弦都是正数。这就是题目要求的两个表达式。

一般规律

比较目前得到的各项，

$$b_2 = \cos\frac{x}{2}, \quad b_3 = \cos\frac{x}{2}\cos\frac{x}{4}, \qquad a_2 = b_2\cos\frac{x}{2}, \quad a_3 = b_3\cos\frac{x}{4}.$$

每个 b 都比前一个多乘一个余弦，角度减半；而每个 a 则是对应的 b 再乘一次它最后的那个余弦。因此我们猜想，对 $n \geq 2$，

$$b_n = \cos\frac{x}{2}\cos\frac{x}{4}\cdots\cos\frac{x}{2^{n-1}} = \prod_{k=1}^{n-1}\cos\frac{x}{2^k},$$
$$a_n = b_n\cos\frac{x}{2^{n-1}} = \cos\frac{x}{2}\cos\frac{x}{4}\cdots\cos\frac{x}{2^{n-2}}\,\cos^2\frac{x}{2^{n-1}}.$$

想一想 要验证一个猜想满足递推关系，究竟需要检查甚么？

我们要检查两件事：猜想与初始值吻合；以及若公式对 a_n 和 b_n 成立，则递推关系会给出 a_{n+1} 和 b_{n+1} 的公式。两者合起来，便由归纳法证明公式对所有 $n \geq 2$ 成立。

初始值。当 $n = 2$ 时，猜想给出 $b_2 = \cos\frac{x}{2}$ 和 $a_2 = \cos^2\frac{x}{2}$，与上面求得的结果吻合。

归纳步骤。假设公式对某个 $n \geq 2$ 成立。设 $\theta = x/2^{n-1}$，则 $x/2^n = \theta/2$。于是

$$a_{n+1} = \tfrac{1}{2}(a_n + b_n) = b_n\cdot\tfrac{1}{2}\big(1+\cos\theta\big) = b_n\cos^2\frac{\theta}{2}.$$

对 b_{n+1} 的猜想是 $b_n\cos\frac{\theta}{2}$（多乘一个余弦，角度为 $x/2^n$），所以上式即 $a_{n+1} = b_{n+1}\cos\frac{x}{2^n}$，正是所猜想的 a_{n+1} 的形式。接著，

$$b_{n+1} = \big(a_{n+1}b_n\big)^{1/2} = \left(b_n^2\cos^2\frac{\theta}{2}\right)^{1/2} = b_n\cos\frac{x}{2^n} = \prod_{k=1}^{n}\cos\frac{x}{2^k},$$

这里取正根，而这样做是合理的，因为 b_n 和 $\cos\frac{x}{2^n}$ 都是正数。这正是所猜想的 b_{n+1} 的形式，所以递推关系的两条等式都得到满足，公式对所有 $n \geq 2$ 成立。

事实上，若约定空积（empty product）等于 1，公式亦给出 $b_1 = 1$ 和 $a_1 = \cos x$，所以公式由 $n = 1$ 起已经成立。

验算

- 取 $x = \frac{\pi}{3}$。直接按递推关系计算，得 $a_3 \approx 0.808013$ 和 $b_3 \approx 0.836516$，而 $\cos\frac{\pi}{6}\cos^2\frac{\pi}{12}$ 和 $\cos\frac{\pi}{6}\cos\frac{\pi}{12}$ 给出相同的数值。

- 在 $(0, \frac{\pi}{2})$ 内随机抽取 2000 个 x，我们把递推关系与乘积公式逐一比较，直至 $n = 24$：两者相差不超过 10^{-15}。
- 当 $x \to 0$ 时，每个余弦都趋向 1，a_1 和 b_1 亦然。所有项都趋向 1，与预期相符。

常见错误

- 写下 $\sqrt{\cos^2\theta} = \cos\theta$ 而没有说明为何 $\cos\theta > 0$。这里等式之所以成立，是因为每个角都位于 $(0, \frac{\pi}{4})$ 内。
- 由 $n = 2$ 和 3 猜出规律后便停笔。题目要求我们验证递推关系，所以归纳步骤必不可少。

STEP 之外

把 b_n 乘以 $\sin\frac{x}{2^{n-1}}$，并反复运用 $2\sin\phi\cos\phi = \sin 2\phi$，可得 $b_n = \dfrac{\sin x}{2^{n-1}\sin(x/2^{n-1})}$。当 $n \to \infty$ 时，它趋向 $\dfrac{\sin x}{x}$（取 $x = 1$，递推关系在数十步之后给出 0.841471）。取 $x = \frac{\pi}{2}$，便得到关于 $2/\pi$ 的韦达（Viète）无穷乘积（infinite product）。这种“先取平均、再取几何平均”的做法，与高斯（Gauss）的算术几何平均（arithmetic–geometric mean）一脉相承，也是学习证明数列（sequence）收敛（converge）并求其极限（limit）的一个很好的入门例子。关于数列极限的严谨论述，可参阅 Abbott, *Understanding Analysis*（一本实分析入门教科书）。

详细解答

STEP I 1998，第 7 题

纯数学 · 等比级数

STEP I 1998，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/98-s1-q7

关键思路

不要逐年追踪结余，而应分开追踪每一笔存款：一笔 £a 的存款在银行里赚取了 k 年利息后，会增长至 £ar^k，而总额就是一个等比级数（geometric series）。至于提款，把同一想法反过来用即可。每一笔提款都是由起始结余中的某一部分支付的，而这部分在提取之前一直在赚取利息，所以我们比较的是同一时刻的价值。

解答

静止期结束时的金额

利率为 $\rho\%$，即每年年终结余乘以 $r = 1 + \frac{\rho}{100}$。

把存款的年份编号为 $1, 2, \ldots, m$。在第 k 年年初存入的款项，会在第 $k, k+1, \ldots, m$ 年的年终赚取利息，共 $m - k + 1$ 次。因此在第 m 年年终，它的价值为 ar^{m-k+1}。对 k 求和，第 m 年年终的结余为

$$\begin{aligned} a\left(r^m + r^{m-1} + \cdots + r\right) &= ar\left(1 + r + \cdots + r^{m-1}\right) \\ &= ar\,\frac{r^m - 1}{r - 1}, \end{aligned}$$

这里用了等比级数的求和公式（由于 $\rho > 0$，故有 $r \neq 1$）。

在其后的 l 年里，既无存款亦无提款，所以这个金额再乘以 r 共 l 次。因此这段时期结束时的总额为

$$£a\,\frac{r^{l+1}(r^m - 1)}{r - 1},$$

正如题目所求。把这个金额记作 P。

提款

第一次提款在紧接静止期之后那一年的年初进行，这正是结余为 P 的时刻。设 B_k 为刚完成第 k 次提款后的结余，则

$$B_1 = P - b, \qquad B_k = rB_{k-1} - b \quad (k = 2, \ldots, n),$$

因为在两次提款之间，这笔钱会赚取一年的利息。把这个递归关系（recurrence）逐层展开，得

$$\begin{aligned} B_n &= r^{n-1}P - b\left(r^{n-1} + r^{n-2} + \cdots + r + 1\right) \\ &= r^{n-1}P - b\,\frac{r^n - 1}{r - 1}. \end{aligned}$$

> **想一想** 为甚么 P 所乘的因子是 r^{n-1}，而不是 r^n？

由第一次提款到第 n 次提款之间，只有 $n-1$ 个年终，所以这笔钱只赚取了 $n-1$ 次利息。最后一次提款之后所派的利息无关重要，因为那时户口已经清空。同样道理，第 j 次提款（由最后一次倒数，最后一次为 $j = 0$）在第 n 次提款之前会乘以 r^j。

第 n 次提款之后户口清空，所以 $B_n = 0$。由此得

$$r^{n-1}P = b\,\frac{r^n - 1}{r - 1}.$$

代入 P：

$$a\,\frac{r^{l+n}(r^m - 1)}{r - 1} = b\,\frac{r^n - 1}{r - 1},$$

于是

$$\boxed{\frac{a}{b} = \frac{r^n - 1}{r^{l+n}\,(r^m - 1)}}$$

要理解这条等式，另一个等价的方法是计算“现值”（present value）。以第一次提款的时刻为准，第 k 次提款的价值为 b/r^{k-1}，而这些价值的总和必须等于 P：

$$P = b\left(1 + r^{-1} + \cdots + r^{-(n-1)}\right).$$

验算

- 若 $m = n$，答案简化为 $a/b = r^{-(l+n)}$，即 $b = ar^{l+n}$。这是合理的：把第 k 笔存款与第 k 次提款配对，两者刚好相隔 $m + l = l + n$ 年，所以每笔存款恰好支付一次提款。
- 当 $r \to 1$（没有利息）时，$\dfrac{r^n - 1}{r^m - 1} \to \dfrac{n}{m}$，所以 $a/b \to n/m$，即 $ma = nb$。存入的钱等于提出的钱。

- 取 $\rho = 5$、$m = n = 10$、$l = 5$ 及 $a = 1000$，逐年模拟得出 $P \approx 16855.58$。公式给出 $a/b = 1.05^{-15} \approx 0.48102$，所以 $b \approx 2078.93$；用这个 b 值，模拟的户口最后刚好归零。此外，用电脑以分数作精确运算，随机选取 500 组 r、l、m 和 n 的值，两条公式亦全部得到确认。

常见错误

- 在第一个和式中把 r^{m+1} 写成 r^m。由于存款是在每年*年初*存入的，即使是最后一笔存款，也会赚取一年的利息。
- 把 $r^{n-1}P$ 写成 r^nP。第 n 年年终的利息，是在户口清空之后才派发的。

STEP 之外

这就是年金（annuity）的数学。把各笔款项折现（discounting）至同一日期（即现值）再作比较，是精算数学和金融数学的基本原则。递归关系 $B_k = rB_{k-1} - b$ 是一条一阶线性差分方程（first-order linear difference equation）。它的解是一个特定的常数加上 r^k 的倍数，与我们求解一阶线性微分方程（differential equation）的方法如出一辙；大学一年级有关差分方程和微分方程的课程会探讨这种对应关系。想做更多这类题目，可参阅 Siklos 的 *Advanced Problems in Mathematics*。

详细解答

STEP I 1998，第 8 题

纯数学 · 二阶微分方程

STEP I 1998，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/98-s1-q8

关键思路

这条题目虽然披着流体力学的外衣，骨子里却是一道纯粹的微分方程（differential equation）题。左边是“导数的导数”，所以只需积分两次，并记住两次积分之间除以 r 会令常数变成对数。接着，物理上的合理性（轴上的速度必须有限）确定其中一个常数，而管壁上的无滑移条件（no-slip condition）则确定另一个。

解答

通解

将方程乘以 r，得

$$\frac{\mathrm{d}}{\mathrm{d}r}\left(r\frac{\mathrm{d}v}{\mathrm{d}r}\right) = -kr.$$

对 r 积分一次，

$$r\frac{\mathrm{d}v}{\mathrm{d}r} = -\tfrac{1}{2}kr^2 + A,$$

其中 A 为常数。两边除以 r（这样做没有问题，因为在管内轴线以外的地方 $r > 0$）：

$$\frac{\mathrm{d}v}{\mathrm{d}r} = -\tfrac{1}{2}kr + \frac{A}{r}.$$

再积分一次，由于 $r > 0$，可写 $\ln r$ 而不必写 $\ln|r|$，

$$\boxed{v = -\tfrac{1}{4}kr^2 + A\ln r + B}$$

其中 A 和 B 为任意常数。二阶方程的通解（general solution）含有两个常数，正如所料。

轴线附近的情况

当 $r \to 0^+$ 时，$-\frac{1}{4}kr^2$ 和 B 两项保持有界（分别趋向 0 和 B），而 $\ln r \to -\infty$。因此若 $A \neq 0$，则 $A \ln r$ 趋向 $-\infty$ 或 $+\infty$（视乎 A 的正负），于是 $|v| \to \infty$。要避免这种情况，唯一的方法是取 $A = 0$，此时 $v \to B$，是有限值。

> **想一想** 为甚么我们要坚持 $A = 0$?

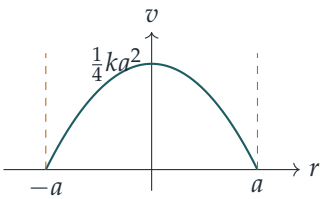
轴线 $r = 0$ 位于流体之内，所以该处的速率必须有限：现实中的流体不会沿着管道中心以无限大的速度流动。$\ln r$ 这个解在 $r > 0$ 时完全满足方程，但在这里从物理角度看却不可接受。（若是两个同轴圆柱之间的流动，轴线并不属于流体的一部分，这个解便可以保留。）

速度分布

取 $A = 0$，便有 $v = B - \frac{1}{4}kr^2$。由条件“在 $r = a$ 处 $v = 0$”得 $B = \frac{1}{4}ka^2$，所以

$$\boxed{v = \tfrac{1}{4}k\left(a^2 - r^2\right).}$$

速度分布（velocity profile）是一条抛物线：速率在轴上最大，为 $\frac{1}{4}ka^2$，到管壁处则降至零。



（图中所示为沿直径切开的截面，虚线代表管壁。）

流量

代入题目给出的公式，

$$\begin{aligned} F &= 2\pi \int_0^a r \cdot \tfrac{1}{4}k\left(a^2 - r^2\right) \mathrm{d}r = \frac{\pi k}{2} \int_0^a \left(a^2 r - r^3\right) \mathrm{d}r \\ &= \frac{\pi k}{2}\left[\frac{a^2 r^2}{2} - \frac{r^4}{4}\right]_0^a = \frac{\pi k}{2} \cdot \frac{a^4}{4}, \end{aligned}$$

所以

$$\boxed{F = \frac{\pi k a^4}{8}.}$$

验算

- 将 $v = \frac{1}{4}k(a^2 - r^2)$ 代回原方程：$r\,v' = -\frac{1}{2}kr^2$，其导数为 $-kr$，再除以 r 得 $-k$，符合要求。
- 平均速率为 $F/(\pi a^2) = \frac{1}{8}ka^2$，恰好是最大速率 $\frac{1}{4}ka^2$ 的一半，这正是抛物线形速度分布广为人知的性质。
- 数值上，取 $k = 2$ 和 $a = 1$，用辛普森法则（Simpson's rule）得 $F \approx 0.785398$，与 $\pi/4$ 吻合。

常见错误

- 第一次积分时漏掉常数，于是失去 $A \ln r$ 一项，令题目中“除非”（unless）的那一部分变得毫无意义。
- 将 $-kr$ 积分得 $-\frac{1}{2}kr^2$ 之后，忘记先除以 r 便再积分。
- 在 F 的积分中漏掉因子 r：通过半径为 r 的薄圆环的流量与其圆周 $2\pi r$ 成正比，而不是与某个常数成正比。

STEP 之外

算子 $\frac{1}{r}\frac{\mathrm{d}}{\mathrm{d}r}\left(r\frac{\mathrm{d}}{\mathrm{d}r}\right)$ 正是拉普拉斯算子（Laplacian）∇^2 在柱坐标（cylindrical coordinates）下作用于只依赖 r 的函数时的形式，所以这条题目其实就是泊松方程（Poisson's equation）$\nabla^2 v = -k$；而 $\ln r$ 这个解，则是点电荷的 $1/r$ 势在二维的对应。$F \propto a^4$ 这个结果称为泊肃叶定律（Poiseuille's law）：管道半径减半，流量便减至原来的十六分之一，这正是动脉收窄影响如此严重的原因。微分方程方面，可参阅 Boyce and DiPrima, *Elementary Differential Equations*（一本微分方程入门教科书）。

详细解答

STEP I 1998，第 9 题

力学 · 摆的碰撞

STEP I 1998，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/98-s1-q9

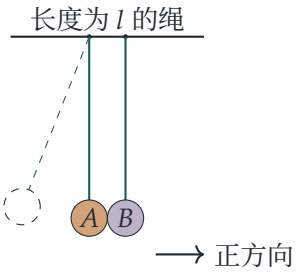
关键思路

由于位移非常小，每个球都像单摆（simple pendulum）一样作简谐运动（simple harmonic motion），而周期 $2\pi\sqrt{l/g}$ 与振幅和速率都无关。两个单摆长度相同，所以在最低点发生任何一次碰撞之后，两个球都会在同一时刻（即半个周期之后）回到最低点，而且速度方向相反。只要知道第二次碰撞在何处、何时发生，每次碰撞都只需写两行：动量守恒和牛顿实验定律（Newton's experimental law）。

解答

模型

我们把 A 和 B 视为质量为 m 的质点，各自系于长度同为 l 的不可伸长轻绳上，忽略空气阻力，并假设碰撞在瞬间完成，冲量（impulse）沿水平的连心线方向。取第一次碰撞时 A 的运动方向（由 A 指向 B）为正方向，并设 $\omega = \sqrt{g/l}$。



为何每次碰撞都在最低点发生

在小幅振动下，不论振幅多大，每个球都作角频率（angular frequency）为 ω 的简谐运动。因此，一个以速率 v 离开最低点的球，其水平位移为 $\frac{v}{\omega}\sin\omega t$，并且在时间 π/ω（半个周期）之后回到最低点，此时速度为 $-v$——对任何 v 都是如此。

第一次碰撞在最低点发生：B 静止地悬在该处，而 A 以速率 u 到达。假设某次碰撞在最低点发生，碰撞后 A 和 B 的速度分别为 $p \le q$（刚碰撞完，两球不可能互相靠近）。从最低点量起，在其后时间 t，两球之间的距离为

$$\frac{q}{\omega}\sin\omega t - \frac{p}{\omega}\sin\omega t = \frac{q-p}{\omega}\sin\omega t,$$

当 $0 < t < \pi/\omega$ 时此值为正，而在 $t = \pi/\omega$ 时为零。所以两球不可能更早相遇，而且恰好在两者都回到最低点时再次相遇。这正是题目的提示：两个单摆周期相同，所以它们总是一同回到最低点。由归纳法（induction）可知，每次碰撞都在最低点发生，而且相邻两次碰撞相隔半个周期。

第一次碰撞

设第一次碰撞刚结束时 A 和 B 的速度分别为 v_A 和 v_B。由动量守恒（conservation of momentum）（在最低点时绳的张力沿铅垂方向，所以不产生水平冲量）和牛顿实验定律，得

$$mv_A + mv_B = mu,$$
$$v_B - v_A = eu.$$

两式相加和相减，得

$$v_A = \tfrac{1}{2}u(1-e), \qquad v_B = \tfrac{1}{2}u(1+e).$$

两者都 ≥ 0，所以两个球都摆向 B 的一方（若 $e = 1$，A 会完全停下，一直悬在最低点）。

想一想 第二次碰撞前一刻的速度是多少？

半个周期之后，每个球都回到最低点，速度方向相反：A 的速度为 $-\frac{1}{2}u(1-e)$，B 的速度为 $-\frac{1}{2}u(1+e)$。B 较快，所以从后面追上 A。它相对于 A 的接近速率为 $\frac{1}{2}u(1+e) - \frac{1}{2}u(1-e) = eu$。

第二次碰撞

设第二次碰撞刚结束时 A 和 B 的速度分别为 w_A 和 w_B。由动量守恒，得

$$mw_A + mw_B = -\tfrac{1}{2}mu(1-e) - \tfrac{1}{2}mu(1+e) = -mu,$$

而由牛顿实验定律（分离速率等于接近速率 eu 的 e 倍），得

$$w_B - w_A = e \cdot eu = e^2 u.$$

两式相减和相加，得

$$w_A = -\tfrac{1}{2}u(1+e^2), \qquad w_B = -\tfrac{1}{2}u(1-e^2).$$

所以紧接第二次碰撞之后，A 的速率为 $\frac{1}{2}u(1+e^2)$，正如题目所求，而 B 的速率为 $\frac{1}{2}u(1-e^2)$，两者都朝着 A 最初被释放的那一方运动。

验算

- 完全弹性，$e=1$：最后 A 的速率为 u，而 B 静止。这是对的：第一次碰撞后 A 停下，B 取得速率 u；其后 B 以速率 u 返回，并把它全数交还给 A。
- $e=0$：两个速率都是 $\frac{1}{2}u$。两球实际上合为一体，第一次碰撞后以速率 $\frac{1}{2}u$ 一起运动，所以这也是对的。
- 第二次碰撞后的动量为 $-\frac{1}{2}mu(1+e^2) - \frac{1}{2}mu(1-e^2) = -mu$，与原来的 mu 方向相反。
- 我们模拟了两个真实（非线性）的单摆，取 $l = 1\,\mathrm{m}$，初始角度为 0.01 rad。第二次碰撞在第一次碰撞之后 1.0001 个半周期发生，位置与最低点相差不到 3×10^{-6} rad；取 $e=0.7$ 时，碰撞后的速率为 $0.745u$ 和 $0.255u$，与 $\frac{1}{2}(1+e^2)$ 和 $\frac{1}{2}(1-e^2)$ 吻合。

常见错误

- 忘记在第二次碰撞之前两个速度都已反向，把情况当作 B 撞向静止的 A。
- 在第二次碰撞时以 eu 代替 $e\cdot$(接近速率)。这里接近速率刚好是 eu，所以分离速率是 e^2u。
- 说碰撞在最低点发生，是“因为球本来就悬在那里”。真正的原因是周期与振幅无关，这也是为何球只可拉开很小的距离。

STEP 之外

等时性（isochronism，即周期与振幅无关）对单摆而言只是近似成立：摆幅较大时周期会增长，其准确值由椭圆积分（elliptic integral）给出，而碰撞位置亦会逐渐偏离最低点。Kleppner 与 Kolenkow 的 *An Introduction to Mechanics*（力学入门教科书）对单摆和碰撞都有严谨细致的论述。

详细解答

STEP I 1998，第 10 题

力学 • 抛体运动

> **STEP I 1998，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/98-s1-q10

关键思路

我们不知道抛射角，也不需要知道：答案只以 v、r 和 t 表示。因此要找一条能消去这个角的关系式。最简洁的做法是用向量（vector）来思考：质点（particle）经过时间 t 后的落点，等于初速度乘以 t，再加上重力造成的下坠 $\frac{1}{2}gt^2$。这三个向量构成一个直角三角形，余下的工夫交给毕氏定理（Pythagoras' theorem）就可以了。

解答

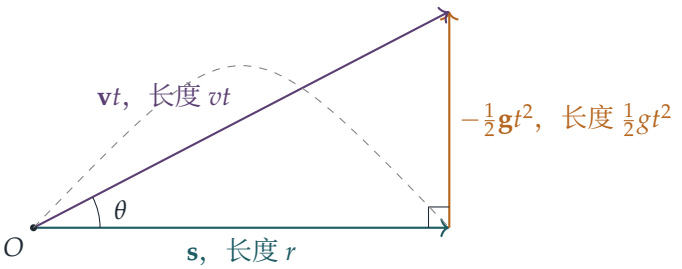
我们把每一粒泥土看成一个质点，只受重力作用而自由运动（不计空气阻力），由位于地面的爆炸中心 O 出发。垂直向上射出的质点会落在 O，所以可取 $r > 0$，并设抛射角（angle of projection）θ 为与水平方向所成的仰角，其中 $0 < \theta < \frac{\pi}{2}$。

著地时间

在恒定加速度 $\mathbf{g}$（垂直向下）之下，经过时间 t 后的位移（displacement）为

$$\mathbf{s} = \mathbf{v}t + \tfrac{1}{2}\mathbf{g}t^2,$$

其中 $\mathbf{v}$ 是初速度，其大小为 v。在著地的一刻，$\mathbf{s}$ 是水平的，长度为 r。移项后得 $\mathbf{v}t = \mathbf{s} - \frac{1}{2}\mathbf{g}t^2$，即一个长度为 r 的水平向量与一个长度为 $\frac{1}{2}gt^2$ 的垂直（向上）向量之和。



这两个向量互相垂直，所以根据毕氏定理，

$$v^2t^2 = r^2 + \left(\tfrac{1}{2}gt^2\right)^2 = r^2 + \tfrac{1}{4}g^2t^4.$$

两边乘以 g^2，并设 $u = \frac{1}{2}g^2t^2$（于是 $u^2 = \frac{1}{4}g^4t^4$ 及 $g^2v^2t^2 = 2v^2u$），上式变成

$$u^2 - 2v^2u + g^2r^2 = 0,$$

这是 u 的二次方程（quadratic）。它的根为

$$\tfrac{1}{2}g^2t^2 = u = v^2 \pm \sqrt{v^4 - g^2r^2},$$

这正是所要求的结果。

想一想 如果不用向量，我们怎样消去 θ?

用分量（components）。水平方向有 $r = vt\cos\theta$；垂直方向，落在地面即表示 $0 = vt\sin\theta - \frac{1}{2}gt^2$，所以 $vt\sin\theta = \frac{1}{2}gt^2$（因为 $t \neq 0$）。把两式平方后相加，再利用 $\cos^2\theta + \sin^2\theta = 1$，便再次得到 $v^2t^2 = r^2 + \frac{1}{4}g^2t^4$。向量图不过是把这个计算画成一个三角形而已。

只有当 $v^4 \geq g^2r^2$，即 $r \leq v^2/g$ 时，两个根才都是实数：这就是大家熟悉的最大射程（maximum range）。当 $r < v^2/g$ 时，两个根都是正数（两根之积为 $g^2r^2 > 0$，两根之和为 $2v^2 > 0$），所以有两个著地时间，一个属于高抛，一个属于平射。由于 $t = 2v\sin\theta/g$ 随 θ 递增，$+$ 号对应较陡的轨迹（trajectory），即 $\theta > \frac{\pi}{4}$，而 $-$ 号则对应较平的轨迹。

最大高度

整段飞行是对称的：垂直速度 $v\sin\theta - gt'$ 在 $t' = \frac{1}{2}t$ 时为零，即飞行的中途。由最高点起，质点（在垂直方向上）由静止下坠，历时 $\frac{1}{2}t$，所以最大高度为

$$H = \tfrac{1}{2}g\left(\tfrac{1}{2}t\right)^2 = \tfrac{1}{8}gt^2 = \frac{1}{4g}\cdot\tfrac{1}{2}g^2t^2.$$

利用第一部分的结果，

$$H = \frac{v^2 \pm \sqrt{v^4 - g^2r^2}}{4g},$$

其中 $+$ 号对应较陡的轨迹，$-$ 号对应较平的轨迹（当 $r < v^2/g$ 时两者皆有可能；当 $r = v^2/g$ 时两者重合，此时 $H = v^2/(4g)$）。

验算

- 垂直发射：$r = 0$ 给出 $\frac{1}{2}g^2t^2 = 2v^2$，所以 $t = 2v/g$，而 $H = v^2/(2g)$，与预期相符。另一个根 $t = 0$、$H = 0$，则是贴着地面掠过的一击。
- 两个高度之和为 $v^2/(2g)$。这是理所当然的：两条轨迹的抛射角互为余角，即 θ 和 $\frac{\pi}{2} - \theta$，而 $\frac{v^2\sin^2\theta}{2g} + \frac{v^2\cos^2\theta}{2g} = \frac{v^2}{2g}$。
- 数值上，取 $g = 9.8$、$v = 20$ 和 $\theta = 1$ 弧度，逐步模拟运动可得 $t \approx 3.4346$ 和 $H \approx 14.450$；取 $+$ 号的公式给出 3.4346 和 14.450。余角的情况则与 $-$ 号吻合（$t \approx 2.2053$，$H \approx 5.958$）。

常见错误

- 只给出一个高度。速率相同的质点可以沿两条不同的路径到达同一点，而 $\pm$ 号会一直带到 H 的式子里。
- 把 H 写成 θ 的式子便停下来。题目只要求以 v、r 和 g 表示，所以藉 $H = \frac{1}{8}gt^2$ 这条关系式连系到第一部分，才是最省力的途径。

STEP 之外

这个关于 t^2 的二次方程，是“轨迹的包络线（envelope）”在时域上的版本：满足 $v^4 = g^2r^2$ 的点（以及离开地面时的抛物面（paraboloid）$z = \frac{v^2}{2g} - \frac{gr^2}{2v^2}$）标示出泥土所能到达的区域的边界，而在这个区域内，每一点都恰好有两条轨迹经过。曲线族的包络线，是大学一年级课程中偏微分（partial differentiation）的一个漂亮应用。Kleppner 与 Kolenkow 所著的 *An Introduction to Mechanics*（力学入门教科书）以恰到好处的程度讲解抛体运动（projectile motion）和向量运动学。

详细解答

STEP I 1998，第 11 题

力学 · 跨过滑轮的连接质点

STEP I 1998，第 11 题

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

step.maths.org/questions/98-s1-q11

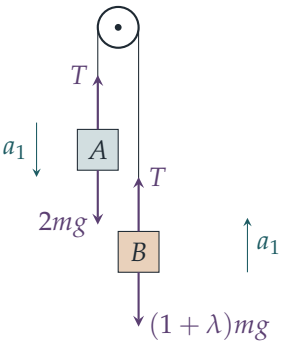
关键思路

每次往返分为两个阶段，而在每个阶段中各质量保持不变，所以加速度是常数，可以使用 $s = \frac{1}{2}at^2$。下行时，盛满水的桶比矿石连桶更重，两者合起来就是一部阿特伍德机（Atwood machine）。回程上行时，桶 A 是空的而且没有质量，所以没有任何东西拉住 B：它只是自由下坠（falls freely）。至于第二部分，“在很长的时间内运上最多矿石”的意思是“每个循环运上的矿石量除以每个循环所需的时间要最大”。

解答

一个循环所需的时间

模型：两个桶连同所盛之物视为质点（particle），链是轻而不可伸长的（light and inextensible），滑轮细小、光滑而且轻。因此链两端的张力（tension）T 相同，而两个桶以相同的速率运动。当 A 在顶部时，B 在底部（链的长度为 l），所以每个阶段中每个桶都移动距离 l。



第 1 阶段：A 下降。桶 A 盛有质量 $2m$ 的水；桶 B 的总质量为 $m+\lambda m=(1+\lambda)m$。设 a_1 为两者共同的加速度（A 向下，B 向上）。对每个桶应用牛顿第二定律（Newton's second law），得

$$\begin{aligned} A\text{（向下）：}&\quad 2mg-T=2ma_1,\\ B\text{（向上）：}&\quad T-(1+\lambda)mg=(1+\lambda)ma_1. \end{aligned}$$

两式相加消去 T：

$$(1-\lambda)mg=(3+\lambda)ma_1 \quad\Longrightarrow\quad a_1=\frac{(1-\lambda)g}{3+\lambda},$$

由于 $\lambda<1$，这个值是正数，所以 A 确实下降。由静止出发，以恒加速度移动距离 l，有 $l=\frac{1}{2}a_1t_1^2$，所以

$$t_1=\sqrt{\frac{2l}{a_1}}=\sqrt{\frac{2l}{g}}\sqrt{\frac{3+\lambda}{1-\lambda}}.$$

想一想 在第 2 阶段，桶 A 完全没有质量。链的张力是多少?

第 2 阶段：B 下降。现在 A 是空的，所以它的质量为零，而 B 的质量为 m。对没有质量的桶 A，牛顿定律给出 $T-0\cdot g=0\cdot a$，所以 $T=0$：一条轻链若另一端系着的东西没有质量，就无法拉动任何物体。因此 B 由静止以加速度 g 自由下坠，经过距离 l，由 $l=\frac{1}{2}gt_2^2$ 得

$$t_2=\sqrt{\frac{2l}{g}}.$$

（用阿特伍德机的公式，取质量为 0 和 m，也得到同样的答案：$a=\frac{m-0}{m+0}g=g$。）这个阶段结束时，A 第一次回到顶部。

总时间。由在顶部释放 A 起，到它第一次回到顶部为止，所需的时间为

$$\boxed{t_1+t_2=\sqrt{\frac{2l}{g}}\left(1+\sqrt{\frac{3+\lambda}{1-\lambda}}\right)=\sqrt{\frac{2l}{g}}\,\frac{(1-\lambda)^{1/2}+(3+\lambda)^{1/2}}{(1-\lambda)^{1/2}}}$$

使运上的矿石量最大

以 $\tau(\lambda)=t_1+t_2$ 表示循环时间。每个循环运上质量为 λm 的矿石，而整个过程不断以完全相同的方式重复。在一段很长的时间 S 内，共有 $\lfloor S/\tau\rfloor$ 个完整循环，所以运上的矿石量介乎 $\lambda m\,(S/\tau-1)$ 与 $\lambda m\,S/\tau$ 之间。当 S 远大于 τ 时，那个未完成的循环所损失的部分可以忽略，运上的矿石量基本上就是

$$\frac{\lambda mS}{\tau(\lambda)}=mS\sqrt{\frac{g}{2l}}\,\frac{\lambda(1-\lambda)^{1/2}}{(1-\lambda)^{1/2}+(3+\lambda)^{1/2}}=mS\sqrt{\frac{g}{2l}}\,\mathrm{f}(\lambda).$$

因子 $mS\sqrt{g/2l}$ 与 λ 无关，所以我们必须选取 $\lambda\in(0,1)$，使 $\mathrm{f}(\lambda)$ 尽可能大。

想一想 为甚么最佳的 λ 必定是驻点（stationary point），而不是端点?

函数 f 在 $(0,1)$ 上可导（differentiable），因为在该区间内 $1-\lambda>0$ 和 $3+\lambda>0$，而且分母是正数。它在 $(0,1)$ 上取正值，并且在两端都趋向 0：当 $\lambda\to 0$ 时，因子 λ 趋向零（没有运载矿石）；当 $\lambda\to 1$ 时，因子 $(1-\lambda)^{1/2}$ 趋向零（两个桶几乎平衡，所以第 1 阶段需时极长）。因此 f 的最大值在区间内部的某一点取得，而可导函数若在内点取得极大值，该处的导数必为零。所以最佳的 λ 满足 $f'(\lambda)=0$，这正是所要证明的。

验算

- 极限情况：当 $\lambda\to 1$ 时，第 1 阶段中两个桶的质量分别为 $2m$ 和 $2m$，所以 $a_1\to 0$ 而 $t_1\to\infty$，与公式吻合。
- 当 $\lambda\to 0$ 时，得 $a_1=g/3$，正是质量为 $2m$ 和 m 的阿特伍德机的惯常数值。
- 数值上，取 $g=9.8$，对随机选取的 λ 和 l 逐步模拟运动，结果与公式一致：例如 $\lambda=0.504$、$l=26.9$ 给出的循环时间为 8.568 s。以数值方法求 f 的最大值，得 $\lambda\approx 0.704$；而在一段很长的时间内计数完整循环的数目，也得到同一个最佳的 λ。

常见错误

- 忘记了桶 B 本身的质量 m，在第 1 阶段用了 λm 而不是 $(1+\lambda)m$。
- 把第 2 阶段当作带有某个张力的阿特伍德机来处理：由于 A 没有质量，B 是自由下坠的。
- 求每个循环运上的矿石量 λm 的最大值，而不是每单位时间运上的矿石量的最大值。载得愈重，循环愈慢。

STEP 之外

把一个速率（每个循环的产出除以每个循环的时间）最优化，正是概率论中*更新报酬定理*（renewal–reward theorem）背后的想法：在那里循环的长度是随机的，而长远的报酬率仍然等于每个循环的平均报酬除以平均循环长度。这里的力学是恒加速度运动，Kleppner 与 Kolenkow 所著的 *An Introduction to Mechanics*（力学入门）由此出发，发展成对质点系统的完整论述。

详细解答

STEP I 1998，第 12 题

概率与统计 · 计数与独立性

STEP I 1998，第 12 题

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

step.maths.org/questions/98-s1-q12

关键思路

这一题从头到尾都是计数（counting）。只要知道解的总数，以及其中有多少个解的 X 等于某个指定的值，每个概率（probability）就是两个数目之比。至于独立性（independence），最快捷的检验方法是考虑一个极端情况：找出一对数值，使联合概率（joint probability）根本不可能等于边际概率（marginal probability）之积。最后一部分中，“XYZ 可被 5 整除”的意思是“X, Y, Z 之中最少有一个可被 5 整除”，这就要用上容斥原理（inclusion–exclusion）。

解答

数算解的数目

若固定 $X = x$（其中 $0 \le x \le 20$），则 Y 可取 $0, 1, \ldots, 20 - x$ 中的任何一个值，而 $Z = 20 - x - Y$ 随之确定，并且是非负数。因此恰好有 $21 - x$ 个解满足 $X = x$，而解的总数为

$$N = \sum_{x=0}^{20} (21 - x) = 21 + 20 + \cdots + 1 = \frac{21 \cdot 22}{2} = 231.$$

（这正是“隔板法”（stars and bars）的数目 $\binom{22}{2}$：把 20 颗星和 2 块隔板排成一行。）每个解的概率均为 $1/231$，而

$$\mathrm{P}(X = x) = \frac{21 - x}{231}, \qquad x = 0, 1, \ldots, 20.$$

根据对称性（symmetry），Y（以及 Z）有相同的分布（distribution）。

X 与 Y 是否独立?

不是。取 $X = 20$ 和 $Y = 20$。这两个事件各自都有可能发生:

$$\mathrm{P}(X = 20) = \mathrm{P}(Y = 20) = \frac{1}{231} > 0,$$

但两者不可能同时发生,因为 $X + Y \leq 20$。所以

$$\mathrm{P}(X = 20,\ Y = 20) = 0 \neq \frac{1}{231^2} = \mathrm{P}(X = 20)\,\mathrm{P}(Y = 20),$$

故 X 与 Y 并不独立。直观地说,X 的值越大,留给 Y 的空间就越少。

> ? **想一想** 只说“$X + Y \leq 20$,所以它们显然是相关的”不就够了吗?

想到这样的限制条件,直觉是对的;但独立性是一个关于概率的精确陈述,所以论证时应该举出一对数值,使得 $\mathrm{P}(X = x, Y = y) \neq \mathrm{P}(X = x)\mathrm{P}(Y = y)$。只需一个反例(counterexample)便已足够,而 $(20, 20)$ 这对极端数值令计算变得轻而易举。

X 可被 5 整除

范围内 5 的倍数为 $X = 0, 5, 10, 15, 20$,对应的解分别有 $21, 16, 11, 6, 1$ 个。所以

$$\mathrm{P}(5 \mid X) = \frac{21 + 16 + 11 + 6 + 1}{231} = \frac{55}{231} = \frac{5}{21},$$

正如题目所求。

XYZ 可被 5 整除

由于 5 是质数(prime),$5 \mid XYZ$ 当且仅当 5 整除 X、Y、Z 之中最少一个。(谨记 0 也算是可被 5 整除。)设 A、B、C 分别为 X、Y、Z 可被 5 整除的解所组成的集合。由上一部分可知 $|A| = |B| = |C| = 55$。

考虑 $A \cap B$:设 $X = 5a$、$Y = 5b$,其中 $a, b \geq 0$ 且 $a + b \leq 4$。这样的数对共有 $5 + 4 + 3 + 2 + 1 = 15$ 个。在每一个情况下,$Z = 20 - 5a - 5b$ 同样是 5 的倍数,所以

$$|A \cap B| = |B \cap C| = |C \cap A| = |A \cap B \cap C| = 15.$$

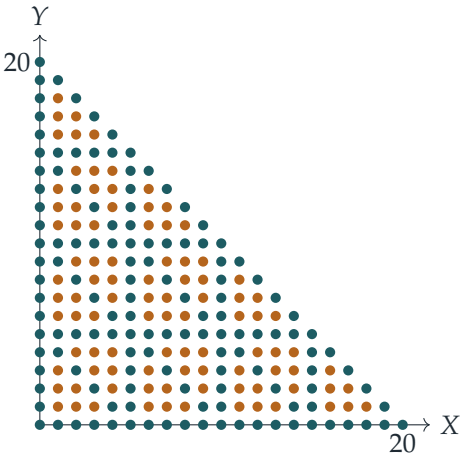
根据容斥原理,

$$\begin{aligned} |A \cup B \cup C| &= 3 \cdot 55 - 3 \cdot 15 + 15 \\ &= 135, \end{aligned}$$

因此

$$\mathrm{P}(5 \mid XYZ) = \frac{135}{231} = \frac{45}{77}.$$

下图把各个解表示为满足 $X+Y \le 20$ 的格点（lattice point）(X,Y)。蓝绿色的点位于直线 $X=5k$、$Y=5k$ 或 $X+Y=20-5k$（即 $Z=5k$）其中之一上，而琥珀色的点则是 X,Y,Z 全都不是 5 的倍数的 96 个解。



验算

- 补集（complement）应有 $231-135=96$ 个解，其中 X,Y,Z 全都不是 5 的倍数：这就是图中琥珀色的点。用电脑列出全部 231 个解，并以精确分数计算，可确认 96、$|A\cap B|=15$、5/21 以及 $135/231=45/77\approx 0.584$ 均正确。
- 对 5/21 作一个合理性检查：X 可被 5 整除的概率比 1/5 稍大，因为 X 取较小的值的可能性最高，而 $X=0$ 更是最有可能出现的值。

常见错误

- 忘记 0 可被 5 整除。漏掉 $X=0$ 会得出 34/231，而不是 5/21。
- 直接计算 $3\times\frac{5}{21}$，这样会把重叠的部分重复计算多次。本题的重叠部分相当大：只要其中两个变量是 5 的倍数，第三个也必然是。
- 论证两者相关时，没有举出一对令乘法法则不成立的具体数值。

STEP 之外

从所有解中均匀随机地选取一个，等同于随机选取 20 分成三部分的一个有序分拆（composition）。若把 20 换成总和 n，并让 n 增大，$(X/n, Y/n)$ 便会趋于在一个三角形上均匀分布，这是狄利克雷分布（Dirichlet distribution）最简单的情况。容斥原理是处理“最少

一个”这类事件的标准工具，它还会在错排（derangement）和欧拉函数（Euler's totient function）中再次出现。Grimmett 与 Welsh 合著的 *Probability: An Introduction*（概率论入门教材）对计数和独立性这两方面的概念均有论述。

详细解答

STEP I 1998，第 13 题

概率与统计 · 放回抽样

> **STEP I 1998，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/98-s1-q13

关键思路

这是一道条件概率（conditional probability）的题目：作为条件的事件是“在 n 次抽取中刚好吃掉一粒软糖”，而我们按这粒软糖是在何时吃掉的，把这个事件拆开。每一部分都是几个简单概率的乘积，因为袋中的内容只会在吃掉红色软糖的那一刻改变。把各部分加起来，就是一个等比级数（geometric series），两个 n 次幂之差正是由此而来。

解答

设 $N = r + b$ 为开始时软糖的总数。我们假设 $r \geq 1$ 及 $b \geq 1$（若 $b = 0$ 而 $n \geq 2$，则不可能刚好只吃掉一粒软糖，也就没有可作条件的事件）。

每次抽取时的袋子。 非红色的软糖会放回袋中，所以袋子只会在吃掉红色软糖时改变。吃掉红色软糖之前，袋中有 N 粒软糖，其中 b 粒非红色；之后袋中有 $N-1$ 粒软糖，其中仍有 b 粒非红色。由于每次抽取前都把袋子彻底摇匀，每次抽取都是从当时袋中的软糖里重新作一次均匀的随机选取，所以我们可以把各次抽取的概率逐一相乘。

按吃掉的时间拆分。 对 $k = 1, 2, \ldots, n$，设 E_k 为以下事件：在 n 次抽取中唯一吃掉的那粒软糖，是在第 k 次抽取时吃掉的。那么第 $1, \ldots, k-1$ 次是从原来的袋中抽到非红色，第 k 次是从原来的袋中抽到红色，而第 $k+1, \ldots, n$ 次是从少了一粒的袋中抽到非红色：

$$\mathrm{P}(E_k) = \left(\frac{b}{N}\right)^{k-1} \frac{r}{N} \left(\frac{b}{N-1}\right)^{n-k} = \frac{r\, b^{n-1}}{N^k\, (N-1)^{n-k}}.$$

事件 $E_1, \ldots, E_n$ 互斥（disjoint），而它们的并集（union）就是“刚好吃掉一粒软糖”这个事件 E。所以

$$\mathrm{P}(E) = r\,b^{n-1} \sum_{k=1}^{n} \frac{1}{N^k (N-1)^{n-k}} = \frac{r\,b^{n-1}}{N^n (N-1)^n} \sum_{k=1}^{n} N^{n-k} (N-1)^k,$$

这里我们把每一项的分子和分母同乘以 $N^n (N-1)^n$。

> ? **想一想** 怎样求 $\sum_{k=1}^{n} N^{n-k}(N-1)^k$ 的和?

这是一个等比级数，首项为 $N^{n-1}(N-1)$，公比（common ratio）为 $\frac{N-1}{N}$。更快捷的做法，是利用大家熟悉的因式分解（factorisation）

$$x^n - y^n = (x-y)\left(x^{n-1} + x^{n-2}y + \cdots + y^{n-1}\right)$$

取 $x = N$、$y = N-1$，于是 $x - y = 1$：

$$\sum_{k=1}^{n} N^{n-k}(N-1)^k = (N-1) \sum_{j=0}^{n-1} N^{n-1-j}(N-1)^j = (N-1)\left(N^n - (N-1)^n\right).$$

因此

$$\mathrm{P}(E) = \frac{r\,b^{n-1}\left(N^n - (N-1)^n\right)}{N^n (N-1)^{n-1}}.$$

条件概率。“我在最后一次抽取时吃掉它，而且只吃了一粒”这个事件就是 E_n，其概率为 $\mathrm{P}(E_n) = \dfrac{r\,b^{n-1}}{N^n}$。由于 $E_n \subseteq E$，

$$\begin{aligned}\mathrm{P}(E_n \mid E) = \frac{\mathrm{P}(E_n)}{\mathrm{P}(E)} &= \frac{r\,b^{n-1}}{N^n} \cdot \frac{N^n (N-1)^{n-1}}{r\,b^{n-1}\left(N^n - (N-1)^n\right)} \\ &= \frac{(N-1)^{n-1}}{N^n - (N-1)^n} = \frac{(r+b-1)^{n-1}}{(r+b)^n - (r+b-1)^n},\end{aligned}$$

这正是所求的结果。

留意 r 和 b 完全消去，只剩下它们的总和。其实上面的计算还说明了更多：已知 E 发生，软糖在第 k 次抽取时被吃掉的概率与 $\left(\frac{N-1}{N}\right)^k$ 成正比，所以较早的抽取比较后的抽取更有可能。这是合理的：早些吃掉红色软糖，袋中非红色软糖所占的比例便较大，余下各次“再没有抽到红色”的机会也就较高。

验算

► $n = 1$：公式给出 $\dfrac{1}{N - (N-1)} = 1$，正确，因为唯一的一次抽取就是最后一次。

- $n = 2$：公式给出 $\dfrac{N-1}{2N-1}$，略小于 $\frac{1}{2}$，与“较早的抽取可能性稍高”一致。
- 取 $r = 3$、$b = 5$、$n = 4$，公式给出 $343/1695 \approx 0.2024$，而若四次抽取的可能性均等，则应为 $\frac{1}{4}$。模拟 4×10^5 次的结果为 0.2028；另外，逐一列举所有抽取序列并追踪袋中内容，所得的精确结果在 $1 \le r, b \le 4$ 及 $1 \le n \le 8$ 的所有情况下都与公式完全吻合。

常见错误

- 在红色软糖被吃掉*之后*，仍然用 b/N 作为抽到非红色的概率。这时袋中只有 $N-1$ 粒软糖，而这正是本题的关键所在：如果由始至终都用 b/N，每个 E_k 的概率便会相同，答案就变成 $1/n$。
- 忘记取条件：单是 $\mathrm{P}(E_n) = rb^{n-1}/N^n$ 并不是答案，因为题目已告诉我们刚好吃掉了一粒软糖。

STEP 之外

这里的袋子是一个简单的马可夫链（Markov chain）：它的状态是余下红色软糖的数目，而状态向下移动的机会只取决于当前的状态。以这类链的最终状态为条件，反过来问它所经过的路径，正是“隐马可夫模型”（hidden Markov models）以及对历史过程作贝叶斯推断（Bayesian inference）背后的想法。Grimmett 与 Welsh 的 *Probability: An Introduction*（概率入门）正好以这个程度讲解条件概率和马可夫链，而 Blitzstein 与 Hwang 的 *Introduction to Probability*（概率导论）则有许多同类风格的题目。

详细解答

STEP I 1998，第 14 题

概率与统计 · 几何概率

STEP I 1998，第 14 题

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

step.maths.org/questions/98-s1-q14

关键思路

硬币落在哪里，完全取决于它的圆心落在哪里。由于棋盘的图案规则地重复，我们可以假设圆心在单独一个 2×2 正方形内均匀分布（uniformly distributed），于是每一种结果都对应该正方形内的一个区域。把“压过一条线”和“盖住一个交点”转化为圆心与各条线及各个角的距离所满足的条件，概率（probability）便变成面积，而这个面积是 a 的二次式，我们可以求其最大值。

解答

税款加倍的概率

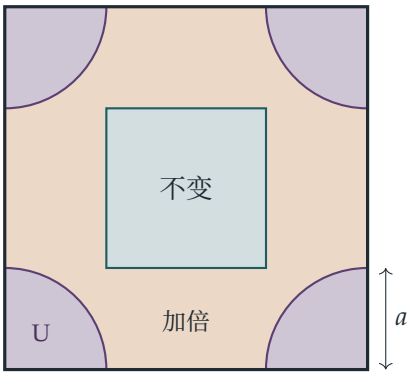
设硬币的圆心均匀随机地落在方格网的其中一个正方形内；棋盘很大，所以边缘效应可以忽略，而且每个正方形看起来都一样。在这个正方形上设立坐标，使它成为 $[0,2] \times [0,2]$，方格线沿着它的四边。

设 U 为圆心到最近的垂直线的距离，V 为圆心到最近的水平线的距离。由于 x 坐标在 $[0,2]$ 上均匀分布，$U = \min(x, 2-x)$ 在 $[0,1]$ 上均匀分布；同理 V 在 $[0,1]$ 上均匀分布，而且因为 x 与 y 独立（independent），U 与 V 亦独立。所以 (U,V) 在单位正方形 $[0,1]^2$ 上均匀分布，而概率就是该正方形内的面积。

- 硬币不压过任何线，当且仅当 $U > a$ 且 $V > a$。其概率为 $(1-a)^2$。
- 硬币盖住一个交点，当且仅当正方形最近的一个角与圆心的距离小于 a，即 $U^2 + V^2 < a^2$。由于 $a < 1$，这个四分之一圆盘完全位于 $[0,1]^2$ 之内，所以概率为 $\frac{1}{4}\pi a^2$。

这两个事件互斥（disjoint）：如果硬币盖住一个交点，它必定压过通过该交点的线。（硬

币刚好碰到一条线之类的边界情况，概率为零，所以用 $<$ 还是 $\leq$ 都没有分别。）



2

图中显示圆心在一个正方形内的可能位置：蓝绿色（没有压过任何线）和紫红色（盖住一个交点，在其中一个角标示为 U）的区域令税款不变；琥珀色的区域令税款加倍。

所以税款加倍的概率为

$$\begin{aligned} p(a) &= 1-(1-a)^2-\tfrac{1}{4}\pi a^2 \\ &= 2a-\left(1+\frac{\pi}{4}\right)a^2. \end{aligned}$$

使税收最大

要征得最多的税，就要令加倍的机会尽可能大，所以我们在 $0<a<1$ 的范围内求 $p(a)$ 的最大值。它是一个首项系数为负的二次函数，而

$$p'(a)=2-2\left(1+\frac{\pi}{4}\right)a=0 \iff a=\left(1+\frac{\pi}{4}\right)^{-1}.$$

这个值是 $\frac{4}{4+\pi}\approx 0.560$，确实位于 $(0,1)$ 之内，而且由于 $p''(a)<0$，它给出最大值，正如题目所求。以 a^* 表示这个值，则有 $(1+\frac{\pi}{4})a^{*2}=a^*$，所以最大值为

$$p_{\max}=2a^*-a^*=a^*=\frac{4}{4+\pi}.$$

n 年后税款的期望值

想一想 为甚么每年的期望增长倍数是 $1+p$ 而不是 $2p$?

在税款没有加倍的年份，税款维持不变，所以它是乘以 1，而不是乘以 0。设 M_k 为第 k 年的倍数：$M_k=2$ 的概率为 p，$M_k=1$ 的概率为 $1-p$。于是

$$\mathrm{E}(M_k)=2p+(1-p)=1+p=1+\frac{4}{4+\pi}=\frac{8+\pi}{4+\pi}.$$

假设每年的投掷互相独立。n 年后的税款为每头 yak $T_n = M_1 M_2 \cdots M_n$ Slihings（由 1 开始），而独立随机变量（random variables）之积的期望值（expectation）等于它们各自期望值之积，所以

$$\mathrm{E}(T_n) = \mathrm{E}(M_1)\,\mathrm{E}(M_2)\cdots\mathrm{E}(M_n) = \left(\frac{8+\pi}{4+\pi}\right)^n,$$

正如题目所求。

验算

- 当 $a \to 0$ 时，硬币几乎从不碰到任何线；当 $a = 1$ 时，它必定压过一条线，所以 $p(1) = 1 - \frac{\pi}{4} \approx 0.215$，即避开角上圆盘的机会。两者都与公式吻合，而最大值位于两者之间。
- 以 400 000 个随机圆心进行模拟，在 $a = 0.56$ 时得出 $p \approx 0.5620$，而公式给出 0.5601。对 a 作精细搜索，得出最大值位于 $a \approx 0.5601 = 4/(4+\pi)$。
- 以 $p = 4/(4+\pi)$ 模拟每年的加倍，得出 3 年后税款的平均值约为 3.805，而 $\left(\frac{8+\pi}{4+\pi}\right)^3 \approx 3.797$。

常见错误

- 从“不压线”的面积中减去“盖住交点”的面积，或者把它计算了两次。这两个事件互斥，两者都要从 1 中减去。
- 在每个角使用整个圆盘的面积 πa^2，忘记了只有四分之一位于正方形之内（等价地说，每个角上的圆盘由四个正方形共用）。
- 把税款的期望值写成 $(2p)^n$，忘记了没有加倍的年份税款维持不变。

STEP 之外

这是蒲丰投针问题（Buffon's needle problem）的近亲：把随机放置的物件化为基本区域（fundamental region）内的一个随机点，概率便成为面积。这个想法是*几何概率*（geometric probability）和积分几何（integral geometry）的起点，由此可得出一些出人意表的结果，例如以投针来估算 π。最后一部分用到独立随机变量的期望值可以相乘这个事实，这正是金融学和种群生物学中乘法增长模型的基础。Grimmett 与 Welsh 合著的 *Probability: An Introduction*（概率入门）涵盖了这两个想法。

详细解答

STEP II 1998，第 1 题

纯数学 · 丢番图方程与奇偶性

STEP II 1998，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/98-s2-q1

关键思路

把立方展开后，大部分的项会互相抵消，剩下一个关于 n 的三次式，其常数项很小。比较等式两边的奇偶性（parity），可知 n 必为偶数。再把这个三次方程写成 $n^2(\text{某个式子}) = 54$ 的形式，问题便化为一个关于因数（divisor）的问题。至于 $(**)$，利用偶数代换 $n = 2m$，便直接回到 $(*)$。

解答

方程 $(*)$

利用 $(a+b)^3 - (a-b)^3 = 6a^2b + 2b^3$，并取 $a = n$，$b = 3$，可得

$$(n+3)^3 - (n-3)^3 = 18n^2 + 54,$$

所以 $(*)$ 等价于

$$n^3 = 18n^2 + 54, \qquad \text{即} \qquad n^2(n-18) = 54. \tag{1}$$

n 是偶数。对于任何整数 n，右边 $18n^2 + 54 = 2(9n^2 + 27)$ 都是偶数。假如 n 是奇数，则 n^3 也是奇数，与 $n^3 = 18n^2 + 54$ 矛盾。因此 n 是偶数。

n^2 整除 54。在 (1) 的第二种写法中，$n - 18$ 是整数，所以 54 等于 n^2 乘以一个整数。因此 n^2 是 54 的因数。

? **想一想** 这两项事实合起来，怎样把所有整数都排除掉?

由于 n 是偶数，可设 $n = 2k$。于是 $n^2 = 4k^2$，它是 4 的倍数（而且 $n \neq 0$，因为把 $n = 0$ 代入 (1) 会得出 $0 = 54$）。但 $54 = 2 \times 27$ 不能被 4 整除，所以除了 0 之外，没有任何 4

的倍数能整除它。这个矛盾说明没有整数 n 满足 $(*)$。

（另一个看法：能整除 $54 = 2 \cdot 3^3$ 的平方数只有 1 和 9，所以 $n \in \{\pm 1, \pm 3\}$，而这些全是奇数。）

方程 $(**)$

现在取 $b = 6$：

$$(n+6)^3 - (n-6)^3 = 36n^2 + 432,$$

所以 $(**)$ 等价于

$$n^3 = 36n^2 + 432.$$

对于任何整数 n，右边都是偶数，所以与之前完全一样，n^3 是偶数。由于奇数的立方是奇数，n 必为偶数。

> ? **想一想** 为甚么不能直接照搬第一部分的因数论证？

我们固然可以写成 $n^2(n-36) = 432$，但 $432 = 2^4 \cdot 3^3$ 有不少偶数的平方因数（4、16、36、144），所以单靠整除性的论证并不能完成证明。我们改为利用偶数这个性质。设 $n = 2m$，其中 m 为整数。于是

$$8m^3 = 36 \cdot 4m^2 + 432 \quad \Longleftrightarrow \quad m^3 = 18m^2 + 54,$$

这正正是整数 m 所满足的方程 (1)。换句话说，m 满足 $(*)$。事实上，把 $n = 2m$ 代入 $(**)$，所得的不过是把 $n = m$ 代入 $(*)$ 后再全式乘以 8：

$$(2m-6)^3 + (2m)^3 - (2m+6)^3 = 8\big[(m-3)^3 + m^3 - (m+3)^3\big].$$

我们已证明没有整数满足 $(*)$，所以这样的 m 并不存在。因此没有整数 n 满足 $(**)$。

验算

- 把 $n = 1$ 代入展开式试试：$(4)^3 - (-2)^3 = 64 + 8 = 72 = 18 + 54$。✓
- 三次式 $n^3 - 18n^2 - 54$ 恰好有一个实根，约为 $n \approx 18.164$，并不是整数。（用电脑直接搜索 $|n| < 10^5$ 的范围，两个方程都找不到解。）
- 对于 $(**)$，实数解为 $n \approx 2 \times 18.164 = 36.33$，与 $n = 2m$ 这个缩放关系吻合。

常见错误

- 忘记 $n = 0$ 需要另外交代一句：0 是偶数，但 0 不是 54 的因数。在这里，(1) 已即时把它排除。

- 把立方完全展开，结果在运算中出错。恒等式 $(a+b)^3-(a-b)^3=6a^2b+2b^3$ 可以省掉一半工夫。
- 在 $(**)$ 中，尝试证明 $n^2 \mid 432$ 是行不通的。“n 是偶数”的真正用处在于代换 $n=2m$。

STEP 之外

最后一步由 $(**)$ 的一个解，得出一条形式相同但数字较小的方程的解，这是费马（Fermat）*递降法*（method of descent）的一个简单形式。反复运用这个方法，可以证明 $x^3+y^3=z^3$ 没有正整数解，这就是费马最后定理（Fermat's Last Theorem）的三次情形，最先由欧拉（Euler）证明。本题的三项构成等差数列（arithmetic progression），令问题容易得多。事实上，$(n-3k)^3+n^3=(n+3k)^3$ 表示 $t=n/k$ 是 t^3-18t^2-54 的根，而根据有理根定理（rational root theorem），这个三次式没有有理根，所以对任何 $k\neq 0$ 都没有整数解。Davenport 的 *The Higher Arithmetic* 是一本介绍整除性、奇偶性论证和丢番图方程（Diophantine equations）的出色入门书。

详细解答

STEP II 1998，第 2 题

纯数学 · 二项式级数近似

STEP II 1998，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/98-s2-q2

关键思路

二项式级数（binomial series）在变量很小时收敛（converge）得很快，所以我们希望把 $\sqrt{2}$（或 $\sqrt[3]{2}$）写成一个简单的有理数乘以 $(1+x)^a$，而其中的 x 极小。就平方根而言，$49/50$ 与 1 只差一点，而且含有因数 2；就立方根而言，我们要找一个几乎等于 $2\times$ 另一个立方数的立方数，而 $128 = 2\times 4^3$ 正好紧贴着 $125 = 5^3$。

解答

2 的平方根

首先，我们把题目给出的式子与 $\sqrt{2}$ 联系起来。由于 $1-\frac{1}{50} = \frac{49}{50}$，

$$\left(1-\frac{1}{50}\right)^{1/2} = \frac{7}{\sqrt{50}} = \frac{7}{5\sqrt{2}} = \frac{7\sqrt{2}}{10}, \qquad \text{所以} \qquad \sqrt{2} = \frac{10}{7}\left(1-\frac{1}{50}\right)^{1/2}.$$

$(1+x)^{1/2}$ 的二项式级数的首几项为

$$(1+x)^{1/2} = 1 + \tfrac{1}{2}x - \tfrac{1}{8}x^2 + \tfrac{1}{16}x^3 - \cdots,$$

对 $|x|<1$ 成立。设 $x = -\frac{2}{100} = -2y$，其中 $y = \frac{1}{100}$。2 的幂刚好可以约去：

$$\begin{aligned}\left(1-\frac{2}{100}\right)^{1/2} &\approx 1 - \tfrac{1}{2}(2y) - \tfrac{1}{8}(2y)^2 - \tfrac{1}{16}(2y)^3 \\ &= 1 - y - \tfrac{1}{2}y^2 - \tfrac{1}{2}y^3 \\ &= 1 - 0.01 - 0.00005 - 0.0000005 \\ &= 0.9899495.\end{aligned}$$

这就是把 $1/50$ 写成 $2/100$ 的好处：每一项都变成 10 的幂乘以 1 或 $\frac{1}{2}$，所以运算全是精确的小数。

因此

$$\sqrt{2} \approx \frac{10 \times 0.9899495}{7} = \frac{9.899495}{7} = 1.41421357\ldots,$$

准确至小数点后六位，$\sqrt{2} \approx 1.414214$，正如题目所求。

2 的立方根

> ? **想一想** 应取哪个 N 和 a？怎样才能令 $(1+N/125)^a$ 与 $\sqrt[3]{2}$ 扯上关系？

由于 $125 = 5^3$，$(125+N)/125$ 的立方根就是 $\sqrt[3]{125+N}/5$。因此我们希望 $125+N$ 是某个完全立方数（perfect cube）的 2 倍，而且 N 要小。从 2×1^3, 2×2^3, 2×3^3, $2\times 4^3 = 2, 16, 54, 128$ 可见，128 与 125 只相差 3。所以取 $N=3$ 和 $a=\frac{1}{3}$：

$$\left(1+\frac{3}{125}\right)^{1/3} = \frac{\sqrt[3]{128}}{5} = \frac{4\sqrt[3]{2}}{5}, \qquad \text{所以} \qquad \sqrt[3]{2} = \frac{5}{4}\left(1+\frac{3}{125}\right)^{1/3}.$$

$(1+x)^{1/3}$ 的二项式级数的首几项为

$$(1+x)^{1/3} = 1 + \tfrac{1}{3}x + \frac{\frac{1}{3}\cdot\left(-\frac{2}{3}\right)}{2!}x^2 + \frac{\frac{1}{3}\cdot\left(-\frac{2}{3}\right)\cdot\left(-\frac{5}{3}\right)}{3!}x^3 + \cdots = 1 + \tfrac{1}{3}x - \tfrac{1}{9}x^2 + \tfrac{5}{81}x^3 - \cdots.$$

和之前一样，作一个代换（substitution）便可令运算变得简洁。设 $x = 3y$，其中 $y = \frac{1}{125} = 0.008$；于是 $\frac{1}{3}x = y$、$\frac{1}{9}x^2 = y^2$ 及 $\frac{5}{81}x^3 = \frac{5}{3}y^3$。所以

$$\begin{aligned}\left(1+\frac{3}{125}\right)^{1/3} &\approx 1 + y - y^2 + \tfrac{5}{3}y^3 \\ &= 1 + 0.008 - 0.000064 + \tfrac{5}{3}(0.000000512) \\ &= 1.007936 + 0.00000085333\ldots \\ &= 1.00793685333\ldots.\end{aligned}$$

乘以 $\frac{5}{4}$，得

$$\sqrt[3]{2} \approx 1.25 \times 1.00793685333\ldots = 1.25992106\ldots,$$

所以准确至小数点后六位，$\sqrt[3]{2} \approx 1.259921$。

验算

计算机给出 $\sqrt{2} = 1.41421356\ldots$ 和 $\sqrt[3]{2} = 1.25992105\ldots$，所以两个近似值都准确至小数点后六位；我们求得的值是 1.41421357 和 1.25992107。我们已用电脑以精确分数核对过这两个级数。从第一个被略去的项可以大致看出误差的大小：$-\frac{5}{128}x^4$ 对 $\sqrt{2}$ 的估值贡献约 -9×10^{-9}，而 $-\frac{10}{243}x^4$ 对 $\sqrt[3]{2}$ 的估值贡献约 -1.7×10^{-8}。两者都远小于足以改变小数点后第六位的 5×10^{-7}。

常见错误

- 忘记还原缩放的倍数：级数给出的是 $\frac{7\sqrt{2}}{10}$ 和 $\frac{4\sqrt[3]{2}}{5}$，而不是 $\sqrt{2}$ 和 $\sqrt[3]{2}$ 本身。
- 级数中的正负号出错。在第一部分中 x 是负数，所以 1 之后的四项全是负的。在第二部分中，正负号则交替出现。
- 太早舍入。在最后一步除法或乘法之前，应至少保留八位小数。

STEP 之外

这道题的背后是带余项的泰勒定理（Taylor's theorem），它把“下一项极小”这个说法变成对误差的严格界限。它解释了为何一个公比约为 0.02 的级数，每多取一项便大约多得一位半至两位小数的准确度。同一个技巧，即找一个有理数 p/q 使 p^n/q^n 接近目标数的某个简单倍数，与数论中的连分数（continued fractions）和有理逼近（rational approximation）息息相关。Davenport 的 *The Higher Arithmetic*（数论入门名著）是阅读这些课题的好去处，而 Abbott 的 *Understanding Analysis*（数学分析入门教材）则仔细讲解了泰勒余项。

详细解答

STEP II 1998，第 3 题

纯数学 · 裂项相消级数

STEP II 1998，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/98-s2-q3

关键思路

通项的分母若是几个连续因子的乘积，自然应该把它拆成部分分式（partial fractions）。拆开后的各项会在求和时互相抵消，即所谓裂项相消（telescoping），最后只剩下首尾两端的几项。在第二部分，题目给出比值 a_n/a_{n-1}，这表示 a_n/a_1 是一连串比值的乘积，而这个乘积同样可以逐项相消。运气好的话，a_n 会变回第一部分的级数（series），而事实正是如此。

解答

和 S_N

设

$$\frac{2n-1}{n(n+1)(n+2)} = \frac{A}{n} + \frac{B}{n+1} + \frac{C}{n+2}.$$

利用掩盖法（cover-up rule）：取 $n=0$，得 $A = \frac{-1}{1\cdot 2} = -\frac{1}{2}$；取 $n=-1$，得 $B = \frac{-3}{(-1)(1)} = 3$；取 $n=-2$，得 $C = \frac{-5}{(-2)(-1)} = -\frac{5}{2}$。因此

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

留意 $A+B+C=0$。正因如此，这个和才可以裂项相消。

设 $H_N = 1 + \frac{1}{2} + \cdots + \frac{1}{N}$。则

$$\sum_{n=1}^{N} \frac{1}{n+1} = H_N - 1 + \frac{1}{N+1}, \qquad \sum_{n=1}^{N} \frac{1}{n+2} = H_N - 1 - \frac{1}{2} + \frac{1}{N+1} + \frac{1}{N+2}.$$

于是

$$\begin{aligned} S_N &= -\tfrac{1}{2}H_N + 3\left(H_N - 1 + \frac{1}{N+1}\right) - \tfrac{5}{2}\left(H_N - \tfrac{3}{2} + \frac{1}{N+1} + \frac{1}{N+2}\right) \\ &= \left(-\tfrac{1}{2} + 3 - \tfrac{5}{2}\right)H_N + \left(-3 + \tfrac{15}{4}\right) + \left(3 - \tfrac{5}{2}\right)\frac{1}{N+1} - \frac{5}{2(N+2)} \\ &= \frac{3}{4} + \frac{1}{2(N+1)} - \frac{5}{2(N+2)} = \frac{1}{2}\left(\frac{3}{2} + \frac{1}{N+1} - \frac{5}{N+2}\right), \end{aligned}$$

这正是所需的结果。由于 $A + B + C = 0$，H_N 的系数为零。

当 $N \to \infty$ 时，$\frac{1}{N+1}$ 和 $\frac{5}{N+2}$ 都趋向 0，所以

$$S_N \to \frac{3}{4}.$$

数列 a_n

对于 $n \geq 2$，把题目给出的比值逐个相乘：

$$\frac{a_n}{a_1} = \prod_{k=2}^{n} \frac{a_k}{a_{k-1}} = \prod_{k=2}^{n} \frac{k-1}{k+2} \times \prod_{k=2}^{n} \frac{2k-1}{2k-3}.$$

想一想 这两个乘积可以怎样化简?

把它们逐项写出来。第二个乘积可以直接相消，因为每一项的分子都与下一项的分母约去：

$$\prod_{k=2}^{n} \frac{2k-1}{2k-3} = \frac{3}{1} \cdot \frac{5}{3} \cdot \frac{7}{5} \cdots \frac{2n-1}{2n-3} = 2n - 1.$$

在第一个乘积中，分子依次为 $1, 2, \ldots, n-1$，分母依次为 $4, 5, \ldots, n+2$。由 4 至 $n-1$ 的所有因子全部约去，所以

$$\prod_{k=2}^{n} \frac{k-1}{k+2} = \frac{(n-1)!}{(n+2)!/3!} = \frac{6}{n(n+1)(n+2)}.$$

这个阶乘（factorial）形式对所有 $n \geq 2$ 都成立，即使 $n < 5$、根本没有因子可约时亦然。因此

$$\frac{a_n}{a_1} = \frac{6(2n-1)}{n(n+1)(n+2)}.$$

当 $n = 1$ 时，右边等于 $\frac{6}{6} = 1$，所以这条公式对所有 $n \geq 1$ 都成立。作为验算，取 $n = 2$ 得 $\frac{18}{24} = \frac{3}{4}$，与题目给出的比值 $\frac{1 \cdot 3}{4 \cdot 1}$ 吻合。

当 $a_1 = \frac{2}{9}$ 时，

$$a_n = \frac{2}{9} \cdot \frac{6(2n-1)}{n(n+1)(n+2)} = \frac{4}{3} \cdot \frac{2n-1}{n(n+1)(n+2)}.$$

所以 a_n 是第一个级数第 n 项的 $\frac{4}{3}$ 倍。它的部分和（partial sums）为 $\frac{4}{3}S_N \to \frac{4}{3} \cdot \frac{3}{4}$，因此

$$\sum_{n=1}^{\infty} a_n = 1.$$

验算

- $N=1$：公式给出 $\frac{1}{2}\left(\frac{3}{2}+\frac{1}{2}-\frac{5}{3}\right)=\frac{1}{6}$，正是首项 $\frac{1}{1\cdot2\cdot3}$。$N=2$ 给出 $\frac{1}{2}\left(\frac{3}{2}+\frac{1}{3}-\frac{5}{4}\right)=\frac{7}{24}=\frac{1}{6}+\frac{1}{8}$。两者皆正确。
- 数值上，首 1999 项之和为 $0.74900\ldots$，从下方趋近 $\frac{3}{4}$，与公式的预测一致（$S_N\approx\frac{3}{4}-\frac{2}{N}$）。
- 取 $a_1=\frac{2}{9}$，由递推关系（recurrence）逐项算出 a_n，首 299 项之和为 $0.9911\ldots$。余下的尾项之和约为 $\frac{4}{3}\cdot\frac{2}{300}\approx0.009$，与总和为 1 相符。

常见错误

- 处理裂项相消的首尾两端时出错。像上面那样用 H_N 来表示，就能把各项算得清清楚楚。
- 阶乘乘积错了一位。约分之前，先把开头和结尾的几个因子写出来。
- 忘记检查 a_n/a_1 的公式在 $n=1$ 时是否也成立，此时乘积是空的。

STEP 之外

裂项相消是微积分基本定理（fundamental theorem of calculus）的离散版本：$\sum(f(n+1)-f(n))=f(N+1)-f(1)$，正如 $\int f'=f(b)-f(a)$。第二部分的数列是一个超几何（hypergeometric）数列，即相邻两项之比是 n 的有理函数（rational function）。Gosper 算法（Gosper's algorithm）能够准确判定这类和何时具有闭合形式（closed form），电脑代数系统正是用这个方法来求级数之和。至于部分和极限的分析，可参阅 Abbott 的 *Understanding Analysis*。

详细解答

STEP II 1998，第 4 题

纯数学 · 分部积分法

STEP II 1998，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/98-s2-q4

关键思路

那个棘手的因子 $\operatorname{cosec}(x/2)$ 无法直接积分，但相邻两个被积函数（integrand）之*差*却好处理得多：正弦的差化积公式会产生一个因子 $\sin(x/2)$，刚好把余割完全约去。剩下的是多项式乘以余弦，用分部积分法（integration by parts）便可按标准做法处理。然后把这些差逐项相消（telescope），最终归结到一个可以直接手算的初始值。

解答

差 $I_n - I_{n-1}$

由于 $nx + x/2 = (n+\frac{1}{2})x$，两个被积函数之差为

$$\left(\tfrac{\pi}{2} - x\right)\left[\sin(n+\tfrac{1}{2})x - \sin(n-\tfrac{1}{2})x\right]\operatorname{cosec}\tfrac{x}{2}.$$

根据和差化积公式（factor formula）$\sin A - \sin B = 2\cos\frac{A+B}{2}\sin\frac{A-B}{2}$，取 $\frac{A+B}{2} = nx$ 及 $\frac{A-B}{2} = \frac{x}{2}$，得

$$\sin(n+\tfrac{1}{2})x - \sin(n-\tfrac{1}{2})x = 2\cos nx\,\sin\tfrac{x}{2}.$$

$\sin\frac{x}{2}$ 与余割互相约去，所以

$$I_n - I_{n-1} = 2\int_0^{\pi}\left(\tfrac{\pi}{2} - x\right)\cos nx\,\mathrm{d}x.$$

? **想一想** 在 $x = 0$ 处 $\operatorname{cosec}(x/2)$ 为无穷大，这样约去是否合理?

合理。每个原来的被积函数在 $x \to 0$ 时都有有限的极限，因为 $\sin(n+\frac{1}{2})x \big/ \sin\frac{x}{2} \to 2n+1$，

所以这些积分都是再普通不过的积分。当 $0 < x \le \pi$ 时，约去只是严格的代数运算，而只在 $x = 0$ 这一点改变被积函数的值，并不会改变积分的值。

现在用分部积分法，取 $u = \frac{\pi}{2} - x$ 及 $\frac{\mathrm{d}v}{\mathrm{d}x} = \cos nx$，于是 $v = \frac{1}{n}\sin nx$（因 $n \ge 1$，故成立）：

$$\begin{aligned}\int_0^{\pi} \left(\tfrac{\pi}{2} - x\right)\cos nx \,\mathrm{d}x &= \left[\left(\tfrac{\pi}{2} - x\right)\frac{\sin nx}{n}\right]_0^{\pi} + \frac{1}{n}\int_0^{\pi} \sin nx \,\mathrm{d}x \\ &= 0 + \frac{1}{n}\left[-\frac{\cos nx}{n}\right]_0^{\pi} = \frac{1-\cos n\pi}{n^2} = \frac{1-(-1)^n}{n^2},\end{aligned}$$

其中边界项为零，因为 $\sin 0 = \sin n\pi = 0$。因此

$$I_n - I_{n-1} = \frac{2\big(1-(-1)^n\big)}{n^2} = \begin{cases} \dfrac{4}{n^2}, & n \text{ 为奇数}, \\ 0, & n \text{ 为偶数}. \end{cases}$$

求 I_n 的值

要逐项相消，我们需要一个初始值。这个积分同样可以用来定义 I_0：当 $n = 0$ 时，被积函数是 $(\frac{\pi}{2} - x)\sin\frac{x}{2}\operatorname{cosec}\frac{x}{2} = \frac{\pi}{2} - x$，所以

$$I_0 = \int_0^{\pi}\left(\tfrac{\pi}{2} - x\right)\mathrm{d}x = \left[\tfrac{\pi}{2}x - \tfrac{1}{2}x^2\right]_0^{\pi} = 0,$$

而上面的差公式对 $n = 1$ 也成立（推导过程只用到 $n \ge 1$）。把 $1, 2, \ldots, n$ 各项的差相加：

$$I_n = I_0 + \sum_{k=1}^{n}\big(I_k - I_{k-1}\big) = \sum_{k=1}^{n}\frac{2\big(1-(-1)^k\big)}{k^2}.$$

只有奇数的 k 才有贡献，每项为 $4/k^2$，所以

$$\boxed{I_n = 4\sum_{\substack{k=1 \\ k \text{ 为奇数}}}^{n}\frac{1}{k^2} = 4\left(1 + \frac{1}{3^2} + \frac{1}{5^2} + \cdots\right),}$$

若 n 为奇数，最后一项是 $1/n^2$；若 n 为偶数，最后一项是 $1/(n-1)^2$。等价地，$I_n = 4\sum_{r=1}^{m}\frac{1}{(2r-1)^2}$，其中 $m = \lceil n/2 \rceil$。

如果你对使用 $n = 0$ 感到不安（题目说明 n 是正数），可以改由 I_1 开始。由于 $\sin\frac{3x}{2} = \sin\frac{x}{2}(1 + 2\cos x)$，利用上面 $n = 1$ 时的分部积分结果，可得 $I_1 = \int_0^{\pi}(\frac{\pi}{2} - x)(1 + 2\cos x)\,\mathrm{d}x = 0 + 2 \cdot 2 = 4$。这与公式吻合。

验算

- $I_1 = I_2 = 4$ 及 $I_3 = I_4 = 4 + \frac{4}{9} = \frac{40}{9} \approx 4.4444$。对原来的积分使用辛普森法则（Simpson's rule）（在 $x = 0$ 处取被积函数的极限 $\frac{\pi}{2}(2n+1)$），算得 I_3 为 $4.444444\ldots$，而且对每个 $n \le 12$，结果与上述和式的误差都在 10^{-8} 之内。

- 对应 n 的被积函数是 $\frac{\pi}{2} - x$ 乘以 $1 + 2\cos x + \cdots + 2\cos nx$，而 $\frac{\pi}{2} - x$ 关于 $x = \frac{\pi}{2}$ 是奇函数。当 k 为偶数时（常数 1 亦然），$\cos kx$ 关于 $x = \frac{\pi}{2}$ 是偶函数，所以乘积是奇函数，在 $[0, \pi]$ 上的积分为零。这就是 $I_0 = 0$ 以及偶数 k 没有任何贡献的原因。

常见错误

- 把正弦差化积公式中的正负号或因子 2 弄错；可代入一组数值检验，例如 $x = \pi, n = 1$。
- 把 $\cos n\pi = 1$ 写了出来，而没有写成 $(-1)^n$，结果令每一个差都变成零。
- 忘记初始值，或把答案写成对所有 k 求和，而没有留意到偶数项全部为零。

STEP 之外

函数 $\sin(n + \frac{1}{2})x \big/ \sin\frac{x}{2} = 1 + 2\sum_{k=1}^{n}\cos kx$ 称为狄利克雷核（*Dirichlet kernel*），是证明傅里叶级数（Fourier series）收敛的核心工具。当 $n \to \infty$ 时，我们的和趋向 $4 \cdot \frac{\pi^2}{8} = \frac{\pi^2}{2}$，这里用到 $\sum_{\text{奇数}k} 1/k^2 = \pi^2/8$，它是欧拉（Euler）的 $\sum 1/k^2 = \pi^2/6$ 的近亲（数值上，$I_{200} \approx 4.925$，而 $\pi^2/2 \approx 4.935$）。关于傅里叶级数和这类和式，Boyce 与 DiPrima 的 *Elementary Differential Equations*（常微分方程入门教材）有很好的讲解；至于收敛性背后的分析学，可参考 Abbott 的 *Understanding Analysis*（数学分析入门教材）。

详细解答

STEP II 1998，第 5 题

纯数学 · 复数的三角不等式

STEP II 1998，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/98-s2-q5

关键思路

$|z_1 - z_2|$ 是阿根图（Argand diagram）上两点之间的距离，所以关于 $|z_1 + z_2|$ 的不等式，其实就是说三角形的任何一边都不会比另外两边之和更长。把这个结果推广到 n 项之后，最后一部分只需用每一项的模（modulus）来估计该项的大小，再认出一个等比级数（geometric series）：$|z|$ 太小的话，左边就小得永远达不到 1。

解答

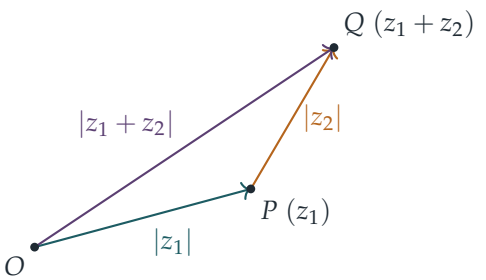
模、距离与三角不等式

若 $z = x + \mathrm{i}y$，其中 x, y 为实数，则 z 的模为

$$|z| = \sqrt{x^2 + y^2},$$

即原点到代表 z 的点 (x, y) 的距离。若 $z_1 = x_1 + \mathrm{i}y_1$ 及 $z_2 = x_2 + \mathrm{i}y_2$，则 $|z_1 - z_2| = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$，这正是代表 z_1 和 z_2 的两点之间的距离。

现在设 O、P 和 Q 分别为代表 0、z_1 和 $z_1 + z_2$ 的点。



根据上述的几何解释，

$$OP = |z_1 - 0| = |z_1|, \qquad PQ = |(z_1 + z_2) - z_1| = |z_2|, \qquad OQ = |z_1 + z_2|.$$

直线段 OQ 是由 O 到 Q 的最短路径，所以它不会比路径 $O \to P \to Q$ 更长；即 $OQ \le OP + PQ$，而等号成立当且仅当 P 位于线段 OQ 上（包括各点重合的退化情况）。因此

$$|z_1 + z_2| \le |z_1| + |z_2|.$$

这就是三角不等式（triangle inequality）。

n 项的不等式

我们用归纳法（induction）证明：对所有 $n \ge 1$ 及所有复数 $z_1, \dots, z_n$，

$$|z_1 + z_2 + \cdots + z_n| \le |z_1| + |z_2| + \cdots + |z_n|.$$

当 $n = 1$ 时，两边都等于 $|z_1|$。假设结果对某个 $n = k \ge 1$ 成立。给定 $z_1, \dots, z_{k+1}$，把两项的不等式应用于 $w = z_1 + \cdots + z_k$ 和 z_{k+1} 这两个复数：

$$\begin{aligned} |z_1 + \cdots + z_k + z_{k+1}| = |w + z_{k+1}| &\le |w| + |z_{k+1}| \\ &\le |z_1| + \cdots + |z_k| + |z_{k+1}|, \end{aligned}$$

其中最后一步用了归纳假设（inductive hypothesis）。所以结果对 $n = k + 1$ 成立，由归纳法可知它对所有 $n \ge 1$ 都成立。

最后一部分

假设 $|z| \le \frac{1}{4}$，并记 $r = |z|$，于是 $0 \le r \le \frac{1}{4}$。每一项都满足 $|a_k z^k| = |a_k|\,|z|^k \le 3r^k$，所以由 n 项的不等式可得

$$\begin{aligned} |a_1 z + a_2 z^2 + \cdots + a_n z^n| &\le |a_1|\, r + |a_2|\, r^2 + \cdots + |a_n|\, r^n \\ &\le 3\big(r + r^2 + \cdots + r^n\big). \end{aligned}$$

> **想一想** 怎样把 $3(r + r^2 + \cdots + r^n)$ 化成可以与 1 比较的形式？又怎样确保不等式是严格（strict）的？

这是一个有限等比级数，而由于 $r \le \frac{1}{4} < 1$，我们可以使用求和公式：

$$3\big(r + r^2 + \cdots + r^n\big) = \frac{3r(1 - r^n)}{1 - r}.$$

若 $r = 0$，上式等于 0。若 $0 < r \le \frac{1}{4}$，则 $0 < 1 - r^n < 1$，所以

$$\frac{3r(1 - r^n)}{1 - r} < \frac{3r}{1 - r} \le \frac{3 \cdot \frac{1}{4}}{1 - \frac{1}{4}} = 1,$$

其中最后一步成立，是因为 $\dfrac{3r}{1-r}$ 在 $[0,1)$ 上随 r 递增（分子递增，而取正值的分母递减）。无论是哪一种情况，都有

$$|a_1 z + a_2 z^2 + \cdots + a_n z^n| < 1,$$

所以左边不可能等于 1。因此方程没有满足 $|z| \le \frac{1}{4}$ 的解。

验算

- 这个界是最佳的：取每个 $a_k = 3$ 及 $z = \frac{1}{4}$，左边就是 $3\left(\frac{1}{4} + \cdots + \frac{1}{4^n}\right) = 1 - 4^{-n}$，例如当 $n = 1, 2, 3$ 时分别为 0.75、0.9375、0.984375。它可以任意接近 1，却永远达不到，这正是严格不等式如此重要的原因。
- 在一次随机测试中，我们选取了 200 000 组数据，其中 $n \le 15$、a_k 满足 $|a_k| \le 3$、z 满足 $|z| \le \frac{1}{4}$，所得 $|a_1 z + \cdots + a_n z^n|$ 的最大值约为 0.996，始终小于 1。

常见错误

- 单靠无穷级数之和 $3r/(1-r) \le 1$，只能得出 $|\text{左边}| \le 1$，这并不足以排除方程有解的可能。严格不等式要靠有限项之和（或者 $r > 0$ 时 $r^n > 0$ 这个事实）才能得到。
- 在归纳法中，两项的不等式必须应用于 $z_1 + \cdots + z_k$，并把它视为单一个复数；这正是该步骤的关键所在。

STEP 之外

三角不等式是度量（metric）的定义性质，而 $|z_1 - z_2|$ 令 $\mathbb{C}$ 成为一个度量空间（metric space）；分析学中几乎每一个关于收敛的论证都建基于它。最后一部分是一类定理的入门版本，这类定理用来确定多项式和幂级数（power series）的零点（zeros）位置：这里 $f(z) = a_1 z + \cdots + a_n z^n - 1$ 在圆盘 $|z| \le \frac{1}{4}$ 内没有零点，与儒歇定理（Rouché's theorem）的精神一脉相承。想了解复数的几何，可读 Needham 的 *Visual Complex Analysis*（以图像阐释复分析）；想了解分析学中的度量概念，可读 Abbott 的 *Understanding Analysis*（实分析入门）。

详细解答

STEP II 1998，第 6 题

纯数学 · 摆线与法线

STEP II 1998，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/98-s2-q6

关键思路

两条曲线都是摆线（cycloid），即滚动的圆的圆周上一点所描出的路径。只要用上半角公式（half-angle formulae）$1+\cos\theta = 2\cos^2\frac{\theta}{2}$、$1-\cos\theta = 2\sin^2\frac{\theta}{2}$ 和 $\sin\theta = 2\sin\frac{\theta}{2}\cos\frac{\theta}{2}$，有关它们的一切都会变得简单。斜率（gradient）会化简成 $\frac{\theta}{2}$ 的单一个正切或余切，而法线（normal）的方程亦会缩成简洁的一行。要证明一条直线是曲线 (2) 的切线（tangent），只需在 (2) 上找出一点，使该点在直线上，而且曲线在该点的斜率等于直线的斜率。

（题目中第二条曲线本应写 y_2 的地方写成了 y_1；这是印刷错误。）

解答

斜率

对于曲线 (1)，$\dfrac{\mathrm{d}x_1}{\mathrm{d}\theta} = 1+\cos\theta$ 及 $\dfrac{\mathrm{d}y_1}{\mathrm{d}\theta} = -\sin\theta$，所以

$$\frac{\mathrm{d}y_1}{\mathrm{d}x_1} = \frac{-\sin\theta}{1+\cos\theta} = \frac{-2\sin\frac{\theta}{2}\cos\frac{\theta}{2}}{2\cos^2\frac{\theta}{2}} = -\tan\frac{\theta}{2},$$

此式在 $\cos\frac{\theta}{2} \neq 0$ 时成立，即 $\theta \neq \pi$。

对于曲线 (2)，$\dfrac{\mathrm{d}x_2}{\mathrm{d}\theta} = 1-\cos\theta$ 及 $\dfrac{\mathrm{d}y_2}{\mathrm{d}\theta} = \sin\theta$，所以

$$\frac{\mathrm{d}y_2}{\mathrm{d}x_2} = \frac{\sin\theta}{1-\cos\theta} = \frac{2\sin\frac{\theta}{2}\cos\frac{\theta}{2}}{2\sin^2\frac{\theta}{2}} = \cot\frac{\theta}{2},$$

此式在 $\theta \neq 0, 2\pi$ 时成立。

当 $\theta = \pi/2$ 时，$\frac{\theta}{2} = \frac{\pi}{4}$：曲线 (1) 的斜率为 -1，曲线 (2) 的斜率为 1。

当 $\theta = 3\pi/2$ 时，$\frac{\theta}{2} = \frac{3\pi}{4}$，此时 $\tan = -1$ 且 $\cot = -1$：曲线 (1) 的斜率为 1，曲线 (2) 的斜率为 -1。

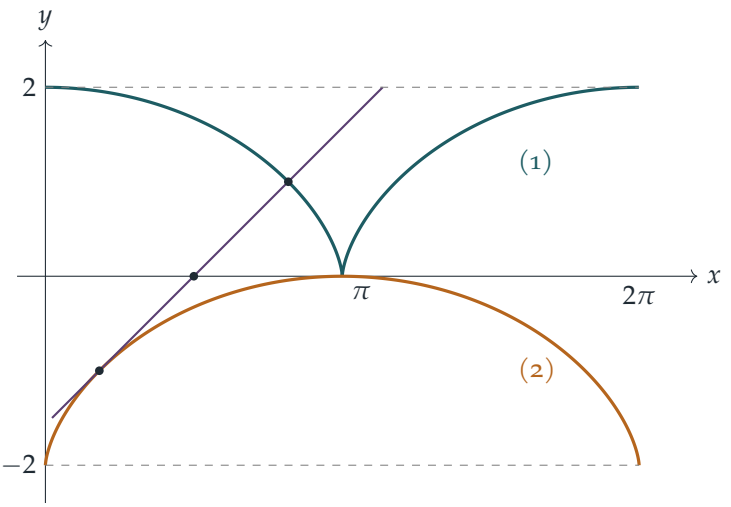
略图

x_1 和 x_2 都是 θ 的非递减函数（它们的导数 $1 \pm \cos\theta$ 永不为负），所以每条曲线都由左至右，从 $x = 0$ 延伸到 $x = 2\pi$。

曲线 (1) 由 $(0, 2)$ 出发，起点处斜率为 0，下降至 $(\pi, 0)$，然后对称地上升至 $(2\pi, 2)$。当 $\theta = \pi$ 时，两个导数同时为零，而 $|\mathrm{d}y_1/\mathrm{d}x_1| = |\tan\frac{\theta}{2}| \to \infty$，所以在 $(\pi, 0)$ 有一个尖点（cusp），该处的切线是铅垂的。由于 $0 \le y_1 \le 2$，曲线位于 $y = 0$ 与 $y = 2$ 之间。

曲线 (2) 在 $(0, -2)$ 有一个尖点（切线铅垂，因为 $\cot\frac{\theta}{2} \to \infty$），上升至 $(\pi, 0)$，该处的斜率为 $\cot\frac{\pi}{2} = 0$，然后下降至尖点 $(2\pi, -2)$。它是一个立于直线 $y = -2$ 上的普通摆线拱。

两条曲线在 $(\pi, 0)$ 相接：曲线 (1) 铅垂地到达该点，曲线 (2) 则水平地越过该点。整幅图关于 $x = \pi$ 对称，因为 $\theta \mapsto 2\pi - \theta$ 会令 $x \mapsto 2\pi - x$，而两条曲线上的 y 都保持不变。图中紫红色的直线是 (1) 在 $\theta = \pi/2$ 处的法线，下文会再提到它。



曲线 (1) 的法线

(1) 的切线斜率为 $-\tan\frac{\theta}{2}$，所以法线的斜率为 $\cot\frac{\theta}{2}$（其中 $\theta \neq 0, \pi, 2\pi$；这些情况留待下文处理）。在点 $(\theta + \sin\theta,\ 1 + \cos\theta)$ 处的法线为

$$y - (1 + \cos\theta) = \frac{\cos\frac{\theta}{2}}{\sin\frac{\theta}{2}}\left(x - \theta - \sin\theta\right).$$

全式乘以 $\sin\frac{\theta}{2}$，并使用半角形式：

$$y\sin\tfrac{\theta}{2} - 2\cos^2\tfrac{\theta}{2}\sin\tfrac{\theta}{2} = (x - \theta)\cos\tfrac{\theta}{2} - 2\sin\tfrac{\theta}{2}\cos^2\tfrac{\theta}{2}.$$

两边最后一项相等，可以互相抵消，剩下

$$\boxed{y\sin\tfrac{\theta}{2} = (x-\theta)\cos\tfrac{\theta}{2}}$$

这就是法线的方程。

这个形式同时涵盖了那些例外的值。当 $\theta = 0$ 时，它变成 $x = 0$，即通过 $(0,2)$ 的铅垂线；这是正确的，因为该处的切线是水平的。当 $\theta = 2\pi$ 时，同理可得 $x = 2\pi$。当 $\theta = \pi$ 时，它变成 $y = 0$，即通过尖点的水平线，与尖点处的铅垂切线互相垂直。

留意法线必定通过 x 轴上的点 $(\theta, 0)$。这正是滚动的圆与它所滚过的直线相切的那一点，是摆线的一个经典性质。

法线是曲线 (2) 的切线

? **想一想** 法线应该与曲线 (2) 相切于哪一点?

(2) 在参数（parameter）为 ϕ 处的切线斜率为 $\cot\frac{\phi}{2}$，而法线的斜率为 $\cot\frac{\theta}{2}$。对于 ϕ, θ 在 $(0, 2\pi)$ 内的情况，两者只在 $\phi = \theta$ 时相等，因为 cot 在 $(0, \pi)$ 上是一对一（one-to-one）的。所以唯一的候选点就是 (2) 上参数同为 θ 的那一点。

该点是 $(x_2, y_2) = (\theta - \sin\theta,\ -(1+\cos\theta))$。代入法线的方程：

$$\begin{aligned}
\text{左方} &= y_2\sin\tfrac{\theta}{2} = -2\cos^2\tfrac{\theta}{2}\,\sin\tfrac{\theta}{2},\\
\text{右方} &= (x_2-\theta)\cos\tfrac{\theta}{2} = -\sin\theta\cos\tfrac{\theta}{2} = -2\sin\tfrac{\theta}{2}\cos^2\tfrac{\theta}{2}.
\end{aligned}$$

两者相等，所以 (1) 在 θ 处的法线通过 (2) 上参数为 θ 的点。在该点，(2) 的斜率为 $\cot\frac{\theta}{2}$，与法线的斜率相同。一条直线若通过曲线上某一点，而且斜率等于曲线在该点的斜率，它就是曲线在该点的切线，所以 (1) 的法线是 (2) 的切线，证明完毕。

特殊值的情况同样成立。当 $\theta = \pi$ 时，法线 $y = 0$ 与 (2) 相切于其顶点 $(\pi, 0)$，该处 (2) 的斜率为 0。当 $\theta = 0$（或 2π）时，法线 $x = 0$（或 $x = 2\pi$）就是 (2) 在尖点处的铅垂切线。

验算

- 当 $\theta = \pi/2$ 时，法线为 $y = x - \frac{\pi}{2}$。它通过 (1) 上的 $(\frac{\pi}{2}+1, 1)$，通过 $(\frac{\pi}{2}, 0)$，亦通过 (2) 上的 $(\frac{\pi}{2}-1, -1)$，而 (2) 在该点的斜率为 1，与第一部分的结果一致。这就是略图中紫红色的直线。
- 以有限差分（finite differences）对斜率作数值验算，结果与 $-1, 1$ 及 $1, -1$ 吻合至小数点后六位。在 1000 个随机选取的 θ 值上，法线都通过两个点，并且与 (2) 的切线平行，误差在 10^{-12} 之内。

常见错误

- 把斜率保留为 $-\sin\theta/(1+\cos\theta)$ 的形式，然后在求法线时硬闯繁复的代数运算。用上半角形式，求法线只需两行计算。
- 只证明了法线与曲线 (2) 相交。相交并不足够：还须证明在交点处两者的斜率相同。

STEP 之外

一条曲线的各个曲率中心(centres of curvature)所描出的曲线,称为它的*渐屈线*(evolute),它同时也是该曲线所有法线的包络线（envelope）。这道题说明了摆线的渐屈线是一条完全相同的摆线，只是平移了半个拱。沿法线由 (1) 到它与 (2) 的切点的距离为 $4|\cos\frac{\theta}{2}|$,刚好等于 (2) 由该点到其顶点的弧长（arc length）。所以 (1) 可以看成是一条由 (2) 上逐渐解开的绳子的末端所描出的曲线。惠更斯（Huygens）利用这一点造出了一个周期与摆幅无关的摆。想了解曲率、渐屈线和渐伸线（involutes），可参阅 Pressley 的 *Elementary Differential Geometry*（初等微分几何）。

详细解答

STEP II 1998，第 7 题

纯数学 · 以导数证明不等式

STEP II 1998，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/98-s2-q7

关键思路

每个不等式（inequality）的证法都一样：函数在 $x = 0$ 处为零，而它的导数（derivative）为正，所以函数由零开始上升。巧妙之处在于这些函数是环环相扣的。g 的导数原来与 f 有关，而 h 的导数原来就是 g 在 $2x$ 处的值。因此每一部分都为下一部分提供所需的正导数。在第 (iv) 部分，取对数（logarithm）可把商（quotient）F 化为和，而 F'/F 的正负号就是第 (iii) 部分那个式子的正负号。

解答

全题所用的工具是：若 ϕ 可微（differentiable），$\phi(0) = 0$，且当 $0 < x < a$ 时 $\phi'(x) > 0$，则 ϕ 在 $[0, a)$ 上严格递增（strictly increasing），所以当 $0 < x < a$ 时 $\phi(x) > \phi(0) = 0$。（根据中值定理（mean value theorem），存在某个 $c \in (0, x)$ 使得 $\phi(x) = \phi(x) - \phi(0) = x\,\phi'(c)$，而这是正数。）

第 (i) 部分

我们有 $\mathrm{f}(0) = \tan 0 - 0 = 0$，而且

$$\mathrm{f}'(x) = \sec^2 x - 1 = \tan^2 x.$$

当 $0 < x < \frac{1}{2}\pi$ 时，$\tan x$ 有定义且为正，所以 $\mathrm{f}'(x) > 0$。因此当 $0 < x < \frac{1}{2}\pi$ 时 $\mathrm{f}(x) > 0$；也就是说，在该范围内 $\tan x > x$。

第 (ii) 部分

同样地，$g(0) = 2 - 2 - 0 = 0$。求导，并对 $x \sin x$ 使用积法则（product rule），

$$\begin{aligned} g'(x) &= 2\sin x - (\sin x + x\cos x) \\ &= \sin x - x\cos x \\ &= \cos x\,(\tan x - x) = \cos x\, f(x). \end{aligned}$$

当 $0 < x < \frac{1}{2}\pi$ 时，我们有 $\cos x > 0$，而由第 (i) 部分知 $f(x) > 0$。所以 $g'(x) > 0$，又因为 $g(0) = 0$，可知当 $0 < x < \frac{1}{2}\pi$ 时 $g(x) > 0$。

第 (iii) 部分

这次同样有 $h(0) = 0$。求导，

$$\begin{aligned} h'(x) &= 2 + \cos 2x - 2x\sin 2x - 3\cos 2x \\ &= 2 - 2\cos 2x - 2x\sin 2x. \end{aligned}$$

> **想一想** 你以前见过这个式子吗?

它就是把 g 中的 x 换成 $2x$:

$$h'(x) = 2 - 2\cos(2x) - (2x)\sin(2x) = g(2x).$$

若 $0 < x < \frac{1}{4}\pi$，则 $0 < 2x < \frac{1}{2}\pi$，所以由第 (ii) 部分知 $g(2x) > 0$。这就是为甚么范围现在是 $\frac{1}{4}\pi$ 而不是 $\frac{1}{2}\pi$。所以在 $0 < x < \frac{1}{4}\pi$ 上 $h'(x) > 0$，再加上 $h(0) = 0$，便得出当 $0 < x < \frac{1}{4}\pi$ 时 $h(x) > 0$。

至于“由此”的部分，用二倍角（double angle）公式改写题目所给的式子。由于 $\sin^2 x + 3\cos^2 x = 1 + 2\cos^2 x = 2 + \cos 2x$ 及 $3\sin x\cos x = \frac{3}{2}\sin 2x$，

$$\begin{aligned} x(\sin^2 x + 3\cos^2 x) - 3\sin x\cos x &= x(2 + \cos 2x) - \tfrac{3}{2}\sin 2x \\ &= 2x + x\cos 2x - \tfrac{3}{2}\sin 2x = h(x). \end{aligned}$$

所以该式子正正就是 $h(x)$，而它在 $0 < x < \frac{1}{4}\pi$ 时为正，这正是所需的结果。

第 (iv) 部分

当 $0 < x < \frac{1}{4}\pi$ 时，x、$\sin x$ 和 $\cos x$ 这几个量全为正，所以 $F(x) > 0$，我们可以取对数:

$$\ln F(x) = \ln x + \tfrac{1}{3}\ln\cos x - \ln\sin x.$$

两边对 x 求导，

$$\frac{F'(x)}{F(x)} = \frac{1}{x} - \frac{\sin x}{3\cos x} - \frac{\cos x}{\sin x}.$$

想一想 怎样看出它的正负号?

把它通分，公分母（common denominator）为 $3x\sin x\cos x$，在这个区间上是正数：

$$\frac{\mathrm{F}'(x)}{\mathrm{F}(x)} = \frac{3\sin x\cos x - x\sin^2 x - 3x\cos^2 x}{3x\sin x\cos x}$$
$$= -\frac{x(\sin^2 x + 3\cos^2 x) - 3\sin x\cos x}{3x\sin x\cos x} = -\frac{\mathrm{h}(x)}{3x\sin x\cos x}.$$

由第 (iii) 部分，分子 $\mathrm{h}(x)$ 为正，分母亦为正，所以 $\mathrm{F}'(x)/\mathrm{F}(x) < 0$。由于 $\mathrm{F}(x) > 0$，两边乘以它便得出当 $0 < x < \frac{1}{4}\pi$ 时 $\mathrm{F}'(x) < 0$，这正是所需的结果。

验算

- 抽查数值：$\mathrm{F}(0.1) \approx 0.999998$、$\mathrm{F}(0.5) \approx 0.99849$ 及 $\mathrm{F}(\frac{1}{4}\pi) \approx 0.98954$，所以 F 是递减的，只是减得很慢。
- 在 0 附近，这些不等式非常紧。展开成麦克劳林级数（Maclaurin series）可得 $\mathrm{h}(x) = \frac{4}{15}x^5 + \cdots$，而 x, x^3 两项互相抵消。这就是为甚么你不能单凭观察就“看出” $\mathrm{h} > 0$，而需要这一连串导数。数值上，$\mathrm{h}(0.01)/0.01^5 \approx 0.26666$。
- 我们在 20 000 个随机点上测试了全部四个不等式，以及恒等式 $\mathrm{g}' = \mathrm{f}\cos x$、$\mathrm{h}'(x) = \mathrm{g}(2x)$ 和 $\mathrm{F}'/\mathrm{F} = -\mathrm{h}/(3x\sin x\cos x)$。

常见错误

- 只论证 $\mathrm{f}'(x) \geq 0$，或忘记指出 $\mathrm{f}(0) = 0$。两项事实都不可缺少：起点为零，以及导数严格为正。
- 在第 (iii) 部分使用第 (ii) 部分的结果，却没有检查 $2x$ 是否落在 $(0, \frac{1}{2}\pi)$ 之内。$x < \frac{1}{4}\pi$ 这个限制正是由此而来。
- 由 $\mathrm{F}'/\mathrm{F} < 0$ 推出 $\mathrm{F}' < 0$，却没有指出 $\mathrm{F} > 0$。

STEP 之外

“起点为零而导数为正，则函数为正”这个原理是中值定理的推论，而中值定理是大学分析（analysis）课程中最早遇到的重要定理之一。像本题这样反复运用这个原理，是证明 $\sin x > x - \frac{1}{6}x^3$ 这类不等式的标准方法，而且它与带余项的泰勒定理（Taylor's theorem）关系密切。Abbott 的 *Understanding Analysis*（分析学入门教科书）对中值定理及其应用有仔细的讲解。

详细解答

STEP II 1998，第 8 题

纯数学 · 向量几何

STEP II 1998，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/98-s2-q8

关键思路

这里的一切，都在于把每条直线写成“点 + 参数（parameter）× 方向”的形式，然后巧妙地选取参数。各式子对 **a**、**b**、**c** 是对称（symmetric）的，所以如果某个参数值给出的答案对这三个向量对称，那么同一点就位于全部三条直线上。这正是题目中“凭观察”（by inspection）一语要我们去找的东西。

解答

三条中线（median）交于 R

BC 的中点（mid-point）的位置向量（position vector）为 $\frac{1}{2}(\mathbf{b}+\mathbf{c})$，所以由 A 到该中点的直线为

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

我们希望三个系数相等，使结果对称：由 $1-s = \frac{s}{2}$ 得 $s = \frac{2}{3}$，而该点为

$$\mathbf{r} = \tfrac{1}{3}(\mathbf{a}+\mathbf{b}+\mathbf{c}).$$

由 B 和由 C 出发的直线，可把 $\mathbf{a}, \mathbf{b}, \mathbf{c}$ 循环置换而得，而同样取 $s = \frac{2}{3}$，便在每条直线上得出同一个对称的点。因此三条直线都通过点 R，其位置向量为

$$\mathbf{r}_R = \tfrac{1}{3}(\mathbf{a}+\mathbf{b}+\mathbf{c}),$$

即三角形的形心（centroid）。（由于 A、B、C 构成一个真正的三角形，这些直线中没有两条重合，所以 R 是它们唯一的公共点。）

三条平行线交于 Q

设 P 的位置向量为 $\mathbf{p}$。三条直线分别是

$$\begin{aligned}
&\text{通过}BC\text{ 的中点，且平行于}PA: \quad \mathbf{r} = \tfrac{1}{2}(\mathbf{b}+\mathbf{c}) + \lambda(\mathbf{a}-\mathbf{p}),\\
&\text{通过}CA\text{ 的中点，且平行于}PB: \quad \mathbf{r} = \tfrac{1}{2}(\mathbf{c}+\mathbf{a}) + \mu(\mathbf{b}-\mathbf{p}),\\
&\text{通过}AB\text{ 的中点，且平行于}PC: \quad \mathbf{r} = \tfrac{1}{2}(\mathbf{a}+\mathbf{b}) + \nu(\mathbf{c}-\mathbf{p}).
\end{aligned}$$

想一想 λ 取甚么值，才会使第一个式子对 $\mathbf{a}$、$\mathbf{b}$、$\mathbf{c}$ 对称?

在第一条直线中，$\mathbf{b}$ 和 $\mathbf{c}$ 的系数各为 $\frac{1}{2}$，而 $\mathbf{a}$ 的系数是 λ；要令它们一致，便需要 $\lambda = \frac{1}{2}$。于是

$$\mathbf{r} = \tfrac{1}{2}(\mathbf{a}+\mathbf{b}+\mathbf{c}-\mathbf{p}).$$

在另外两条直线中取 $\mu = \frac{1}{2}$ 和 $\nu = \frac{1}{2}$，得到的是完全相同的向量。因此三条直线都通过点 Q，其中

$$\mathbf{q} = \tfrac{1}{2}(\mathbf{a}+\mathbf{b}+\mathbf{c}-\mathbf{p}).$$

PQ 与平面 ABC 交于 R

首先，$Q \neq P$：假如 $\mathbf{q} = \mathbf{p}$，则 $\mathbf{a}+\mathbf{b}+\mathbf{c} = 3\mathbf{p}$，所以 $\mathbf{p} = \mathbf{r}_R$，$P$ 便会是点 R，而这一点位于平面 ABC 上。所以 PQ 确实是一条直线，即

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

要消去含 $\mathbf{p}$ 的项，我们需要 $1 - \frac{3}{2}t = 0$，即 $t = \frac{2}{3}$，这时

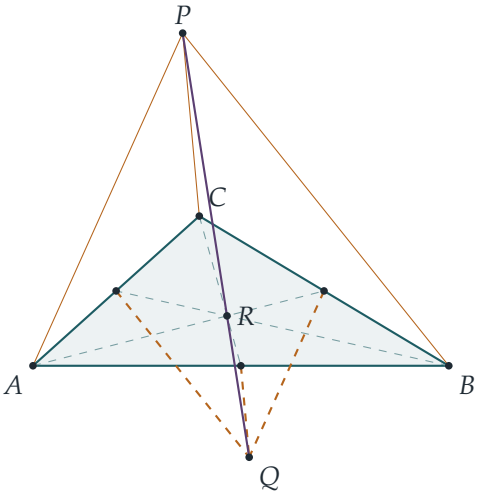
$$\mathbf{r} = \mathbf{p} + \tfrac{1}{3}(\mathbf{a}+\mathbf{b}+\mathbf{c}) - \mathbf{p} = \tfrac{1}{3}(\mathbf{a}+\mathbf{b}+\mathbf{c}) = \mathbf{r}_R.$$

所以 R 位于直线 PQ 上。它同时位于平面 ABC 上（它在由 A 出发的中线上，而这条中线位于该平面内）。

想一想 PQ 会不会还在别处与这个平面相交?

不会。一条直线与一个平面或者没有交点，或者恰好有一个交点，或者整条位于平面之内。直线 PQ 并不在平面 ABC 内，因为 P 在这条直线上，而 P 不在该平面上。所以它与平面最多只有一个交点，而这个交点就是 R。这就证明了 PQ 与平面 ABC 交于 R。

额外的收获是，$t = \frac{2}{3}$ 告诉我们 R 把 PQ 按比 $PR : RQ = 2:1$ 分割；等价地，$\mathbf{q} - \mathbf{r}_R = -\frac{1}{2}(\mathbf{p} - \mathbf{r}_R)$。



验算

- 每个声称的交点都可以代回去检验：例如 $\mathbf{q} - \frac{1}{2}(\mathbf{b}+\mathbf{c}) = \frac{1}{2}(\mathbf{a}-\mathbf{p})$，显然平行于 PA。
- 取 $A=(0,0,0)$、$B=(3,0,0)$、$C=(0,3,0)$ 和 $P=(0,0,3)$，我们得到 $R=(1,1,0)$ 和 $Q=(\frac{3}{2},\frac{3}{2},-\frac{3}{2})$；由 P 到 Q 的三分之二处的点是 $(1,1,0)=R$，位于平面 $z=0$ 上。
- 我们还用精确的有理数运算，对 2000 组随机选取的 A、B、C、P 检验了所有结果：直线 PQ 每次都与平面 ABC 交于 R，参数为 $t=\frac{2}{3}$。

常见错误

- 解方程求出其中两条直线上的公共点后，忘记检查它是否也在第三条直线上。
- 证明了 R 位于 PQ 上，却没有解释为何它就是 PQ 与平面唯一的交点；“P 不在平面上”这个条件正是用在这里。

STEP 之外

这个结果背后有一个简洁的几何原因。以 R 为中心、比例因子（scale factor）为 $-\frac{1}{2}$ 的放大（enlargement）（即映射 $\mathbf{x} \mapsto \mathbf{r}_R - \frac{1}{2}(\mathbf{x}-\mathbf{r}_R)$）把 A、B、C 分别送到 BC、CA、AB 的中点，并把任何直线送到与之平行的直线。所以它把交于 P 的直线 PA、PB、PC 变成那三条新直线，因而这三条新直线必定交于 P 的像；这个像就是 Q，而 P、R、Q 共线（collinear），因为放大的中心与每一点及其像都共线。通过变换（仿射映射（affine map）及其不动点（fixed point））来思考几何，是大学一年级线性代数（linear algebra）的一

个核心思想；Strang 的 *Introduction to Linear Algebra*（线性代数入门教材）是了解这个思想如何展开的好去处。

详细解答

STEP II 1998，第 9 题

力学 · 桁架的静力学

> **STEP II 1998，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/98-s2-q9

关键思路

框架是轻质的（light），所以只有两个外力作用在它上面：A 处的支承力和 C 处的重量。两个力要达致平衡（equilibrium），必须大小相等、方向相反，并且沿同一直线作用；这就同时决定了支承力和六边形的悬挂方式。之后便是节点法（method of joints）：从只有两枝杆相交的节点入手，分解（resolve）各力，然后沿六边形逐个节点推进，每次都移到最多只有两个未知力的节点。

解答

支承力

把六边形的各边和三枝对角杆视为轻杆（light rod），在六个顶点处以光滑铰接（smoothly pinned）相连，并以 W 表示 1 kg 质量的重量，所以 $W = g \approx 9.8\,\mathrm{N}$。各杆都是轻质的，因此作用于整个框架（framework）的外力只有 A 处支承所施的力 $\mathbf{R}$，以及在 C 处铅垂向下作用的重量 W。

> ? **想一想** 框架可以随我们喜欢以任何角度悬挂吗？

不可以。就整个框架而言，各力之和必须为零，而且总力矩（moment）也必须为零。对 A 取力矩，支承力没有力矩，所以 C 处的重量对 A 的力矩也必须为零。这只有在 C 位于 A 正下方时才会发生。因此框架悬挂时，短对角线 AC 是铅垂的，而分解各力可得

$$\mathbf{R} = W \text{ 铅垂向上}, \qquad |\mathbf{R}| = g \approx 9.8\,\mathrm{N},$$

即 1 kg 的重量，沿 CA 作用。

各杆中的力

取边长为 1（各力与边长无关）。当 AC 铅垂时，B 位于 AC 的一半高度处，在它的一侧，而六边形的其余部分则在另一侧：F 与 A 同高，E 与 B 同高，D 与 C 同高。所以 AF、BE 和 CD 是水平的，边 AB、BC、DE、EF 与水平线成 $60°$，而短对角线 BD、BF 与水平线成 $30°$。

设 T_{XY} 为杆 XY 中的张力（tension）；在杆的每一端，它都把节点拉向另一端。张力为负，表示该杆受推力（thrust）。

想一想 共有九个未知的杆力。我们应从哪里入手?

从只有两枝杆相交的节点入手，因为这样两条方程（水平和铅垂方向）便可以立即解出。节点 C 只连着 CB 和 CD，而且负载就挂在这里。

节点 C。杆 CB 指向水平线以上 $60°$，而 CD 是水平的，背向 B 的一侧。分解各力，

$$\begin{aligned} &\text{铅垂方向:} \quad T_{BC}\sin 60° = W \;\Rightarrow\; T_{BC} = \tfrac{2W}{\sqrt{3}}, \\ &\text{水平方向:} \quad T_{CD} = T_{BC}\cos 60° = \tfrac{W}{\sqrt{3}}. \end{aligned}$$

节点 D。现在只有 DE 和 DB 是未知的。从 D 看，杆 DC 沿水平方向指向 B 的一侧，DE 指向水平线以上 $60°$ 并背向该侧，而 DB 指向水平线以上 $30°$ 并朝向该侧。分解各力，

$$\begin{aligned} &\text{铅垂方向:} \quad T_{DE}\sin 60° + T_{BD}\sin 30° = 0 \;\Rightarrow\; T_{BD} = -\sqrt{3}\,T_{DE}, \\ &\text{水平方向:} \quad T_{CD} - T_{DE}\cos 60° + T_{BD}\cos 30° = 0 \;\Rightarrow\; T_{CD} - \tfrac{1}{2}T_{DE} - \tfrac{3}{2}T_{DE} = 0, \end{aligned}$$

所以 $T_{DE} = \frac{1}{2}T_{CD} = \frac{W}{2\sqrt{3}}$，而 $T_{BD} = -\frac{W}{2}$。

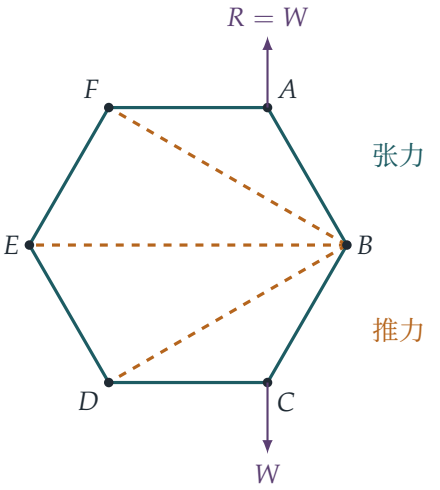
节点 E。杆 ED 指向水平线以下 $60°$，EF 指向水平线以上 $60°$，而 EB 是水平的，三者都朝向 B 的一侧。沿铅垂方向分解，得 $T_{EF} = T_{DE} = \frac{W}{2\sqrt{3}}$。沿水平方向分解，

$$\tfrac{1}{2}T_{DE} + \tfrac{1}{2}T_{EF} + T_{BE} = 0 \;\Rightarrow\; T_{BE} = -\tfrac{W}{2\sqrt{3}}.$$

节点 F。这是节点 D 对直线 BE 的镜像。沿铅垂方向分解，$T_{BF} = -\sqrt{3}\,T_{EF} = -\frac{W}{2}$；沿水平方向，$T_{FA} = 2T_{EF} = \frac{W}{\sqrt{3}}$。

节点 A。水平方向，$T_{AB}\cos 60° = T_{FA}$，所以 $T_{AB} = \frac{2W}{\sqrt{3}}$。铅垂方向，$R = T_{AB}\sin 60° = W$，与第一部分的结果一致。

节点 B 还会给出两条方程，但它们自动成立：整个框架本身已经处于平衡。验证这两条方程，是检查运算有否出错的好方法。



总结如下，其中 $W = g$：

- 六条边全部受张力：AB 和 BC 承受 $\frac{2W}{\sqrt{3}}$，CD 和 FA 承受 $\frac{W}{\sqrt{3}}$，DE 和 EF 承受 $\frac{W}{2\sqrt{3}}$；
- 三枝内杆全部受推力：BD 和 BF 承受 $\frac{W}{2}$，而 BE 承受 $\frac{W}{2\sqrt{3}}$。

所以 BE 中的力是推力，大小为

$$\frac{g}{2\sqrt{3}} = \frac{\sqrt{3}g}{6} \approx 2.83\,\text{N}.$$

验算

- 对称性：对直线 BE 作反射，会把 $A \leftrightarrow C$ 和 $F \leftrightarrow D$ 互换，并把 A 处向上的力 W 变成 C 处向下的力 W。所以负载是对称的，答案也应该对称：$T_{AB} = T_{BC}$，$T_{FA} = T_{CD}$，$T_{EF} = T_{DE}$，$T_{BF} = T_{BD}$。结果确实如此。
- 节点 B：水平方向，两条边向一方给出 $2 \cdot \frac{2W}{\sqrt{3}} \cdot \frac{1}{2}$，而 BD、BF、BE 向另一方给出 $2 \cdot \frac{W}{2} \cdot \frac{\sqrt{3}}{2} + \frac{W}{2\sqrt{3}} = \frac{2W}{\sqrt{3}}$；铅垂方向，$AB$ 和 BC 互相抵消，BD 和 BF 亦然。
- 我们取 $g = 9.8$，以数值方法解出全部十二条节点方程：它们互相一致，给出 $R = 9.8\,\text{N}$ 向上，以及 $|T_{BE}| \approx 2.829\,\text{N}$ 的推力；而若 AC 倾斜，方程便无解。

常见错误

- 把六边形画成有一条边或一条长对角线是铅垂的。悬挂位置是被迫确定的：C 在 A 的下方。

- 在不同节点之间混淆正负号的约定。一开始就定下“张力把节点拉向杆的另一端”，然后让负的答案告诉你哪些杆受推力。

STEP 之外

一个有 j 个节点的框架，当它恰好有 $2j-3$ 枝杆时，便是刚性（rigid）而且静定（statically determinate）的：这里 $2\cdot 6-3=9$，这就是为何单靠静力学便可求出各力。若杆的数目更多，结构便有“冗余”，这时就需要考虑杆的弹性；这是结构力学的起点，也是图的刚性理论这门数学的起点。想了解背后的向量力学，可读 Kleppner and Kolenkow, *An Introduction to Mechanics*（力学入门教材）。

详细解答

STEP II 1998，第 10 题

力学 · 移动楔上的质点

STEP II 1998，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/98-s2-q10

关键思路

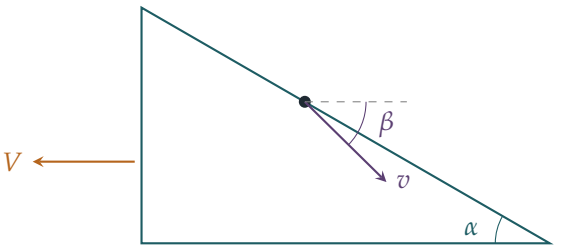
整道题由两个事实推动。由于质点（particle）一直留在斜面上，它相对于楔块（*wedge*）的速度必定沿着斜面的方向。另外，由于系统以外没有任何力在水平方向推动它，而且除重力外没有其他力做功，所以水平动量（momentum）和能量都守恒。(i)–(iii) 每一部分，都是把其中一个事实写成方程。之后第 (ii) 部分说明 β 始终不变，于是最后一部分就变成匀加速运动。

解答

模型。质点视为质点；楔块是只滑动而不转动的刚体（rigid body）；斜面和地面都是光滑的；质点一直与斜面保持接触。取斜面向右下方倾斜，于是质点向右下方运动，而楔块被推向左方，如图所示。

取水平分量向右为正、垂直分量向上为正，

$$\text{质点的速度} = (v\cos\beta,\ -v\sin\beta), \qquad \text{楔块的速度} = (-V,\ 0).$$



第 (i) 部分

质点相对于楔块的速度是

$$(v\cos\beta,\ -v\sin\beta)-(-V,\ 0)=(v\cos\beta+V,\ -v\sin\beta).$$

质点在斜面上时，这个相对速度（relative velocity）沿着斜面，方向在水平线以下 α 角（指向右下方）。所以

$$\frac{v\sin\beta}{v\cos\beta+V}=\tan\alpha=\frac{\sin\alpha}{\cos\alpha}.$$

交叉相乘，

$$\begin{aligned} & v\sin\beta\cos\alpha=v\cos\beta\sin\alpha+V\sin\alpha \\ \Longrightarrow\quad & V\sin\alpha=v(\sin\beta\cos\alpha-\cos\beta\sin\alpha)=v\sin(\beta-\alpha), \end{aligned}$$

这正是所要证明的结果。这纯粹是运动学（kinematics）：不需要任何力学原理，只需要“质点不离开斜面”这个事实。

第 (ii) 部分

原理：水平方向的线动量守恒（conservation of linear momentum）。把质点和楔块视为一个系统。作用在系统上的外力是两者的重量和地面的法向反作用力（normal reaction），由于地面光滑，这些力全都是垂直的。（质点与斜面之间的反作用力是内力。）所以水平总动量是常数，而由于一切都从静止开始，它起初是零：

$$mv\cos\beta-MV=0,\qquad \text{所以}\qquad V=\frac{m}{M}v\cos\beta.$$

代入 (i)，并展开 $\sin(\beta-\alpha)$，

$$\frac{m}{M}v\cos\beta\sin\alpha=v(\sin\beta\cos\alpha-\cos\beta\sin\alpha).$$

当 $t>0$ 时质点正在运动，所以 $v\neq 0$。另外 $\cos\alpha\neq 0$，而且 $\cos\beta\neq 0$：假如 β 是 90°，由动量可得 $V=0$，于是 (i) 给出 $\sin(\beta-\alpha)=0$，所以 $\beta=\alpha<90^\circ$，矛盾。两边除以 $v\cos\alpha\cos\beta$，

$$\frac{m}{M}\tan\alpha=\tan\beta-\tan\alpha\quad\Longrightarrow\quad\tan\beta=\left(1+\frac{m}{M}\right)\tan\alpha.$$

> ? **想一想** 为甚么 β 不是直接等于 α?

因为斜面本身正在向左退开。相对于楔块，质点以 α 角沿斜面下滑，但楔块在后退，所以相对于地面，质点下降得更陡。留意 (ii) 的右边是一个常数：β 在每一刻都相同，所以质点沿直线运动。如果 $M\to\infty$，楔块就留在原位，而 $\beta\to\alpha$，正如所料。

第 (iii) 部分

原理：能量守恒（conservation of energy）。所有接触面都是光滑的，所以对整个系统做功的力就只有重量。（质点与斜面之间的法向反作用力会分别对两个物体做功，但这两份功互相抵消，因为相对速度沿着斜面，所以与反作用力垂直。地面的反作用力则与楔块的运动方向垂直。）楔块的质心（centre of mass）在垂直方向没有移动，所以势能（potential energy）的损失是 mgy。由静止开始，

$$mgy = \tfrac{1}{2}mv^2 + \tfrac{1}{2}MV^2.$$

利用 (ii) 中的 $V = \dfrac{m}{M}v\cos\beta$，

$$mgy = \tfrac{1}{2}mv^2 + \frac{m^2v^2\cos^2\beta}{2M}.$$

两边乘以 $2/m$，

$$2gy = v^2\left(1 + \frac{m\cos^2\beta}{M}\right) = \frac{v^2\left(M + m\cos^2\beta\right)}{M},$$

这正是所要证明的结果。

关于 y 的微分方程

质点下降的速率就是它速度的向下分量：

$$\frac{\mathrm{d}y}{\mathrm{d}t} = v\sin\beta.$$

把 $v = \dot{y}/\sin\beta$ 代入 (iii)（注意由于 $\beta > \alpha > 0$，所以 $\sin\beta \neq 0$）。这就得到微分方程（differential equation）

$$\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2 = k\,y, \qquad \text{其中} \qquad k = \frac{2gM\sin^2\beta}{M + m\cos^2\beta}.$$

关键在于 k 是一个常数，因为由 (ii) 可知 β 是常数。

要解这个方程，先对 t 求导：

$$2\dot{y}\,\ddot{y} = k\,\dot{y}.$$

当 $t > 0$ 时质点正在下降，所以 $\dot{y} > 0$，我们可以除以它：$\ddot{y} = \tfrac{1}{2}k$。质点的垂直运动有恒定的加速度

$$\ddot{y} = \frac{gM\sin^2\beta}{M + m\cos^2\beta}.$$

积分两次，并利用 $t = 0$ 时 $y = 0$ 及 $\dot{y} = 0$（它由静止开始），

$$y = \frac{gMt^2\sin^2\beta}{2\left(M + m\cos^2\beta\right)},$$

这就是所求的结果。

（另一个做法是取正平方根 $\dot{y} = \sqrt{ky}$，然后分离变量（separation of variables）：$\int y^{-1/2}\,\mathrm{d}y = \sqrt{k}\,t$ 给出 $2\sqrt{y} = \sqrt{k}\,t$，得到相同的答案。先求导的好处，是不必顾虑 $y \equiv 0$ 这个解；它同样满足 $\dot{y}^2 = ky$ 和 $y(0) = 0$，却不是这里的运动。）

验算

- 固定的楔块。令 $M \to \infty$。这时 $\beta \to \alpha$，而 $y \to \frac{1}{2}gt^2\sin^2\alpha$，这是正确的：在固定的光滑斜面上，沿斜面的加速度是 $g\sin\alpha$，而它的垂直分量是 $g\sin^2\alpha$。
- 一个数值例子。取 $M = 2$、$m = 1$、$\alpha = 30°$，由 (ii) 得 $\tan\beta = \frac{3}{2}\cdot\frac{1}{\sqrt{3}} = \frac{\sqrt{3}}{2}$，所以 $\sin^2\beta = \frac{3}{7}$、$\cos^2\beta = \frac{4}{7}$，而公式给出 $y = \frac{1}{6}gt^2$。取 $g = 9.8$ 和 $t = 1$，即 $y \approx 1.6333$。直接对楔块和质点作逐步模拟（用牛顿定律（Newton's laws），并由接触条件求出法向反作用力），得到相同的数值；而且在 2000 个随机例子中，(i)–(iii) 都在舍入误差范围内成立。

常见错误

- 说质点的运动方向与水平线成 α 角。这只是相对于楔块而言才成立。
- 沿斜面方向或垂直方向使用动量守恒。地面的反作用力是垂直方向的外力，所以只有水平分量守恒。
- 在能量方程中漏掉楔块的动能（kinetic energy）$\frac{1}{2}MV^2$。

STEP 之外

被约束在运动曲面上的质点，是拉格朗日力学（Lagrangian mechanics）的经典试金石，你会在大学一年级或二年级学到这门学问。在那里，你选取两个坐标（例如楔块的位置和质点沿斜面下滑的距离），约束力（constraint forces）就会从方程中完全消失，所以只需几行就能得出那些恒定的加速度。关于带约束的牛顿力学，以及许多正是这一类的题目，可参阅 Kleppner and Kolenkow, *An Introduction to Mechanics*（一本力学入门教科书）。

详细解答

STEP II 1998，第 11 题

力学 • 抛体运动

STEP II 1998，第 11 题

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

step.maths.org/questions/98-s2-q11

本题在 *Beyond STEP* 第 11.11 节外野手如何接球中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch11/fielders.html

关键思路

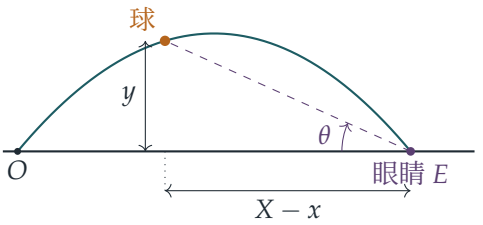
把原点设在击球点，并取其高度为眼睛的高度，然后把 $\tan\theta$ 写成“球的高度除以球与外野手眼睛之间的水平距离”。一个由眼睛高度出发、最后又回到眼睛高度的抛体（projectile），其高度可以因式分解为 $\frac{1}{2}gt(T-t)$，其中 T 是飞行时间（time of flight）；而如果外野手最后身处正确的位置，水平距离同样是 $T-t$ 的倍数。因子 $T-t$ 互相约去，余下的 $\tan\theta$ 便与 t 成正比。

解答

我们把球视为只受重力作用（没有空气阻力（air resistance））的质点（particle），并把外野手的眼睛视为一个与击球点等高的点。取 x 轴为水平方向，由击球点 O 指向外野手，y 轴铅垂向上。设球的初速度的水平分量为 $U>0$，铅垂分量为 $W>0$。那么在时刻 t

$$x=Ut, \qquad y=Wt-\tfrac{1}{2}gt^2=\tfrac{1}{2}gt\,(T-t), \qquad \text{其中} T=\frac{2W}{g}$$

是球回到眼睛高度的时刻。



如果外野手的眼睛位于水平位置 X（球在飞行期间有 $X > x$），视线（line of sight）便与水平线成仰角 θ，其中

$$\tan\theta = \frac{y}{X-x}.$$

站立不动的外野手

“位置恰到好处”的意思是球刚好飞到她的眼睛：她的眼睛位于球回到眼睛高度的那一点，即 $X = UT$。于是当 $0 < t < T$ 时

$$\tan\theta = \frac{\frac{1}{2}gt(T-t)}{UT-Ut} = \frac{\frac{1}{2}gt(T-t)}{U(T-t)} = \frac{g}{2U}t,$$

这里约去了 $T-t$，因为球在飞行期间它不等于零。因此

$$\frac{\mathrm{d}}{\mathrm{d}t}(\tan\theta) = \frac{g}{2U},$$

这是一个常数，正是所求。（在接球前的一刻，$\tan\theta \to gT/(2U) = W/U$：她看见球飞来的角度，恰好就是球被击出时的角度，这与抛物线的对称性吻合。）

奔跑中的外野手

现在设下一球的速度分量为 U、W（所以同样有 $y = \frac{1}{2}gt(T-t)$，其中 $T = 2W/g$），并设外野手在球被击出的一刻与 O 相距 D，然后以恒速 v 跑向击球手。她的眼睛位于 $X = D - vt$，所以

$$\tan\theta = \frac{\frac{1}{2}gt(T-t)}{D-(U+v)t}. \qquad (*)$$

她能接到球，当且仅当在时刻 T 她的眼睛正好位于球回到眼睛高度的那一点，即

$$D - vT = UT \quad\Longleftrightarrow\quad D = (U+v)T.$$

我们要证明，这种情况发生当且仅当 $\frac{\mathrm{d}}{\mathrm{d}t}(\tan\theta)$ 为常数，因此“选择 v 使 $\frac{\mathrm{d}}{\mathrm{d}t}(\tan\theta)$ 保持不变”确是正确的法则。

如果她选对了速率，则 $D = (U+v)T$，而 $(*)$ 变成

$$\tan\theta = \frac{\frac{1}{2}gt(T-t)}{(U+v)(T-t)} = \frac{g}{2(U+v)}t, \qquad \text{所以} \qquad \frac{\mathrm{d}}{\mathrm{d}t}(\tan\theta) = \frac{g}{2(U+v)},$$

这是一个常数，与第一部分完全一样，只是 U 换成了球相对于她的速率 $U+v$。

想一想 这不是只完成了一半吗？会不会有另一个 v 同样令变化率保持不变？

不会。假设在整个飞行过程中 $\frac{\mathrm{d}}{\mathrm{d}t}(\tan\theta)=k$ 为常数。在 $t=0$ 时球位于眼睛高度，所以 $\tan\theta=0$，因而 $\tan\theta=kt$。把这个结果代入 $(*)$ 并除以 t（其中 $0<t<T$），得

$$\tfrac{1}{2}g\,(T-t)=k\big(D-(U+v)t\big)\quad \text{对所有}t\in(0,T).$$

两个关于 t 的线性函数如果在一个区间上相等，它们的系数必定相同，所以

$$\tfrac{1}{2}g=k(U+v)\qquad \text{及}\qquad \tfrac{1}{2}gT=kD.$$

由第一式可知 $k\neq 0$；把第二式除以第一式便得 $D=(U+v)T$，这正是接到球的条件。

所以，能把她带到球那里的速率 $v=D/T-U$，恰好就是令 $\frac{\mathrm{d}}{\mathrm{d}t}(\tan\theta)$ 保持不变的那个速率。实际上，球还在空中时她无法计算 $D/T-U$，但她可以观察：如果球在她眼中上升得愈来愈快，她便站得太后，应该跑快一点；如果球看来上升得愈来愈慢，她便应该放慢脚步。（由于她是朝着击球手跑的，球的落点比她的起始位置近，即 $D>UT$，所以 $v>0$。）

验算

- 在第二部分取 $v=0$，便回到第一部分的结果：$D=UT$，变化率为 $g/(2U)$。
- 单位：g/U 的单位是 $\mathrm{m\,s^{-2}}/\mathrm{m\,s^{-1}}=\mathrm{s^{-1}}$，正好是一个纯数的变化率应有的单位。
- 一个数值例子：球以 $20\ \mathrm{m\,s^{-1}}$ 的速率、45° 的仰角被击出，取 $g=9.8\ \mathrm{m\,s^{-2}}$，则 $T\approx 2.886\ \mathrm{s}$，射程为 $40.82\ \mathrm{m}$；站在该处的外野手会看见 $\tan\theta$ 以 $g/(2U)\approx 0.346\ \mathrm{s^{-1}}$ 的速率增长。如果她改为从 $49.47\ \mathrm{m}$ 外起步，并以 $3\ \mathrm{m\,s^{-1}}$ 的速率奔跑，变化率则为 $0.286\ \mathrm{s^{-1}}$。在多个时刻对 $(*)$ 作数值求导，可确认这两个变化率都是常数，而选错了 v 则会得到一个不断改变的变化率。

常见错误

- 从击球手而不是从外野手的眼睛量度 θ：$\tan\theta$ 中的水平距离是 $X-x$，而不是 x。
- 只证明了正确的 v 会给出恒定的变化率，却没有证明恒定的变化率必然导致正确的 v；正是这个逆命题（converse）才令这条法则真正可用。

STEP 之外

这就是接球心理学中的“光学加速度抵消”（optical acceleration cancellation）概念：外野手似乎是这样跑的——让球的影像以稳定的速率上升，而从不需要预测球会落在哪里。这是反馈控制律（feedback control law）的一个例子：把一个观察到的量保持不变，而不是去计算轨迹。关于抛体运动、相对速度（relative velocity）以及这种思考方式的入门，可参阅 Kleppner and Kolenkow, *An Introduction to Mechanics*（力学入门教材）。

详细解答

STEP II 1998，第 12 题

概率与统计 · 贝叶斯定理

> **STEP II 1998，第 12 题**
>
> 继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：
>
> step.maths.org/questions/98-s2-q12

本题在 *Beyond STEP* 第 12.2 节基本比率谬误中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch12/base-rate.html

关键思路

这一题是贝叶斯定理（Bayes' theorem）的两次应用。第一部分是经典的“罕见疾病”意外：当疾病很罕见时，即使检测相当准确，大部分阳性结果仍可能是假阳性（false positive）。至于第二部分，窍门在于看出接受 STEP 检测的人并不是整体人口的随机样本。他们是 AL 阳性的那一群人，所以第一部分的答案就成为第二次检测的先验概率（prior）。

解答

设 D 为某人患有 mathematitis 的事件，A 为 AL 检测呈阳性的事件，B 为 STEP 检测呈阳性的事件。已知

$$\mathrm{P}(D) = \tfrac{1}{100}, \qquad \mathrm{P}(A \mid D) = \tfrac{9}{10}, \qquad \mathrm{P}(A \mid D') = \tfrac{1}{11}.$$

Frankie 的概率

根据全概率公式（law of total probability），

$$\begin{aligned}\mathrm{P}(A) &= \mathrm{P}(A \mid D)\,\mathrm{P}(D) + \mathrm{P}(A \mid D')\,\mathrm{P}(D') \\ &= \frac{9}{10}\cdot\frac{1}{100} + \frac{1}{11}\cdot\frac{99}{100} = \frac{9}{1000} + \frac{9}{100} = \frac{99}{1000}.\end{aligned}$$

再根据贝叶斯定理，

$$\mathrm{P}(D \mid A) = \frac{\mathrm{P}(A \mid D)\,\mathrm{P}(D)}{\mathrm{P}(A)} = \frac{9/1000}{99/1000} = \frac{1}{11}.$$

所以即使 AL 检测结果呈阳性，Frankie 是患者的概率也只有 $\frac{1}{11}$。在 1000 人当中，约有 9 名患者检测呈阳性，但 990 名非患者中也约有 90 人呈阳性，假阳性的数目远远盖过了真阳性。

求 p

现在只考虑 AL 检测呈阳性的人。在这一群人之中，

$$\mathrm{P}(D \mid A) = \tfrac{1}{11}, \qquad \mathrm{P}(D' \mid A) = \tfrac{10}{11}.$$

我们把题目所给有关 STEP 的资料理解为 $\mathrm{P}(B \mid D \cap A) = 0.9$ 和 $\mathrm{P}(B \mid D' \cap A) = p$：STEP 检测的表现只取决于该人是否患者，而与 AL 的结果无关。

> **想一想** 为甚么我们不从 $\mathrm{P}(D) = \frac{1}{100}$ 重新开始?

因为只有 AL 呈阳性的人才会接受 STEP 检测，而在这一群人之中，患者所占的比例已不再是 1%，而是 $\frac{1}{11}$，这正是第一部分求得的结果。第一次检测的后验概率（posterior）成为第二次检测的先验概率。

在 AL 阳性的一群人之中，

$$\mathrm{P}(D \cap B \mid A) = \frac{1}{11} \cdot \frac{9}{10} = \frac{9}{110},$$
$$\mathrm{P}(D' \cap B \mid A) = \frac{10}{11} \cdot p = \frac{10p}{11}.$$

题目告诉我们，两项检测皆呈阳性的人当中有一半是患者，即 $\mathrm{P}(D \mid A \cap B) = \frac{1}{2}$。这表示上面两个概率相等：

$$\frac{9}{110} = \frac{10p}{11} \quad \Longrightarrow \quad p = \frac{9}{100} = 0.09.$$

验算

- 赔率形式（odds form）：患病的先验赔率为 $1 : 99$。AL 把它乘以 $\frac{0.9}{1/11} = 9.9$，得 $9.9 : 99 = 1 : 10$，与 $\frac{1}{11}$ 吻合。要使赔率达到 $1 : 1$，STEP 必须把它乘以 10，所以 $\frac{0.9}{p} = 10$，即 $p = 0.09$。
- 模拟二百万人先后接受两项检测（取 $p = 0.09$），结果 AL 阳性者当中患者所占的比例为 0.0909，而两项检测皆呈阳性者当中患者所占的比例为 0.5003。

常见错误

- 以 $\mathrm{P}(D) = 0.01$ 作为 STEP 检测的先验概率，这样会得出 $p = 0.0009$，比正确答案小了一百倍。
- 把 $\mathrm{P}(A \mid D) = 0.9$ 与 $\mathrm{P}(D \mid A)$ 混淆。第一部分的重点正正在于这两者相差甚远。

STEP 之外

把贝叶斯定理写成赔率形式，即后验赔率等于先验赔率乘以似然比（likelihood ratio），连续检测的计算就变得近乎机械化：每一项独立的证据只不过是把赔率乘以它自己的因子。这是贝叶斯推论（Bayesian inference）的基础，也是现实中设计医学筛查计划的依据。我们所作的假设，即在给定患病状况下各项检测条件独立（conditionally independent），与“朴素贝叶斯”（naive Bayes）分类器背后的假设相同。Blitzstein 与 Hwang 的 *Introduction to Probability*（概率论入门教科书）对条件概率和贝叶斯法则有极佳的讲解，并附有许多与本题类似的例子。

详细解答

STEP II 1998，第 13 题

概率与统计 · 指数分布

STEP II 1998，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/98-s2-q13

关键思路

第一个结果是指数分布（exponential distribution）的无记忆性（memoryless property）：如果一次服务已经持续了时间 t，余下所需的时间与一次全新服务的时间有相同的分布。有了这个结果，店铺这个情境几乎不用再作任何计算。就概率而言，服务到一半的顾客与刚刚开始接受服务的顾客并无分别。因此，在每一个关键时刻，仍在接受服务的人之间的竞赛，都是相同而且独立（independent）的指数时间之间的竞赛。

解答

无记忆性

对于 $x \geq 0$，

$$\mathrm{P}(X > x) = \int_x^\infty \lambda \mathrm{e}^{-\lambda u}\,\mathrm{d}u = \Big[-\mathrm{e}^{-\lambda u}\Big]_x^\infty = \mathrm{e}^{-\lambda x}.$$

取 $s, t \geq 0$。事件 $\{X > s+t\}$ 包含于 $\{X > t\}$ 之内，所以两者的交集就是 $\{X > s+t\}$。根据条件概率（conditional probability）的定义，

$$\begin{aligned}\mathrm{P}(X > s+t \mid X > t) &= \frac{\mathrm{P}(X > s+t \text{ 及 } X > t)}{\mathrm{P}(X > t)} = \frac{\mathrm{P}(X > s+t)}{\mathrm{P}(X > t)} \\ &= \frac{\mathrm{e}^{-\lambda(s+t)}}{\mathrm{e}^{-\lambda t}} = \mathrm{e}^{-\lambda s} = \mathrm{P}(X > s),\end{aligned}$$

这就是所要求的结果。（我们可以作除法，是因为 $\mathrm{P}(X > t) = \mathrm{e}^{-\lambda t} > 0$。）

两位顾客都仍在接受服务

模型：两个服务时间互相独立，而且各自服从题目所给的指数分布。设 A 和 B 为我到达时正在接受服务的两位顾客。

> **想一想** 顾客 A 已经接受了一段长度未知的服务。这难道不会改变他还需要多少时间的分布吗?

不会，而这正是第一部分告诉我们的。假设 A 的服务至今已持续了 u。那么它最少再持续 t 的概率是

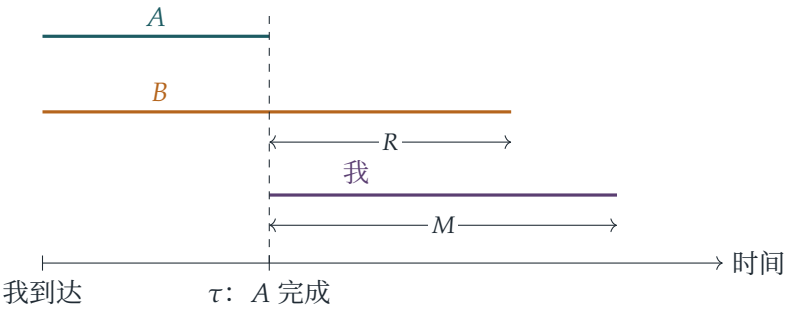
$$\mathrm{P}(X > u + t \mid X > u) = \mathrm{P}(X > t) = \mathrm{e}^{-\lambda t},$$

不论 u 是多少。所以，由我走进店舖的一刻起计，A 的剩余服务时间服从同一个参数（parameter）为 λ 的指数分布，B 的情况亦然。由于完整的服务时间互相独立，这两个剩余时间也互相独立。因此

$$\mathrm{P}(\text{两人均仍在接受服务，于时刻}t) = \mathrm{e}^{-\lambda t} \cdot \mathrm{e}^{-\lambda t} = \mathrm{e}^{-2\lambda t}.$$

另一位顾客完成时，我是否仍在接受服务?

假设 A 先完成，时刻为我到达后的 τ，而我随即到 A 的店员那里接受服务。



在时刻 τ，设 R 为 B 仍需要的时间，并设 M 为我的服务时间。

- M 服从参数为 λ 的指数分布，因为我的服务在 τ 才重新开始。它与之前发生的一切独立，因为不同顾客的服务时间互相独立。
- R 同样服从参数为 λ 的指数分布。在时刻 τ，B 已经接受了一段时间的服务（我到达之前的时间再加上 τ），而再次根据无记忆性，剩余时间服从原来的分布。它与 M 独立。

B 完成时我仍在接受服务，当且仅当 $M > R$。现在 M 和 R 互相独立，并且有相同的连续分布 (continuous distribution)，所以根据对称性 (symmetry)，$\mathrm{P}(M > R) = \mathrm{P}(R > M)$。

此外，对于连续随机变量（continuous random variables），$\mathrm{P}(M = R) = 0$，所以这两个概率之和为 1。因此

$$\mathrm{P}(\text{另一位顾客完成时我仍在接受服务}) = \tfrac{1}{2}.$$

如果你宁愿用计算而不用对称性的论证，可以以 R 为条件：

$$\mathrm{P}(M > R) = \int_0^\infty \lambda \mathrm{e}^{-\lambda r}\,\mathrm{P}(M > r)\,\mathrm{d}r = \int_0^\infty \lambda \mathrm{e}^{-2\lambda r}\,\mathrm{d}r = \tfrac{1}{2}.$$

验算

- 当 $t = 0$ 时，$\mathrm{e}^{-2\lambda t} = 1$：我到达时，两位顾客必定都正在接受服务。当 $t \to \infty$ 时，概率趋向 0，这也是应有的结果。
- 答案 $\frac{1}{2}$ 与 λ 无关，这正如我们所料，因为改变 λ 只是改变了时间的单位。
- 以 $\lambda = 0.7$ 和 $t = 0.9$ 进行模拟，得出两位顾客都仍在接受服务的比例为 0.2831，而 $\mathrm{e}^{-1.26} \approx 0.2837$。在另一个模拟中，两位顾客已经各自接受了随机长度的服务，结果我在 0.5000 比例的试验中仍在接受服务。

常见错误

- 以为正在接受服务的顾客“快要完成”，所以他们的剩余时间会较短。对指数分布而言，已经发生的事情不会造成任何分别。
- 把 $\mathrm{P}(X > s + t \text{ 及 } X > t)$ 写成 $\mathrm{P}(X > s + t)\,\mathrm{P}(X > t)$。这两个事件并不独立；其中一个包含另一个。

STEP 之外

指数分布是唯一具有无记忆性的连续分布。这个事实是*泊松过程*（Poisson process）和*排队论*（queueing theory）的基础；在排队论中，一间有 c 位店员、服务时间服从指数分布的店舖，就是经典的 M/M/c 队列。两位顾客的计算可以推广至 n 位：n 个参数为 λ 的独立指数变量之中的最小值，服从参数为 $n\lambda$ 的指数分布（把尾概率相乘即可），而根据对称性，每个变量成为最小者的可能性均等。Grimmett 与 Welsh 所著的 *Probability: An Introduction*（概率入门）以这个程度讲解指数变量和泊松过程。

详细解答

STEP II 1998，第 14 题

概率与统计 · 二项分布与泊松分布的期望值

STEP II 1998，第 14 题

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

step.maths.org/questions/98-s2-q14

关键思路

虽然资料库把这题另作归类，但它其实是一道概率（probability）题：只要 Doom 博士在新方案下的期望薪金高于固定薪金 A，她就应该接受新方案。不合格人数服从二项分布（binomial distribution），所以我们需要求 $\mathrm{E}\left[1/(1+X)\right]$。棘手的因子 $\frac{1}{k+1}$ 可以用恒等式 $\frac{1}{k+1}\binom{N}{k} = \frac{1}{N+1}\binom{N+1}{k+1}$ 吸收掉，于是该和式就变成 $(p+q)^{N+1}$ 几乎完整的二项展开式。同样的技巧配合 $\frac{\lambda^k}{(k+1)!}$，便可处理泊松（Poisson）模型。

解答

何时接受新方案

设 X 为 Doom 博士的学生中不合格的人数，并记 $q = 1-p$。各学生是否不合格互相独立（independent），每人不合格的概率均为 p，所以 $X \sim \mathrm{B}(N,p)$：

$$\mathrm{P}(X=k) = \binom{N}{k} p^k q^{N-k}, \qquad k = 0,1,\ldots,N.$$

我们假设 Doom 博士愿意以期望薪金来衡量两个方案。在新方案下，她的薪金是随机变量（random variable）$S = B/(1+X)$，所以

$$\mathrm{E}(S) = B\sum_{k=0}^{N} \frac{1}{k+1}\binom{N}{k} p^k q^{N-k}.$$

想一想 怎样消去和式中的 $\frac{1}{k+1}$？

利用阶乘（factorial）：

$$\frac{1}{k+1}\binom{N}{k} = \frac{N!}{(k+1)!\,(N-k)!} = \frac{1}{N+1}\cdot\frac{(N+1)!}{(k+1)!\,\big((N+1)-(k+1)\big)!}$$
$$= \frac{1}{N+1}\binom{N+1}{k+1}.$$

把这个结果代入，并同时乘以和除以 p，使 p 的幂次与新的二项式系数（binomial coefficient）相配，

$$\begin{aligned}\mathrm{E}(S) &= \frac{B}{(N+1)p}\sum_{k=0}^{N}\binom{N+1}{k+1}p^{k+1}q^{N-k}\\ &= \frac{B}{(N+1)p}\sum_{j=1}^{N+1}\binom{N+1}{j}p^{j}q^{N+1-j}\\ &= \frac{B}{(N+1)p}\Big[(p+q)^{N+1}-q^{N+1}\Big]\end{aligned}$$

其中先设 $j = k+1$, 再运用二项式定理 (binomial theorem), 因为该和式正是 $(p+q)^{N+1}$ 的完整展开式，只欠 $j=0$ 的一项 q^{N+1}。由于 $p+q=1$，

$$\mathrm{E}(S) = \frac{B\big(1-(1-p)^{N+1}\big)}{(N+1)p}.$$

若 $\mathrm{E}(S) > A$，她就应该接受。两边同乘 $(N+1)p$，而此数为正（公式要有意义，须有 $p>0$；若 $p=0$，她只需直接比较 B 与 A），便得

$$A(N+1)p < B\big(1-(1-p)^{N+1}\big),$$

这就是所求的条件。

泊松模型

当 N 很大而 p 很小，且 Np 大小适中时，二项分布 $\mathrm{B}(N,p)$ 可以用平均值为 $\lambda = Np$ 的泊松分布（Poisson distribution）作良好的近似。因此，如果 Doom 博士任教的班人数众多，每名学生不合格的机会都很低，而且各人是否不合格仍然互相独立，她的学生的不合格人数就可以用泊松分布作模型。（更一般地说，若不合格的情况逐一出现、互相独立，且平均发生率恒定，便适合采用泊松模型。）

设 $X \sim \mathrm{Po}(\lambda)$，$\lambda = Np$，同样的想法依然可行，这次用的是 $\frac{1}{k+1}\cdot\frac{1}{k!} = \frac{1}{(k+1)!}$：

$$\mathrm{E}\Big(\frac{B}{1+X}\Big) = B\sum_{k=0}^{\infty}\frac{1}{k+1}\cdot\frac{\mathrm{e}^{-\lambda}\lambda^{k}}{k!} = \frac{B\mathrm{e}^{-\lambda}}{\lambda}\sum_{k=0}^{\infty}\frac{\lambda^{k+1}}{(k+1)!}$$
$$= \frac{B\mathrm{e}^{-\lambda}}{\lambda}\big(\mathrm{e}^{\lambda}-1\big) = \frac{B\big(1-\mathrm{e}^{-\lambda}\big)}{\lambda},$$

因为最后的和式正是指数级数（exponential series），只欠首项 1。所以在这个模型下，Doom 博士的期望薪金为

$$\frac{B\left(1-\mathrm{e}^{-Np}\right)}{Np} \text{ 英镑.}$$

验算

- $N=1$：直接计算得 $\mathrm{E}(S)=B\left(q+\frac{1}{2}p\right)=B\left(1-\frac{1}{2}p\right)$，而公式给出 $B\,\frac{1-q^2}{2p}=B\,\frac{1+q}{2}=B\left(1-\frac{1}{2}p\right)$。
- $p\to 0$：二项分布的表达式趋向 B（没有人不合格），而当 $\lambda\to 0$ 时，泊松分布的表达式亦然。
- 在近似理应准确的情况下，两个答案相当吻合：取 $N=1000$，$p=0.003$，由二项分布公式得 $\mathrm{E}(S)/B$ 为 0.31655，由泊松公式则得 0.31674。取 $N=20$，$p=0.3$ 时，逐项精确计算与模拟的结果均为 $\mathrm{E}(S)/B\approx 0.1586$。

常见错误

- 写成 $\mathrm{E}\left(\frac{1}{1+X}\right)=\frac{1}{1+\mathrm{E}(X)}=\frac{1}{1+Np}$。期望值（expectation）不能直接穿过非线性函数；事实上 $\frac{1}{1+x}$ 是凸函数（convex function），所以根据延森不等式（Jensen's inequality），真正的值比这个大。
- 忘记了重新编排指标后的和式欠了 $j=0$ 的一项，结果得出 $\mathrm{E}(S)=B/((N+1)p)$。

STEP 之外

这里的技巧其实是一个特例：计算 $\mathrm{E}\left(1/(1+X)\right)$ 时可以利用 $\frac{1}{1+x}=\int_0^1 t^x\,\mathrm{d}t$，从而把它与概率生成函数（probability generating function）$G(t)=\mathrm{E}(t^X)$ 联系起来：答案就是 $\int_0^1 G(t)\,\mathrm{d}t$。试以 $G(t)=(q+pt)^N$ 和 $G(t)=\mathrm{e}^{\lambda(t-1)}$ 验证一下。关于生成函数，以及二项分布收敛（convergence）至泊松分布，Grimmett 与 Welsh 所著的 *Probability: An Introduction*（概率论入门教材）有很好的讲解。

详细解答

STEP III 1998，第 1 题

纯数学 · 曲线描绘

> **STEP III 1998，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/98-s3-q1

关键思路

把所有式子都用 $\cos x$ 表示：利用 $\sin^2 x = 1 - \cos^2 x$，函数 f 便成为 $\cos x$ 的二次式。配方（completing the square）之后，f 的值域（range）一目了然，而最后一部分正是由这个值域决定。至于 g，用商法则（quotient rule）求得的导数（derivative）是 $\mathrm{f}'(x)$ 的常数倍再除以一个平方。

解答

描绘 $y = \mathrm{f}(x)$ 的图像

利用 $\sin^2 x = 1 - \cos^2 x$，

$$\mathrm{f}(x) = 2 + 2\cos x - \cos^2 x = 3 - (1 - \cos x)^2.$$

由于 $0 \le 1 - \cos x \le 2$，可得 $-1 \le \mathrm{f}(x) \le 3$。当 $\cos x = 1$ 时取得值 3，当 $\cos x = -1$ 时取得值 -1。

求导得

$$\mathrm{f}'(x) = -2(1 - \cos x)\sin x = 2\sin x\,(\cos x - 1).$$

当 $\sin x = 0$ 或 $\cos x = 1$ 时，上式为零；在 $[0, 2\pi]$ 上，即 $x = 0,\ \pi,\ 2\pi$。驻点（stationary point）为

$$(0,\ 3), \qquad (\pi,\ -1), \qquad (2\pi,\ 3).$$

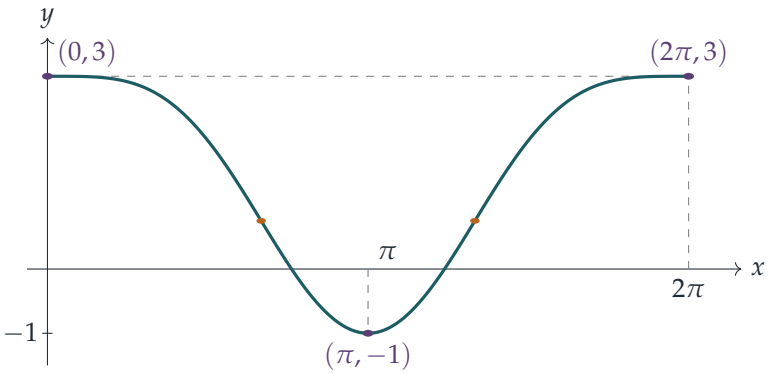
在 $(0, \pi)$ 上，$\sin x > 0$ 而 $\cos x - 1 < 0$，所以 $\mathrm{f}' < 0$，f 递减。在 $(\pi, 2\pi)$ 上，$\sin x < 0$ 而 $\cos x - 1$ 仍然是负数，所以 $\mathrm{f}' > 0$，f 递增。因此 $(\pi, -1)$ 是极小点（minimum），而端点 $(0, 3)$ 和 $(2\pi, 3)$ 是极大点（maximum）。

想一想 曲线在 $x = 0$ 附近是甚么样子？像抛物线吗？

不像，它比抛物线更平坦。在 $x = 0$ 附近，$1 - \cos x \approx \frac{1}{2}x^2$，所以 $\mathrm{f}(x) \approx 3 - \frac{1}{4}x^4$。曲线在顶部非常平坦，这与 $\mathrm{f}''(0) = 0$ 吻合。在 $x = \pi$ 附近，设 $x = \pi + t$，则 $1 - \cos x = 1 + \cos t \approx 2 - \frac{1}{2}t^2$，所以 $\mathrm{f} \approx -1 + 2t^2$：这是一个普通的抛物线形极小点。

描绘图像时还有以下几个有用的特征：

- 曲线关于 $x = \pi$ 对称，因为 $\cos(2\pi - x) = \cos x$。
- 曲线与 x 轴相交于 $(1 - \cos x)^2 = 3$ 之处，即 $\cos x = 1 - \sqrt{3}$。由此得 $x = \arccos(1 - \sqrt{3}) \approx 2.39$ 和 $x = 2\pi - \arccos(1 - \sqrt{3}) \approx 3.89$。
- $\mathrm{f}''(x) = 2\cos 2x - 2\cos x = 2(2\cos x + 1)(\cos x - 1)$。它在 $\cos x = -\frac{1}{2}$ 处变号，所以拐点（point of inflection）位于 $\left(\frac{2\pi}{3}, \frac{3}{4}\right)$ 和 $\left(\frac{4\pi}{3}, \frac{3}{4}\right)$。



琥珀色的点是拐点。

g 的驻点

根据商法则，凡是 $c\,\mathrm{f}(x) + d \neq 0$ 之处，

$$\mathrm{g}'(x) = \frac{a\mathrm{f}'(x)\big(c\mathrm{f}(x) + d\big) - c\mathrm{f}'(x)\big(a\mathrm{f}(x) + b\big)}{\big(c\mathrm{f}(x) + d\big)^2} = \frac{(ad - bc)\,\mathrm{f}'(x)}{\big(c\mathrm{f}(x) + d\big)^2}.$$

由于 $ad \neq bc$，在 g 有定义的每一点上，$\mathrm{g}'(x) = 0$ 当且仅当 $\mathrm{f}'(x) = 0$。所以除了 f 的驻点之外，g 没有其他驻点。

我们还需要 g 在 $x = 0, \pi, 2\pi$ 处有定义，另外两个条件正是为此而设。在 $x = 0$ 和 $x = 2\pi$ 处，分母是 $3c + d$，因为 $d \neq -3c$，所以它不等于零。在 $x = \pi$ 处，分母是 $d - c$，因为 $d \neq c$，所以它不等于零。因此 $y = \mathrm{g}(x)$ 的驻点恰好位于 $x = 0,\ \pi,\ 2\pi$，正如题目所求，

并且

$$\mathrm{g}(0)=\mathrm{g}(2\pi)=\frac{3a+b}{3c+d},\qquad \mathrm{g}(\pi)=\frac{b-a}{d-c}.$$

$|\mathrm{g}|$ 何时有界

假设 $d/c<-3$ 或 $d/c>1$。特别地，$c\neq 0$，所以可以写成

$$c\,\mathrm{f}(x)+d=c\Big(\mathrm{f}(x)+\frac{d}{c}\Big).$$

由于 $\mathrm{f}(x)$ 在 $[-1,3]$ 内取值，括号内的式子落在 $\left[\frac{d}{c}-1,\ \frac{d}{c}+3\right]$ 之内。

- 若 $d/c>1$，括号内的式子至少是 $\frac{d}{c}-1>0$。
- 若 $d/c<-3$，括号内的式子至多是 $\frac{d}{c}+3<0$。

无论哪种情况，都存在常数 $k>0$，使得对所有 x 都有 $\left|c\,\mathrm{f}(x)+d\right|\geq|c|\,k$。换句话说，分母始终与零保持一段距离。分子是有界（bounded）的，因为 $|a\mathrm{f}(x)+b|\leq 3|a|+|b|$。因此对 $[0,2\pi]$ 内的每个 x，都有

$$|\mathrm{g}(x)|\leq\frac{3|a|+|b|}{|c|\,k}$$

所以 $|\mathrm{g}(x)|$ 不可能任意大。

> ? **想一想** 如果 $-3<d/c<1$，会出甚么问题?

这时 $-d/c$ 严格介乎 -1 与 3 之间。由于 f 连续（continuous），并且取遍 $[-1,3]$ 内的每一个值，所以存在某个 x 使得 $\mathrm{f}(x)=-d/c$，分母在该处等于零。分子在同一点却不等于零：假如两者同时为零，便有 $a(-d/c)+b=0$，即 $ad=bc$。所以 g 有一条铅垂渐近线（vertical asymptote），$|\mathrm{g}|$ 无界。由此可见，最后一部分的条件恰到好处，不多也不少。

验算

- 以数值计算，在 $[0,2\pi]$ 上取一个细密的网格，f 的最小值是 -1，最大值是 3。此外 $3-\mathrm{f}(0.1)\approx 2.496\times10^{-5}$，而 $\frac{1}{4}(0.1)^4=2.5\times10^{-5}$，印证了曲线在 $x=0$ 处的四次平坦性。
- 一个特殊情况：取 $a=1$、$b=0$、$c=0$、$d=1$，则 $\mathrm{g}=\mathrm{f}$，而公式给出 $\mathrm{g}(0)=3$ 和 $\mathrm{g}(\pi)=-1$，与预期相符。
- 随机选取 a,b,c,d,以有限差分(finite difference)计算 g 的导数,结果与 $(ad-bc)\mathrm{f}'/(c\mathrm{f}+d)^2$ 一致。

常见错误

- 遗漏端点 $x = 0$ 和 $x = 2\pi$ 这两个驻点。f′ 在这两处确实等于零。
- 忘记检查 g 在 f 的驻点处是否有定义。条件 $d \neq -3c$ 和 $d \neq c$ 正是为此而设。
- 在 $x = 0$ 处把顶部画成尖削的抛物线形状。曲线在那里其实平坦得多。

STEP 之外

当 $ad \neq bc$ 时，映射 $t \mapsto \dfrac{at + b}{ct + d}$ 称为莫比乌斯变换（*Möbius transformation*）。它是可逆（invertible）的，而且它的导数 $(ad - bc)/(ct + d)^2$ 永不为零，所以与它复合既不会产生新的驻点，也不会令原有的驻点消失。莫比乌斯映射在复分析（complex analysis）和几何中占有核心地位，并且在复合运算下构成一个群（group）。想看一本以图像为主、赏心悦目的介绍，可参阅 Needham 的 *Visual Complex Analysis*（《复分析：可视化方法》）。

详细解答

STEP III 1998，第 2 题

纯数学 • Beta 积分的递推关系

> **STEP III 1998，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/98-s3-q2

关键思路

这个积分其实是乔装了的贝塔函数（beta function）。每一部分都只需一个标准手法。第 (i) 部分用反射 $t \mapsto 1-t$，把 t 与 $1-t$ 的角色对调。第 (ii) 部分把被积函数（integrand）乘以 $1 = t + (1-t)$。第 (iii) 部分用分部积分法（integration by parts），把 $(1-t)$ 的一个幂次转移到 t 上。重复使用第 (iii) 部分，直至 $(1-t)$ 的幂次变成零，剩下的积分便可立即算出。

解答

第 (i) 部分

作代换（substitution）$u = 1-t$，则 $\mathrm{d}t = -\mathrm{d}u$，而积分限 $t = 0, 1$ 变成 $u = 1, 0$：

$$\mathrm{I}(a,b) = \int_1^0 (1-u)^a u^b \,(-\mathrm{d}u) = \int_0^1 u^b (1-u)^a \,\mathrm{d}u = \mathrm{I}(b,a).$$

$\mathrm{d}t = -\mathrm{d}u$ 带来的负号，刚好在把积分限换回 0 和 1 时抵销。

第 (ii) 部分

由于 $t + (1-t) = 1$，被积函数可以写成

$$t^a(1-t)^b = t^a(1-t)^b\big(t + (1-t)\big) = t^{a+1}(1-t)^b + t^a(1-t)^{b+1}.$$

把每一项在 $[0,1]$ 上积分，便得 $\mathrm{I}(a,b) = \mathrm{I}(a+1,b) + \mathrm{I}(a,b+1)$，正是所求。

第 (iii) 部分

我们对 $\mathrm{I}(a,b)$ 作分部积分，对 $(1-t)^b$ 求导，对 t^a 积分：

$$\begin{aligned}\mathrm{I}(a,b) &= \left[\frac{t^{a+1}}{a+1}(1-t)^b\right]_0^1 - \int_0^1 \frac{t^{a+1}}{a+1}\cdot\left(-b(1-t)^{b-1}\right)\mathrm{d}t \\ &= \left[\frac{t^{a+1}}{a+1}(1-t)^b\right]_0^1 + \frac{b}{a+1}\int_0^1 t^{a+1}(1-t)^{b-1}\,\mathrm{d}t.\end{aligned}$$

> ? **想一想** 为甚么边界项（boundary term）会消失？哪里需要用到 a 和 b 是正数？

在 $t=1$ 处，因为 $b>0$，因子 $(1-t)^b$ 等于 0。在 $t=0$ 处，因为 $a+1>0$，因子 t^{a+1} 等于 0。所以方括号内的项为零，两边同乘以 $a+1$ 便得

$$(a+1)\,\mathrm{I}(a,b) = b\,\mathrm{I}(a+1,b-1),$$

这就是所求的结果。（若 $0<b<1$，最后那个积分中 $(1-t)$ 的幂次是负数。但由于 $b-1>-1$，它仍然收敛（converge）。至于下面的整数情形，反正总有 $b-1\geq 0$。）

对整数 a,b 计算 $\mathrm{I}(a,b)$。把结果写成 $\mathrm{I}(a,b)=\dfrac{b}{a+1}\mathrm{I}(a+1,b-1)$，然后反复使用。每一步都令第一个参数增加 1，第二个参数减少 1。当第二个参数降至 0 时便停止：

$$\begin{aligned}\mathrm{I}(a,b) &= \frac{b}{a+1}\mathrm{I}(a+1,b-1) = \frac{b}{a+1}\cdot\frac{b-1}{a+2}\mathrm{I}(a+2,b-2) = \cdots \\ &= \frac{b(b-1)\cdots 1}{(a+1)(a+2)\cdots(a+b)}\mathrm{I}(a+b,0).\end{aligned}$$

每一步都是合理的，因为每次使用公式时，第二个参数都是正整数。现在 $\mathrm{I}(a+b,0)=\int_0^1 t^{a+b}\,\mathrm{d}t=\dfrac{1}{a+b+1}$。另外 $(a+1)(a+2)\cdots(a+b)=\dfrac{(a+b)!}{a!}$，所以

$$\boxed{\mathrm{I}(a,b)=\frac{a!\,b!}{(a+b+1)!}}$$

这条公式对 a 和 b 是对称的，与第 (i) 部分的要求一致。当 a 或 b 等于 0 时，公式同样成立。

验算

- $\mathrm{I}(1,1)=\int_0^1(t-t^2)\,\mathrm{d}t=\frac{1}{2}-\frac{1}{3}=\frac{1}{6}$，而公式给出 $\frac{1!\,1!}{3!}=\frac{1}{6}$。
- 公式满足第 (ii) 部分：$\frac{(a+1)!\,b!}{(a+b+2)!}+\frac{a!\,(b+1)!}{(a+b+2)!}=\frac{a!\,b!\,(a+b+2)}{(a+b+2)!}=\frac{a!\,b!}{(a+b+1)!}$。
- 数值方面，用辛普森法则（Simpson's rule）得 $\mathrm{I}(2,3)\approx 0.0166667$，与 $\frac{2!\,3!}{6!}=\frac{1}{60}$ 吻合。我们亦随机取非整数的 a,b，以数值积分（quadrature）验算了第 (i) 至 (iii) 部分，结果吻合至约 10^{-11}。

常见错误

- 分部积分时把方向弄反。对 t^a 求导会令 $(1-t)$ 的幂次升高，方向错了。
- 忘记说明边界项为何消失。它之所以为零，全因 $a+1>0$ 及 $b>0$。
- 递推（recursion）早了一步停下，或者漏掉来自 $\mathrm{I}(a+b,0)$ 的最后一个因子 $\frac{1}{a+b+1}$。

STEP 之外

$\mathrm{I}(a,b)$ 就是贝塔函数 $B(a+1,b+1)$。对所有正实数参数，它与伽玛函数（gamma function）有以下关系：$B(x,y)=\Gamma(x)\Gamma(y)/\Gamma(x+y)$。我们的答案正是 $\Gamma(n+1)=n!$ 这个特殊情形。在概率论中，$t^a(1-t)^b/\mathrm{I}(a,b)$ 是贝塔分布（beta distribution）的密度函数（density），而贝塔分布在贝叶斯统计学（Bayesian statistics）中占有核心地位。Blitzstein 与 Hwang 合著的 *Introduction to Probability*（概率论入门）对此有很精彩的论述，当中包括一个计数论证（counting argument），解释了为甚么答案是 $a!\,b!/(a+b+1)!$。

详细解答

STEP III 1998，第 3 题

纯数学 · 线性递推关系

> **STEP III 1998，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/98-s3-q3

关键思路

这是一个带常数项的一阶线性递推关系（first-order linear recurrence），可说是微分方程（differential equation）$y' = cy - d$ 的离散版本。处理手法与常微分方程（ODE）相同：先找出常数解（即“平衡”（equilibrium）解），然后留意到与平衡解之间的偏差每天只是乘以同一个因子。至于极限，真正要花工夫的只有一件事：弄清楚 $\big((1+c)^N - 1\big)/c$ 在 $c \to 0$ 时的变化，而二项式定理（binomial theorem）正好解决这个问题。

解答

$c > 0$ 时求解

我们尝试对每个 T 设 $V_T = Ak^T + B$。代入递推关系，得

$$Ak^{N+1} + B = (1+c)\big(Ak^N + B\big) - d \quad \Longleftrightarrow \quad Ak^N\big(k - (1+c)\big) = cB - d.$$

右边与 N 无关，而左边除非恒为零，否则必随之变化（我们要求 $A \neq 0$ 及 $k \neq 0$，否则这个形式只是一个常数）。因此对每个 N，两边都必须为零：

$$k = 1 + c, \qquad B = \frac{d}{c},$$

这里可以除以 c，因为 $c > 0$。取这些值后，无论常数 A 是甚么，$V_T = A(1+c)^T + d/c$ 都满足递推关系。代入 $T = 0$ 得 $V_0 = A + d/c$，所以 $A = V_0 - d/c$，于是

$$\boxed{V_N = \Big(V_0 - \frac{d}{c}\Big)(1+c)^N + \frac{d}{c}}$$

想一想 我们猜了一个形式，结果行得通。但怎样知道没有其他解呢?

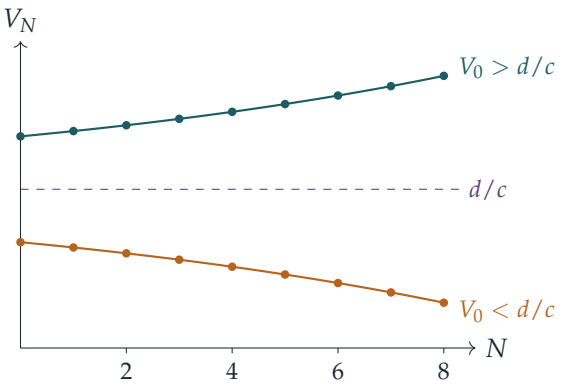
递推关系由 V_0 决定 V_1，再由 V_1 决定 V_2，如此类推，所以给定 V_0 后，恰好只有一个数列符合条件。我们的公式满足递推关系，而且 V_0 的值也正确，所以它就是那个数列。（如果你想要一个直接的论证，可设 $U_N = V_N - d/c$；递推关系变成 $U_{N+1} = (1+c)U_N$，所以 $U_N = (1+c)^N U_0$，答案相同。）

把答案改写成以下形式，稍后会很有用：

$$V_N = V_0(1+c)^N - d\,\frac{(1+c)^N - 1}{c}. \qquad (*)$$

这个式子有一个很好的解读：第一项是 V_0 以每天 c 的比率增长，第二项是 $d\big(1+(1+c)+\cdots+(1+c)^{N-1}\big)$，即每天的扣减额各自按余下的日数增长后，以等比级数（geometric series）求和的结果。

常数 d/c 是平衡值。若 $V_0 > d/c$，债券的价值会无止境地增长；若 $V_0 = d/c$，价值维持不变；若 $V_0 < d/c$，价值会愈跌愈快：



（图中取 $c = 0.1$ 及 $d/c = 5$，并分别取 $V_0 = 6$ 和 $V_0 = 4$。）

$c = 0$ 的情形

这时递推关系是 $V_{N+1} = V_N - d$：价值每天减少 d，是一个等差数列（arithmetic progression）。所以

$$\boxed{V_N = V_0 - Nd}$$

留意 $c > 0$ 时的公式不能直接代入 $c = 0$ 求值，因为式中有 d/c；这正是题目要问极限的原因。

$c \to 0$ 时的极限

固定 N，并使用 $(*)$。第一项趋向 V_0，因为对固定的 N 有 $(1+c)^N \to 1$。至于第二项，用二项式定理展开：

$$(1+c)^N - 1 = Nc + \binom{N}{2}c^2 + \cdots + c^N,$$

$$\text{所以} \quad \frac{(1+c)^N - 1}{c} = N + \binom{N}{2}c + \cdots + c^{N-1}.$$

由于 N 是固定的，这是一个关于 c 的多项式，只有有限多项，而且第一项之后的每一项都含有 c 的正数次幂。因此当 $c \to 0$ 时它趋向 N，于是

$$\lim_{c\to 0} V_N = V_0 - Nd,$$

这正是 $c = 0$ 时的解，符合题目要求。（等价地说，$\big((1+c)^N - 1\big)/c$ 是 x^N 在 $x = 1$ 处的差商（difference quotient），所以它的极限就是导数 N。）

“对固定的 N”这几个字很重要：若 c 很小但 N 很大，像 $\binom{N}{2}c$ 这样的项不一定小，而指数增长最终会胜过线性变化。

验算

- $N = 1$：公式给出 $(V_0 - d/c)(1+c) + d/c = (1+c)V_0 - d$，这就是递推关系本身。
- $V_0 = d/c$ 时，对所有 N 都有 $V_N = d/c$，即平衡状态。
- 数值上，取 $V_0 = 7$、$d = 2$、$N = 12$：公式在 $c = 0.01$ 时给出 $V_{12} \approx -17.48$，在 $c = 10^{-4}$ 时给出 -17.005，逐渐接近 $V_0 - 12d = -17$。在许多随机选取的有理数例子中，这个闭合公式也与直接迭代的结果完全吻合。

常见错误

- 取 $B = -d/c$：正负号来自 $cB - d = 0$，所以 $B = +d/c$。
- 用 $A = V_0$ 而不是 $A = V_0 - d/c$：常数 B 在 $T = 0$ 时同样有贡献。
- 试图在 $c > 0$ 的公式中直接代入 $c = 0$，但该式此时没有定义；必须取极限。

STEP 之外

这类递推关系称为*线性差分方程*（linear difference equations），其理论与线性常微分方程并行：通解（general solution）等于一个特解（particular solution，这里是常数解）加

上齐次方程（homogeneous equation）$U_{N+1} = (1+c)U_N$ 的通解。在金融学中，这就是定额还款的贷款或年金（annuity）的公式，而 $c \to 0$ 的极限则是由复利变化过渡到单利变化。Boyce 与 DiPrima 的 *Elementary Differential Equations*（初等微分方程教科书）在常微分方程理论之外，另有一节讲述一阶差分方程。

详细解答

STEP III 1998，第 4 题

纯数学 · 极坐标玫瑰线

> **STEP III 1998，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/98-s3-q4

关键思路

要理解极坐标曲线（polar curve），方程简单时最容易的做法是把它化为直角坐标（Cartesian）形式；方程不简单时，则追踪 r 的正负号。整道题的关键在于以下约定：$r<0$ 的点要画在相反方向 $\theta+\pi$ 上、与原点距离 $|r|$ 之处。对于 $r=\cos k\theta$，这个约定决定了负的回圈是落在新的位置（形成新的花瓣），还是叠在已画好的花瓣之上。

解答

圆 $r=\cos\theta$

两边同乘 r，并利用 $x=r\cos\theta$、$y=r\sin\theta$、$r^2=x^2+y^2$：

$$\begin{aligned} r^2=r\cos\theta \quad &\Longrightarrow \quad x^2+y^2=x \\ &\Longleftrightarrow \quad \left(x-\tfrac{1}{2}\right)^2+y^2=\tfrac{1}{4}. \end{aligned}$$

因此曲线上每一点都在圆心为 $(\frac{1}{2},0)$、半径为 $\frac{1}{2}$ 的圆上。

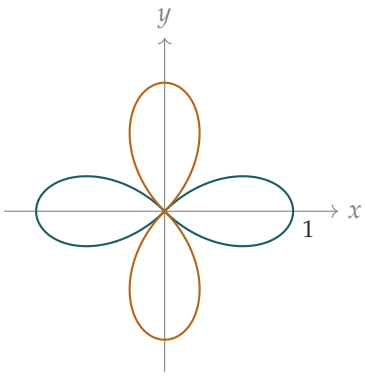
反过来，取这个圆上原点以外的任意一点 (x,y)。由于 $x=x^2+y^2>0$，它的极角满足 $-\frac{1}{2}\pi<\theta<\frac{1}{2}\pi$，而把 $r^2=x=r\cos\theta$ 除以 $r>0$ 便得 $r=\cos\theta$。至于原点，则在 $\theta=\pm\frac{1}{2}\pi$ 时到达。所以这条曲线正正就是这个圆，当 θ 由 $-\frac{1}{2}\pi$ 走到 $\frac{1}{2}\pi$ 时恰好描画一次；在这个范围内始终有 $r=\cos\theta\geq 0$，故此毋须用到正负号的约定。

曲线 $r=\cos 2\theta$

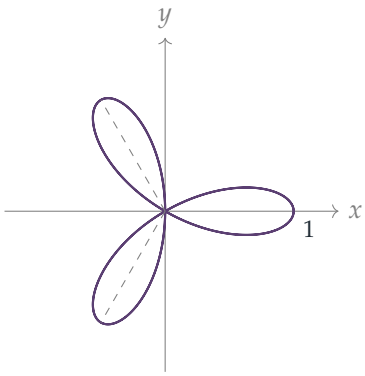
现在 r 会变号，而当 $r<0$ 时，我们把 (r,θ) 画在方向 $\theta+\pi$ 上、距离为 $|r|$ 之处。

- 当 $-\frac{1}{4}\pi < \theta < \frac{1}{4}\pi$ 时，$r > 0$，由 0 上升至 $\theta = 0$ 时的 1，再回落到 0：得出一片沿正 x 轴的花瓣。
- 当 $\frac{1}{4}\pi < \theta < \frac{3}{4}\pi$ 时，$r < 0$，所以这些点画在 $\frac{5}{4}\pi$ 至 $\frac{7}{4}\pi$ 的方向上：得出一片沿负 y 轴的花瓣。
- 同理，$\frac{3}{4}\pi < \theta < \frac{5}{4}\pi$（$r > 0$）给出一片沿负 x 轴的花瓣，而 $\frac{5}{4}\pi < \theta < \frac{7}{4}\pi$（$r < 0$）给出一片沿正 y 轴的花瓣。

结果是一条四叶玫瑰线（four-leaved rose），每片花瓣长度为 1，关于两条坐标轴都对称，而且花瓣之间只在原点相接。在草图中，蓝绿色的花瓣来自 $r > 0$，琥珀色的花瓣来自 $r < 0$。



$r = \cos 2\theta$



$r = \cos 3\theta$

曲线 $r = \cos 2n\theta$

由于 $\cos(-2n\theta) = \cos 2n\theta$，$n$ 与 $-n$ 给出同一条曲线，所以取 $n \geq 1$（$n = 0$ 的情况稍后再谈）。

> ? **想一想** 当 $r < 0$ 时，那些回圈是叠在已画好的回圈之上，还是落在新的位置?

在以 $\theta = k\pi/n = 2k \cdot \frac{\pi}{2n}$（$k = 0, 1, \ldots, 2n-1$）为中心、长度为 $\pi/(2n)$ 的各个区间上，我们有 $r > 0$。每个区间给出一片指向 $2k \cdot \frac{\pi}{2n}$ 方向的花瓣，这个方向是 $\frac{\pi}{2n}$ 的偶数倍。

在以 $\theta = (2k+1)\frac{\pi}{2n}$ 为中心的各个区间上，我们有 $r < 0$，而这些点画在方向 $\theta + \pi$ 上，所以花瓣指向

$$(2k+1)\frac{\pi}{2n} + \pi = (2k+1+2n)\frac{\pi}{2n},$$

即 $\frac{\pi}{2n}$ 的奇数倍。因此负的回圈落在新的位置，也就是正花瓣之间的空隙。

所以 $r = \cos 2n\theta$ 是一条有 $4n$ 片全等（congruent）花瓣的玫瑰线，每片花瓣长度为 1，分别指向 $m\pi/(2n)$（$m = 0, 1, \ldots, 4n-1$）的方向，每片花瓣占的角为 $\pi/(2n)$，而相邻的

花瓣只在原点相接。（当 $n=1$ 时，这就是上面的四叶玫瑰线。）各花瓣所占的角合共为 $4n\cdot\frac{\pi}{2n}=2\pi$，所以它们刚好填满一整个周角，互不重叠。

至于退化情况 $n=0$，曲线是 $r=1$，即单位圆。

面积与 n 无关

由于花瓣互不重叠，当 θ 由 0 走到 2π 时，所围区域的每一部分都恰好被扫过一次，于是面积公式 $\frac{1}{2}\int r^2\,\mathrm{d}\theta$（它只涉及 r^2，所以 r 的正负号并无影响）给出：当 $n\geq 1$ 时，

$$\begin{aligned}A&=\frac{1}{2}\int_0^{2\pi}\cos^2 2n\theta\,\mathrm{d}\theta=\frac{1}{4}\int_0^{2\pi}(1+\cos 4n\theta)\,\mathrm{d}\theta\\&=\frac{1}{4}\left[\theta+\frac{\sin 4n\theta}{4n}\right]_0^{2\pi}=\frac{\pi}{2},\end{aligned}$$

因为 $\sin 8n\pi=0$。这个结果与 n 无关，正如题目所求：花瓣愈多，每片花瓣就愈窄，两者恰好互相抵销。（被排除的情况 $n=0$，即单位圆，面积为 π；公式中 $\sin 4n\theta/(4n)$ 一项正是用到 $n\neq 0$ 的地方。）

曲线 $r=\cos 3\theta$

> **想一想** 为甚么这条曲线有三片花瓣，而不是六片？

现在 $r>0$ 的区间以 $\theta=0,\frac{2}{3}\pi,\frac{4}{3}\pi$ 为中心，给出指向这些方向的花瓣。负的回圈以 $\theta=\frac{1}{3}\pi,\pi,\frac{5}{3}\pi$ 为中心，并画在方向 $\frac{4}{3}\pi$、2π 和 $\frac{8}{3}\pi\equiv\frac{2}{3}\pi$ 上：正正就是正花瓣的方向。事实上，把 θ 换成 $\theta+\pi$，便得到相反方向上的

$$r=\cos(3\theta+3\pi)=-\cos 3\theta$$

，那是同一点。所以 $r=\cos 3\theta$ 是一条三叶玫瑰线，花瓣长度为 1，沿 $\theta=0,\frac{2}{3}\pi,\frac{4}{3}\pi$ 的方向，每片占的角为 $\frac{1}{3}\pi$，而当 θ 由 0 走到 2π 时，整条曲线被描画了*两次*（$0\leq\theta\leq\pi$ 时描画一次）。这就是上面右边的草图。

同样的论证说明：对任何奇数 k，$r=\cos k\theta$ 是一条描画两次的 k 叶玫瑰线；而对偶数 k，它是一条描画一次的 $2k$ 叶玫瑰线。

验算

- 在 $r=\cos\theta$ 中代入 $\theta=0$：点 $(1,0)$ 在 $(x-\frac{1}{2})^2+y^2=\frac{1}{4}$ 上，原点（在 $\theta=\pm\frac{1}{2}\pi$ 时）也在其上。
- 圆 $r=\cos\theta$ 的面积为 $\pi(\frac{1}{2})^2=\frac{1}{4}\pi$，而 $\frac{1}{2}\int_{-\pi/2}^{\pi/2}\cos^2\theta\,\mathrm{d}\theta=\frac{1}{4}\pi$ 与此吻合。同样，对于三叶玫瑰线，$\frac{1}{2}\int_0^{\pi}\cos^2 3\theta\,\mathrm{d}\theta=\frac{1}{4}\pi$。若在这里用 0 至 2π，便会重复计算，得出 $\frac{1}{2}\pi$。

- 数值验证：对 $n = 1, \ldots, 5$ 绘画 $r = \cos 2n\theta$，可找到 $4n$ 个花瓣尖端，而以蒙地卡罗方法（Monte Carlo）估算所围面积，每次都得出 $1.57 \approx \pi/2 = 1.5708$。

常见错误

- 忽略负的 r，结果 $r = \cos 2\theta$ 只画出两片花瓣。
- 以为 $r = \cos 3\theta$ 有六片花瓣，理由是 r 有六个回圈：其实负的回圈叠在正的回圈之上。
- 对描画两次的曲线（例如 $r = \cos 3\theta$）使用 $\frac{1}{2}\int_0^{2\pi} r^2 \,\mathrm{d}\theta$，令面积变成两倍。

STEP 之外

这些曲线称为 *rhodonea*，即玫瑰线（rose curve）。把一点写成复数（complex number）$z = r\,\mathrm{e}^{\mathrm{i}\theta}$，玫瑰线 $r = \cos k\theta$ 就是 $z = \frac{1}{2}\left(\mathrm{e}^{\mathrm{i}(k+1)\theta} + \mathrm{e}^{-\mathrm{i}(k-1)\theta}\right)$，即两个方向相反的圆周运动之和。这个观点把玫瑰线与本轮（epicycle）及傅里叶级数（Fourier series）联系起来，而面积公式 $\frac{1}{2}\int r^2 \,\mathrm{d}\theta$ 则是向量微积分（vector calculus）中格林定理（Green's theorem）的特殊情况。Needham 的 *Visual Complex Analysis*（《复分析：可视化方法》）把这种以几何眼光看待复数的方法发挥得淋漓尽致。

详细解答

STEP III 1998，第 5 题

纯数学 · 矩阵指数

> **STEP III 1998，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/98-s3-q5

关键思路

矩阵（matrix）的无穷级数（infinite series）看来吓人，但这里两个矩阵的幂都按非常简单的规律重复：$\mathbf{M}$ 的幂像 i 的幂一样循环，而 $\mathbf{N}$ 的幂由平方起全部为零。一旦知道每一个幂，级数便会分拆成熟悉的纯量级数（或者干脆终止），于是可以直接读出相应的变换（transformation）。最后一部分则是直接的矩阵计算，而它的几何意义，来自追问每个变换会令哪些直线保持不变。

解答

矩阵 $\mathbf{M}$

乘开，得

$$\mathbf{M}^2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = -\mathbf{I},$$

正如所求。因此，对每个整数 $k \geq 0$，

$$\mathbf{M}^{2k} = (\mathbf{M}^2)^k = (-1)^k\mathbf{I}, \qquad \mathbf{M}^{2k+1} = (-1)^k\mathbf{M}.$$

把级数按 $\theta\mathbf{M}$ 的偶次幂和奇次幂分开，

$$\begin{aligned}\exp(\theta\mathbf{M}) &= \sum_{k=0}^{\infty}\frac{\theta^{2k}}{(2k)!}\mathbf{M}^{2k}+\sum_{k=0}^{\infty}\frac{\theta^{2k+1}}{(2k+1)!}\mathbf{M}^{2k+1}\\ &= \Big(\sum_{k=0}^{\infty}\frac{(-1)^k\theta^{2k}}{(2k)!}\Big)\mathbf{I}+\Big(\sum_{k=0}^{\infty}\frac{(-1)^k\theta^{2k+1}}{(2k+1)!}\Big)\mathbf{M}\\ &= \cos\theta\,\mathbf{I}+\sin\theta\,\mathbf{M}=\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}.\end{aligned}$$

（这样分拆是合理的，因为矩阵级数的每个元素都是一个实数级数，其各项的大小以 $|\theta|^r/r!$ 为界，所以它绝对收敛（converges absolutely），可以重新组合。）

> ? **想一想** 我在哪里见过 $\cos\theta\,\mathbf{I}+\sin\theta\,\mathbf{M}$?

由于 $\mathbf{M}^2=-\mathbf{I}$，矩阵 $\mathbf{M}$ 的表现与 i 完全一样，而这个计算正是 $e^{i\theta}=\cos\theta+i\sin\theta$ 的矩阵版本。从几何上看，$\exp(\theta\mathbf{M})$ 把 $(1,0)$ 送到 $(\cos\theta,\sin\theta)$，把 $(0,1)$ 送到 $(-\sin\theta,\cos\theta)$：它是**平面绕原点逆时针旋转（rotation）角** θ。（事实上 $\mathbf{M}$ 本身就是转角为 $\frac{\pi}{2}$ 的旋转，正如乘以 i 一样。）

矩阵 N

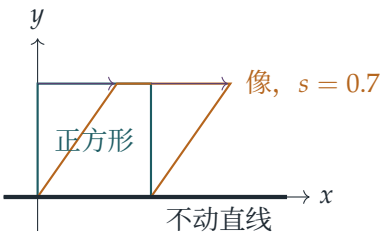
这里

$$\mathbf{N}^2=\begin{pmatrix}0&1\\0&0\end{pmatrix}\begin{pmatrix}0&1\\0&0\end{pmatrix}=\begin{pmatrix}0&0\\0&0\end{pmatrix},$$

所以对所有 $r\geq 2$ 都有 $\mathbf{N}^r=\mathbf{0}$，级数在两项之后便终止：

$$\exp(s\mathbf{N})=\mathbf{I}+s\mathbf{N}=\begin{pmatrix}1&s\\0&1\end{pmatrix}.$$

它把 (x,y) 送到 $(x+sy,\,y)$。每一点都平行于 x 轴移动，移动的距离与它的高度 y 成正比；x 轴上的点则完全不动。所以 $\exp(s\mathbf{N})$ 是一个**平行于 x 轴的错切（shear）**，以 x 轴为其不动点（fixed point）所成的直线，并把单位正方形送到一个面积相同的平行四边形（行列式（determinant）为 1）。



它们何时可交换？

记 $c = \cos\theta$ 及 $\sigma = \sin\theta$。则

$$\exp(s\mathbf{N})\exp(\theta\mathbf{M}) = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}\begin{pmatrix} c & -\sigma \\ \sigma & c \end{pmatrix} = \begin{pmatrix} c+s\sigma & -\sigma+sc \\ \sigma & c \end{pmatrix},$$

$$\exp(\theta\mathbf{M})\exp(s\mathbf{N}) = \begin{pmatrix} c & -\sigma \\ \sigma & c \end{pmatrix}\begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} c & sc-\sigma \\ \sigma & s\sigma+c \end{pmatrix}.$$

两式相减，

$$\exp(s\mathbf{N})\exp(\theta\mathbf{M}) - \exp(\theta\mathbf{M})\exp(s\mathbf{N}) = s\sin\theta\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

这个矩阵对每个 s 都是零矩阵，当且仅当 $\sin\theta = 0$（若 $\sin\theta \neq 0$，取 $s = 1$ 即可）。所以这两个矩阵对所有 s 都可交换（commute），恰好是在

$$\theta = n\pi, \qquad n \in \mathbb{Z}.$$

的时候。对这些值，$\exp(\theta\mathbf{M}) = (-1)^n\mathbf{I}$，即不是单位矩阵（identity matrix）就是 $-\mathbf{I}$。

? **想一想** 为甚么与这个错切可交换的旋转，必定是半周旋转或者根本不转？

设 R 为该旋转，S 为该错切（其中 $s \neq 0$），并假设 $SR = RS$。在 x 轴上任取一点 P，则 $S(P) = P$。于是

$$S\big(R(P)\big) = R\big(S(P)\big) = R(P),$$

所以 $R(P)$ 也是错切的不动点。但错切所固定的点只有 x 轴上的点。所以这个旋转必须把 x 轴映射到它自身，而绕原点的旋转之中能做到这一点的，只有转角为 0（恒等变换）或 π（半周旋转 $(x,y) \mapsto (-x,-y)$）的两个。反过来，这两个变换与每一个线性映射（linear map）都可交换，因为它们就是 $\pm\mathbf{I}$：把平面转半周和对它作错切，不论先后次序，结果都一样。任何其他旋转都会令错切的不动直线倾斜，所以先错切后旋转与先旋转后错切会得出不同的结果。

验算

- 当 $\theta = 0$ 时旋转矩阵是 $\mathbf{I}$，当 $s = 0$ 时错切是 $\mathbf{I}$，与级数的要求一致。
- 两个行列式都是 1（分别为 $\cos^2\theta + \sin^2\theta$ 和 $1\cdot 1$），旋转和错切都保持面积不变，理应如此。
- 数值上，取 $\theta = 1$ 并把级数的首 80 项相加，所得的元素与 $\cos 1 \approx 0.5403$ 和 $\sin 1 \approx 0.8415$ 吻合。取 $\theta = 1$、$s = 1$ 时，交换子（commutator）的对角线元素为 $\pm\sin 1 \approx \pm 0.8415$，与 $s\sin\theta\,\mathrm{diag}(1,-1)$ 吻合。

常见错误

- 以为对矩阵也有 $\exp(\mathbf{A})\exp(\mathbf{B}) = \exp(\mathbf{A}+\mathbf{B})$。这条法则需要 $\mathbf{AB} = \mathbf{BA}$，而这正是此处不成立的条件，除非 $\sin\theta = 0$。
- 只给出 $\theta = 0$（或 $\theta = 0, \pi$），而没有给出 π 的所有整数倍。
- 把 $\exp(s\mathbf{N})$ 称为拉伸（stretch）或放大（enlargement）：它保持面积不变，并且逐点固定 x 轴，这正是它成为错切的原因。

STEP 之外

矩阵指数（matrix exponential）是李代数（*Lie algebra*）与李群（*Lie group*）之间的桥梁：$\mathbf{M}$ 生成旋转群，$\mathbf{N}$ 生成错切群，而 $\exp(t\mathbf{A})$ 是线性微分方程（linear differential equation）$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} = \mathbf{A}\mathbf{x}$ 的解。对于不可交换的矩阵，$\exp(\mathbf{A})\exp(\mathbf{B}) = \exp(\mathbf{A}+\mathbf{B})$ 不成立，其偏差的程度由交换子 $\mathbf{AB} - \mathbf{BA}$ 来量度，这个想法你在最后一部分已经遇过。Strang 的 *Introduction to Linear Algebra*（线性代数入门）涵盖矩阵指数及其在微分方程中的应用。

详细解答

STEP III 1998，第 6 题

纯数学 · 多面体的体积

> **STEP III 1998，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/98-s3-q6

关键思路

把正方体放在坐标轴上，使它的顶点（vertex）就是坐标全为 0 或 a 的那些点。这样，距离可由三维的毕氏定理（Pythagoras' theorem）求得，而那些不易处理的立体，则可以从正方体切去简单的小块来量度。正四面体（regular tetrahedron）就是正方体切去四个角后剩下的部分，而正八面体（regular octahedron）则是两个底对底拼合的正方形底角锥（pyramid）。

解答

第 (i) 部分

取正方体 $[0,a]^3$。它的顶点是八个点 (x,y,z)，其中每个坐标为 0 或 a；两个顶点相邻（adjacent，即由一条棱（edge）相连）当且仅当它们恰好有一个坐标不同。

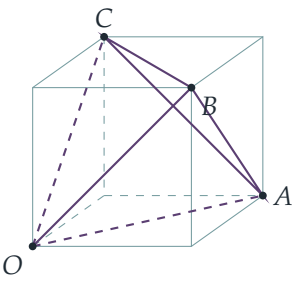
可以选哪四个顶点？ 按顶点的坐标中有多少个等于 a，把顶点称为偶顶点或奇顶点。沿一条棱移动会改变一个坐标，所以必定令奇偶互换。假设我们的四个顶点中包括一个偶顶点 P 和一个奇顶点 Q。它们并不相邻，所以它们有三个坐标不同：Q 就是与 P 相对的顶点。其余每个顶点都与 P 或 Q 相邻（三个奇顶点与 P 相邻，三个偶顶点与 Q 相邻），所以无法再加入第三个顶点。因此这四个顶点全是偶顶点，

$$O=(0,0,0),\quad A=(a,a,0),\quad B=(a,0,a),\quad C=(0,a,a),$$

或者全是奇顶点（即这组顶点对平面 $x=\frac{1}{2}a$ 的反射，所得的立体与原来的全等（congruent））。

正则性。 O,A,B,C 中任意两点都恰好有两个坐标不同，每个相差 a，所以六个距离全都

是 $\sqrt{a^2+a^2}=\sqrt{2}\,a$。因此四个面都是边长为 $\sqrt{2}\,a$ 的等边三角形（equilateral triangle），而 $OABC$ 是一个正四面体。它的每条棱都是正方体某一个面的对角线。



体积。从正方体中移走这个四面体，剩下四个角块，每个奇顶点处各有一个。例如，位于 $(a,0,0)$ 的角块，就是以该顶点及其三个相邻顶点 O、A、B 为顶点的四面体。它在 $(a,0,0)$ 处的三条棱互相垂直，长度均为 a，所以它是一个角锥，底是面积为 $\frac{1}{2}a^2$ 的直角三角形，高为 a，体积为 $\frac{1}{3}\cdot\frac{1}{2}a^2\cdot a=\frac{1}{6}a^3$。因此

$$V_{\text{tet}}=a^3-4\cdot\tfrac{1}{6}a^3=\tfrac{1}{3}a^3.$$

要令棱长为单位长度，我们需要 $\sqrt{2}\,a=1$，即 $a=1/\sqrt{2}$ 及 $a^3=1/(2\sqrt{2})$。所以棱长为单位长度的正四面体的体积是

$$\frac{1}{3}\cdot\frac{1}{2\sqrt{2}}=\frac{1}{6\sqrt{2}}=\frac{\sqrt{2}}{12}.$$

? **想一想** 为甚么四个角块和这个四面体恰好填满正方体，而且互不重叠?

四面体的每个面，例如 OAB，都恰好切去正方体的一个顶点，这里是 $(a,0,0)$：通过 O、A、B 的平面 $x-y-z=0$ 把 $(a,0,0)$ 严格地分在一侧，而其余七个顶点则在另一侧或在平面上。所以正方体分成四个角块（每个面之外各有一个）和位于中间的四面体。

第 (ii) 部分

棱长为单位长度的正八面体，可以把六个顶点放在

$$(\pm h,0,0),\quad (0,\pm h,0),\quad (0,0,\pm h),\qquad h=\tfrac{1}{\sqrt{2}},$$

其中每个顶点都与不和它相对的四个顶点相连。两个这样的顶点，例如 $(h,0,0)$ 和 $(0,h,0)$，相距 $\sqrt{2}\,h=1$。所以十二条棱的长度全是 1，而八个面（每个卦限（octant）各有一个）都是等边三角形。位于平面 $z=0$ 上的四个顶点组成一个边长为 1 的正方形。这个八面体就是以这个正方形为底的两个正方形底角锥，顶点为 $(0,0,\pm h)$，高各为 h。因此

$$V_{\text{oct}}=2\cdot\tfrac{1}{3}\cdot 1^2\cdot\frac{1}{\sqrt{2}}=\frac{\sqrt{2}}{3}.$$

更一般地，把长度放大 e 倍，体积就放大 e^3 倍，所以棱长为 e 的正八面体的体积是 $\frac{\sqrt{2}}{3}e^3$。

第 (iii) 部分

现在把边长为 a 的正方体的中心放在原点会较为简洁，即取 $[-\frac{1}{2}a, \frac{1}{2}a]^3$。它各个面的中心是

$$(\pm\tfrac{1}{2}a, 0, 0), \quad (0, \pm\tfrac{1}{2}a, 0), \quad (0, 0, \pm\tfrac{1}{2}a).$$

这正是第 (ii) 部分的布局，其中 $h = \frac{1}{2}a$。相对两面的中心相距 a。相邻两面的中心相距 $\sqrt{(\frac{1}{2}a)^2 + (\frac{1}{2}a)^2} = a/\sqrt{2}$。以这六点为顶点的凸（convex）立体有八个面，每个卦限各有一个：例如，在卦限 $x, y, z \geq 0$ 内的面是以 $(\frac{1}{2}a, 0, 0)$、$(0, \frac{1}{2}a, 0)$、$(0, 0, \frac{1}{2}a)$ 为顶点的三角形。每个面都是边长为 $a/\sqrt{2}$ 的等边三角形，所以各面的中心组成一个棱长为 $e = a/\sqrt{2}$ 的正八面体。

根据第 (ii) 部分，它的体积是

$$\frac{\sqrt{2}}{3}\left(\frac{a}{\sqrt{2}}\right)^3 = \frac{\sqrt{2}}{3} \cdot \frac{a^3}{2\sqrt{2}} = \frac{a^3}{6}.$$

第 (i) 部分的四面体，其顶点是同一个正方体的其中四个顶点，体积为 $\frac{1}{3}a^3$。由于 $\frac{1}{6}a^3 = \frac{1}{2} \cdot \frac{1}{3}a^3$，八面体的体积是四面体体积的一半，正如题目所求。

验算

- 利用“底面积乘高”的公式，以底面积 $\frac{\sqrt{3}}{4}$ 和高 $\sqrt{2/3}$ 计算，得出单位正四面体的体积为 $\frac{1}{3} \cdot \frac{\sqrt{3}}{4} \cdot \sqrt{\frac{2}{3}} = \frac{\sqrt{2}}{12} \approx 0.11785$，与第 (i) 部分吻合。
- 棱长相等时，$V_{\text{oct}} = 4V_{\text{tet}}$，因为 $\frac{\sqrt{2}}{3} = 4 \cdot \frac{\sqrt{2}}{12}$。这与一个经典的分割相符：把棱长为 2 的四面体分割成四个棱长为 1 的四面体（在它的四个角）和一个棱长为 1 的八面体（在中间）；以单位四面体为量度单位，就是 $8 = 4 + 4$。
- 数值上，对所有四个互不相邻的正方体顶点的选法作行列式（determinant）计算，得出单位正方体的情况下体积为 $\frac{1}{3}$；而对区域 $|x| + |y| + |z| \leq 1/\sqrt{2}$ 作蒙地卡罗（Monte Carlo）估算，得出 0.4715，对照 $\sqrt{2}/3 \approx 0.47140$。

常见错误

- 在第 (i) 部分把正方体的边长取为 1：这时四面体的棱长是 $\sqrt{2}$ 而不是 1，它的体积须按 $(1/\sqrt{2})^3$ 的比例重新缩放。
- 减去错误的角块：角块共有四个（每个没有用到的顶点处各有一个），每个体积为 $\frac{1}{6}a^3$，而不是八个。

► 在第 (iii) 部分，把正方体的边长当作八面体的棱长。相邻两面的中心相距 $a/\sqrt{2}$。

STEP 之外

正方体内的两个四面体，以及由各面中心组成的八面体，是柏拉图立体（Platonic solids）及其对称群（symmetry group）这个故事的开端：正方体的旋转之中，保持其中一个内接四面体不变的那些旋转组成一个阶为 12 的群，是正方体全部 24 个旋转的一半。正方体与八面体互为对偶（dual）（其中一个的各面中心就是另一个的顶点），而这类立体的体积可以由行列式简洁地求出，因为行列式量度的正是线性映射（linear map）把体积缩放多少。Strang 的 *Introduction to Linear Algebra*（线性代数入门）把行列式视为体积来处理，而 Armstrong 的 *Groups and Symmetry*（群与对称）则研究这些立体的对称群。

详细解答

STEP III 1998，第 7 题

纯数学 · 曲线描绘与积分

> **STEP III 1998，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/98-s3-q7

关键思路

第一个函数并非随意安排的热身题：把 $\mathrm{b}(x)$ 求导并令导数为零时，所得的方程正好是 $\mathrm{f}(x/T)=0$。因此，我们对 f 的一切认识（只有一个正根，且位于 2 与 3 之间）都告诉我们 b 在何处转向。至于 B 如何依赖 T，则来自自然的代换（substitution）$x=Tu$，它把 T 从被积函数（integrand）中抽走。

解答

绘画 f 的草图并求 m

我们有 $\mathrm{f}(0)=1\cdot(-3)+3=0$，而且

$$\mathrm{f}'(s)=\mathrm{e}^{s}(s-3)+\mathrm{e}^{s}=\mathrm{e}^{s}(s-2).$$

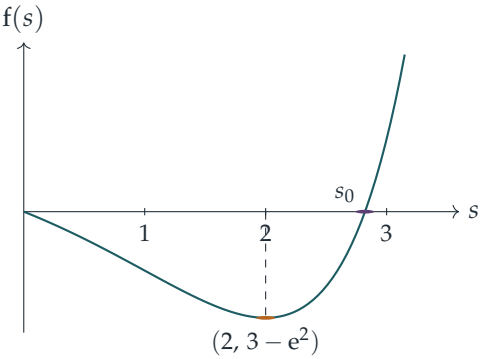
由于 $\mathrm{e}^{s}>0$，f 在 $[0,2]$ 上递减，在 $[2,\infty)$ 上递增，并有极小值 $\mathrm{f}(2)=3-\mathrm{e}^{2}\approx 3-7.29=-4.29$。当 $s\to\infty$ 时，$\mathrm{e}^{s}(s-3)\to\infty$，所以 $\mathrm{f}(s)\to\infty$。

因为 $\mathrm{f}(0)=0$ 且 f 在 $[0,2]$ 上严格递减（strictly decreasing），所以当 $0<s\le 2$ 时 $\mathrm{f}(s)<0$。在 $[2,\infty)$ 上，它由一个负值严格递增至 $+\infty$，故此在该处恰好穿过零一次，设在某个 $s_0>2$。因此 s_0 是 f 唯一的正根。

取 $\mathrm{e}\approx 2.7$：

$$\mathrm{f}(1)=3-2\mathrm{e}\approx -2.4,\qquad \mathrm{f}(2)\approx -4.29,\qquad \mathrm{f}(3)=0+3=3.$$

所以满足 $\mathrm{f}(m)>0$ 的最小正整数是 $m=3$，而且 $2<s_0<3$（事实上 $s_0\approx 2.82$）。



b 的唯一转向点

设 $u = x/T$。根据商法则（quotient rule），

$$\mathrm{b}'(x) = \frac{3x^2(\mathrm{e}^{x/T}-1) - x^3\cdot\frac{1}{T}\mathrm{e}^{x/T}}{(\mathrm{e}^{x/T}-1)^2} = \frac{x^2\left(3\mathrm{e}^{u}-3-u\,\mathrm{e}^{u}\right)}{(\mathrm{e}^{u}-1)^2} = -\frac{x^2\,\mathrm{f}(u)}{(\mathrm{e}^{u}-1)^2}.$$

当 $x > 0$ 时，我们有 $x^2 > 0$ 及 $(\mathrm{e}^u - 1)^2 > 0$，所以 $\mathrm{b}'(x) = 0$ 当且仅当 $\mathrm{f}(x/T) = 0$。由第一部分可知，这只会在 $x/T = s_0$ 时发生，所以 b 在 $0 < x < \infty$ 内只有一个转向点（turning point），位于

$$x = s_0 T \approx 2.82\,T.$$

此外，当 $0 < u < s_0$ 时 $\mathrm{f}(u) < 0$，当 $u > s_0$ 时 $\mathrm{f}(u) > 0$，所以在这一点之前 $\mathrm{b}' > 0$，之后 $\mathrm{b}' < 0$：它是一个极大点。

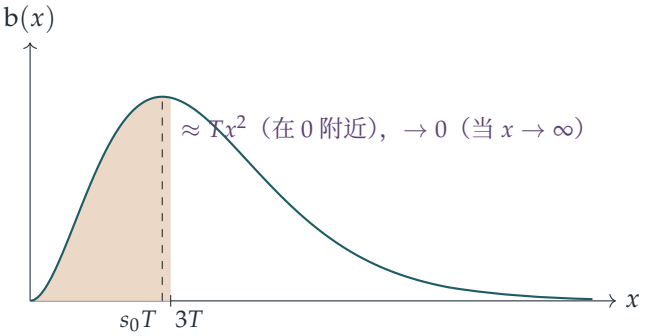
? **想一想** 在 $x = 0$ 附近，分子和分母同时趋于零，b 的图像是甚么样子的?

当 x 很小时，$\mathrm{e}^{x/T} - 1 \approx x/T$，所以

$$\mathrm{b}(x) \approx \frac{x^3}{x/T} = Tx^2,$$

图像离开原点时形如抛物线，斜率为零。（所以 $\mathrm{b}(0) = 0$ 是自然的取值。）当 x 很大时，指数函数压倒任何幂函数：$\mathrm{b}(x) \approx x^3\mathrm{e}^{-x/T} \to 0$，所以 x 轴是渐近线（asymptote）。由于对所有 $x > 0$ 都有 $\mathrm{b}(x) > 0$，图像由原点上升至位于 $x = s_0 T$ 的单一个峰，然后向横轴衰减。

利用第一部分，峰的高度是 $\mathrm{b}(s_0T) = T^3 s_0^3/(\mathrm{e}^{s_0}-1) \approx 1.42\,T^3$，而 $\mathrm{b}(3T) = 27T^3/(\mathrm{e}^3-1)$ 只是略小一点。



（草图以 $T = 1$ 绘画；阴影区域是最后一部分所用的区域。）

B 的形式

代入 $x = Tu$，则 $\mathrm{d}x = T\,\mathrm{d}u$，而由于 $T > 0$，积分上下限 $0, \infty$ 不变：

$$B = \int_0^\infty \frac{T^3u^3}{\mathrm{e}^u - 1}\,T\,\mathrm{d}u = T^4\int_0^\infty \frac{u^3}{\mathrm{e}^u - 1}\,\mathrm{d}u.$$

最后这个积分不含 T，所以它是一个常数 K（我们可假设它是有限的）。因此 $B = KT^4$，其中 $n = 4$。

估算 K

用同一个代换，并取 $m = 3$，

$$KT^4 = B \approx 2\int_0^{3T} \mathrm{b}(x)\,\mathrm{d}x = 2T^4\int_0^3 \frac{u^3}{\mathrm{e}^u - 1}\,\mathrm{d}u, \qquad \text{所以} \qquad K \approx 2\int_0^3 \frac{u^3}{\mathrm{e}^u - 1}\,\mathrm{d}u.$$

这是草图中阴影面积的两倍（取 $T = 1$）。从图像读出高度，并取 $\mathrm{e} \approx 2.7$：

$$\mathrm{b}(0) = 0, \quad \mathrm{b}(1) \approx \frac{1}{1.7} \approx 0.59, \quad \mathrm{b}(2) \approx \frac{8}{6.29} \approx 1.27, \quad \mathrm{b}(3) \approx \frac{27}{18.7} \approx 1.45.$$

以阔度为 1 的长条使用梯形法则（trapezium rule），得

$$\int_0^3 \approx \tfrac{1}{2}(0) + 0.59 + 1.27 + \tfrac{1}{2}(1.45) \approx 2.58,$$

所以

$$K \approx 5.$$

更粗略的估算是把阴影区域当作底为 3、高约 1.45 的三角形，得面积约 2.2，即 $K \approx 4.3$。无论用哪种方法，K 大约是 4 至 5。（三角形的估算必定偏低，因为曲线位于由原点至 $(3, \mathrm{b}(3))$ 的直线之上。）

验算

- 以数值二分法（bisection）求得 f 的根为 $s_0 \approx 2.8214$；取 $T = 1.7$ 对 b′ 作数值扫描，恰好找到一次变号，位于 $x \approx 4.797 \approx 1.7 s_0$。
- 用辛普森法则（Simpson's rule）计算，当 $T = 0.5$、1 及 2.3 时都得 $B/T^4 \approx 6.4939$：每次的值相同，印证了 $n = 4$。这与已知的精确值 $K = \pi^4/15 \approx 6.494$ 吻合。
- 积分 $\int_0^3 u^3/(\mathrm{e}^u - 1)\,\mathrm{d}u \approx 2.552$，所以题目容许的近似给出 $K \approx 5.10$。我们估算的 5 大致正确。与 6.49 之间的差距，是题目的近似 $B \approx 2\int_0^{3T}$ 本身的误差，而不是我们计算上的误差。

常见错误

- 忘记 $\mathrm{f}(0) = 0$：$s = 0$ 是 f 的一个根，但它对应 $x = 0$，不在 $0 < x < \infty$ 之内，所以转向点仍然只有一个。
- 在 B 中作变量代换时，忘记由 $\mathrm{d}x = T\,\mathrm{d}u$ 而来的因子 T，结果得出 $n = 3$。

STEP 之外

函数 b 就是黑体辐射（black-body radiation）的普朗克定律（Planck's law），其中 x 担当频率的角色，T 则是温度。峰的位置 $x = s_0 T$ 是维恩位移定律（Wien's displacement law），而 $B \propto T^4$ 是斯特藩–玻尔兹曼定律（Stefan–Boltzmann law）。精确值 $K = \pi^4/15$ 来自展开式 $1/(\mathrm{e}^u - 1) = \sum_{k\ge1} \mathrm{e}^{-ku}$，由此得 $K = 6\sum_{k\ge1} 1/k^4 = 6\zeta(4)$。论证求和与积分可以这样互换，是典型的大学一年级分析（analysis）题目；可参阅 Abbott 的 *Understanding Analysis*。

详细解答

STEP III 1998，第 8 题

纯数学 · 球面的向量几何

> **STEP III 1998，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/98-s3-q8

关键思路

这里的一切都归结为同一个做法：把“点 $\mathbf{r}$ 位于球面上”写成纯量方程（scalar equation）$|\mathbf{r}-\mathbf{c}|^2 = a^2$，用点积（dot product）展开，然后读出所需的结果。对于 (i) 中的直线，这样会得到一个关于 λ 的二次方程，交点的数目由其判别式（discriminant）决定。对于 (ii)，我们需要球心到平面的距离；而对于 (iii)，把两个球面方程相减，得到的正是“垂直”所问及的那个点积。

解答

第 (i) 部分

点 $\mathbf{b}+\lambda\mathbf{m}$ 位于球面上，当且仅当 $(\mathbf{b}+\lambda\mathbf{m})\cdot(\mathbf{b}+\lambda\mathbf{m}) = a^2$。展开，并利用 $\mathbf{m}\cdot\mathbf{m} = 1$（因为 $\mathbf{m}$ 是单位向量（unit vector）），得

$$\lambda^2 + 2(\mathbf{b}\cdot\mathbf{m})\lambda + \mathbf{b}\cdot\mathbf{b} - a^2 = 0.$$

每一个实根 λ 给出一个交点，而不同的 λ 值给出不同的点（因为 $\mathbf{m} \neq \mathbf{0}$）。所以，有两个交点当且仅当这个二次方程有两个相异实根，即其判别式为正：

$$4(\mathbf{b}\cdot\mathbf{m})^2 - 4(\mathbf{b}\cdot\mathbf{b} - a^2) > 0 \iff a^2 > \mathbf{b}\cdot\mathbf{b} - (\mathbf{b}\cdot\mathbf{m})^2,$$

这就是所要求的条件。

当判别式为零时，恰好有一个交点：

$$a^2 = \mathbf{b}\cdot\mathbf{b} - (\mathbf{b}\cdot\mathbf{m})^2.$$

这时重根为 $\lambda = -\mathbf{b}\cdot\mathbf{m}$（即 λ 的系数的一半再取负号），所以该点为

$$\mathbf{p} = \mathbf{b} - (\mathbf{b}\cdot\mathbf{m})\,\mathbf{m}, \qquad \text{而} \qquad \mathbf{m}\cdot\mathbf{p} = \mathbf{b}\cdot\mathbf{m} - (\mathbf{b}\cdot\mathbf{m})(\mathbf{m}\cdot\mathbf{m}) = 0,$$

这里再次用到 $\mathbf{m}\cdot\mathbf{m} = 1$。所以切线（tangent line）垂直于通往切点的半径，这正是应有的结果。

第 (ii) 部分

垂直于 $\mathbf{n}$ 的平面，其方程为 $\mathbf{r}\cdot\mathbf{n} = k$，其中 k 为常数；由于 $\mathbf{n}$ 是单位向量，该平面与原点的距离为 $|k|$。选取 $\mathbf{n}$ 的正负号使 $k \geq 0$（即 $\mathbf{n}$ 由原点指向平面），则平面为

$$\mathbf{r}\cdot\mathbf{n} = l.$$

想一想 球心 $\mathbf{d}$ 与这个平面相距多远?

由 $\mathbf{d}$ 出发沿法线（normal）方向走：点 $\mathbf{d} + \mu\mathbf{n}$ 位于平面上，当 $\mathbf{d}\cdot\mathbf{n} + \mu = l$，所以 $\mu = l - \mathbf{d}\cdot\mathbf{n}$。这一点就是由 $\mathbf{d}$ 所作垂线的垂足（foot of the perpendicular），所以球心到平面的距离为 $|\mu| = |l - \mathbf{d}\cdot\mathbf{n}|$。

平面与球面相切，当且仅当这个距离等于半径（距离较小时，平面与球面相交于一个圆；距离较大时，两者不相交）。所以条件是

$$|\mathbf{d}\cdot\mathbf{n} - l| = a, \qquad \text{等价于} \qquad (\mathbf{d}\cdot\mathbf{n} - l)^2 = a^2,$$

即 $\mathbf{d}\cdot\mathbf{n} = l \pm a$。（若 $\mathbf{n}$ 改为背向平面，则平面的方程为 $\mathbf{r}\cdot\mathbf{n} = -l$，而条件变成 $|\mathbf{d}\cdot\mathbf{n} + l| = a$。）

第 (iii) 部分

第一个球面的球心为 O，第二个的球心为 D（位置向量（position vector）为 $\mathbf{d}$），两者半径均为 a。点 $\mathbf{r}$ 同时位于两个球面上，当

$$\mathbf{r}\cdot\mathbf{r} = a^2 \qquad \text{及} \qquad (\mathbf{r} - \mathbf{d})\cdot(\mathbf{r} - \mathbf{d}) = a^2.$$

把第二式展开再减去第一式，得 $-2\,\mathbf{r}\cdot\mathbf{d} + \mathbf{d}\cdot\mathbf{d} = 0$，所以每一个交点都满足

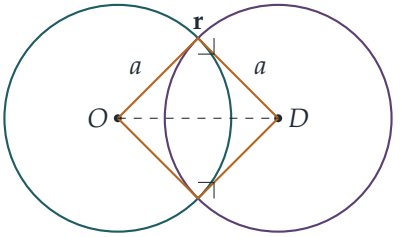
$$\mathbf{r}\cdot\mathbf{d} = \tfrac{1}{2}\,\mathbf{d}\cdot\mathbf{d}.$$

通往点 $\mathbf{r}$ 的两条半径是向量 $\mathbf{r}$（由 O 出发）和 $\mathbf{r} - \mathbf{d}$（由 D 出发）。两者垂直，当

$$\mathbf{r}\cdot(\mathbf{r} - \mathbf{d}) = \mathbf{r}\cdot\mathbf{r} - \mathbf{r}\cdot\mathbf{d} = a^2 - \tfrac{1}{2}\,\mathbf{d}\cdot\mathbf{d}$$

等于零。右边的式子与我们所选的交点无关，所以若 $\mathbf{d}\cdot\mathbf{d} = 2a^2$，则在每一个交点处两条半径都互相垂直，这正是我们要证明的。（这个论证也可以倒过来：只要在其中一个公共点处两条半径互相垂直，便有 $\mathbf{d}\cdot\mathbf{d} = 2a^2$。）

在这个情况下，两个球面确实相交：$|\mathbf{d}| = \sqrt{2}\,a$ 严格介乎 0 与 $2a$ 之间，所以两者相交于一个圆。通过 O 和 D 的截面（cross-section）显示了其中的直角。



在平面上，两个半径为 a 的圆若在公共点处的半径互相垂直，则 OD 是一个等腰直角三角形的斜边，所以 $OD^2 = a^2 + a^2 = 2a^2$：用毕氏定理（Pythagoras' theorem）得到同一答案。

验算

- 在 (i) 中，取 $\mathbf{b} = \mathbf{0}$：直线通过球心，条件变成 $a^2 > 0$，而这样的直线确实总与球面相交于两点。
- 在 (ii) 中，若球心在原点（$\mathbf{d} = \mathbf{0}$），条件就是 $l = a$：平面与球心的距离为 a 时，便与球面相切。
- 数值上，取 $a = 1$ 及 $|\mathbf{d}| = 1.2$（而非 $\sqrt{2}$），其中一个交点为 $(0.6, 0.8, 0)$，而 $\mathbf{r}\cdot(\mathbf{r}-\mathbf{d}) = 1 - 0.72 = 0.28 \neq 0$，与 $a^2 - \frac{1}{2}\mathbf{d}\cdot\mathbf{d}$ 吻合。对全部三个部分所作的随机测试，结果的误差均在 10^{-13} 以内。

常见错误

- 忘记 $\mathbf{m}\cdot\mathbf{m} = 1$，以致 λ^2 带着一个不必要的系数。
- 把 (ii) 中的相切条件只写成 $\mathbf{d}\cdot\mathbf{n} - l = a$：球心可以在平面的任何一侧。

STEP 之外

$\mathbf{b}\cdot\mathbf{b} - (\mathbf{b}\cdot\mathbf{m})^2$ 这个量就是 $|\mathbf{b}\times\mathbf{m}|^2$，即原点到直线的距离的平方；而 $\mathbf{b} - (\mathbf{b}\cdot\mathbf{m})\mathbf{m}$ 是 $\mathbf{b}$ 除去 $\mathbf{m}$ 方向分量后的正交投影（orthogonal projection），这个概念正是线性代数（linear algebra）的核心。以直角相交的球面称为正交（orthogonal）球面，它们在反演几何（inversive geometry）和双曲几何（hyperbolic geometry）的模型中都占有中心地位。Strang 的 *Introduction to Linear Algebra*（线性代数入门）讲述投影；Needham 的 *Visual Complex Analysis*（复分析的视觉化讲解）则展示正交圆的实际应用。

详细解答

STEP III 1998，第 9 题

力学 · 转动惯量

> **STEP III 1998，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/98-s3-q9

关键思路

实心圆锥可以看成由一叠薄圆盘堆叠而成，圆盘的半径随着与顶点的距离线性增长。只要知道薄圆盘的转动惯量（moment of inertia）（绕其轴线的，以及绕一条直径的），便可以用一个积分把各薄片的贡献加起来。至于第二条轴，我们需要垂直轴定理（perpendicular axis theorem）（把圆盘绕轴线的转动惯量化为绕直径的转动惯量）和平行轴定理（parallel axis theorem）（把该轴平移到顶点）。之后的摆动就是一个复摆（compound pendulum），其周期所需要的正是绕顶点的转动惯量。

解答

建立薄片

把顶点置于原点，圆锥的轴线沿 z 轴，于是底面就是平面 $z = h$。设 ρ 为（恒定的）密度。在高度 z 处的横截面是一个圆盘，由相似三角形可知其半径为

$$r = \frac{az}{h},$$

厚度为 δz 的薄片是一个薄圆盘，质量为 $\delta m = \rho\pi r^2\,\delta z = \rho\pi\dfrac{a^2z^2}{h^2}\,\delta z$。积分得

$$m = \int_0^h \rho\pi\frac{a^2z^2}{h^2}\,\mathrm{d}z = \tfrac{1}{3}\rho\pi a^2 h, \qquad \text{所以} \qquad \rho\pi = \frac{3m}{a^2h}.$$

绕轴线的转动惯量

质量为 M、半径为 r 的均匀圆盘，绕其轴线的转动惯量为 $\frac{1}{2}Mr^2$。（把它分割成半径为 s、质量为 $\frac{2Ms}{r^2}\,\delta s$ 的圆环；每个圆环的贡献是其质量的 s^2 倍，而 $\int_0^r \frac{2M}{r^2}s^3\,\mathrm{d}s = \frac{1}{2}Mr^2$。）

圆锥的每一块薄片的中心都在轴线上，所以各薄片的贡献可以直接相加：

$$\begin{aligned} I_{\text{轴}} &= \int_0^h \tfrac{1}{2} r^2 \, \rho\pi r^2 \, \mathrm{d}z = \tfrac{1}{2}\rho\pi \frac{a^4}{h^4} \int_0^h z^4 \, \mathrm{d}z = \tfrac{1}{10}\rho\pi a^4 h \\ &= \frac{1}{10} \cdot \frac{3m}{a^2 h} \cdot a^4 h = \frac{3}{10} m a^2, \end{aligned}$$

正如所求。

绕过顶点的轴的转动惯量

我们要用到两个定理。

- **垂直轴定理**。对于位于某平面内的薄片（lamina），绕一条垂直于该平面的轴的转动惯量，等于绕平面内任意两条互相垂直、并与该轴交于同一点的轴的转动惯量之和。
- **平行轴定理**。若质量为 M 的物体绕一条通过其质心（centre of mass）G 的轴的转动惯量为 I_G，则它绕一条与之平行、相距 d 的轴的转动惯量为 $I_G + Md^2$。

对于薄圆盘，由对称性可知两条直径是等价的，所以垂直轴定理给出 $2I_{\text{直径}} = \frac{1}{2}Mr^2$，即 $I_{\text{直径}} = \frac{1}{4}Mr^2$。

取过顶点且平行于底面的轴 ℓ（例如沿 x 轴）。在高度 z 处的薄片，其质心在 z 轴上，与 ℓ 相距 z，而它本身平行于 ℓ 的那条直径正通过该质心。由平行轴定理，这块薄片的贡献为 $\left(\frac{1}{4}r^2 + z^2\right)\delta m$。因此

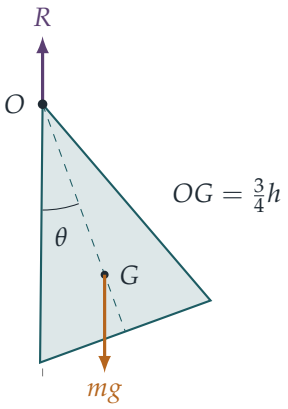
$$\begin{aligned} I_{\text{顶点}} &= \int_0^h \left(\frac{a^2 z^2}{4h^2} + z^2\right) \rho\pi \frac{a^2 z^2}{h^2} \, \mathrm{d}z = \rho\pi \frac{a^2}{h^2} \left(\frac{a^2}{4h^2} + 1\right) \frac{h^5}{5} \\ &= \frac{3m}{a^2 h} \cdot \frac{a^2 h^3}{5} \cdot \frac{a^2 + 4h^2}{4h^2} = \frac{3m}{20}\left(a^2 + 4h^2\right). \end{aligned}$$

> ? **想一想** 为甚么不能一次过对整个圆锥应用平行轴定理，由质心平移过去？

其实可以，但我们暂时还不知道绕过 G 的平行轴的转动惯量；那正是需要计算的东西。分割成薄片就可以避开这一点，因为每个圆盘绕自身直径的转动惯量是已知的，而且每个圆盘的中心都在轴线上，与 ℓ 的距离是已知的 z。（顺带可得 $I_G = I_{\text{顶点}} - m\left(\frac{3}{4}h\right)^2 = \frac{3m}{80}(4a^2 + h^2)$。）

绕顶点的小幅摆动

把圆锥视为刚体，可绕一条通过其顶点、垂直于运动平面的光滑水平轴自由转动。这样的轴平行于底面，所以绕它的转动惯量就是 $I_{\text{顶点}}$。设 θ 为圆锥轴线与铅垂向下方向之间的夹角。



轴上的反作用力对 O 没有力矩，所以唯一的力矩来自重力；重力作用于 G，与 O 相距 $\frac{3}{4}h$。它对 O 的力矩为 $mg \cdot \frac{3}{4}h\sin\theta$，其作用是使 θ 减小。对该轴取力矩，

$$I_{\text{顶点}}\,\ddot{\theta} = -\tfrac{3}{4}mgh\sin\theta.$$

对于小幅摆动，$\sin\theta \approx \theta$，所以

$$\ddot{\theta} = -\omega^2\theta, \qquad \omega^2 = \frac{\frac{3}{4}mgh}{\frac{3m}{20}(a^2+4h^2)} = \frac{5gh}{4h^2+a^2}.$$

这是简谐运动（simple harmonic motion），周期为

$$\frac{2\pi}{\omega} = 2\pi\sqrt{\frac{4h^2+a^2}{5gh}},$$

正如所求。

验算

- 单位：$\frac{4h^2+a^2}{gh}$ 的单位是 $\mathrm{m}^2/(\mathrm{m\,s^{-2}}\cdot\mathrm{m}) = \mathrm{s}^2$，与周期的平方应有的单位一致。
- 细长圆锥（$a \to 0$）：全部质量都集中在轴线上，而 $\int z^2\,\mathrm{d}m = \frac{3}{5}mh^2$，与 $I_{\text{顶点}} \to \frac{3}{5}mh^2$ 吻合。周期变成 $2\pi\sqrt{4h/(5g)}$，即长度为 $\frac{4}{5}h$ 的单摆（simple pendulum）的周期；这个长度比 $\frac{3}{4}h$ 长，对于质量沿长度方向分布的物体来说，理应如此。
- 数值验算：取 $m = 1$、$a = 0.7$、$h = 2.3$，用辛普森法则（Simpson's rule）对薄片求和，以及在圆锥内以蒙地卡罗（Monte Carlo）方法抽样取点，两者都得出 $I_{\text{轴}} = 0.147$ 和 $I_{\text{顶点}} \approx 3.2475$，与公式吻合；取 $g = 9.81$，由小振幅出发对 $I\ddot{\theta} = -\frac{3}{4}mgh\sin\theta$ 逐步作数值计算，得出周期为 2.7525 s，与公式的预测一致。

常见错误

- 在第二部分对每块薄片使用 $\frac{1}{2}r^2\,\delta m$：绕一条平行于底面的轴时，圆盘是绕其直径转动的，贡献为 $\frac{1}{4}r^2\,\delta m$，而不是 $\frac{1}{2}r^2\,\delta m$。
- 由顶点而不是由质心出发应用平行轴定理：只有当两条轴的其中一条通过质心时，该定理才成立。
- 把圆锥当作长度为 $\frac{3}{4}h$ 的单摆；决定周期的是转动惯量，而不仅仅是 G 的位置。

STEP 之外

绕过某一点的任意轴的转动惯量，全部记录在一个对称的 3×3 矩阵之中，这个矩阵称为惯性张量（inertia tensor）；垂直轴定理和平行轴定理，正是它在旋转和平移下如何变化的特殊情况。它的特征向量（eigenvectors）称为主轴（principal axes），可以解释为甚么旋转的物体绕某些轴转动时稳定，绕另一些轴时却会摇晃。Kleppner 与 Kolenkow 合著的 *An Introduction to Mechanics*（力学入门）以这个程度讲述刚体转动和复摆。

详细解答

STEP III 1998，第 10 题

力学 · 斜碰撞

> **STEP III 1998，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/98-s3-q10

关键思路

由于两球质量相等，质量在两条守恒定律中都可以约去，剩下两个简洁的向量关系：速度之和不变，以及速率平方之和不变。比较 $|\mathbf{a}+\mathbf{b}|^2$ 与 $|\mathbf{a}|^2+|\mathbf{b}|^2$，便可分离出纯量积（scalar product）$\mathbf{a}\cdot\mathbf{b}$，而整条题目的每一部分都由这一个量带动。在最后一部分，只要知道第二个球的速度，动量便决定了第一个球的速度，于是能量方程变成只含一个未知速率的二次方程，而有实根的二次方程，其判别式（discriminant）必为非负。

解答

把两球看作质量同为 m 的光滑球体，在光滑桌面上运动，因此碰撞期间水平方向的力只有两球互相施加的力。

守恒定律与相对速率

由动量（momentum）和动能（kinetic energy）守恒，得

$$m\mathbf{v}_1+m\mathbf{v}_2=m\mathbf{v}_1'+m\mathbf{v}_2',$$
$$\tfrac{1}{2}m|\mathbf{v}_1|^2+\tfrac{1}{2}m|\mathbf{v}_2|^2=\tfrac{1}{2}m|\mathbf{v}_1'|^2+\tfrac{1}{2}m|\mathbf{v}_2'|^2,$$

即

$$\mathbf{v}_1+\mathbf{v}_2=\mathbf{v}_1'+\mathbf{v}_2' \quad (1), \qquad |\mathbf{v}_1|^2+|\mathbf{v}_2|^2=|\mathbf{v}_1'|^2+|\mathbf{v}_2'|^2 \quad (2).$$

把 (1) 的两边各自与自身取纯量积：

$$|\mathbf{v}_1|^2+|\mathbf{v}_2|^2+2\,\mathbf{v}_1\cdot\mathbf{v}_2=|\mathbf{v}_1'|^2+|\mathbf{v}_2'|^2+2\,\mathbf{v}_1'\cdot\mathbf{v}_2'.$$

减去 (2)，余下

$$\mathbf{v}_1 \cdot \mathbf{v}_2 = \mathbf{v}_1' \cdot \mathbf{v}_2'. \qquad (3)$$

相对速率（relative speed）的平方是

$$|\mathbf{v}_1 - \mathbf{v}_2|^2 = |\mathbf{v}_1|^2 + |\mathbf{v}_2|^2 - 2\,\mathbf{v}_1 \cdot \mathbf{v}_2,$$

而根据 (2) 和 (3)，右边每一项在碰撞前后都相同。所以 $|\mathbf{v}_1 - \mathbf{v}_2|^2 = |\mathbf{v}_1' - \mathbf{v}_2'|^2$，又因为速率为非负，故 $|\mathbf{v}_1 - \mathbf{v}_2| = |\mathbf{v}_1' - \mathbf{v}_2'|$：相对速率守恒。

其中一球起初静止

设 $\mathbf{v}_2 = \mathbf{0}$。则由 (3) 得

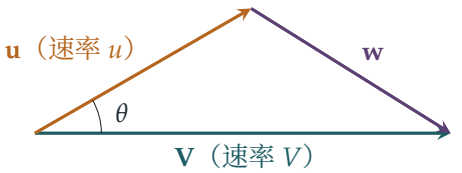
$$\mathbf{v}_1' \cdot \mathbf{v}_2' = \mathbf{v}_1 \cdot \mathbf{0} = 0.$$

由于碰撞后两球都在运动，$\mathbf{v}_1' \neq \mathbf{0}$ 且 $\mathbf{v}_2' \neq \mathbf{0}$，所以纯量积为零即表示两个末速度互相垂直。（若没有这个条件，纯量积为零可能只是因为其中一个速度为零，例如正面碰撞（head-on collision）的情况，此时两球只是互换速度。）

损失能量的碰撞

现在 (1) 仍然成立，但 (2) 不再成立。把静止的球称为 A，运动的球称为 B。设 B 的初速度为 $\mathbf{V}$，其中 $|\mathbf{V}| = V$，末速度为 $\mathbf{u}$，速率为 u，与 $\mathbf{V}$ 成角 θ。设 A 的末速度为 $\mathbf{w}$。

由动量守恒得 $\mathbf{V} = \mathbf{u} + \mathbf{w}$，所以三个速度构成一个三角形：



因此 $\mathbf{w} = \mathbf{V} - \mathbf{u}$，而由于 $\mathbf{V} \cdot \mathbf{u} = Vu\cos\theta$，

$$|\mathbf{w}|^2 = V^2 - 2Vu\cos\theta + u^2$$

（即三角形中的余弦定理（cosine rule））。最终动能是初始动能的 $1-k$ 倍。约去 $\frac{1}{2}m$，

$$u^2 + |\mathbf{w}|^2 = (1-k)V^2 \quad \Longrightarrow \quad 2u^2 - 2Vu\cos\theta + V^2 = (1-k)V^2,$$

所以第二个球的末速率 u 满足

$$\boxed{2u^2 - 2Vu\cos\theta + kV^2 = 0.}$$

想一想 题目只要求我们证明 $k \le \frac{1}{2}$。我们手上只有一条方程，不等式从何而来？

速率 u 是一个实际出现的实数，所以这条关于 u 的二次方程确实有实根。因此它的判别式必为非负：

$$4V^2\cos^2\theta - 8kV^2 \ge 0 \quad \Longrightarrow \quad k \le \tfrac{1}{2}\cos^2\theta \le \tfrac{1}{2},$$

当中我们除以 $8V^2 > 0$（第二个球在运动），并用了 $\cos^2\theta \le 1$。所以 $k \le \frac{1}{2}$，正是所求。

等号成立需要 $\cos^2\theta = 1$，并且有重根 $u = \frac{1}{2}V\cos\theta$。由于 $u \ge 0$，这表示 $\theta = 0$ 及 $u = \frac{1}{2}V$，于是 $\mathbf{w} = \frac{1}{2}\mathbf{V}$ 亦然：两球以一半的速率一同离开，这正是完全非弹性（perfectly inelastic）的正面碰撞。动能最多只能损失一半。

验算

- $k = 0$ 时得 $2u^2 = 2Vu\cos\theta$，所以 $u = V\cos\theta$（或 $u = 0$）。此时 $\mathbf{w} = \mathbf{V} - \mathbf{u}$ 与 $\mathbf{u}$ 垂直，因为 $\mathbf{u}\cdot(\mathbf{V}-\mathbf{u}) = Vu\cos\theta - u^2 = 0$。这与第二部分吻合。
- 恢复系数（coefficient of restitution）为 e 的正面碰撞给出 $k = \frac{1}{2}(1-e^2)$，其值介乎 0 与 $\frac{1}{2}$ 之间。取 $e = 0.3$，得 $k = 0.455$。
- 数值方面，我们模拟了 10^5 次质量相等的光滑球体的随机弹性碰撞（elastic collision）。动量、动能、$\mathbf{v}_1\cdot\mathbf{v}_2$ 和相对速率全部吻合，误差在 10^{-13} 以内。在 2×10^5 个随机的能量损失结果中，该二次方程成立，误差在 10^{-14} 以内，而所得最大的 k 为 0.49999。

常见错误

- 把“相对速率”当作 $\big||\mathbf{v}_1| - |\mathbf{v}_2|\big|$。它其实是相对速度的大小，即 $|\mathbf{v}_1 - \mathbf{v}_2|$。
- 在第二部分，忘记使用两球都在运动这一事实。只有当两个向量都不是零向量时，纯量积为零才表示互相垂直。
- 在最后一部分，再次使用 (2) 或 (3)。它们来自动能守恒，而动能守恒已不再成立。只有动量守恒仍然有效。

STEP 之外

这个纯量积论证，其实是一个标准技巧的不依赖参考系的版本：在零动量参考系（zero-momentum frame）中，弹性碰撞只是把两个速度旋转，而大小保持不变。最大损失为 $\frac{1}{2}$，反映的是一个普遍事实：只有相对于质心（centre of mass）的动能才可以损失，在这里就是 $\frac{1}{2}\cdot\frac{1}{2}mV^2$。这些想法通往散射理论（scattering theory），以及狭义相对论（special

relativity）中的碰撞。Kleppner 与 Kolenkow 所著的 *An Introduction to Mechanics*（力学入门教科书）对质心参考系中的碰撞有仔细的论述。

详细解答

STEP III 1998，第 11 题

力学 · 大振幅摆

STEP III 1998，第 11 题

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

step.maths.org/questions/98-s3-q11

本题在 *Beyond STEP* 第 11.3 节*真实的摆*中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch11/real-pendulum.html

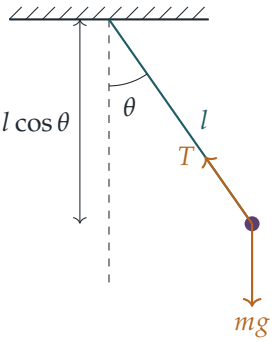
关键思路

能量守恒（energy conservation）给出摆在每个角度的角速率（angular speed），而周期（period）等于四分之一次摆动所需时间的四倍，$\int \mathrm{d}t = \int \mathrm{d}\theta/\dot{\theta}$。半角恒等式（half-angle identity）$\cos\theta = 1 - 2\sin^2(\theta/2)$ 把能量方程化成被积函数（integrand）的形式。题目建议的代换（substitution）消去了在 $\theta = \gamma$ 处那个棘手的无穷大，剩下的被积函数形如 $(1-x)^{-1/2}$，其中 x 很小，这正是二项展开式（binomial expansion）大派用场的地方。

解答

能量方程

把摆看成一个质量为 m 的质点，系在一条长度为 l 的不可伸长轻绳上，在铅垂平面内摆动，不计空气阻力。角 θ 由向下的铅垂线量起。



摆锤（bob）位于支点下方 $l\cos\theta$ 的深度。张力（tension）沿绳的方向作用，与摆锤的速度（沿圆弧方向）垂直，所以张力不作功；只有重力作功，因此机械能（mechanical energy）守恒。摆锤的速率为 $l\dot{\theta}$。

在最大摆角 $\theta=\gamma$ 处，摆锤瞬时静止。由该处到一般的角度 θ，摆锤下降的高度为 $l\cos\theta - l\cos\gamma$，所以

$$\tfrac{1}{2}ml^2\dot{\theta}^2 = mgl(\cos\theta-\cos\gamma) \quad\Longrightarrow\quad \tfrac{1}{2}l\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2 = g(\cos\theta-\cos\gamma),$$

这就是第一个结果。

周期

对两个余弦都使用 $\cos\alpha = 1-2\sin^2(\alpha/2)$，得

$$\cos\theta-\cos\gamma = 2\big(\sin^2(\gamma/2)-\sin^2(\theta/2)\big),$$

于是能量方程变成

$$\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2 = \frac{4g}{l}\big(\sin^2(\gamma/2)-\sin^2(\theta/2)\big).$$

根据对称性（运动是可逆的，而且能量方程在 $\theta\mapsto-\theta$ 之下不变），四段四分之一摆动 $0\to\gamma\to0\to-\gamma\to0$ 所需的时间相等。在由 $\theta=0$ 到 $\theta=\gamma$ 的那四分之一摆动中，θ 递增，所以取正平方根：

$$\frac{\mathrm{d}t}{\mathrm{d}\theta} = \frac{1}{2}\sqrt{\frac{l}{g}}\left(\sin^2(\gamma/2)-\sin^2(\theta/2)\right)^{-1/2}.$$

由 0 积分至 γ，再乘以 4，得

$$P = 2\sqrt{\frac{l}{g}}\int_0^{\gamma}\left(\sin^2(\gamma/2)-\sin^2(\theta/2)\right)^{-1/2}\mathrm{d}\theta,$$

这正是所求。（被积函数在 $\theta=\gamma$ 处趋向无穷大，因为摆锤在该处静止，但积分是收敛（converge）的；下面的代换会清楚显示这一点。）

小振幅近似

记 $k=\sin(\gamma/2)$，其中 $0<\gamma<\pi$，因而 $0<k<1$，并设 $\sin(\theta/2)=k\sin\phi$。当 θ 由 0 变到 γ 时，$\sin\phi$ 由 0 变到 1，所以 ϕ 由 0 变到 $\pi/2$。求导后得

$$\tfrac{1}{2}\cos(\theta/2)\,\mathrm{d}\theta = k\cos\phi\,\mathrm{d}\phi.$$

由于 $0\le\theta/2<\pi/2$ 及 $0\le\phi\le\pi/2$，两个余弦都非负，所以

$$\cos(\theta/2)=\sqrt{1-k^2\sin^2\phi},\qquad \left(k^2-k^2\sin^2\phi\right)^{1/2}=k\cos\phi.$$

> **想一想** 在这个代换之下，那个麻烦的因子 $(\sin^2(\gamma/2)-\sin^2(\theta/2))^{-1/2}$ 会变成甚么?

它变成 $1/(k\cos\phi)$，而 $\mathrm{d}\theta$ 中的 $k\cos\phi$ 恰好把它完全约去：

$$\frac{\mathrm{d}\theta}{\left(k^2-\sin^2(\theta/2)\right)^{1/2}}=\frac{1}{k\cos\phi}\cdot\frac{2k\cos\phi\,\mathrm{d}\phi}{\sqrt{1-k^2\sin^2\phi}}=\frac{2\,\mathrm{d}\phi}{\sqrt{1-k^2\sin^2\phi}}.$$

因此

$$P=4\sqrt{\frac{l}{g}}\int_0^{\pi/2}\left(1-k^2\sin^2\phi\right)^{-1/2}\mathrm{d}\phi,$$

这个积分的被积函数性质十分良好。（到此为止，一切都是精确的。）

当 γ 很小时，$x=k^2\sin^2\phi$ 满足 $0\le x\le k^2$，而这个上界很小，所以由二项展开式 $(1-x)^{-1/2}=1+\frac{1}{2}x+\frac{3}{8}x^2+\cdots$ 得

$$\left(1-k^2\sin^2\phi\right)^{-1/2}\approx 1+\tfrac{1}{2}k^2\sin^2\phi,$$

误差的阶为 k^4。由于 $\int_0^{\pi/2}\sin^2\phi\,\mathrm{d}\phi=\int_0^{\pi/2}\frac{1}{2}(1-\cos 2\phi)\,\mathrm{d}\phi=\frac{\pi}{4}$，

$$P\approx 4\sqrt{\frac{l}{g}}\left(\frac{\pi}{2}+\frac{k^2}{2}\cdot\frac{\pi}{4}\right)=2\pi\sqrt{\frac{l}{g}}\left(1+\frac{k^2}{4}\right).$$

最后，当 γ 很小时，$k=\sin(\gamma/2)\approx\gamma/2$（误差的阶为 γ^3），所以 $k^2\approx\gamma^2/4$，只相差阶为 γ^4 的项，于是

$$P\approx 2\pi\sqrt{\frac{l}{g}}\left(1+\frac{\gamma^2}{16}\right),$$

这正是所求。首项是大家熟悉的小幅振荡周期；修正项则说明摆幅愈大，摆动所需的时间略长。

验算

- 当 $\gamma \to 0$ 时，我们重新得到 $2\pi\sqrt{l/g}$，即由线性化方程 $\ddot{\theta} = -(g/l)\theta$ 所得的周期。
- 量纲（dimensions）：$\sqrt{l/g}$ 是时间，而 γ 是无量纲的。
- 数值上，取 $l/g = 1.7/9.81$ 及 $\gamma = 0.5$，以步长很小的龙格－库塔（Runge–Kutta）方法逐步求解方程 $\ddot{\theta} = -(g/l)\sin\theta$，得 $P = 2.65705$，与精确的 ϕ 积分在所示的各位数字上完全吻合；近似公式则给出 2.65646，相对误差约为 2×10^{-4}。该误差与 γ^4 成比例（约为 $0.0036\gamma^4$），由此确认 γ^2 项是正确的。

常见错误

- 忘记因子 4：由 0 到 γ 的积分只是四分之一个周期。
- 在展开之前，就在代换中使用 $\cos(\theta/2) \approx 1$。这样会丢掉一个 $k^2\sin^2\phi$ 项，而它与你要找的修正项同阶；应把 $\sqrt{1 - k^2\sin^2\phi}$ 保持精确，直到使用二项展开的那一步。
- 停在 $1 + k^2/4$，没有把 $k = \sin(\gamma/2)$ 换成 $\gamma/2$。

STEP 之外

精确积分 $\int_0^{\pi/2}(1 - k^2\sin^2\phi)^{-1/2}\,\mathrm{d}\phi$ 就是**第一类完全椭圆积分**（complete elliptic integral of the first kind）$K(k)$，所以精确的周期是 $4\sqrt{l/g}\,K(\sin(\gamma/2))$。把二项级数继续展开下去，可得 $1 + \frac{1}{16}\gamma^2 + \frac{11}{3072}\gamma^4 + \cdots$，而高斯（Gauss）发现，利用算术几何平均（arithmetic–geometric mean）可以极快地计算 K。周期取决于振幅这一事实，是非线性振荡（nonlinear oscillation）最简单的例子。Kleppner 与 Kolenkow 合著的 *An Introduction to Mechanics*（力学入门教材）以大学一年级的程度讨论了摆与振荡。

详细解答

STEP III 1998，第 12 题

概率与统计 · 网络可靠度

STEP III 1998，第 12 题

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

step.maths.org/questions/98-s3-q12

关键思路

道路只有六条，所以道路开通与封闭的组合只有 $2^6 = 64$ 种，而最简洁的做法是数算坏的组合：即各村庄分裂成若干组、组与组之间无法互通的那些组合。把坏的组合按仍然开通的道路数目分类，数算就变得很短，因为四条村庄之间若有四条或以上的道路开通，网络（network）绝不可能不连通。

动笔之前先提醒一点：试卷上印出的公式，括号前面是 p^2，但正确的因子是 p^3。印出的版本不可能正确，我们会在“验算”一节说明。我们要证明的是修正后的结果

$$P = 1 - p^3(6p^3 - 12p^2 + 3p + 4).$$

解答

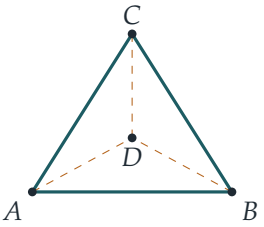
设 $q = 1 - p$ 为某一条道路开通的概率（probability）。由于各条道路的情况互相独立（independent），任何一个有 k 条道路开通、$6 - k$ 条道路封闭的特定组合，其概率为 $q^k p^{6-k}$。设 D 为网络不连通（disconnected）这一事件，即有某条村庄无法到达另一条村庄。于是 $P = 1 - \mathbb{P}(D)$，而 $\mathbb{P}(D)$ 就是对所有不连通的组合把 $q^k p^{6-k}$ 加起来的总和。我们按开通道路的数目 k 来数算这些组合。

四条或以上道路开通：绝不会不连通。 如果网络不连通，各村庄便分裂成至少两组，组与组之间没有开通的道路。可能的分法只有 $1+3$（一条孤立的村庄加上其余三条）和 $2+2$ 两种。所有开通的道路都在组内。在 $1+3$ 的分法中，三条村庄的那一组只有 3 条道路；在 $2+2$ 的分法中，每一对只有 1 条道路，合共 2 条。无论哪一种，最多只能有 3 条道路开通。所以凡是有 $k \geq 4$ 条道路开通的组合都是连通（connected）的。

恰好三条道路开通。

? **想一想** 有三条道路开通时，哪些组合无法把四条村庄全部连接起来?

根据上面的论证，有三条开通道路而又不连通的组合，必定是 $1+3$ 的分法，而且三条村庄那一组之内的三条道路全部开通：即一个三角形，第四条村庄被隔绝。孤立的村庄有 4 种选法，所以这样的组合有 4 个。(其余 $\binom{6}{3}-4=16$ 个组合是四面体的生成树 (spanning tree)，它们都是连通的。)



开通的道路 AB, BC, CA（实线）；
通往 D 的道路封闭（虚线）：
D 被隔绝

两条或以下道路开通：必定不连通。 要把四条村庄连接起来，至少需要三条道路（每条道路最多只能把一条新的村庄，接到从 A 出发已能到达的那一组)。所以有两条道路开通的全部 $\binom{6}{2}=15$ 个组合、有一条开通的全部 6 个，以及完全没有道路开通的那 1 个，都是不连通的。

加起来。 因此

$$\mathbb{P}(D)=4p^3q^3+15p^4q^2+6p^5q+p^6=p^3\left(4q^3+15pq^2+6p^2q+p^3\right).$$

现在以 $q=1-p$ 展开括号：

$$\begin{aligned}4(1-p)^3&=4-12p+12p^2-4p^3,\\15p(1-p)^2&=15p-30p^2+15p^3,\\6p^2(1-p)&=6p^2-6p^3,\\p^3&=p^3.\end{aligned}$$

相加后，常数项是 4，p 的各项给出 $-12p+15p=3p$，p^2 的各项给出 $12-30+6=-12$，而 p^3 的各项给出 $-4+15-6+1=6$。所以

$$\mathbb{P}(D)=p^3\left(6p^3-12p^2+3p+4\right),\qquad\text{因此}\qquad P=1-p^3\left(6p^3-12p^2+3p+4\right).$$

验算

▶ $p=0$（没有道路会封闭）给出 $P=1$，而 $p=1$ 给出 $P=1-(6-12+3+4)=0$。两者都正确，但 p^2 的版本同样满足这两点，所以它们无法分辨两个版本。

- p 很小时就能把两者区分开来。网络要不连通，至少要有三条道路封闭（例如连接同一条村庄的那三条），所以当 p 很小时，$\mathbb{P}(D)$ 应该大约是 $4p^3$。我们的公式给出 $\mathbb{P}(D) \approx 4p^3$；印出的公式则会给出 $4p^2$，这是不可能的。
- 当 $p = \frac{1}{3}$ 时，我们的公式给出 $P = 1 - \frac{1}{27} \cdot \frac{105}{27} = \frac{208}{243} \approx 0.856$。用电脑在 21 个 p 值上枚举全部 64 个组合，结果与 p^3 的公式完全吻合；另外在 $p = 0.4$ 时模拟 200 000 场降雪，得到 0.767，而公式的值是 0.7655。

常见错误

- 只数算有孤立村庄的情况，而忘记检查 $2+2$ 的分法。在这一题它们没有带来新的组合：$2+2$ 的分法最多只留下两条开通的道路，所以那些组合已经包括在“两条或以下”的数算之中。但你仍然需要把这一点说明。
- 把互相重叠的事件（例如“A 被隔绝”和“B 被隔绝”）的概率直接相加，而没有修正重叠的部分。以整个组合为单位来数算，就可避免重复计算。
- 相信印出的 p^2，并硬把代数运算凑成那个样子。如果老老实实算出来的结果与“show that”题目所给的不符，就用一个特殊情况来检验。

STEP 之外

这是网络可靠性（network reliability）的一个小例子：图（graph）的每条边各自独立地失效，而我们想求余下的图仍然连通的概率。答案必定是 p 的多项式，其系数数算的是连通生成子图（connected spanning subgraph）的数目（这里对应 $6, 5, 4, 3$ 条开通道路的是 $1, 6, 15, 16$）。对于大型的图，这个多项式在计算上很难求出，并且与 Tutte 多项式（Tutte polynomial）和渗流理论（percolation theory）密切相关。Grimmett 与 Welsh 的 *Probability: An Introduction*（概率论入门）是打好概率基础的好选择；Bollobás 的 *Modern Graph Theory*（现代图论）则涵盖随机图（random graph）和 Tutte 多项式。

详细解答

STEP III 1998，第 13 题

概率与统计 · 最大似然估计

STEP III 1998，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/98-s3-q13

关键思路

这里的未知数是 n，而它是整数，所以微积分（calculus）派不上用场：我们无法对 n 求导。要求一个数列（sequence）的最大值，标准的工具是考察相邻两项的比值（ratio）。比值大于 1 时数列上升，一旦跌到 1 以下数列便下降；而出现并列的情况（即有两个最大似然估计，maximum likelihood estimate，简称 MLE），正正就是比值在某处等于 1 的时候。

解答

掷一枚公平硬币（fair coin）n 次，所得正面的次数服从二项分布（binomial distribution），所以

$$\mathrm{P}(k \text{ 次正面，共掷} n \text{ 次}) = \binom{n}{k}\frac{1}{2^n}, \qquad n \geq k,$$

而若 $n < k$，这个概率是 0。以下一律取 $k \geq 1$（那位朋友至少见过一次正面），所以只有 $n \geq k$ 才是可能的答案。

公平硬币的一般计算

对 $n \geq k$，记 $L(n) = \binom{n}{k}2^{-n}$，把它看成 k 固定时 n 的函数。对 $n \geq k$，

$$\frac{L(n+1)}{L(n)} = \frac{\binom{n+1}{k}}{\binom{n}{k}} \cdot \frac{1}{2} = \frac{n+1}{n+1-k} \cdot \frac{1}{2},$$

其中用到 $\binom{n+1}{k} = \frac{n+1}{n+1-k}\binom{n}{k}$。由于 $n+1-k > 0$，我们可以消去分母而无须改变不等号的方向：

$$\frac{L(n+1)}{L(n)} \geq 1 \iff n+1 \geq 2(n+1-k) \iff n \leq 2k-1,$$

等号恰好在 $n = 2k-1$ 时成立。因此

$$L(k) < L(k+1) < \cdots < L(2k-1) = L(2k) > L(2k+1) > L(2k+2) > \cdots .$$

想一想 比值怎样告诉我们最大值在哪里?

当 $n < 2k-1$ 时, 比值大于 1, 所以每一项都比前一项大: 数列严格递增, 直至 $n = 2k-1$。在 $n = 2k-1$ 时，比值恰好是 1，所以 $L(2k) = L(2k-1)$。当 $n > 2k-1$ 时，比值小于 1，数列自此严格递减。因此最大值恰好在两个 n 的值上取得。

一次正面

取 $k = 1$，两个 MLE 是 $n = 2k-1 = 1$ 和 $n = 2k = 2$。直接计算,

$$L(1) = \tfrac{1}{2}, \quad L(2) = 2 \cdot \tfrac{1}{4} = \tfrac{1}{2}, \quad L(3) = 3 \cdot \tfrac{1}{8} = \tfrac{3}{8},$$

而上面的论证表明 $L(n)$ 在 $n = 2$ 之后一直递减。所以有两个 MLE，即 1 和 2，正如题目所要求。

k 次正面

根据一般计算，掷硬币次数的两个 MLE 是

$$n = 2k-1 \quad 及 \quad n = 2k.$$

偏差硬币

现在设 $0 < p < 1$ 及 $q = 1-p$，于是

$$L(n) = \binom{n}{k} p^k q^{n-k}, \qquad n \geq k.$$

对 $n \geq k$，因子 p^k 互相抵消，并多出一个因子 q:

$$\frac{L(n+1)}{L(n)} = \frac{(n+1)q}{n+1-k}.$$

同样有 $n+1-k > 0$，所以

$$\frac{L(n+1)}{L(n)} \gtreqless 1 \iff (n+1)(1-p) \gtreqless n+1-k \iff k \gtreqless (n+1)p \iff n+1 \lesseqgtr \frac{k}{p}.$$

因此 L 在 $n+1 < k/p$ 时递增，在 $n+1 = k/p$ 时持平，而一旦 $n+1 > k/p$ 便递减。留意 $k/p > k$，因为 $p < 1$，所以转折点落在 $n \geq k$ 这个范围之内。

- 若 k/p 不是整数，设 $m = \lfloor k/p \rfloor$。那么 $n+1 < k/p$ 当且仅当 $n+1 \le m$，所以 $L(k) < \cdots < L(m)$，而 $L(m+1) < L(m)$，其后继续递减。此时只有一个 MLE，
$$n = \left\lfloor \frac{k}{p} \right\rfloor.$$
- 若 $k/p = N$ 是整数，则 L 一直递增至 $L(N-1)$，并有 $L(N) = L(N-1)$，然后递减。此时有两个 MLE，
$$n = \frac{k}{p} - 1 \quad \text{及} \quad n = \frac{k}{p}.$$
两者都不小于 k：由于 N 是整数且 $N > k$，我们有 $N - 1 \ge k$。

所以 MLE 唯一**当且仅当** k/p **不是整数**。对公平硬币而言，$k/p = 2k$ 总是整数，这就是前面各部分总是得到两个 MLE，即 $2k-1$ 和 $2k$ 的原因。

验算

- 答案合乎情理：若有 k 次正面，而每次掷出正面的概率是 p，我们预计大约掷了 k/p 次，而 MLE 正是 k/p 向下取整（若它是整数，则是其左右并列的两个值）。
- 取 $p = \frac{1}{3}$，$k = 1$，于是 $k/p = 3$。那么 $L(2) = 2 \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{4}{9}$ 而 $L(3) = 3 \cdot \frac{1}{3} \cdot \frac{4}{9} = \frac{4}{9}$，两者并列，而 $L(4) = 4 \cdot \frac{1}{3} \cdot \frac{8}{27} = \frac{32}{81} < \frac{4}{9}$。
- 我们用电脑以分数精确计算 $L(n)$，并找出取得最大值的位置：公平硬币取 $k = 1, \ldots, 39$（结果总是 $\{2k-1, 2k\}$），另外还有 924 对 (p,k)，其中 p 是分母不超过 12 的分数，且 $k \le 14$。每一个情况都与上述规则吻合。

常见错误

- 对 n 求导。这里 n 是整数，而若没有额外的工具，$\binom{n}{k}$ 对其他的 n 值并无定义；比值判别法（ratio test）才是合适的工具。
- 把不等号的方向弄反，或者忘记了正是比值等于 1 才产生第二个 MLE。
- 在所有情况下都答 $\lfloor k/p \rfloor$，而遗漏了 k/p 是整数时的并列情况。

STEP 之外

最大似然法（maximum likelihood）是统计学中最核心的估计方法：选取令观察所得数据出现的概率最大的参数值。参数通常是连续的，我们会用微积分求对数似然（log-likelihood）的最大值。估计像 n 这样的整数（即“试验次数”问题），则是这个做法失效

的经典例子，而本题更显示 MLE 甚至不一定唯一。Blitzstein 与 Hwang 的 *Introduction to Probability*（概率入门）对二项分布讲解得很好，而 Grimmett 与 Welsh 的 *Probability: An Introduction*（概率导论）则涵盖了相关的概率背景知识。

详细解答

STEP III 1998，第 14 题

概率与统计 · 无偏估计量

STEP III 1998，第 14 题

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

step.maths.org/questions/98-s3-q14

本题在 *Beyond STEP* 第 13.3 节德国坦克问题中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch13/german-tanks.html

关键思路

Z_2 是若干独立（independent）部分之和，所以它的平均值和方差（variance）可直接由单一个均匀分布（uniform distribution）变量的平均值和方差得出。Z_1 取决于最大值，而处理最大值最自然的方法，是从它的分布函数（distribution function）入手：各个 X_i 中最大的一个不超过 m，当且仅当每一个 X_i 都不超过 m。这样便化成若干概率的乘积，再求导就得到密度函数（density）。

解答

Z_2 的平均值

每个 X_i 在 $[0, N]$ 上的密度为 $1/N$，所以

$$\mathrm{E}(X_i) = \frac{N}{2}, \qquad \mathrm{E}(X_i^2) = \frac{1}{N}\int_0^N x^2\,\mathrm{d}x = \frac{N^2}{3}, \qquad \mathrm{Var}(X_i) = \frac{N^2}{3} - \frac{N^2}{4} = \frac{N^2}{12}.$$

由期望值的线性性质（linearity of expectation），

$$\mathrm{E}(Z_2) = \frac{2}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{2}{n}\cdot n\cdot\frac{N}{2} = N.$$

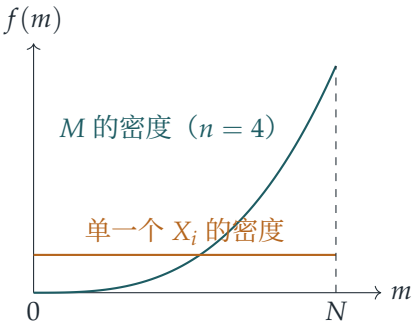
最大值的分布

设 $M = \max\{X_1, \ldots, X_n\}$。对于 $0 \le m \le N$，事件 $M \le m$ 发生，当且仅当对每个 i 都有 $X_i \le m$；而各个 X_i 互相独立，所以

$$\mathrm{P}(M \le m) = \prod_{i=1}^{n} \mathrm{P}(X_i \le m) = \left(\frac{m}{N}\right)^n.$$

求导后，得 M 的密度为

$$f(m) = \frac{n\, m^{n-1}}{N^n}, \qquad 0 \le m \le N.$$



最大值集中在 N 附近，所以只需乘上一个很小的修正因子，便可成为无偏（unbiased）估计量。它的矩（moments）为

$$\mathrm{E}(M) = \int_0^N m \cdot \frac{n\, m^{n-1}}{N^n}\, \mathrm{d}m = \frac{n}{N^n} \cdot \frac{N^{n+1}}{n+1} = \frac{n}{n+1} N,$$
$$\mathrm{E}(M^2) = \int_0^N m^2 \cdot \frac{n\, m^{n-1}}{N^n}\, \mathrm{d}m = \frac{n}{n+2} N^2.$$

Z_1 的平均值

$$\mathrm{E}(Z_1) = \frac{n+1}{n}\,\mathrm{E}(M) = \frac{n+1}{n} \cdot \frac{n}{n+1} N = N.$$

因此两个估计量（estimator）的平均值都是 N，正如题目所求。

方差

对于 Z_1，首先

$$\mathrm{Var}(M) = \frac{n}{n+2}N^2 - \frac{n^2}{(n+1)^2}N^2 = \frac{n\left[(n+1)^2 - n(n+2)\right]}{(n+2)(n+1)^2}N^2 = \frac{n}{(n+2)(n+1)^2}N^2,$$

因为 $(n+1)^2 - n(n+2) = 1$。乘以 $\left(\frac{n+1}{n}\right)^2$ 后，

$$\boxed{\mathrm{Var}(Z_1) = \frac{(n+1)^2}{n^2} \cdot \frac{n}{(n+2)(n+1)^2} N^2 = \frac{N^2}{n(n+2)}.}$$

对于 Z_2，各个 X_i 互相独立，所以它们的方差可以相加：

$$\boxed{\operatorname{Var}(Z_2) = \frac{4}{n^2}\sum_{i=1}^{n}\operatorname{Var}(X_i) = \frac{4}{n^2}\cdot n\cdot\frac{N^2}{12} = \frac{N^2}{3n}.}$$

想一想 哪一个较小？是否永远都是同一个较小？

两者之比为

$$\frac{\operatorname{Var}(Z_1)}{\operatorname{Var}(Z_2)} = \frac{3n}{n(n+2)} = \frac{3}{n+2}.$$

当 $n = 1$ 时，这个比等于 1（此时 $Z_1 = Z_2 = 2X_1$，两个估计量其实是同一个）；而对每个 $n \geq 2$，它都小于 1。

应选哪一个

两者都是无偏的，所以我们比较它们的离散程度。我们选择 Z_1：

- 当 $n \geq 2$ 时，它的方差严格较小，而且 n 越大，优势越明显：$\operatorname{Var}(Z_1)$ 以 $1/n^2$ 的速度下降，而 $\operatorname{Var}(Z_2)$ 只以 $1/n$ 的速度下降。例如当 $n = 10$ 时，Z_1 的标准差（standard deviation）约为 $0.09N$，而 Z_2 的则约为 $0.18N$。
- Z_1 必定不小于所见到的最大识别号码，而 N 亦必须如此。Z_2 却可能比我们实际观察到的某个号码还要小（例如读数为 $1, 2, 30$ 时，得 $Z_2 = 22$），给出一个我们明知不可能的估计值。

验算

- 当 $n = 1$ 时，两条公式都给出 $N^2/3$；这是必然的，因为 $Z_1 = Z_2 = 2X_1$，而 $\operatorname{Var}(2X_1) = 4\cdot N^2/12$。
- 以 $N = 10$ 进行 200 000 次模拟试验，两个估计量的样本平均值与 10 相差都不超过 0.01；当 $n = 4$ 时，样本方差分别为 4.16 和 8.33，而理论值为 $\frac{100}{24} \approx 4.17$ 和 $\frac{100}{12} \approx 8.33$。

常见错误

- 试图直接处理最大值的密度。应先求 $\mathrm{P}(M \leq m)$；由于互相独立，它只是一个简单的乘积。
- 忘记把常数平方：$\operatorname{Var}(cY) = c^2\operatorname{Var}(Y)$，所以两个因子分别是 $\left(\frac{n+1}{n}\right)^2$ 和 $\frac{4}{n^2}$。

STEP 之外

这就是经典的“德国坦克问题”（German tank problem）：第二次世界大战期间，盟军的统计学家利用坦克的序号估计德国的坦克产量，结果准确得惊人。在这里，最大值是一个充分统计量（sufficient statistic）：一旦知道了它，样本的其余部分就再不能提供关于 N 的任何额外信息。莱曼–谢费定理（Lehmann–Scheffé theorem）进一步指出，在所有无偏估计量之中，Z_1 的方差最小。充分性和有效估计量（efficient estimator）是大学一年级统计学的课题；若想认识顺序统计量（order statistics），Blitzstein 与 Hwang 合著的 *Introduction to Probability*（概率论入门）是很好的起点。

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

1998 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。

