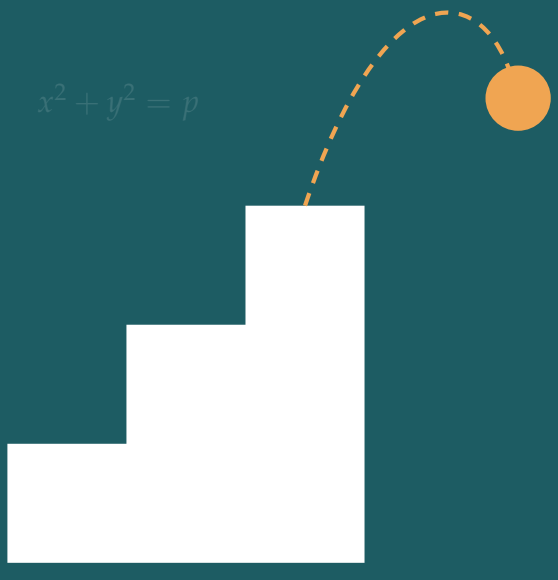
STEP 备试 · 大学数学

Beyond STEP

STEP I, II 及 III

2002 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

2002 年详细解答

Keith McNulty

Beyond STEP：以历届 STEP 试题为门，走进大学数学

2002 年详细解答

2002 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 圆与椭圆 纯数学 . . . 3

第 2 题 曲线描绘 纯数学 . . . 6

第 3 题 以微积分证明不等式 纯数学 . . . 10

第 4 题 切线与不等式 纯数学 . . . 14

第 5 题 多项式的整数根 纯数学 . . . 18

第 6 题 角锥的几何 纯数学 . . . 21

第 7 题 三角函数积分 纯数学 . . . 25

第 8 题 复利贷款 纯数学 . . . 29

第 9 题 斜面上的倾倒 力学 . . . 32

第 10 题 弹性碰撞 力学 . . . 36

第 11 题 弹性斜碰撞 力学 . . . 40

第 12 题 几何概率 概率与统计 . . . 43

第 13 题 条件期望值 概率与统计 . . . 47

第 14 题 概率中的等比级数 概率与统计 . . . 50

STEP II

第 1 题 代换积分法 纯数学 . . . 53

第 2 题 倒数多项式方程 纯数学 . . . 56

第 3 题 费马数与质数 纯数学 . . . 60

第 4 题 由积分推导不等式 纯数学 . . . 63

第 5 题 逻辑斯谛映射的迭代 纯数学 . . . 66

第 6 题 斜面上的直线 纯数学 . . . 70

第 7 题 直线的向量几何 纯数学 . . . 73

第 8 题 一阶微分方程 纯数学 . . . 77

第 9 题 抛体运动 力学 81
第 10 题 动学与最优化 力学 85
第 11 题 摩擦力与极限平衡 力学 89
第 12 题 顺态近似与泊松近似 概率与统计 93
第 13 题 变换后的分布函数 概率与统计 97
第 14 题 混合分布与贝叶斯定理 概率与统计 100

STEP III

第 1 题 广义积分与体积 纯数学 104
第 2 题 反正切级数 纯数学 107
第 3 题 曲线描绘 纯数学 110
第 4 题 丢番图不等式 纯数学 114
第 5 题 四次方程的因式分解 纯数学 117
第 6 题 非线性微分方程 纯数学 120
第 7 题 坐标几何恒等式 纯数学 124
第 8 题 托勒密定理 纯数学 128
第 9 题 质心与倾倒 力学 132
第 10 题 滑及摩擦力的转动 力学 136
第 11 题 墙壁的斜碰撞 力学 140
第 12 题 条件化求期望值 概率与统计 145
第 13 题 整数分布变量的最小值 概率与统计 149
第 14 题 方差与协方差 概率与统计 152

详细解答

STEP I 2002，第 1 题

纯数学 · 圆与椭圆

STEP I 2002，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/02-s1-q1

关键思路

先找出两个椭圆（ellipse）实际相交之处。两条方程都只通过 y^2 含有 y，所以消去 y^2 后便得到一条 x 的二次方程，而结果显示两个椭圆只相交于两个实点。通过两个已知点的圆，其圆心必在这两点的垂直平分线（perpendicular bisector）上，这样便把圆的方程确定为一个单参数族（one-parameter family）。

解答

求交点

展开后，两个椭圆的方程为

$$x^2 + 4x + 2y^2 = 14,$$
$$9x^2 - 18x + 16y^2 = 16.$$

把第一式乘以 8，再减去第二式，便可消去 y^2：

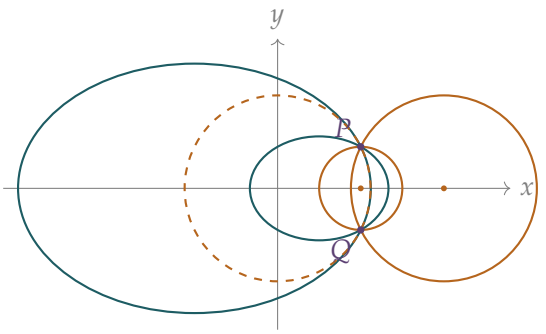
$$8x^2 + 32x - 9x^2 + 18x = 112 - 16 \quad \Longrightarrow \quad x^2 - 50x + 96 = 0 \quad \Longrightarrow \quad (x-2)(x-48) = 0.$$

? **想一想** 为甚么不能把两个根都保留?

每个根仍须给出实数的 y。由第一个椭圆得 $2y^2 = 14 - x^2 - 4x$。

- $x = 48$：$2y^2 = 14 - 2304 - 192 < 0$，所以没有实点。（无论如何，第一个椭圆只涵盖 $-2 - 3\sqrt{2} \le x \le -2 + 3\sqrt{2}$。）
- $x = 2$：$2y^2 = 14 - 4 - 8 = 2$，所以 $y = \pm 1$。

因此两个椭圆恰好相交于两点，即 $P = (2,1)$ 和 $Q = (2,-1)$。（两点亦满足第二条方程：$9 + 16 = 25$。）



图中所示为两个椭圆（蓝绿色），以及通过 P 和 Q 的其中三个圆（琥珀色），圆心分别为 $(0,0)$、$(2,0)$ 和 $(4,0)$。

通过 P 和 Q 的圆

任何通过 P 和 Q 的圆，其圆心与这两点等距，所以圆心位于 PQ 的垂直平分线上。线段 PQ 是铅垂的，中点为 $(2,0)$，所以它的垂直平分线就是 x 轴。因此圆心为 $(a,0)$，其中 a 为某个实数，而半径满足

$$r^2 = (2-a)^2 + 1^2.$$

于是圆的方程为

$$\begin{aligned}(x-a)^2 + y^2 &= (2-a)^2 + 1\\ x^2 - 2ax + a^2 + y^2 &= 4 - 4a + a^2 + 1\\ x^2 - 2ax + y^2 &= 5 - 4a,\end{aligned}$$

这正是所求的形式。

反过来说，对每个实数 a，这条方程都表示一个真正的圆，圆心为 $(a,0)$，半径为 $\sqrt{(a-2)^2+1} > 0$，而且它通过 $(2,\pm 1)$，因为 $4 - 4a + 1 = 5 - 4a$。所以这个圆族恰好就是通过两个交点的所有圆所组成的集合。

想一想 那么，把两条方程组合起来这个惯用技巧又如何?

取 $7\times$（第一式）$-$（第二式），可令 x^2 和 y^2 的系数相等，得出 $x^2 - 23x + y^2 + 41 = 0$，这是我们的圆族中 $a = \frac{23}{2}$ 的那一个成员。但这只是一个圆。两条方程的组合，是通过全部四个（复数）公共点的圆锥曲线（conic），当中包括 $x = 48$ 的两个非实点；而一个圆只需通过那两个实点。这就是本题必须采用直接方法的原因。

验算

- $a = 2$ 给出 $x^2 - 4x + y^2 = -3$，即 $(x-2)^2 + y^2 = 1$：这是以 PQ 为直径的圆，正如所料。
- $a = 0$ 给出 $x^2 + y^2 = 5$，而确实 $2^2 + 1^2 = 5$。
- 在 $-6 \le x \le 6$ 的范围内以数值方法搜索同时位于两个椭圆上的点，只找到 $x = 2$。在 $(2,1)$ 处，两个椭圆的法线（normal）方向分别为 $(8,4)$ 和 $(18,32)$，两者并不平行，所以两个椭圆在该点是真正相交，而不是相切。

常见错误

- 保留 $x = 48$，却没有检查 y 是否实数。
- 由两条椭圆方程的组合得出单独一个圆后便停下来：题目问的是通过这些点的*任何*圆。

STEP 之外

两条圆锥曲线的组合 $\lambda E_1 + \mu E_2 = 0$ 称为*曲线束*（pencil）：它由通过两者四个公共点的圆锥曲线组成，而这些公共点是在复射影平面（complex projective plane）上计算的。这道题目正好提醒我们，这些点当中有些可能是非实点，而圆是特殊的（所有圆都通过两个固定的复数点，即“无穷远圆点”（circular points at infinity））。这些概念属于射影几何（projective geometry）和代数几何（algebraic geometry）的范畴；若想以这种观点初步认识平面曲线和圆锥曲线，可参阅 Reid 的 *Undergraduate Algebraic Geometry*（一本平易近人的入门书），而 Siklos 的 *Advanced Problems in Mathematics* 则有更多 STEP 风格的练习。

详细解答

STEP I 2002，第 2 题

纯数学 · 曲线描绘

> **STEP I 2002，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/02-s1-q2

关键思路

多项式 $x^m(x-1)^n$ 在 $x=0$ 和 $x=1$ 处为零，所以曲线在两者之间必定要转向。用乘积法则（product rule）求导，并且把所有公因式（*common factor*）都提出来：导数（derivative）于是分解成几个简单的部分，各部分的正负号一目了然。整个草图由奇偶性（parity）决定：某个幂是偶数还是奇数，决定了曲线是与轴相切还是穿过轴，也决定了曲线在轴的哪一侧。

解答

驻点

由乘积法则，

$$\begin{aligned} \mathrm{f}'(x) &= mx^{m-1}(x-1)^n + nx^m(x-1)^{n-1} \\ &= x^{m-1}(x-1)^{n-1}\big[m(x-1)+nx\big] \\ &= x^{m-1}(x-1)^{n-1}\big[(m+n)x-m\big]. \end{aligned}$$

所以 $\mathrm{f}'(x)=0$ 当且仅当 $x=0$、$x=1$ 或

$$x = x_0 = \frac{m}{m+n}.$$

由于 $m,n>0$，我们有 $0<m<m+n$，所以 $0<x_0<1$。这就是所求的驻点（stationary point）。（由于 $m,n\geq 2$，幂 $m-1$ 和 $n-1$ 都至少是 1，所以 $x=0$ 和 $x=1$ 也是驻点。画草图时会用到它们。）

驻点的性质

把导数写成 $\mathrm{f}'(x) = g(x)\,h(x)$，其中 $g(x) = x^{m-1}(x-1)^{n-1}$，$h(x) = (m+n)x - m$。那么

$$\mathrm{f}''(x) = g'(x)h(x) + g(x)h'(x) = g'(x)h(x) + (m+n)g(x).$$

> ? **想一想** 我们真的需要把 g 完整地求导吗?

不需要。在 x_0 处我们有 $h(x_0) = 0$，所以无论 g' 是甚么，第一项都会消失：

$$\mathrm{f}''(x_0) = (m+n)\,x_0^{m-1}\,(x_0-1)^{n-1}.$$

这里 $m+n > 0$，$x_0^{m-1} > 0$，但 $x_0 - 1 < 0$。所以 $\mathrm{f}''(x_0)$ 不为零，其正负号与 $(-1)^{n-1}$ 相同：

- 若 n 是偶数，则 $n-1$ 是奇数，$\mathrm{f}''(x_0) < 0$，x_0 是**极大点（maximum）**；
- 若 n 是奇数，则 $n-1$ 是偶数，$\mathrm{f}''(x_0) > 0$，x_0 是**极小点（minimum）**。

因为 $\mathrm{f}''(x_0) \neq 0$，二阶导数判别法（second-derivative test）可以下定论。这正是我们要证明的。

草图

四种情况由 m 和 n 的奇偶性决定。要画出每一种情况，我们先整理以下事实。

*f 的正负号。*当 $x > 1$ 时，两个因式都是正的，所以 $\mathrm{f} > 0$。当 $0 < x < 1$ 时，$x^m > 0$，而 $(x-1)^n$ 的正负号与 $(-1)^n$ 相同。当 $x < 0$ 时，正负号是 $(-1)^m(-1)^n = (-1)^{m+n}$。

*在根（root）附近的表现。*在 $x = 0$ 附近，我们有 $\mathrm{f}(x) \approx (-1)^n x^m$。若 m 是偶数，f 在 0 的两侧保持同号，所以曲线在该处与轴相切（一个转向点（turning point））。若 m 是奇数，f 变号，又因为 $\mathrm{f}'(0) = 0$，曲线会水平地穿过轴（一个驻拐点（stationary point of inflection））。把同样的推理用于 n，便得到 $x = 1$ 处的情况。

*两端。*当 $|x|$ 很大时，$\mathrm{f}(x)$ 的表现与 x^{m+n} 相似，所以当 $x \to +\infty$ 时 $\mathrm{f} \to +\infty$。当 $x \to -\infty$ 时，若 $m+n$ 是偶数则 $\mathrm{f} \to +\infty$，若 $m+n$ 是奇数则 $\mathrm{f} \to -\infty$。

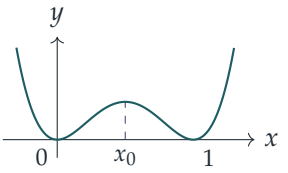
*没有其他驻点。*驻点只有 $x = 0$、x_0 和 $x = 1$。

综合以上各点：

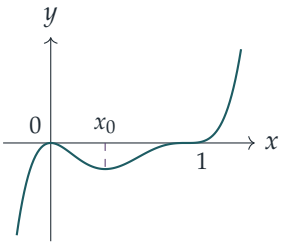
- **m 为偶数，n 为偶数：**处处皆有 $\mathrm{f} \geq 0$。在 0 和 1 处有位于轴上的极小点，在 x_0 处有一个极大点：呈 W 形。
- **m 为偶数，n 为奇数：**曲线由 $-\infty$ 上升，在 0 处从下方与轴相切（一个极大点），下降至 x_0 处的极小点，然后上升，经过 1 处的驻拐点，一直升向 $+\infty$。

- **m 为奇数，n 为偶数：**曲线由 $-\infty$ 上升，经过 0 处的驻拐点，到达 x_0 处的极大点，下降至 1 处与轴相切（一个极小点），然后升向 $+\infty$。
- **m 为奇数，n 为奇数：**曲线由 $+\infty$ 下降，经过 0 处的驻拐点，到达 x_0 处的极小点，再回升经过 1 处的驻拐点，一直升向 $+\infty$。

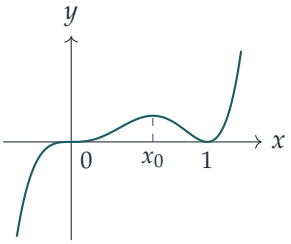
以下草图采用最小的情况 $(m,n) = (2,2),(2,3),(3,2),(3,3)$，每幅图各有自己的纵轴比例。在 x_0 处的值都很小：分别是 $\frac{1}{16}$、$-\frac{108}{3125}$、$\frac{108}{3125}$ 和 $-\frac{1}{64}$。



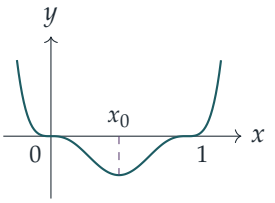
m 为偶数，n 为偶数



m 为偶数，n 为奇数



m 为奇数，n 为偶数



m 为奇数，n 为奇数

验算

- 取 $m = n = 2$：$\mathrm{f}(x) = \big(x(x-1)\big)^2$ 关于 $x = \frac{1}{2}$ 对称，而 $x_0 = \frac{1}{2}$ 确实是极大点，其值为 $\frac{1}{16}$。
- 代换 $x \mapsto 1-x$ 把 $x^m(x-1)^n$ 变成 $(-1)^{m+n}x^n(x-1)^m$。所以把 m 和 n 对调，相当于把图像沿 $x = \frac{1}{2}$ 反射（若 $m+n$ 是奇数，还要再沿 x 轴反射）。这就是两幅奇偶性相异的草图互为镜像的原因。
- 我们把 f' 与数值导数作了比对，又把 $\mathrm{f}''(x_0)$ 的正负号与有限差分（finite difference）作了比对，范围涵盖所有 $2 \le m,n \le 8$。例如当 $(m,n) = (2,3)$ 时，$\mathrm{f}''(\frac{2}{5}) = 5 \cdot \frac{2}{5} \cdot \frac{9}{25} = 0.72 > 0$，是极小点，与预测相符。

常见错误

- 把 $f''(x)$ 完整展开。在 x_0 处，因式 $(m+n)x - m$ 为零，所以只有一项留下来。
- 忘记 $x = 0$ 和 $x = 1$ 也是驻点（因为 $m, n \geq 2$），结果把曲线画成在该处斜斜地穿过轴。
- 弄错 $x < 0$ 时的正负号。它取决于 $m + n$，而不是单单取决于 m。

STEP 之外

多项式在重数（multiplicity）为 k 的根附近，局部形状与 $c(x-a)^k$ 相似，而 k 的奇偶性决定了曲线是相切还是穿过。这是迈向泰勒定理（Taylor's theorem）和驻点的高阶导数判别法的第一步：若驻点处第一个不为零的导数是 k 阶的，则当 k 是偶数时该点是转向点，当 k 是奇数时该点是拐点。Abbott 的 *Understanding Analysis* 为这些想法建立了严格的基础。

详细解答

STEP I 2002，第 3 题

纯数学 · 以微积分证明不等式

> **STEP I 2002，第 3 题**
>
> 继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：
>
> step.maths.org/questions/02-s1-q3

关键思路

把两个平方根记作 $\sqrt{P}$ 和 $\sqrt{Q}$。根号下的两个量关系密切：$P+Q=a^2+b^2$ 是常数，而且它们的导数（derivative）大小相等、正负相反。因此 $\mathrm{d}y/\mathrm{d}\theta$ 可以干净利落地因式分解（factorise），其正负只由一个三角因子决定。只要知道 y 的最大值和最小值，那个双重不等式（double inequality）不过就是“最小值 $\le y \le$ 最大值”，而第一部分则告诉我们两端的大小次序正确。

解答

第一个不等式

展开得

$$2a^2+2b^2-(a+b)^2=a^2-2ab+b^2=(a-b)^2\ge 0,$$

所以 $(a+b)^2\le 2a^2+2b^2$，等号成立当且仅当 $a=b$。

驻点

设 $c=\cos\theta$，$s=\sin\theta$，并且

$$P=a^2c^2+b^2s^2,\qquad Q=a^2s^2+b^2c^2,\qquad y=\sqrt{P}+\sqrt{Q}.$$

先看两个特殊情况。若 $a^2=b^2$，则对每一个 θ 都有 $P=Q=a^2$，所以 $y=2|a|$ 是常数，每一点都（平凡地）是驻点（stationary point）；这个情况暂且搁下。若 $a=0$ 或 $b=0$（但不同时为零），则 P 或 Q 会在某些 θ 处等于零，而 y 在该处不可微（not differentiable）；这个情况留待最后处理。所以现在先假设 $ab\neq 0$ 且 $a^2\neq b^2$。于是 $P\ge \min(a^2,b^2)(c^2+s^2)>0$，

同理 $Q>0$，所以 y 处处可微。

求导得

$$\frac{\mathrm{d}P}{\mathrm{d}\theta}=-2a^2cs+2b^2sc=(b^2-a^2)\sin 2\theta,\qquad \frac{\mathrm{d}Q}{\mathrm{d}\theta}=(a^2-b^2)\sin 2\theta.$$

因此

$$\frac{\mathrm{d}y}{\mathrm{d}\theta}=\frac{(b^2-a^2)\sin 2\theta}{2\sqrt{P}}-\frac{(b^2-a^2)\sin 2\theta}{2\sqrt{Q}}=\frac{(b^2-a^2)\sin 2\theta\left(\sqrt{Q}-\sqrt{P}\right)}{2\sqrt{PQ}}.$$

要看出 $\sqrt{Q}-\sqrt{P}$ 的正负，可以把它有理化（rationalise）：

$$\sqrt{Q}-\sqrt{P}=\frac{Q-P}{\sqrt{Q}+\sqrt{P}},\qquad Q-P=(a^2-b^2)(s^2-c^2)=(b^2-a^2)\cos 2\theta.$$

代入上式，并利用 $2\sin 2\theta\cos 2\theta=\sin 4\theta$，得

$$\frac{\mathrm{d}y}{\mathrm{d}\theta}=\frac{(a^2-b^2)^2\sin 4\theta}{4\sqrt{PQ}\left(\sqrt{P}+\sqrt{Q}\right)}.$$

想一想 为甚么值得花工夫把 $\sqrt{Q}-\sqrt{P}$ 有理化?

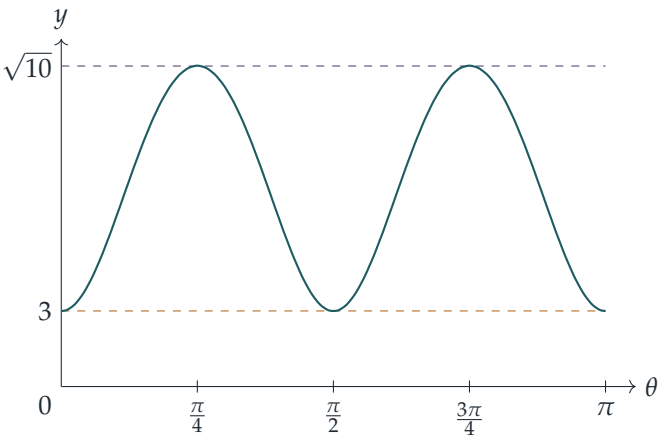
因为这样一来，除了 $\sin 4\theta$ 之外，其余各项一眼就看得出是正数：$(a^2-b^2)^2>0$，而 P、Q 都是正数。所以 $\mathrm{d}y/\mathrm{d}\theta$ 的正负与 $\sin 4\theta$ 完全相同，这既告诉我们驻点在哪里，也告诉我们它们属于哪一类。

驻点出现在 $\sin 4\theta=0$ 之处，即 $\theta=\frac{1}{4}n\pi$，其中 n 为整数。

- 当 n 为偶数时，$\theta=\frac{1}{2}k\pi$，所以 $\{c^2,s^2\}=\{1,0\}$，$\{P,Q\}=\{a^2,b^2\}$。此时 $y=|a|+|b|$（记住 $\sqrt{a^2}=|a|$）。
- 当 n 为奇数时，$\theta=\frac{1}{4}\pi+\frac{1}{2}k\pi$，所以 $c^2=s^2=\frac{1}{2}$，$P=Q=\frac{1}{2}(a^2+b^2)$。此时 $y=2\sqrt{\frac{1}{2}(a^2+b^2)}=\sqrt{2a^2+2b^2}$。

驻点的性质。当 θ 递增经过 $\frac{1}{2}k\pi$ 时，$\sin 4\theta$ 由负变正，所以 y 由递减转为递增：这些点是**极小点（minima）**。当 θ 递增经过 $\frac{1}{4}\pi+\frac{1}{2}k\pi$ 时，$\sin 4\theta$ 由正变负，所以这些点是**极大点（maxima）**。

下图显示 $a=2$、$b=1$ 的情况，纵轴的比例经过拉长：y 在 3 与 $\sqrt{10}\approx 3.162$ 之间振荡，周期（period）为 $\frac{1}{2}\pi$。



最后的不等式

把 θ 换成 $\theta+\frac{1}{2}\pi$，c^2 与 s^2 便会对调，于是 P 与 Q 也对调，所以 y 的周期为 $\frac{1}{2}\pi$。在一个周期 $[0,\frac{1}{2}\pi]$ 上，y 在 $[0,\frac{1}{4}\pi]$ 上递增，在 $[\frac{1}{4}\pi,\frac{1}{2}\pi]$ 上递减（由 $\sin 4\theta$ 的正负可知），所以它的最小值是 $|a|+|b|$，在两个端点取得；最大值是 $\sqrt{2a^2+2b^2}$，在 $\frac{1}{4}\pi$ 取得。根据周期性，这两个值就是对所有 θ 而言的最小值和最大值，所以

$$|a|+|b| \;\le\; y \;\le\; \sqrt{2a^2+2b^2}.$$

这两个界并无矛盾：把第一部分的结果应用于 $|a|$ 和 $|b|$，得 $(|a|+|b|)^2 \le 2a^2+2b^2$，所以下界确实不会超过上界。

先前搁下的情况。若 $a^2=b^2$，则 $y=2|a|$，而 $|a|+|b|=2|a|$，$\sqrt{2a^2+2b^2}=\sqrt{4a^2}=2|a|$，所以不等式处处以等号成立。若 $b=0$，则 $y=|a|\big(|\cos\theta|+|\sin\theta|\big)$。由于每一项都不超过 1，故 $|\cos\theta|+|\sin\theta| \ge \cos^2\theta+\sin^2\theta=1$，而由第一部分可知 $(|\cos\theta|+|\sin\theta|)^2 \le 2$；所以 $|a| \le y \le \sqrt{2}\,|a|$，这正是所求的不等式。$a=0$ 的情况相同，只需把 a 与 b 对调；而 $a=b=0$ 的情况是平凡的。这就证明了结果对所有 a、b 和 θ 都成立。 ■

验算

- 取 $a=2$、$b=1$：$y(0)=2+1=3$，$y(\frac{1}{4}\pi)=2\sqrt{5/2}=\sqrt{10}\approx 3.1623$；以数值方法对大量 θ 值扫描 y，结果都落在这两个值之间。
- 导数公式与 $\mathrm{d}y/\mathrm{d}\theta$ 的有限差分（finite-difference）估计值相符，在随机选取的 a、b、θ 下误差约为 10^{-9}。
- 当 $a=b$ 时，两个界都退化为 $2|a|$，与 y 为常数的情况吻合。

常见错误

- 写成 $\sqrt{a^2} = a$：最小值是 $|a| + |b|$，这正是答案中出现绝对值（modulus）的原因。
- 只做到 $\sin 2\theta = 0$ 就停下来，漏掉了位于 $\frac{1}{4}\pi$ 的奇数倍之处的驻点，在那里 $\sqrt{P} = \sqrt{Q}$。
- 毫无理由就断定某些点是极大点或极小点；利用 $\sin 4\theta$ 的正负，一行就能给出理由。

STEP 之外

从几何上看，$\sqrt{P}$ 是椭圆（ellipse）$x = a\cos\phi$，$y = b\sin\phi$ 的中心到椭圆上某一点的距离，而 $\sqrt{Q}$ 是中心到参数再转四分之一圈之后那一点的距离，所以这道题探讨的是这两条“共轭”半径如何摊分固定的总和 $a^2 + b^2$。这两个界同时也是向量 (a, b) 的 ℓ^1 范数（norm）和（经缩放的）ℓ^2 范数，而像这样比较范数，是大学一年级分析（analysis）和线性代数（linear algebra）中的常用工具。想进一步了解如何利用导数证明不等式，可参阅 Abbott 的 *Understanding Analysis*（数学分析入门教材）。

详细解答

STEP I 2002，第 4 题

纯数学 · 切线与不等式

> **STEP I 2002，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/02-s1-q4

关键思路

一条通过 $(0,1)$ 并在较远处与曲线相切的直线，在这两点之间必定位于曲线的下方，因为曲线在该处是凹的（concave），其后才变得平缓。直线下方的区域是一个梯形（trapezium），把它绕 y 轴旋转会得到一个圆柱体和一个圆锥体，所以毋须用微积分也能轻易求出它的面积和体积。把它们与曲线下方的准确面积和体积比较，便得到那两个不等式。

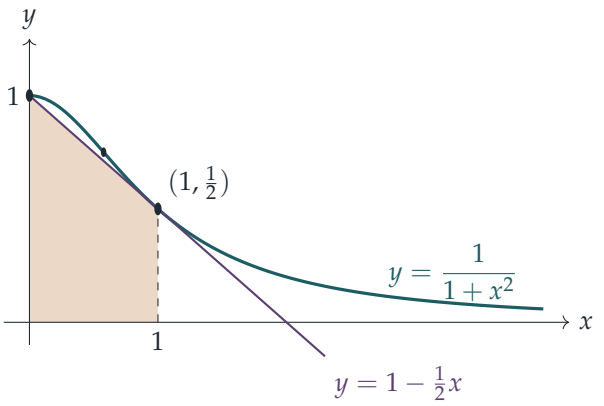
解答

略图

设 $f(x)=\dfrac{1}{1+x^2}$。则 $f(0)=1$，$f(x)>0$，而且当 $x\to\infty$ 时 $f(x)\to 0$，所以 x 轴是渐近线（asymptote）。求导得

$$f'(x)=-\frac{2x}{(1+x^2)^2},\qquad f''(x)=\frac{6x^2-2}{(1+x^2)^3}.$$

所以 $f'(0)=0$（曲线在 $(0,1)$ 处以水平方向出发），而当 $x>0$ 时 $f'(x)<0$，故曲线一直递减。此外，当 $0\le x<1/\sqrt{3}$ 时 $f''(x)<0$，当 $x>1/\sqrt{3}$ 时 $f''(x)>0$，所以在 $\left(1/\sqrt{3},\tfrac{3}{4}\right)$ 有一个拐点（point of inflection）：曲线在 y 轴附近是凹的（向下弯），在较远处则是凸的（convex）。



阴影部分的梯形是 $0 \le x \le 1$ 时直线下方的区域；它位于曲线下方的区域之内。

直线

直线通过 $(0, f(0)) = (0,1)$, 所以它的方程是 $y = 1 + mx$, 其中 m 是某个斜率 (gradient)。假设它在 $x = a > 0$ 处与曲线相切。在该处必须满足两个条件。

它与曲线相交：$1 + ma = \dfrac{1}{1+a^2}$，所以 $ma = -\dfrac{a^2}{1+a^2}$，而由于 $a \neq 0$,

$$m = -\frac{a}{1+a^2}.$$

它的斜率与曲线的斜率相同：$m = f'(a) = -\dfrac{2a}{(1+a^2)^2}$。

令两式相等，再除以 $-a/(1+a^2)$（因为 $a > 0$，这样做是容许的），得 $1 = \dfrac{2}{1+a^2}$，所以 $a^2 = 1$，而由于 $a > 0$，故 $a = 1$。于是 $m = -\frac{1}{2}$，直线是

$$y = 1 - \tfrac{1}{2}x,$$

它在 $(1, \frac{1}{2})$ 处与曲线相切。

再无其他交点

把直线减去曲线：

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

这个式子只在 $x = 0$ 或 $x = 1$ 时等于零。所以直线与曲线相交于 $x = 0$ 和切点 $x = 1$（重因式 $(x-1)^2$ 正是相切在代数上的标志），此外再无其他交点：在 $x \ge 0$ 时没有，事实上在 x 为负数时也没有。

我们还可以从中读出下一步需要的结果。当 $x > 0$ 且 $x \neq 1$ 时，这个式子严格小于零，所以

$$0 < 1 - \tfrac{1}{2}x < \frac{1}{1+x^2} \qquad \text{当} 0 < x < 1. \tag{$*$}$$

（左边的不等式成立，是因为在 $[0,1]$ 上 $1 - \frac{1}{2}x \geq \frac{1}{2}$。）所以在 $[0,1]$ 上，除了端点之外，直线严格位于曲线下方，正如略图所示。

$\pi > 3$

把 $(*)$ 在 $[0,1]$ 上积分。两个函数都是连续的，而且在开区间上曲线严格位于直线上方，所以积分之后不等式仍然是严格不等式：

$$\int_0^1 \frac{1}{1+x^2}\,\mathrm{d}x > \int_0^1 \Big(1 - \frac{x}{2}\Big)\mathrm{d}x.$$

左边是 $\frac{\pi}{4}$（即 $\big[\arctan x\big]_0^1$）。右边是阴影梯形的面积，梯形的两条平行边长为 1 和 $\frac{1}{2}$，宽为 1，所以面积是 $\frac{1}{2}\big(1 + \frac{1}{2}\big) = \frac{3}{4}$。因此 $\frac{\pi}{4} > \frac{3}{4}$，即

$$\pi > 3.$$

$\ln 2 > 2/3$

现在把同样的两个区域（$0 \leq x \leq 1$）绕 y 轴旋转 2π。根据 $(*)$，梯形位于曲线下方的区域之内，所以它的旋转体（solid of revolution）位于另一个旋转体之内，体积也较小。

想一想 如果曲线是以 x 表示 y 的形式给出，我们怎样求绕 y 轴旋转的旋转体体积?

使用圆柱壳法（cylindrical shells）：位于 x 处、宽度为 δx、高度为 y 的窄条，旋转后扫出一个体积约为 $2\pi x\, y\, \delta x$ 的薄壳，所以体积是 $\int_0^1 2\pi x\, y\, \mathrm{d}x$。（另一个方法是沿水平方向切片；见“验算”。）

对于曲线，

$$V_{\text{curve}} = \int_0^1 \frac{2\pi x}{1+x^2}\,\mathrm{d}x = \pi\big[\ln(1+x^2)\big]_0^1 = \pi \ln 2,$$

因为分子 $2x$ 是分母的导数。

对于直线，旋转体是一个半径为 1、高为 $\frac{1}{2}$ 的圆柱体（来自梯形在 $y = \frac{1}{2}$ 以下的部分），上面再加一个半径为 1、高为 $\frac{1}{2}$ 的圆锥体（来自其上方的三角形）。所以

$$V_{\text{line}} = \pi \cdot 1^2 \cdot \tfrac{1}{2} + \tfrac{1}{3}\pi \cdot 1^2 \cdot \tfrac{1}{2} = \frac{\pi}{2} + \frac{\pi}{6} = \frac{2\pi}{3}.$$

用圆柱壳法验算：$\int_0^1 2\pi x\big(1 - \frac{1}{2}x\big)\mathrm{d}x = 2\pi\big(\frac{1}{2} - \frac{1}{6}\big) = \frac{2\pi}{3}$。

被积函数 $2\pi x\left(\frac{1}{1+x^2}-1+\frac{x}{2}\right)$ 是连续的，在 $[0,1]$ 上非负，而且根据 $(*)$ 在 $(0,1)$ 上严格为正，所以 $V_{\text{curve}} > V_{\text{line}}$：

$$\pi \ln 2 > \frac{2\pi}{3} \quad \Longrightarrow \quad \ln 2 > \frac{2}{3},$$

这正是所要证明的。

验算

- 以数值计算，$\pi/4 \approx 0.7854 > 0.75$，$\ln 2 \approx 0.6931 > 0.6667$，所以两个不等式都成立，而且两者都不算宽松。
- 改为沿水平方向切片：当 $\frac{1}{2} \le y \le 1$ 时，曲线满足 $x^2 = \frac{1}{y} - 1$，所以体积是圆柱体的 $\pi \cdot 1^2 \cdot \frac{1}{2}$ 加上 $\int_{1/2}^{1} \pi\left(\frac{1}{y}-1\right)\mathrm{d}y = \pi\left(\ln 2 - \frac{1}{2}\right)$，同样得到 $\pi \ln 2$。
- 快速检验因式分解：在 $x = 2$ 时，直线给出 0，曲线给出 $\frac{1}{5}$，两者相差 $-\frac{1}{5}$；公式给出 $-\frac{2\cdot 1}{2\cdot 5} = -\frac{1}{5}$。

常见错误

- 求出一般点上的切线，却忘记了切线须通过 $(0,1)$ 这个条件（或者反过来）：两个条件都要用上，才能确定 a。
- 止步于“$\ge$”：要得到严格不等式，直线必须在某个区间上严格位于曲线下方，而这正是因式 $x(x-1)^2$ 所提供的。
- 误把区域绕 x 轴旋转，这样会得到 $\int \pi y^2\,\mathrm{d}x$ 和不同的数值。

STEP 之外

这个论证是一个比较原理（comparison principle）：如果在某个区间上 $g \le h$，在某处有严格不等式成立，而且两者都连续，那么 $\int g < \int h$。把“在某处严格成立”说得精确，是大学第一门分析课程中的典型习题，在那里积分是通过上和与下和（upper and lower sums）来定义的。以简单图形的面积来估计 π 的上下界，可以追溯至阿基米德（Archimedes），而这道题还有精密得多的版本（例如 $\int_0^1 x^4(1-x)^4/(1+x^2)\,\mathrm{d}x = \frac{22}{7} - \pi$），值得查阅一下。至于分析方面，可以试读 Abbott 的 *Understanding Analysis*（分析学入门教材）。

详细解答

STEP I 2002，第 5 题

纯数学 • 多项式的整数根

> **STEP I 2002，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/02-s1-q5

关键思路

写成一次因式（linear factor）乘积的多项式（polynomial），在特殊点上很容易求值：代入 $x = 0$、$x = 1$ 或 $x = -1$，因式分解后的形式就变成 k_i、$k_i + 1$ 或 $k_i - 1$ 的乘积，而展开后的形式则给出系数（coefficient）之和。对于题中的四次多项式（quartic），这些值都是固定的整数，而根（root）是整数就意味着每个 $k_i \pm 1$ 都必须是某个已知数的因数（divisor）。把这些数分解成质因数之后，可供选择的余地就所剩无几。

解答

三个恒等式

由于 $f(x)$ 的两个表达式是同一个多项式，它们在 x 的每一个值上都相等。

代入 $x = 0$：展开式给出 $f(0) = a_n$，因为其余各项都含有因子 x。因式分解后的形式给出 $f(0) = k_1 k_2 \cdots k_n$。因此

$$k_1 k_2 \cdots k_n = a_n.$$

代入 $x = 1$：x 的每个幂都变成 1，所以展开式给出 $f(1) = 1 + a_1 + a_2 + \cdots + a_n$，而因式分解后的形式给出 $(1 + k_1)(1 + k_2) \cdots (1 + k_n)$。因此

$$(k_1 + 1)(k_2 + 1) \cdots (k_n + 1) = 1 + a_1 + a_2 + \cdots + a_n,$$

正如所求。

代入 $x = -1$：因式分解后的形式给出 $(k_1 - 1)(k_2 - 1) \cdots (k_n - 1)$。在展开式中，$a_j x^{n-j}$ 一项变成 $(-1)^{n-j} a_j = (-1)^n (-1)^j a_j$（这里用到 $(-1)^{-j} = (-1)^j$），所以

$$(k_1 - 1)(k_2 - 1) \cdots (k_n - 1) = (-1)^n \left(1 - a_1 + a_2 - a_3 + \cdots + (-1)^n a_n\right).$$

当 n 为偶数时，这就是 $1-a_1+a_2-\cdots+a_n$；当 n 为奇数时，每个符号都要反转。

四次多项式

把这个四次多项式写成 $f(x) = (x+k_1)(x+k_2)(x+k_3)(x+k_4)$，于是它的根是 $-k_1,\ldots,-k_4$，而且都是整数。取 $n=4$ 应用上述恒等式：

$$k_1k_2k_3k_4 = 576 = 2^6\cdot 3^2,$$
$$(k_1+1)(k_2+1)(k_3+1)(k_4+1) = 1+22+172+552+576 = 1323 = 3^3\cdot 7^2,$$
$$(k_1-1)(k_2-1)(k_3-1)(k_4-1) = 1-22+172-552+576 = 175 = 5^2\cdot 7.$$

每个 k_i 都至少是 2。所有系数都是正数，所以对每个 $x\geq 0$ 都有 $f(x)>0$；因此每个根都是负数，每个 k_i 都是正整数。此外，没有任何 k_i 等于 1，否则第三个乘积会是 0 而不是 175。所以对每个 i 都有 $k_i\geq 2$，而 k_i-1 和 k_i+1 全都是正整数。

? **想一想** 怎样把三个乘积化为一张简短的清单，列出 k_i 的可能值?

每个 k_i-1 都是 175 的正因数，每个 k_i+1 都是 1323 的正因数。175 的因数是 $1,5,7,25,35,175$，由此得

$$k_i\in\{2,6,8,26,36,176\},\qquad k_i+1\in\{3,7,9,27,37,177\}.$$

在 k_i+1 的这些值当中，只有 $3,7,9,27$ 能整除 $1323=3^3\cdot 7^2$（因为 37 是质数，而 $177=3\cdot 59$）。所以每个 k_i 都是 $2,6,8,26$ 其中之一，每个 k_i+1 都是 $3,7,9,27$ 其中之一。

现在看乘积 $\prod(k_i+1)=3^3\cdot 7^2$。

- 因子 7^2 只能来自等于 7 的因子，所以 k_i+1 当中恰好有两个等于 7：有两个 k_i 是 6。
- 另外两个因子是取自 $\{3,9,27\}$ 的 3 的幂，乘积为 $3^3=27$。两者都至少是 3，所以不可能取 27（否则另一个就要是 1），唯一的选择是 3×9。所以余下的 k_i 是 2 和 8。

于是 $\{k_1,k_2,k_3,k_4\}=\{2,6,6,8\}$。用第一个乘积验证：$2\cdot 6\cdot 6\cdot 8=576$，正如所求。展开亦可确认：

$$(x+2)(x+8)=x^2+10x+16,\qquad (x+6)^2=x^2+12x+36,$$

以及

$$(x^2+10x+16)(x^2+12x+36)=x^4+22x^3+172x^2+552x+576.$$

所以各根是

$$x=-2,\quad x=-6\text{（重根）},\quad x=-8.$$

验算

- 各 k_i 之和是 $2+6+6+8=22$，与 x^3 的系数吻合。
- $f(-6)=1296-4752+6192-3312+576=0$，而 $f'(x)=4x^3+66x^2+344x+552$ 在 $x=-6$ 处同样等于零，确认 -6 是二重根。
- 用数值快速检验这些恒等式：对 $f(x)=(x+2)(x-3)=x^2-x-6$，我们得到 $(2-1)(-3-1)=-4$ 和 $1-(-1)+(-6)=-4$。

常见错误

- 在 $x=-1$ 的结果中弄错符号：当 n 为奇数时，因子 $(-1)^n$ 不能忽略。
- 忘记重根是容许的：一旦在 1323 中见到 7^2，就必定有两个 k_i 相等。
- 没有先证明每个 $k_i \geq 2$ 就只列出正因数；正是这一步保证了所有 $k_i \pm 1$ 都是正数。

STEP 之外

这里的恒等式是韦达定理（Vieta's formulae）的特殊情况；韦达定理把多项式的系数表示为其根的基本对称函数（elementary symmetric functions）。至于因数的论证，则与有理根定理（rational root theorem）一脉相承，也与高斯（Gauss）关于整系数多项式因式分解的引理（Gauss's lemma）相近，后者你会在环（ring）论的入门课程中遇到。Allenby 的 *Rings, Fields and Groups*（《环、域与群》）对这些想法有细致的阐述。

详细解答

STEP I 2002，第 6 题

纯数学 · 角锥的几何

STEP I 2002，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：

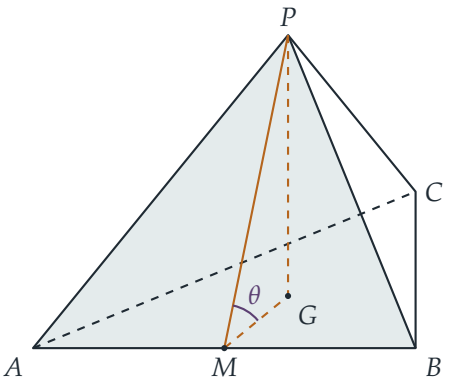
step.maths.org/questions/02-s1-q6

关键思路

根据对称性，顶点位于等边底面中心的正上方，因此只需用一次毕氏定理（Pythagoras' theorem）便可求得高。知道体积之后，我们根本不必直接求新的高：无论哪一个面放在地上，体积都是一样的，所以 $V = \frac{1}{3} \times (\text{新底面的面积}) \times h$。最后一部分所求的角，是两个面沿公共棱所成的夹角；要看清楚这个角，最简洁的方法是考虑通过顶点并垂直于该棱的截面（cross-section）。

解答

把底面记为 ABC，边长为 a，顶点记为 P，且 $PA = PB = PC = b$。设 G 为 ABC 的中心，M 为 AB 的中点。



体积

设 H 为 P 在底面之上的高度，F 为由 P 所作垂线的垂足。直角三角形 PFA、PFB、PFC 有公共边 $PF = H$，而斜边均为 b，所以

$$FA = FB = FC = \sqrt{b^2 - H^2}.$$

因此 F 与三个顶点等距：它是 ABC 的外心（circumcentre），而等边三角形的外心就是其中心 G。ABC 的中线（median）长度为 $\frac{\sqrt{3}}{2}a$，而 G 位于中线上由顶点起计三分之二的位置，所以 $GA = \frac{2}{3}\cdot\frac{\sqrt{3}}{2}a = \frac{a}{\sqrt{3}}$。在三角形 PGA 中运用毕氏定理，得

$$H^2 = b^2 - \frac{a^2}{3} = \frac{3b^2 - a^2}{3}.$$

底面的面积为 $\frac{1}{2}a^2\sin 60^\circ = \frac{\sqrt{3}}{4}a^2$，所以

$$V = \frac{1}{3}\cdot\frac{\sqrt{3}}{4}a^2\cdot\frac{\sqrt{3b^2 - a^2}}{\sqrt{3}} = \frac{1}{12}a^2\left(3b^2 - a^2\right)^{\frac{1}{2}},$$

正是所要求的结果。（这种形状的棱锥只在 $3b^2 > a^2$ 时才存在，而这正好是上式为正实数的条件。）

新的高

现在把面 PAB 放在地上。它是等腰三角形，$PA = PB = b$ 而 $AB = a$，所以 PM 垂直于 AB；在三角形 PMA 中运用毕氏定理，得

$$PM^2 = b^2 - \frac{a^2}{4} = \frac{4b^2 - a^2}{4}, \qquad \text{面积}PAB = \frac{1}{2}a\cdot PM = \frac{a}{4}\sqrt{4b^2 - a^2}.$$

想一想 要求 h，我们真的需要找出新的最高顶点的位置吗?

不需要。立体本身没有改变，所以它的体积仍然是 V；而体积公式对任何一个面作为底面都成立，只要高是垂直于该面量度的。所以

$$V = \frac{1}{3}\cdot\frac{a}{4}\sqrt{4b^2 - a^2}\cdot h.$$

与第一部分的结果比较，

$$\frac{a}{12}\sqrt{4b^2 - a^2}\,h = \frac{a^2}{12}\sqrt{3b^2 - a^2} \quad\Longrightarrow\quad h = \frac{a\sqrt{3b^2 - a^2}}{\sqrt{4b^2 - a^2}},$$

两边平方得

$$h^2 = \frac{a^2(3b^2 - a^2)}{4b^2 - a^2},$$

正是所要求的结果。这里 $4b^2 - a^2 > 0$，因为 $4b^2 > 3b^2 > a^2$。

等边面的倾角

在新的位置，地面就是 PAB 所在的平面，而等边面 ABC 与它沿棱 AB 相交。两个平面之间的夹角，是指在两个平面上各取一条直线，于同一点垂直于公共棱，这两条直线所成的角。在 M 点，面 PAB 上的直线 MP 垂直于 AB（因为 PAB 是等腰三角形），面 ABC 上的直线 MC 亦垂直于 AB（因为 ABC 是等边三角形），而且 G 在 MC 上。所以所求的角是 $\theta = \angle PMC = \angle PMG$。

三角形 PMG 在 G 处为直角，因为 PG 垂直于整个底面。此外，GM 是中线的三分之一，$GM = \frac{1}{3}\cdot\frac{\sqrt{3}}{2}a = \frac{a}{2\sqrt{3}}$。因此

$$\tan\theta = \frac{PG}{GM} = \frac{\sqrt{3b^2-a^2}}{\sqrt{3}}\cdot\frac{2\sqrt{3}}{a} = \frac{2\sqrt{3b^2-a^2}}{a},$$

所以等边三角形与水平面之间的夹角为

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

这个角是锐角，因为 G 位于底面之内，所以它确实是两个平面之间的夹角，而不是其补角。

另一个等价的做法是利用新的高：顶点 C 在地面之上的高度为 h，而沿着该面量度，它与棱 AB 的距离为 $MC = \frac{\sqrt{3}}{2}a$，所以 $\sin\theta = h\big/\frac{\sqrt{3}}{2}a$，由此得

$$\sin\theta = \frac{2\sqrt{3b^2-a^2}}{\sqrt{3(4b^2-a^2)}}, \qquad \cos\theta = \frac{a}{\sqrt{3(4b^2-a^2)}},$$

与上面 $\tan\theta$ 的值吻合。

验算

对于正四面体（regular tetrahedron），$a = b = 1$：公式给出 $V = \frac{\sqrt{2}}{12}$，$h^2 = \frac{2}{3}$（与原来的高相同，根据对称性亦必然如此），以及 $\cos\theta = \frac{1}{3}$，即大家熟悉的正四面体二面角（dihedral angle）$\theta \approx 70.53^\circ$。当 $b \to a/\sqrt{3}$ 时，棱锥变得扁平，$\theta \to 0$；当 $b \to \infty$ 时，它变成又高又尖的锥体，$\theta \to 90^\circ$。我们亦用坐标为 20,000 组随机数对 (a, b) 建构了这个棱锥，并以三重积计算 V，以 C 到平面 PAB 的距离计算 h，以法向量计算 θ；三者均与公式吻合，误差约为 10^{-13}。例如 $a = 2$、$b = 3$ 给出 $h \approx 1.6956$ 及 $\theta \approx 78.22^\circ$。

常见错误

- 把 GA 当作中线的一半，或在需要用外接圆半径（circumradius）$\frac{a}{\sqrt{3}}$ 的地方用了内切圆半径（inradius）$\frac{a}{2\sqrt{3}}$（在最后一部分则刚好相反）。

► 量度该面与地面之间的夹角时，用了一条不垂直于公共棱的直线，例如 AC 或 PC。

STEP 之外

以不同的面作底来计算同一个体积，让我们初次窥见这样一个想法：体积并不取决于我们怎样切割立体或怎样摆放它。在大学一年级的课程中，四面体的体积会写成 $\frac{1}{6}|\mathbf{u}\cdot(\mathbf{v}\times\mathbf{w})|$，即一个纯量三重积（scalar triple product，亦即一个 3×3 行列式（determinant）），而两个面之间的夹角就是它们的法向量（normal vector）之间的夹角，我们的数值验算正是这样做的。Strang 的 *Introduction to Linear Algebra*（线性代数入门）把行列式作为体积来阐述。

详细解答

STEP I 2002，第 7 题

纯数学 · 三角函数积分

STEP I 2002，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/02-s1-q7

关键思路

单独来看，I 和 J 都不易处理，但两者之和与两者之差却都很好计算。把它们相加，分母（denominator）便出现在分子（numerator）上，于是被积函数（integrand）就是 1。把它们相减，分子上出现的是分母的导数（derivative），于是积分是一个对数（logarithm）。后面两个积分用的是同一技巧的较一般形式：把分子拆成分母的某个倍数，加上分母的导数的某个倍数。

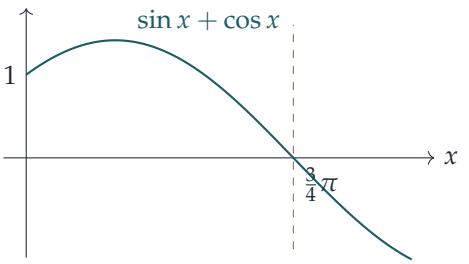
解答

主要结果

首先，为甚么要对 a 加上限制？我们可以写成

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

它在 $-\frac{\pi}{4} < x < \frac{3\pi}{4}$ 时为正，并在 $x = \frac{3}{4}\pi$ 处首次等于零。因此当 $0 \le a < \frac{3}{4}\pi$ 时，分母在整个 $[0, a]$ 上都是正数。这样两个积分都是普通的积分，而下面取对数的也是一个正数。



把两个积分相加，

$$I+J=\int_0^a \frac{\cos x+\sin x}{\sin x+\cos x}\,\mathrm{d}x=\int_0^a 1\,\mathrm{d}x=a.$$

把它们相减，

$$I-J=\int_0^a \frac{\cos x-\sin x}{\sin x+\cos x}\,\mathrm{d}x.$$

分子 $\cos x-\sin x$ 正好是分母 $\sin x+\cos x$ 的导数，所以被积函数具有 $f'(x)/f(x)$ 的形式，其中在 $[0,a]$ 上 $f>0$。因此

$$I-J=\Big[\ln(\sin x+\cos x)\Big]_0^a=\ln(\sin a+\cos a)-\ln 1=\ln(\sin a+\cos a).$$

把两个结果相加，得

$$2I=a+\ln(\sin a+\cos a),$$

这正是所要求的结果。（把两者相减，还可以顺带得到 $2J=a-\ln(\sin a+\cos a)$。）

第 (i) 部分

由于 $p,q>0$，而 $\sin x$ 和 $\cos x$ 在 $[0,\frac{\pi}{2}]$ 上都非负，且不会同时为零，所以分母 $D(x)=p\sin x+q\cos x$ 在该区间上是正数。它的导数是 $D'(x)=p\cos x-q\sin x$。

想一想 当分母带有系数 p 和 q 时，“考虑 $I+J$ 与 $I-J$”这个做法应该怎样推广？

先前取和与差之所以奏效，是因为两个分子相加得出分母，相减则得出分母的导数。所以现在我们要找常数 A 和 B，使得

$$\cos x=A\,(p\sin x+q\cos x)+B\,(p\cos x-q\sin x)\qquad \text{对所有}x.$$

比较 $\sin x$ 和 $\cos x$ 的系数：

$$pA-qB=0,\qquad qA+pB=1.$$

由第一式得 $B=pA/q$；代入第二式得 $qA+p^2A/q=1$，所以

$$A=\frac{q}{p^2+q^2},\qquad B=\frac{p}{p^2+q^2}.$$

现在积分拆成容易的一部分和对数的一部分：

$$\begin{aligned}\int_0^{\pi/2}\frac{\cos x}{p\sin x+q\cos x}\,\mathrm{d}x&=\int_0^{\pi/2}\left(A+B\,\frac{D'(x)}{D(x)}\right)\mathrm{d}x\\&=\frac{\pi}{2}A+B\Big[\ln D(x)\Big]_0^{\pi/2}=\frac{\pi}{2}A+B\ln\frac{p}{q},\end{aligned}$$

其中用到 $D(\frac{\pi}{2})=p$ 和 $D(0)=q$，两者皆为正数。因此

$$\int_0^{\pi/2}\frac{\cos x}{p\sin x+q\cos x}\,\mathrm{d}x=\frac{\frac{1}{2}\pi q+p\ln(p/q)}{p^2+q^2}.$$

第 (ii) 部分

现在分母是 $D(x) = 3\sin x + 4\cos x + 25$，而 $D'(x) = 3\cos x - 4\sin x$。分母是正数，因为 $|3\sin x + 4\cos x| \le 5 < 25$。$D$ 中有常数 25，表示我们还须令常数项相符，所以要找 A 和 B，使得

$$\cos x + 4 = A\,(3\sin x + 4\cos x + 25) + B\,(3\cos x - 4\sin x).$$

比较系数：

$$\sin x:\ 3A - 4B = 0, \qquad \cos x:\ 4A + 3B = 1, \qquad \text{常数项}:\ 25A = 4.$$

由前两式得 $A = \frac{4}{25}$ 和 $B = \frac{3}{25}$，而常数项的方程 $25 \cdot \frac{4}{25} = 4$ 也同样成立。三条方程只有两个未知数，这并非必然成立：分子中的“+4”是特意选来令它成立的。所以

$$\begin{aligned}\int_0^{\pi/2} \frac{\cos x + 4}{3\sin x + 4\cos x + 25}\,\mathrm{d}x &= \int_0^{\pi/2} \left(\frac{4}{25} + \frac{3}{25}\,\frac{D'(x)}{D(x)}\right)\mathrm{d}x \\ &= \frac{4}{25}\cdot\frac{\pi}{2} + \frac{3}{25}\Big[\ln D(x)\Big]_0^{\pi/2} \\ &= \frac{2\pi}{25} + \frac{3}{25}\ln\frac{28}{29},\end{aligned}$$

因为 $D(\frac{\pi}{2}) = 3 + 25 = 28$ 而 $D(0) = 4 + 25 = 29$。

验算

- 取 $p = q = 1$ 时，第 (i) 部分应与主要结果在 $a = \frac{\pi}{2}$ 时的情况一致；后者给出 $2I = \frac{\pi}{2} + \ln 1$，所以 $I = \frac{\pi}{4}$。公式给出 $\frac{\pi/2+0}{2} = \frac{\pi}{4}$。两者吻合。
- 当 $a \to \frac{3}{4}\pi$ 时，$\ln(\sin a + \cos a) \to -\infty$。这与以下事实相符：在分母快要变成零之前，$I$ 的被积函数是绝对值很大的负数，因为在那里 $\cos x < 0$。
- 以数值计算，第 (ii) 部分是 $\frac{2\pi}{25} + \frac{3}{25}\ln\frac{28}{29} \approx 0.247116$，而对该积分使用辛普森法则（Simpson's rule）所得的结果与它吻合至小数点后十二位。主要结果和第 (i) 部分也分别对 200 个随机的 a 值，以及随机的 p 和 q 值，与数值积分（quadrature）的结果吻合。

常见错误

- 在第 (i) 部分得出 $\ln(q/p)$ 而不是 $\ln(p/q)$。在上限 $\frac{\pi}{2}$ 处，留下来的是 sin 那一项，所以得出 p。
- 在第 (ii) 部分忘记常数项，只比较 $\sin x$ 和 $\cos x$ 的系数。这样做碰巧行得通，但你应该加以验证，而不是假定它成立。

STEP 之外

把分子写成“A 乘以分母，加上 B 乘以分母的导数”，其实是一小段线性代数（linear algebra）。函数 $\sin x$ 和 $\cos x$ 张成一个二维空间，而只要 $p^2 + q^2 \neq 0$，D 和 D' 就构成这个空间的另一组基（basis）。同一个想法，即找出合适的基来展开，也是部分分式（partial fractions）和求解微分方程（differential equations）的待定系数法（method of undetermined coefficients）背后的原理，你会在大学一年级的数学方法课程中学到这些内容。Siklos 的 *Advanced Problems in Mathematics* 中还有更多可以用这类对称技巧解决的积分。

详细解答

STEP I 2002，第 8 题

纯数学 • 复利贷款

STEP I 2002，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/02-s1-q8

关键思路

先写下一年之内发生的事：债务先乘以因子 $1+\alpha$，然后减去还款额。把这个一步递推关系（recurrence）逐层展开，便可看出每笔还款 R_k 到第 n 年年底时“值”$(1+\alpha)^{n-k}R_k$，因为它令这么多的利息无从累积。其余的一切都只是等比级数（geometric series）。至于最后的近似值，用二项式定理（binomial theorem）展开 $1+\alpha$ 的幂，保留刚好足够的项即可。

解答

C_{n+1} 的一般公式

在第 k 年年底，加上利息后债务 C_k 变成 $(1+\alpha)C_k$，然后偿还 R_k。余下的就是第 $k+1$ 年年初的债务：

$$C_{k+1} = (1+\alpha)C_k - R_k.$$

重复应用这条关系式，

$$\begin{aligned} C_{n+1} &= (1+\alpha)C_n - R_n \\ &= (1+\alpha)^2 C_{n-1} - (1+\alpha)R_{n-1} - R_n \\ &\vdots \\ &= (1+\alpha)^n C_1 - (1+\alpha)^{n-1}R_1 - (1+\alpha)^{n-2}R_2 - \cdots - R_n, \end{aligned}$$

因此，由于 $C_1 = C$，

$$\boxed{C_{n+1} = (1+\alpha)^n C - \sum_{k=1}^{n}(1+\alpha)^{n-k}R_k}.$$

（要令这个论证严谨，可验证它在 $n=1$ 时成立，此时它说的是 $C_2=(1+\alpha)C-R_1$；并验证若它对 n 成立，则乘以 $1+\alpha$ 再减去 R_{n+1} 便得出 $n+1$ 的情况。）

还款额 $R_k=(1+\alpha)^k r$

在 N 年内还清贷款，即是 $C_{N+1}=0$。当 $R_k=(1+\alpha)^k r$ 时，和式中的每一项都相同：

$$(1+\alpha)^{N-k}R_k=(1+\alpha)^{N-k}(1+\alpha)^k r=(1+\alpha)^N r.$$

因此

$$0=C_{N+1}=(1+\alpha)^N C-N(1+\alpha)^N r=(1+\alpha)^N\,(C-Nr).$$

由于 $(1+\alpha)^N\neq 0$，我们得到 $C=Nr$，即 $r=C/N$，正是所求。

这个结果合乎情理：第 k 笔还款是“以今天的币值计”的 C/N，再加上 k 年的利息增长，所以你合共偿还的是本金（principal）C，另加恰好是它所生的利息。

等额还款 R

现在对每个 k 都有 $R_k=R$，而 $C_{N+1}=0$ 给出

$$(1+\alpha)^N C=R\sum_{k=1}^{N}(1+\alpha)^{N-k}=R\big(1+(1+\alpha)+\cdots+(1+\alpha)^{N-1}\big).$$

括号内是一个等比级数，首项为 1，公比为 $1+\alpha$，共 N 项，所以（因为 $\alpha>0$，公比不等于 1）它等于

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

于是 $(1+\alpha)^N C=R\,\dfrac{(1+\alpha)^N-1}{\alpha}$，整理后得

$$\frac{R}{C}=\frac{\alpha(1+\alpha)^N}{(1+\alpha)^N-1},$$

这就是所求的结果。

$\alpha=\frac{1}{50}$、$N=4$ 的情况

> ? **想一想** 二项展开式应该保留多少项？为甚么分子需要的项数比分母少？

分母 $(1+\alpha)^4-1$ 由 α 的一阶项开始，而分子 $\alpha(1+\alpha)^4$ 则是 α 乘以一个由 1 开始的式子。所以约去因子 α 之后，分子是 $(1+\alpha)^4$，分母是 $\big((1+\alpha)^4-1\big)/\alpha$。要令两者都准确至 α 的一阶，分子需要 $(1+\alpha)^4\approx 1+4\alpha$，但分母需要 $(1+\alpha)^4\approx 1+4\alpha+6\alpha^2$（除以 α 之后，$6\alpha^2$ 变成 6α）。因此

$$\frac{R}{C}=\frac{(1+\alpha)^4}{\big((1+\alpha)^4-1\big)/\alpha}=\frac{1+4\alpha+6\alpha^2+\ldots}{4+6\alpha+4\alpha^2+\alpha^3}\approx\frac{1+4\alpha}{4+6\alpha}.$$

代入 $\alpha = \frac{1}{50}$，

$$\frac{R}{C} \approx \frac{1 + \frac{4}{50}}{4 + \frac{6}{50}} = \frac{54}{206} = \frac{27}{103},$$

正是所求。我们舍去的项，相对大小约为 α^2，在这里只是百分之一的一小部分。

验算

- 当 $N = 1$ 时，公式给出 $R/C = 1 + \alpha$：借入 C，连同一年的利息一并偿还。正确。
- 当 $\alpha \to 0$ 时，$(1+\alpha)^N - 1 \approx N\alpha$，所以 $R/C \to 1/N$：没有利息时，每年只需偿还 C/N。这亦与 $r = C/N$ 吻合。
- 准确值是 $R/C = \dfrac{6765201}{25760050} \approx 0.26262$，而 $27/103 \approx 0.26214$；两者相差约 0.2%。取 $R = 0.26262\,C$，把债务逐年推算下去，四年后（在舍入误差之内）再无欠款。

常见错误

- 弄错和式中的指数：R_k 是在第 k 年年底偿还的，所以它少计的利息是往后的 $n-k$ 年，而不是 $n-k+1$ 年。
- 在分母中也用 $(1+\alpha)^4 \approx 1 + 4\alpha$ 作近似：这样会得出 $R/C \approx 27/100$，是一个较粗略的答案，与题目所求不符。

STEP 之外

$\sum(1+\alpha)^{-k}R_k$ 这个量是各笔还款的现值（present value），而贷款刚好还清，正是还款的现值等于借款额之时；R 的公式就是标准的年金（annuity，即按揭）公式。像 $C_{k+1} = (1+\alpha)C_k - R_k$ 这样的递推关系是一阶线性差分方程（first-order linear difference equation），可说是微分方程（differential equation）$y' = \alpha y - R(t)$ 的离散版近亲，两者都可用同一个“积分因子”（integrating factor）的想法求解。如想多做一些这类把代数与建模（modelling）结合起来的题目，可参阅 Siklos 的 *Advanced Problems in Mathematics*（进阶数学难题集）。

详细解答

STEP I 2002，第 9 题

力学 · 斜面上的倾倒

STEP I 2002，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：

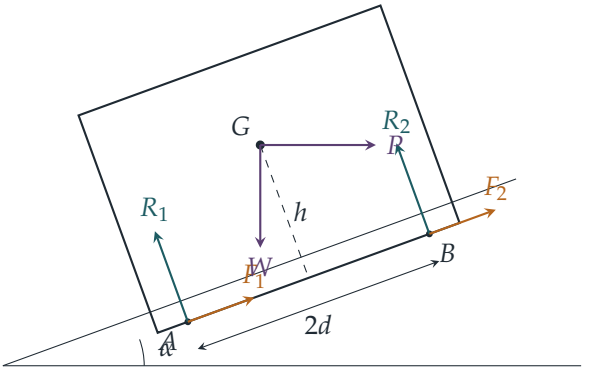
step.maths.org/questions/02-s1-q9

关键思路

货车是处于平衡（equilibrium）状态的刚体（rigid body），所以我们有三条方程：沿斜面方向分解力、沿垂直于斜面的方向分解力，以及取力矩（moments）。绕 G 取力矩是聪明的选择，因为重量和 P 都通过 G，两者一并消失，只剩下车轮上的力。这一条方程把两个法向反作用力（normal reaction）之差与总摩擦力（friction）联系起来，几乎包办了全部工作。于是“不翻倒”的意思不过是：两个法向反作用力都不是负数。

解答

我们把货车视为刚体，其车轮与斜面在两点接触，即 A（较低）和 B（较高），两点相距 $2d$。设 A 处的法向反作用力为 R_1，B 处的为 R_2，并设摩擦力为 F_1 和 F_2，两者均以沿斜面向上为正（任何一个最后都可能是负数）。我们假设摩擦力足以防止滑动；题目只关乎翻倒。如图所示，P 是水平的，指向斜面，所以它有沿斜面向上的分量 $P\cos\alpha$，以及压向斜面的分量 $P\sin\alpha$。



沿垂直于斜面的方向和沿斜面方向分解力（以沿斜面向上为正）：

$$R_1 + R_2 = W\cos\alpha + P\sin\alpha, \tag{1}$$

$$F_1 + F_2 = W\sin\alpha - P\cos\alpha. \tag{2}$$

现在绕 G 取力矩。重量和 P 的作用线都通过 G，所以它们没有任何贡献。每个法向反作用力与 G 的垂直距离为 d，而每个摩擦力沿斜面作用，与 G 的垂直距离为 h。以逆时针方向为正（如图所示，斜面向右上升），R_2 使货车绕 G 逆时针转动，R_1 使它顺时针转动，而沿斜面向上的摩擦力作用于 G 的下方，使它逆时针转动。所以

$$d(R_2 - R_1) + h(F_1 + F_2) = 0. \tag{3}$$

第 (i) 部分

若 $R_1 = R_2$，方程 (3) 变成 $h(F_1 + F_2) = 0$。由于 $h > 0$，可得 $F_1 + F_2 = 0$：摩擦力之和为零，正是所要证明的。

（由 (2) 可知，这种情况恰好在 $P = W\tan\alpha$ 时出现：P 沿斜面向上的部分刚好抵销重量沿斜面向下的部分，所以斜面无须提供任何净摩擦力。）

第 (ii) 部分

把 (2) 代入 (3)，

$$R_2 - R_1 = \frac{h}{d}(P\cos\alpha - W\sin\alpha),$$

再与 (1) 结合，

$$2R_1 = W\cos\alpha + P\sin\alpha - \frac{h}{d}(P\cos\alpha - W\sin\alpha),$$

$$2R_2 = W\cos\alpha + P\sin\alpha + \frac{h}{d}(P\cos\alpha - W\sin\alpha).$$

? **想一想** 用这些方程来说，“不翻倒”是甚么意思?

地面可以推车轮，却不能拉住它，所以每个法向反作用力都必须满足 $R \geq 0$。假如 R_2 的公式算出来是负数，较高的车轮就要靠外力压住才行；实际上它会离地，货车绕 A 转动。同样，$R_1 < 0$ 表示绕 B 转动。所以货车保持直立，当且仅当 $R_1 \geq 0$ 且 $R_2 \geq 0$。

以下我们一直使用 $\frac{d}{h} = \tan\beta$，其中 $0 < \beta < \frac{\pi}{2}$，并且 $0 < \alpha < \frac{\pi}{2}$。

较高的车轮不离地（$R_2 \geq 0$）。把 $2R_2 \geq 0$ 乘以 $d/h > 0$，

$$\tan\beta\,(W\cos\alpha + P\sin\alpha) + P\cos\alpha - W\sin\alpha \geq 0.$$

乘以 $\cos\beta > 0$ 并整理各项，

$$P(\cos\alpha\cos\beta + \sin\alpha\sin\beta) \geq W(\sin\alpha\cos\beta - \cos\alpha\sin\beta),$$

即

$$P\cos(\alpha-\beta)\geq W\sin(\alpha-\beta).$$

由于 $-\frac{\pi}{2}<\alpha-\beta<\frac{\pi}{2}$，我们有 $\cos(\alpha-\beta)>0$，所以可以除以它而无须把不等式（inequality）的方向倒转：

$$P\geq W\tan(\alpha-\beta).$$

较低的车轮不离地（$R_1\geq 0$）。同样地，由 $2R_1\geq 0$ 可得

$$\tan\beta\,(W\cos\alpha+P\sin\alpha)-P\cos\alpha+W\sin\alpha\geq 0,$$

再乘以 $\cos\beta>0$，

$$P(\cos\alpha\cos\beta-\sin\alpha\sin\beta)\leq W(\sin\alpha\cos\beta+\cos\alpha\sin\beta),$$

即

$$P\cos(\alpha+\beta)\leq W\sin(\alpha+\beta).$$

只要 $\alpha+\beta<\frac{\pi}{2}$（题目给出的答案其实暗中假设了这一点，否则 $\tan(\alpha+\beta)$ 是负数或者没有定义），便有 $\cos(\alpha+\beta)>0$，相除后得

$$P\leq W\tan(\alpha+\beta).$$

若 $\alpha+\beta\geq\frac{\pi}{2}$，则当 $P\geq 0$ 时左边 ≤ 0，而右边是正数，所以指向斜面的水平推力无论多大，都不能令货车绕较高的车轮翻倒。

把两个条件合起来，货车不翻倒当且仅当

$$W\tan(\alpha-\beta)\leq P\leq W\tan(\alpha+\beta),$$

这就是所要求的结果。推力太小，货车会绕 A 向斜面下方翻倒；推力太大，它会绕 B 被推得向斜面上方翻倒。

验算

- 几何图像。重量和 P 合成为通过 G 的单一个力，它与铅垂线成角 θ，其中 $\tan\theta=P/W$。当这个合力（resultant）通过车轮的接触点时，货车便处于即将翻倒的状态。直线 GA 与斜面的法线成角 β，而法线与铅垂线成角 α，所以合力在 $\theta=\alpha-\beta$ 时通过 A，在 $\theta=\alpha+\beta$ 时通过 B。这正好就是区间的两个端点。
- 没有斜坡。取 $\alpha=0$，条件变成 $|P|\leq Wd/h$：这正是大家熟悉的条件，即平地上的箱子在高度 h 处受横向推力而不翻倒的条件。
- 数值。取 $W=3$、$\alpha=0.4$、$d=1$、$h=1.5$，区间为 $-0.571\leq P\leq 4.551$。对 20 000 个随机配置直接检查 $R_1,R_2\geq 0$ 并与这个区间比对，结果完全吻合，而且在下端有 $R_2=0$，在上端有 $R_1=0$。

常见错误

- 绕车轮而不是绕 G 取力矩。这样做也行得通，但重量和 P 的力臂（moment arm）就变得很麻烦，正负号也容易出错。
- 忘记摩擦力对 G 有力矩。摩擦力作用于斜面上，在 G 下方距离 h 处，h 正是由此进入答案的。
- 没有检查 $\cos(\alpha \pm \beta)$ 的正负号就除以它，而不等式的方向正是由这个正负号决定的。

STEP 之外

车轮与地面之间的摩擦力是*静不定*（statically indeterminate）的：平衡条件只能确定 $F_1 + F_2$，永远不能分别确定 F_1 和 F_2，要知道两者如何分配，就需要一个描述轮胎如何变形的模型。工程师经常遇到这种情况，例如支承数目多于严格所需的结构。上面的几何验算，用的就是*合力作用线*（line of action of the resultant）和支承多边形（support polygon）的想法，步行机械人和叠起的积木的稳定性，背后也是同一个道理。Kleppner 与 Kolenkow 的 *An Introduction to Mechanics*（力学入门教科书）仔细讲解了刚体静力学（statics）。

详细解答

STEP I 2002，第 10 题

力学 · 弹性碰撞

STEP I 2002，第 10 题

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

step.maths.org/questions/02-s1-q10

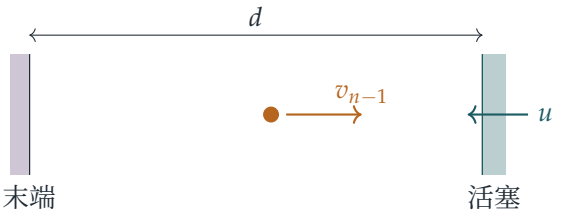
本题在 *Beyond STEP* 第 10.4 节一个分子与一个活塞中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch10/piston.html

关键思路

质点与一面远比它重的墙发生完全弹性碰撞（perfectly elastic collision）时，相对速度会反向，所以最简单的做法是在活塞（piston）的参考系（reference frame）中思考：在那里，质点只是从一面静止的墙反弹回来。因此，每次与活塞碰撞，速率都会增加 $2u$，而与固定端的碰撞则不会带来任何改变。至于距离，留意在两次与活塞碰撞之间，质点和活塞合共走过的长度有固定的规律，由此可得一条以 d_n 表示 d_{n+1} 的线性方程。

解答

我们把空气模拟为一个质点（particle），把活塞视为一面以恒定速率 u 移动、不受碰撞影响的墙，并把圆筒的末端视为一面固定的墙。所有运动都沿着轴线进行，所以我们取由圆筒末端指向活塞的方向为正方向。



速率

假设质点以速率 w 趋近活塞。活塞的速度为 $-u$，所以质点相对于活塞的速度为 $w+u$（朝向活塞）。在与一面不会被撞偏的墙发生完全弹性碰撞时，根据牛顿实验定律（Newton's experimental law），取 $e=1$，相对速度会反向，所以碰撞后质点相对于活塞的速度为 $-(w+u)$。它的实际速度于是为

$$-u-(w+u)=-(w+2u),$$

所以它以速率 $w+2u$ 离开活塞。

在圆筒的固定端，同一条定律（此时墙是静止的）只会令速度反向：速率不变。所以在相邻两次与活塞碰撞之间，质点的速率保持不变。

第一次碰撞之前，质点以速率 v 朝活塞运动，所以 $v_1=v+2u$。刚完成第 n 次碰撞后，它的速率为 v_n；它从末端反弹而速率不变，再以速率 v_n 遇上活塞，离开时速率为 $v_{n+1}=v_n+2u$。由归纳法（induction）可得

$$v_n=v+2nu,$$

正是所求。

距离

在第 n 次与第 $(n+1)$ 次碰撞之间（历时 t_n），活塞以速率 u 向内移动，所以

$$d_n-d_{n+1}=u\,t_n.$$

在同一段时间内，质点以恒定速率 v_n 运动，由活塞走到末端（距离为 d_n，因为末端与活塞原来的位置相距 d_n），再返回活塞（距离为 d_{n+1}，即活塞此刻所在的位置）。所以

$$d_n+d_{n+1}=v_nt_n.$$

> ? **想一想** 怎样才能消去 t_n？

把一条方程除以另一条：由于 $t_n>0$，所以

$$\frac{d_n+d_{n+1}}{d_n-d_{n+1}}=\frac{v_n}{u}.$$

交叉相乘，得 $ud_n+ud_{n+1}=v_nd_n-v_nd_{n+1}$，所以

$$(v_n+u)\,d_{n+1}=(v_n-u)\,d_n.$$

由 $v_n=v+2nu$，可得 $v_n-u=v+(2n-1)u$ 及 $v_n+u=v+(2n+1)u$，所以

$$d_{n+1}=\frac{v+(2n-1)u}{v+(2n+1)u}\,d_n,$$

这就是所求的结果。

d_n 的公式

重复应用这个递推关系（recurrence），

$$d_n = d_1 \prod_{k=1}^{n-1} \frac{v+(2k-1)u}{v+(2k+1)u} = d_1 \cdot \frac{v+u}{v+3u} \cdot \frac{v+3u}{v+5u} \cdots \frac{v+(2n-3)u}{v+(2n-1)u}.$$

这个乘积可以逐项相消（telescope）：每个分母都与下一个分子约去，余下

$$d_n = \frac{v+u}{v+(2n-1)u}\, d_1.$$

当 $n = 1$ 时这条式同样成立，此时它就是 $d_1 = d_1$。

$v = u$ 的情况

代入 $v = u$，得

$$d_n = \frac{2u}{2nu}\, d_1 = \frac{d_1}{n}.$$

然后，由 $ut_n = d_n - d_{n+1}$，

$$ut_n = \frac{d_1}{n} - \frac{d_1}{n+1} = \frac{d_1}{n(n+1)},$$

正是所求。

验算

- 当 $u = 0$ 时，活塞是一面固定的墙：公式给出对所有 n 都有 $v_n = v$ 及 $d_n = d_1$，与预期相符。
- 取 $u = 1$、$v = 3$、$d_1 = 75$。公式预测 $d_2 = 50$、$d_3 = 37.5$、$d_4 = 30$，而速率为 $5, 7, 9, 11$。以有理数运算逐步精确模拟反弹过程，所得结果与此完全相同；随机选取 200 组 u、v 和起始位置，结果亦同样吻合。
- 当 $v = u$ 时，$\sum ut_n = d_1 \sum \left(\frac{1}{n} - \frac{1}{n+1}\right) = d_1$：活塞合共恰好移动了 d_1，到达圆筒的末端，这与 $d_n = d_1/n \to 0$ 一致。

常见错误

- 错用 $d_n - d_{n+1}$，例如把质点来回一趟的路程当作 $2d_n$：活塞在这段时间内已经移动了，所以回程只有 d_{n+1}。
- 每次碰撞只加上 u 而不是 $2u$。相对速度会反向，所以质点速度的改变量是活塞所带来的趋近速率的两倍。

STEP 之外

每次碰撞都令质点的速率改变一个固定的量，但随着间隙收窄，碰撞会愈来愈频密，而当 n 很大时，乘积 $v_n d_n$ 大致保持不变（这里当 n 很大时 $v_n d_n \approx (v+u)d_1$）。这种缓慢变化的守恒量称为绝热不变量（adiabatic invariant）；对大量分子取平均后，它解释了为何气体被迅速压缩时会变热：对于一维的缓慢压缩，$T \propto 1/L^2$。Kleppner 与 Kolenkow 合著的 *An Introduction to Mechanics*（力学入门）是进一步学习碰撞和参考系的好选择。

详细解答

STEP I 2002，第 11 题

力学 · 弹性斜碰撞

STEP I 2002，第 11 题

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

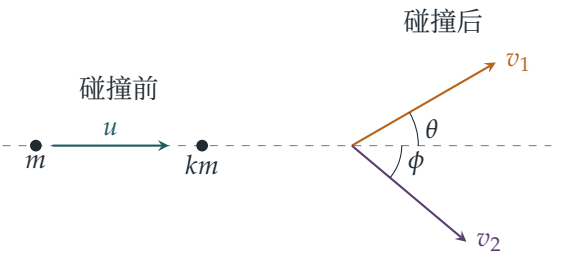
step.maths.org/questions/02-s1-q11

关键思路

一次碰撞给我们两条守恒定律：动量（momentum）守恒（动量是向量，所以有两条分量方程），以及动能（kinetic energy）守恒（因为没有能量损失）。两条动量方程对于碰撞后的未知速率而言是线性的，所以我们可以解出这些速率，并以角度表示。再代入能量方程，便可消去所有速率，只剩下角度与 k 之间的关系式。

解答

我们把 P_1 和 P_2 看成在平面上运动的质点（particle），并假设碰撞期间没有外力作用。设 P_1 以速率 $u > 0$ 沿 x 轴趋近，并设碰撞后的速率分别为 v_1（P_1 的速率，方向在轴的上方，与轴成角 θ）和 v_2（P_2 的速率，方向在轴的下方，与轴成角 ϕ），其中 $0 < \theta < \pi$ 及 $0 < \phi < \pi$。



恒等式

沿原来的运动方向及与之垂直的方向应用动量守恒，（约去 m 后）得

$$v_1\cos\theta + kv_2\cos\phi = u,$$
$$v_1\sin\theta - kv_2\sin\phi = 0,$$

其中出现负号，是因为 P_2 向轴的另一侧离开。由动能守恒，约去 $\frac{1}{2}m$ 后得

$$v_1^2 + kv_2^2 = u^2. \qquad (*)$$

两条动量方程是关于两个未知量 v_1 和 kv_2 的线性方程。把第一式乘以 $\sin\phi$，第二式乘以 $\cos\phi$，然后相加：含 kv_2 的项互相抵消，于是

$$v_1(\sin\theta\cos\phi + \cos\theta\sin\phi) = u\sin\phi, \qquad \text{所以} \qquad v_1 = \frac{u\sin\phi}{\sin(\theta+\phi)}.$$

同样地，把第一式乘以 $\sin\theta$，第二式乘以 $\cos\theta$，然后相减，得

$$kv_2 = \frac{u\sin\theta}{\sin(\theta+\phi)}.$$

想一想 我们可以除以 $\sin(\theta+\phi)$ 吗?

我们需要 $\theta+\phi \neq \pi$（它严格介乎 0 与 2π 之间）。若 $\theta+\phi=\pi$，则 $\sin\phi = \sin\theta$ 及 $\cos\phi = -\cos\theta$。第二条动量方程变成 $(v_1 - kv_2)\sin\theta = 0$，而 $\sin\theta \neq 0$，所以 $v_1 = kv_2$。这时第一条方程变成 $(v_1 - kv_2)\cos\theta = u$，即 $0 = u$，这是不可能的。因此 $\sin(\theta+\phi) \neq 0$，这个除法并无问题。

现在把结果代入 $(*)$，并把 kv_2^2 写成 $(kv_2)^2/k$：

$$\frac{u^2\sin^2\phi}{\sin^2(\theta+\phi)} + \frac{u^2\sin^2\theta}{k\sin^2(\theta+\phi)} = u^2.$$

全式乘以 $k\sin^2(\theta+\phi)/u^2$（此数不等于零），得

$$\sin^2\theta + k\sin^2\phi = k\sin^2(\theta+\phi),$$

这正是所要证明的结果。

（从几何角度看，动量方程表示动量 mv_1、kmv_2 和 mu 构成一个三角形，其内角为 ϕ、θ 和 $\pi-(\theta+\phi)$，而上面两个表达式不过是这个三角形的正弦定理（sine rule）。）

两条路径互相垂直

若两条路径互相垂直，则 $\theta+\phi = \frac{\pi}{2}$，所以 $\sin^2(\theta+\phi) = 1$ 及 $\sin\theta = \cos\phi$。恒等式变成

$$\cos^2\phi + k\sin^2\phi = k, \qquad \text{即} \qquad \cos^2\phi = k(1-\sin^2\phi) = k\cos^2\phi.$$

这里 $\cos\phi \neq 0$：否则 $\phi = \frac{\pi}{2}$，于是 $\theta = 0$，但题目已说明 θ 不等于零。两边除以 $\cos^2\phi$，得 $k = 1$，这正是所要证明的结果。

验算

- 桌球中为人熟知的事实：两个质量相等的球作弹性碰撞（elastic collision），而其中一个原本静止，碰撞后两球的运动方向互相垂直。我们的结果正是它的逆命题（converse）。
- 一个数值例子：取 $k = 2$ 及 $\phi = 30°$。沿该直线的光滑弹性碰撞会使 P_2 以速率 $\frac{2}{3}u\cos 30°$ 离开，而 P_1 则沿 $\theta = 90°$ 的方向运动。恒等式两边都等于 $\frac{3}{2}$。
- 我们亦模拟了 2×10^5 次随机弹性碰撞（随机选取 k、u 和碰撞线）。恒等式以及 v_1 和 kv_2 的两条公式均成立，误差在 10^{-12} 以内。

常见错误

- 把垂直方向的动量方程写成 $v_1\sin\theta + kv_2\sin\phi = 0$：两个质点向相反的两侧离开，所以两个分量的符号相反。
- 在能量方程中漏掉 k，或在最后一步没有说明为何 $\cos\phi \neq 0$。

STEP 之外

在随质心（centre of mass）运动的参考系中，弹性碰撞只是把两个速度旋转，而大小保持不变。所有这些角度关系，都是把这个由各种可能结果组成的圆映射回实验室参考系而得出的。这个观点是处理散射（scattering）的标准方法，由桌球以至粒子物理学皆然。在粒子物理学中，人们利用量度所得的散射角来推算相撞粒子的质量比。Kleppner 与 Kolenkow 所著的 *An Introduction to Mechanics*（力学入门）仔细阐述了以质心处理碰撞的方法。

详细解答

STEP I 2002，第 12 题

概率与统计 • 几何概率

STEP I 2002，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/02-s1-q12

关键思路

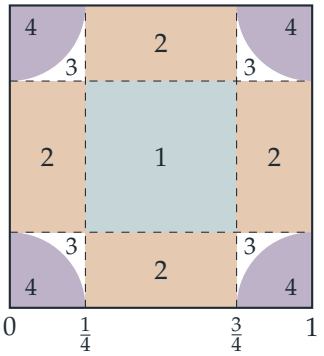
方格网不断重复，所以“随机放下马蹄”的意思，就是马蹄的中心在一个单位正方形上均匀分布（uniformly distributed）。马蹄碰到多少个方格，只取决于中心离该正方形的边和角有多近，因此每个答案都对应正方形内的一个区域，而它的概率就是该区域的面积。求出四个概率之后，余下的工作就是把这些概率及其平均值，与 Harry 的实际表现作比较。

解答

Harry 答案的分布

设马蹄的中心 C 位于单位正方形 $0 \le x \le 1$、$0 \le y \le 1$ 之内。若 Harry 随机放下马蹄，则由方格网的周期性（periodicity）可知 C 在这个正方形上均匀分布，所以 C 落在某个区域内的概率就是该区域的面积。设 N 为被马蹄部分覆盖的方格数目。

马蹄的半径为 $\frac{1}{4}$，小于 $\frac{1}{2}$，所以它最多只能跨过一条垂直格线和一条水平格线。它跨过垂直格线，当且仅当 $x < \frac{1}{4}$ 或 $x > \frac{3}{4}$；它跨过水平格线，当且仅当 $y < \frac{1}{4}$ 或 $y > \frac{3}{4}$。（与格线的距离刚好等于 $\frac{1}{4}$ 的位置面积为零，所以可以不理。）



“跨过垂直格线”和“跨过水平格线”这两个事件的概率各为 $\frac{1}{2}$，而且两者独立（independent），因为 x 和 y 是互相独立的均匀坐标。所以：

- $N = 1$ 即马蹄没有跨过任何格线：此时 C 位于中央边长为 $\frac{1}{2}$ 的正方形内，所以 $\mathrm{P}(N = 1) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$。
- $N = 2$ 即马蹄刚好跨过一条格线：$\mathrm{P}(N = 2) = 2 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}$。
- 当马蹄跨过垂直和水平格线各一条时，C 位于四个边长为 $\frac{1}{4}$ 的角落正方形其中之一，总概率为 $\frac{1}{4}$。这两条格线相交于方格网的一个格点，而马蹄覆盖该格点周围全部四个方格的一部分，当且仅当它覆盖该格点本身。

想一想 在角落正方形内，为甚么答案取决于与格点的距离？

考虑最接近原点的角落正方形，即 $0 < x < \frac{1}{4}$ 及 $0 < y < \frac{1}{4}$。马蹄必定碰到 C 所在的方格、它左边的方格和它下面的方格。至于对角的方格 $x < 0$、$y < 0$，其中最接近的点是原点，与 C 的距离为 $\sqrt{x^2 + y^2}$，所以马蹄进入这个方格，当且仅当 $\sqrt{x^2 + y^2} < \frac{1}{4}$。因此，当 C 与某个格点的距离小于 $\frac{1}{4}$ 时 $N = 4$，否则 $N = 3$。

单位正方形内与其中一个顶点的距离小于 $\frac{1}{4}$ 的点，组成四个半径为 $\frac{1}{4}$ 的四分之一圆盘，总面积为 $\pi\left(\frac{1}{4}\right)^2 = \frac{\pi}{16}$。因此

$$\mathrm{P}(N = 1) = \tfrac{1}{4}, \qquad \mathrm{P}(N = 2) = \tfrac{1}{2}, \qquad \mathrm{P}(N = 3) = \tfrac{1}{4} - \tfrac{\pi}{16}, \qquad \mathrm{P}(N = 4) = \tfrac{\pi}{16}.$$

目前的证据与“随机的马”是否吻合？

当正确答案是 $4 \div 4 = 1$ 时，一匹随机的马答对的概率是 $\mathrm{P}(N = 1) = \frac{1}{4}$。当正确答案是 $2 + 2 = 4$ 时，牠答对的概率是 $\mathrm{P}(N = 4) = \frac{\pi}{16} \approx 0.196$。这正正就是观察所得的成功率，所以答案是肯定的：这些证据与我的怀疑完全吻合，即 Harry 只是随机放下马蹄。

答案的期望值，以及换一匹新马

若 Harry 随机放下马蹄，他的答案便与题目无关，所以他的答案的期望值（expected value）就是 $\mathrm{E}(N)$（各个正确答案出现的频率相同这一点，在这里不起作用）：

$$\begin{aligned}\mathrm{E}(N) &= 1\cdot\tfrac{1}{4} + 2\cdot\tfrac{1}{2} + 3\left(\tfrac{1}{4} - \tfrac{\pi}{16}\right) + 4\cdot\tfrac{\pi}{16}\\ &= \tfrac{1}{4} + 1 + \tfrac{3}{4} + \tfrac{\pi}{16}\\ &= 2 + \tfrac{\pi}{16} \approx 2.20.\end{aligned}$$

把观察所得的平均值 2 与两个假设作比较。

- *Harry 计算正确*。正确答案是 $1, 2, 3, 4$，各自出现的次数相同，所以平均值应为 $\frac{5}{2}$。平均值 2 远低于此，所以 Harry 肯定答错了很多题。事实上，每个错误答案最多比正确答案小 3，而平均少了 $\frac{1}{2}$，所以他的答案中至少有 $\frac{1}{6}$ 是错的。
- *Harry 是随机的*。平均值应接近 $2 + \frac{\pi}{16} \approx 2.20$。单一答案的方差（variance）为 $\mathrm{E}(N^2) - \mathrm{E}(N)^2 = \frac{9}{2} + \frac{7\pi}{16} - \left(2 + \frac{\pi}{16}\right)^2 \approx 1.05$，所以做了 n 道题之后，平均值的标准差（standard deviation）约为 $1.02/\sqrt{n}$。题目“很多”的时候（例如 $n = 400$，标准差约为 0.05），平均值 2 低了大约四个标准差，纯粹出于偶然的可能性极低。

所以，做了很多道题之后，平均值 2 即使对一匹随机放下马蹄的马来说也太低：Harry 似乎偏好较小的答案，仿佛他下蹄很轻，或者瞄准方格的中央。无论如何，他并没有在做计算，而且表现比一匹随机的马还要差。是的：是时候换一匹新马了（不过也许该找一匹会做算术的）。

验算

- 四个概率之和为一：$\frac{1}{4} + \frac{1}{2} + \frac{1}{4} - \frac{\pi}{16} + \frac{\pi}{16} = 1$，而 $\mathrm{P}(N = 3) = \frac{1}{4} - \frac{\pi}{16} \approx 0.054$ 为正数，这是必须的。
- 模拟（simulation）400 000 次随机放下马蹄，并数出与中心的距离小于 $\frac{1}{4}$ 的方格数目，所得的相对频率为 0.2495、0.4990、0.0547、0.1968，平均值为 2.199，而理论值为 $2 + \frac{\pi}{16} \approx 2.196$。
- 合理性检查：若马蹄非常小，答案几乎总是 1，平均值会趋向 1。马蹄愈大，答案愈大，而我们的情况介乎两者之间。

常见错误

- 以为马蹄只要跨过两条格线，就一定覆盖四个方格。其实它还要碰到两条格线相交的格点，而 π 正是因此而出现。

► 因为平均值 2“低于 2.5”，就把它当作随机的证据。随机的假设所预测的平均值约为 2.2，所以必须与这个数值比较，并考虑平均值的离散程度。

STEP 之外

这就是几何概率（geometric probability）：当一点在某个区域上均匀分布时，概率就是面积。同样的想法可以用来处理蒲丰投针问题（Buffon's needle），其答案同样涉及 π，十分著名。最后一部分让你初尝假设检验（hypothesis testing）的滋味：利用中心极限定理（central limit theorem）来判断在某个模型之下，观察所得的平均值是否合理。Grimmett 与 Welsh 的 *Probability: An Introduction*（概率论入门）涵盖了连续均匀分布，以及这类论证背后的极限定理。

详细解答

STEP I 2002，第 13 题

概率与统计 • 条件期望值

STEP I 2002，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/02-s1-q13

关键思路

(U, V) 只有六个可能的组合，所以整件事归结为一个小小的联合概率（joint probability）表，再对每种情况做少许代数运算。对于二次方程，V 的正负号决定一切，因为 U^2 最多是 1。对于三次方程，各项系数与 U 和 V 纠缠得令人起疑；这强烈暗示它可以因式分解（factorise），而一旦分解出来，各个根便一目了然。

解答

首先建立联合分布（joint distribution）。由于 $\mathrm{P}(U = u,\, V = v) = \mathrm{P}(U = u)\,\mathrm{P}(V = v \mid U = u)$，而对每个 u 都有 $\mathrm{P}(U = u) = \frac{1}{3}$，

	$U = 1$	$U = 0$	$U = -1$
$V = 1$	$\frac{1}{9}$	$\frac{1}{6}$	$\frac{2}{9}$
$V = -1$	$\frac{2}{9}$	$\frac{1}{6}$	$\frac{1}{9}$

六个数值加起来是 1，理应如此。

第 (i) 部分

$x^2 + Ux + V = 0$ 的两个根皆为实数，当且仅当判别式（discriminant）非负：$U^2 - 4V \geq 0$。

- 若 $V = -1$，判别式为 $U^2 + 4 > 0$，所以无论 U 取甚么值，两个根都是实数（而且相异）。
- 若 $V = 1$，由于 $U^2 \leq 1$，判别式为 $U^2 - 4 \leq 1 - 4 < 0$。所以两个根永远不是实数。

因此，两个根皆为实数当且仅当 $V=-1$，而

$$\mathrm{P}(\text{两根皆为实数})=\mathrm{P}(V=-1)=\tfrac{2}{9}+\tfrac{1}{6}+\tfrac{1}{9}=\tfrac{4+3+2}{18}=\tfrac{1}{2},$$

正是所求。

第 (ii) 部分

根据第 (i) 部分，以“两根皆为实数”为条件，等同于以 $V=-1$ 为条件。把表中最下一行除以 $\mathrm{P}(V=-1)=\frac{1}{2}$，得

$$\mathrm{P}(U=1\mid V=-1)=\tfrac{4}{9},\qquad \mathrm{P}(U=0\mid V=-1)=\tfrac{1}{3},\qquad \mathrm{P}(U=-1\mid V=-1)=\tfrac{2}{9}.$$

想一想 为甚么这里不能直接用 $\mathrm{P}(U=u)=\frac{1}{3}$?

因为 U 与 V 并非独立（independent）：知道 $V=-1$ 之后，$U=1$ 的可能性便会提高（因为当 $U=1$ 时，$V=-1$ 的概率是 $\frac{2}{3}$）。我们要求的是条件期望值，所以权重必须是条件概率（conditional probability）。

当 $V=-1$ 时，方程为 $x^2+Ux-1=0$，其较大的根是

$$R=\frac{-U+\sqrt{U^2+4}}{2}.$$

所以当 $U=1$ 时 $R=\frac{1}{2}(\sqrt{5}-1)$，当 $U=0$ 时 $R=1$，当 $U=-1$ 时 $R=\frac{1}{2}(\sqrt{5}+1)$。因此

$$\begin{aligned}\mathrm{E}(R\mid\text{两根为实数})&=\tfrac{4}{9}\cdot\frac{\sqrt{5}-1}{2}+\tfrac{1}{3}\cdot 1+\tfrac{2}{9}\cdot\frac{\sqrt{5}+1}{2}\\&=\frac{4\sqrt{5}-4+6+2\sqrt{5}+2}{18}\\&=\frac{6\sqrt{5}+4}{18}=\frac{3\sqrt{5}+2}{9}.\end{aligned}$$

第 (iii) 部分

常数项是 U，这提示我们有因式 $(x+U)$。果然

$$\begin{aligned}(x+U)(x^2-2Vx+1)&=x^3-2Vx^2+x+Ux^2-2UVx+U\\&=x^3+(U-2V)x^2+(1-2UV)x+U,\end{aligned}$$

这正是题目给出的三次式。所以方程的根是 $x=-U$，连同 $x^2-2Vx+1=0$ 的根。由于 $V=\pm1$ 给出 $V^2=1$，这个二次方程的判别式为 $4V^2-4=0$。所以它有重根（repeated root）$x=V$，而三次方程就是

$$(x+U)(x-V)^2=0,$$

其根为 $-U$、V、V。

三个根全为正数，当且仅当 $-U>0$ 且 $V>0$，即 $U=-1$ 且 $V=1$。（当 $U=0$ 时，其中一个根是 0，并非正数。）从表中可得

$$\mathrm{P}(\text{所有根皆为正数})=\mathrm{P}(U=-1,\,V=1)=\tfrac{1}{3}\cdot\tfrac{2}{3}=\tfrac{2}{9}.$$

验算

- 条件概率 $\frac{4}{9},\frac{1}{3},\frac{2}{9}$ 加起来是 1。
- 第 (ii) 部分的答案约为 0.967。它必须介乎较大根的最小可能值 $\frac{1}{2}(\sqrt{5}-1)\approx 0.618$ 与最大可能值 $\frac{1}{2}(\sqrt{5}+1)\approx 1.618$ 之间，而且应该低于中间值 1，因为最小的根所占的权重最大。结果确是如此。
- 在第 (iii) 部分，试代入 $U=-1$、$V=1$：三次式变成 $x^3-3x^2+3x-1=(x-1)^3$，根为 $1,1,1$，全是正数。再试 $U=1$、$V=1$：三次式变成 $x^3-x^2-x+1=(x+1)(x-1)^2$，它有负根 -1。
- 我们以分数对六种情况逐一精确枚举，并在每种情况下以数值方法求出三次方程的根，从而确认了全部三个答案。

常见错误

- 在第 (ii) 部分使用无条件的权重 $\frac{1}{3}$，结果得出 $\frac{1}{3}\left(\sqrt{5}+1\right)\approx 1.079$。
- 在第 (iii) 部分把 $U=0$ 也计算在内：等于 0 的根并非正数。
- 试图用判别式或正负号变化来分析三次方程，而不去寻找因式分解。

STEP 之外

这道题的核心是条件期望值（conditional expectation）$\mathrm{E}(X\mid A)$，计算方法是以条件概率重新加权。到了大学，这个概念会发展成本身就是一个随机变量（random variable）的条件期望值 $\mathrm{E}(X\mid Y)$，以及塔式法则（tower law）$\mathrm{E}\big(\mathrm{E}(X\mid Y)\big)=\mathrm{E}(X)$，后者能令许多困难的期望值计算变得轻而易举。Grimmett 与 Welsh 所著的 *Probability: An Introduction*（概率论入门教材）对此有清晰的阐述。

详细解答

STEP I 2002，第 14 题

概率与统计 • 概率中的等比级数

STEP I 2002，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/02-s1-q14

关键思路

以我总共按了多少个号码作为条件 (conditioning)。这个数目一旦确定，警钟响起与我取得款项之间相隔的时间也随之确定，于是警察先到的概率就是对应该段时间的 $1-\mathrm{e}^{-\lambda t}$。再对尝试次数取平均，便得出答案。在两种方案中，这个平均都会化为以 $q=\mathrm{e}^{-\lambda/r}$ 为公比的等比级数（geometric series），所以值得及早为 q 命名。

解答

顺序按键

设 K 为我的号码在 $1,2,\ldots,n$ 这列数中的位置。那么 K 就是我尝试的次数，而且对 $k=1,\ldots,n$ 有 $P(K=k)=\frac{1}{n}$。

假设 $K=k$。我每分钟按 r 个号码，所以相邻两个号码相隔 $1/r$ 分钟。时间由我按下第一个号码的一刻起计。

- 若 $k=1$，第一个号码就是正确的，我从未输入过错误号码，警察也不会接到通知。警察先到的概率是 $0=1-\mathrm{e}^{0}$。
- 若 $k\geq 2$，第一个号码是错的，所以警钟在时间 0 响起，而正确号码是第 k 个，在时间 $(k-1)/r$ 输入。如果警察在 $t=(k-1)/r$ 分钟内到达，他们就会在我取得款项之前到达，其概率为 $1-\mathrm{e}^{-\lambda(k-1)/r}$。

因此公式 $1-\mathrm{e}^{-\lambda(k-1)/r}$ 适用于所有 k，包括 $k=1$。设 A 为“警察在我取得款项之前到

达”这个事件。根据全概率公式（law of total probability），

$$P(A)=\sum_{k=1}^{n}P(K=k)\,P(A\mid K=k)=\sum_{k=1}^{n}\frac{1-\mathrm{e}^{-\lambda(k-1)/r}}{n},$$

正如题目所求。

为了化简，设 $q=\mathrm{e}^{-\lambda/r}$。由于 $\lambda,r>0$，我们有 $0<q<1$，而且 $\mathrm{e}^{-\lambda(k-1)/r}=q^{k-1}$。所以

$$\begin{aligned}P(A)&=1-\frac{1}{n}\sum_{k=1}^{n}q^{k-1}=1-\frac{1-q^n}{n(1-q)}\\&=1-\frac{1-\mathrm{e}^{-\lambda n/r}}{n\left(1-\mathrm{e}^{-\lambda/r}\right)},\end{aligned}$$

这里用了等比级数求和，因为 $q\neq 1$，所以这一步成立。

随机按键

现在每次尝试都是从 $1,\ldots,n$ 中均匀地重新抽出一个号码，与之前的尝试独立（independent）（我不记得自己试过哪些号码）。所以每次尝试正确的概率都是 $\frac{1}{n}$，而且各次互相独立。若 N 是直至按中正确号码为止（包括该次）的尝试次数，则 N 服从几何分布（geometric distribution）：

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

想一想 当号码可以重复时，关于时间的论证是否仍然成立?

成立。警察在第一个错误号码出现时接到通知，而若 $N=k\geq 2$，那正是第一次尝试，发生在时间 0。款项在第 k 次尝试时到手，即时间 $(k-1)/r$。重复号码只会改变 N 的分布，不会改变时间的计算。所以与之前完全一样，$P(A\mid N=k)=1-q^{k-1}$，而这对 $k=1$ 也是对的。

记 $p=\frac{n-1}{n}$。那么

$$\begin{aligned}P(A)&=\sum_{k=1}^{\infty}\frac{1}{n}\,p^{k-1}\left(1-q^{k-1}\right)=\frac{1}{n}\sum_{k=1}^{\infty}p^{k-1}-\frac{1}{n}\sum_{k=1}^{\infty}(pq)^{k-1}\\&=\frac{1}{n}\cdot\frac{1}{1-p}-\frac{1}{n}\cdot\frac{1}{1-pq}.\end{aligned}$$

把和式拆开是可以的，因为两个级数都收敛（converge）：$0\leq p<1$ 而且 $0<pq<1$。第一项是 $\frac{1}{n}\cdot n=1$，这只不过是说 N 必定是有限的。至于第二项，

$$n(1-pq)=n-(n-1)q,$$

所以

$$P(A) = 1 - \frac{1}{n - (n-1)\mathrm{e}^{-\lambda/r}},$$

正如题目所求。

验算

- 当 $n = 1$ 时，永远不会出现错误号码，而两条公式都给出 0：第一条是因为 $1 - q^1 = 1 - q$，第二条是因为分母等于 1。
- 当 $\lambda \to \infty$（警察即时到达）时，$q \to 0$，两个答案都趋向 $1 - \frac{1}{n}$。这正是第一次尝试出错的概率，理应如此。
- 随机按键应该较差。取 $n = 5$、$\lambda = 0.7$、$r = 3$，顺序按键得 0.3382，随机按键得 0.4543。两种方案各做 2×10^5 次模拟试验，结果分别是 0.339 和 0.456。

常见错误

- 把时间写成 k/r 而不是 $(k-1)/r$。计时由第一个（错误的）号码开始，所以第 k 个号码是在 $k - 1$ 个间隔之后出现。
- 在随机的情况下，给每个 k 的概率都是 $\frac{1}{n}$，或者求和只加到 n 为止。既然容许重复，N 服从几何分布，而且没有上界。

STEP 之外

这里警察的到达时间是一个指数随机变量（exponential random variable），而整个计算就是对随机的 X 求 $P(T < X)$，方法是以 X 为条件。当 X 服从几何分布时，答案有简洁的闭合形式（closed form），因为 $E\left(q^X\right)$ 正是 X 的概率生成函数（probability generating function）在 q 处的值。事实上，两个答案都是 $1 - E\left(q^{N-1}\right)$。生成函数和指数分布是大学一年级概率论的核心内容。Grimmett 与 Welsh 合著的 *Probability: An Introduction*（概率论入门）对两者都有很好的讲解。

详细解答

STEP II 2002，第 1 题

纯数学 · 代换积分法

STEP II 2002，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/02-s2-q1

关键思路

倍角公式（double-angle formula）把 $1-\cos 2\theta$ 化为 $2\sin^2\theta$，而 $\operatorname{cosec}^2\theta$ 是一眼就能积分的函数。题目建议的代换（substitution）$x=\sin 2\theta$ 是特意设计的，使 $\sqrt{1-x^2}$ 变成 $\cos 2\theta$，从而把第二个积分（integral）拉回第一个积分。至于最后一个积分，积分限（limits）1 和 $2/\sqrt{3}$ 正是积分限 1 和 $\sqrt{3}/2$ 的倒数（reciprocal），这强烈暗示应该试用 $y=1/x$。

解答

第一个积分

由于 $\cos 2\theta = 1-2\sin^2\theta$，我们有 $1-\cos 2\theta = 2\sin^2\theta$，它在 $[\frac{1}{6}\pi, \frac{1}{4}\pi]$ 上为正，所以被积函数（integrand）在该区间上性质良好。利用 $\frac{\mathrm{d}}{\mathrm{d}\theta}(\cot\theta) = -\operatorname{cosec}^2\theta$，

$$\begin{aligned}\int_{\pi/6}^{\pi/4} \frac{1}{1-\cos 2\theta}\,\mathrm{d}\theta &= \frac{1}{2}\int_{\pi/6}^{\pi/4} \operatorname{cosec}^2\theta\,\mathrm{d}\theta = \frac{1}{2}\Big[-\cot\theta\Big]_{\pi/6}^{\pi/4} \\ &= \frac{1}{2}\left(\cot\tfrac{\pi}{6} - \cot\tfrac{\pi}{4}\right) = \frac{1}{2}\left(\sqrt{3}-1\right) = \frac{\sqrt{3}}{2} - \frac{1}{2},\end{aligned}$$

正如所求。

第二个积分

设 $x=\sin 2\theta$，则 $\mathrm{d}x = 2\cos 2\theta\,\mathrm{d}\theta$。当 $\theta=\frac{1}{6}\pi$ 时，$x=\sin\frac{1}{3}\pi=\frac{\sqrt{3}}{2}$；当 $\theta=\frac{1}{4}\pi$ 时，$x=\sin\frac{1}{2}\pi=1$。当 θ 走遍 $[\frac{1}{6}\pi, \frac{1}{4}\pi]$ 时，x 由 $\frac{\sqrt{3}}{2}$ 稳定地递增至 1。

想一想 $\sqrt{1-x^2}$ 等于 $\cos 2\theta$，还是等于 $-\cos 2\theta$?

我们有 $\sqrt{1-x^2} = \sqrt{\cos^2 2\theta} = |\cos 2\theta|$。在我们的范围内 $2\theta \in [\frac{1}{3}\pi, \frac{1}{2}\pi]$，此处 $\cos 2\theta \geq 0$，所以 $\sqrt{1-x^2} = \cos 2\theta$。因此

$$\int_{\sqrt{3}/2}^{1} \frac{1}{1-\sqrt{1-x^2}}\,\mathrm{d}x = \int_{\pi/6}^{\pi/4} \frac{2\cos 2\theta}{1-\cos 2\theta}\,\mathrm{d}\theta.$$

为了引入第一个积分，把分子用分母来表示：

$$\frac{\cos 2\theta}{1-\cos 2\theta} = \frac{1-(1-\cos 2\theta)}{1-\cos 2\theta} = \frac{1}{1-\cos 2\theta} - 1.$$

于是，利用第一个结果，

$$\begin{aligned}\int_{\sqrt{3}/2}^{1} \frac{1}{1-\sqrt{1-x^2}}\,\mathrm{d}x &= 2\int_{\pi/6}^{\pi/4} \frac{1}{1-\cos 2\theta}\,\mathrm{d}\theta - 2\int_{\pi/6}^{\pi/4} 1\,\mathrm{d}\theta \\ &= 2\left(\frac{\sqrt{3}}{2} - \frac{1}{2}\right) - 2\left(\frac{\pi}{4} - \frac{\pi}{6}\right) \\ &= \sqrt{3} - 1 - \frac{\pi}{6},\end{aligned}$$

正如所求。

最后一个积分

设 $y = 1/x$，则 $\mathrm{d}y = -\mathrm{d}x/x^2$。当 $y = 1$ 时，$x = 1$；当 $y = 2/\sqrt{3}$ 时，$x = \sqrt{3}/2$。在这个范围内 $x > 0$，所以

$$\sqrt{y^2-1} = \sqrt{\frac{1-x^2}{x^2}} = \frac{\sqrt{1-x^2}}{x},$$

$$y\left(y - \sqrt{y^2-1}\right) = \frac{1}{x}\cdot\frac{1-\sqrt{1-x^2}}{x} = \frac{1-\sqrt{1-x^2}}{x^2}.$$

想一想 $\mathrm{d}y = -\mathrm{d}x/x^2$ 中的负号到哪里去了?

这个代换把积分限的次序颠倒了，把它们调换回来时正好抵销了负号：

$$\begin{aligned}\int_{1}^{2/\sqrt{3}} \frac{1}{y\left(y-\sqrt{y^2-1}\right)}\,\mathrm{d}y &= \int_{1}^{\sqrt{3}/2} \frac{x^2}{1-\sqrt{1-x^2}}\cdot\left(-\frac{1}{x^2}\right)\mathrm{d}x \\ &= \int_{\sqrt{3}/2}^{1} \frac{1}{1-\sqrt{1-x^2}}\,\mathrm{d}x \\ &= \sqrt{3} - 1 - \frac{\pi}{6}.\end{aligned}$$

x^2 这个因子刚好完全约去，这正是分母中要多放一个因子 y 的原因。

验算

- 第二个和第三个积分的被积函数都是正的，所以答案必须为正。事实上 $\sqrt{3}-1-\frac{\pi}{6}\approx 0.73205-0.52360=0.20845$。
- 对三个积分分别使用辛普森法则（Simpson's rule），第一个得 $0.366025\ldots$（与 $\frac{\sqrt{3}}{2}-\frac{1}{2}$ 吻合），第二个和第三个得 $0.208452\ldots$（与 $\sqrt{3}-1-\frac{\pi}{6}$ 吻合）。
- 粗略估算：在第二个积分中，区间的长度为 $1-\frac{\sqrt{3}}{2}\approx 0.134$，而被积函数由 2（在 $x=\frac{\sqrt{3}}{2}$ 处）下降至 1（在 $x=1$ 处），所以积分值应介乎 0.134 与 0.268 之间。结果确实如此。

常见错误

- 没有检查 $\cos 2\theta$ 在该范围内的正负号，就写下 $\sqrt{1-\sin^2 2\theta}=\cos 2\theta$。
- 忘记更改积分限，或在 $y=1/x$ 令积分限次序颠倒时漏掉负号。
- 把 $\left[-\cot\theta\right]$ 算成 $\cot\frac{\pi}{4}-\cot\frac{\pi}{6}$，这样得出的数值大小正确，但正负号错了。

STEP 之外

最后一部分是一个普遍原则的小例子：代换是一种变量变换（change of variables），可以把一个积分准确地对应到另一个积分上，而留意到积分限互为倒数，就是 $y=1/x$ 能够奏效的可靠提示。在大学的第一门分析（analysis）课程中，代换法会由链式法则（chain rule）和微积分基本定理（fundamental theorem of calculus）严格地推导出来，并要小心留意代换在何处单调（monotonic）、被积函数在何处连续（continuous），这些正是我们在这里检查过的条件。Abbott 的 *Understanding Analysis*（一本分析学入门教科书）对这些运算背后的理论有清晰的阐述。

详细解答

STEP II 2002，第 2 题

纯数学 · 倒数多项式方程

STEP II 2002，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/02-s2-q2

关键思路

由外至内观察各项系数：先是 1 和 1，然后是 5 和 -5，最后是中间的 4。系数成对相等或成对异号，表示把方程除以 z 的中间次幂之后，只会剩下 $z^k + z^{-k}$ 和 $z^k - z^{-k}$ 的组合，而这些式子每一个都是 $w = z - z^{-1}$ 的多项式（polynomial）。这样次数（degree）便减半。那个八次方程也有相同的结构，而且化简之后，竟然再一次出现这种结构，所以这个技巧要用两次。

解答

四次方程

由于 $z = 0$ 不是根（root）（常数项是 1），我们可以除以 z^2：

$$z^2 + 5z + 4 - 5z^{-1} + z^{-2} = 0, \qquad \text{即} \qquad \left(z^2 + z^{-2}\right) + 5\left(z - z^{-1}\right) + 4 = 0.$$

设 $w = z - z^{-1}$，则 $w^2 = z^2 - 2 + z^{-2}$，所以 $z^2 + z^{-2} = w^2 + 2$。代入得

$$(w^2 + 2) + 5w + 4 = 0 \quad \Longleftrightarrow \quad w^2 + 5w + 6 = 0,$$

正如题目所求。此式可因式分解（factorise）为 $(w + 2)(w + 3) = 0$，所以 $w = -2$ 或 $w = -3$。

要回到 z，把 $z - z^{-1} = w$ 两边乘以 z：

$$z^2 - wz - 1 = 0.$$

▶ $w = -2$：$z^2 + 2z - 1 = 0$，所以 $z = -1 \pm \sqrt{2}$。

► $w=-3$：$z^2+3z-1=0$，所以 $z=\frac{1}{2}\left(-3\pm\sqrt{13}\right)$。

这四个互不相同的值就是该四次方程的根：

$$z=-1\pm\sqrt{2},\qquad z=\frac{-3\pm\sqrt{13}}{2}.$$

八次方程

同样，$z=0$ 不是根，所以我们除以 z^4，并把系数以中间项为中心对称地配对：

$$2\left(z^4+z^{-4}\right)-3\left(z^3-z^{-3}\right)-12\left(z^2+z^{-2}\right)+12\left(z-z^{-1}\right)+22=0.$$

（核对各对系数：z^8 和 z^0 的系数都是 2；z^7 的系数是 -3，z^1 的系数是 $+3$；z^6 和 z^2 的系数都是 -12；z^5 的系数是 12，z^3 的系数是 -12。）

我们需要把每个括号都以 $w=z-z^{-1}$ 表示。已知 $z^2+z^{-2}=w^2+2$。取立方，

$$w^3=z^3-3z+3z^{-1}-z^{-3}\quad\Longrightarrow\quad z^3-z^{-3}=w^3+3w,$$

再把 z^2+z^{-2} 平方，

$$z^4+z^{-4}=\left(z^2+z^{-2}\right)^2-2=(w^2+2)^2-2=w^4+4w^2+2.$$

代入得

$$\begin{aligned}0&=2(w^4+4w^2+2)-3(w^3+3w)-12(w^2+2)+12w+22\\&=2w^4-3w^3-4w^2+3w+2.\end{aligned}$$

想一想 现在剩下一个 w 的四次方程。我们有甚么得着吗？

有：看看它的系数 2, -3, -4, 3, 2。它们的规律与之前完全一样（最外面的一对相等，其次的一对异号），所以同一个技巧可以在下一层再用一次。由于 $w=0$ 不是根，除以 w^2：

$$2\left(w^2+w^{-2}\right)-3\left(w-w^{-1}\right)-4=0.$$

设 $v=w-w^{-1}$，于是 $w^2+w^{-2}=v^2+2$：

$$2(v^2+2)-3v-4=0\quad\Longleftrightarrow\quad 2v^2-3v=0\quad\Longleftrightarrow\quad v(2v-3)=0.$$

所以 $v=0$ 或 $v=\frac{3}{2}$。现在利用 $w^2-vw-1=0$ 逐层还原：

► $v=0$：$w^2=1$，所以 $w=\pm1$。

► $v=\frac{3}{2}$：$2w^2-3w-2=0$，即 $(2w+1)(w-2)=0$，所以 $w=2$ 或 $w=-\frac{1}{2}$。

最后利用 $z^2 - wz - 1 = 0$ 再还原一次：

- $w = 1$：$z^2 - z - 1 = 0$，所以 $z = \frac{1}{2}\left(1 \pm \sqrt{5}\right)$。
- $w = -1$：$z^2 + z - 1 = 0$，所以 $z = \frac{1}{2}\left(-1 \pm \sqrt{5}\right)$。
- $w = 2$：$z^2 - 2z - 1 = 0$，所以 $z = 1 \pm \sqrt{2}$。
- $w = -\frac{1}{2}$：$2z^2 + z - 2 = 0$，所以 $z = \frac{1}{4}\left(-1 \pm \sqrt{17}\right)$。

每个二次方程 $z^2 - wz - 1 = 0$ 的判别式（discriminant）都是 $w^2 + 4 > 0$，所以这些根全是实数，而且这八个值互不相同。由于八次多项式最多只有八个根，这些就是全部的根：

$$z = \frac{1 \pm \sqrt{5}}{2}, \qquad z = \frac{-1 \pm \sqrt{5}}{2}, \qquad z = 1 \pm \sqrt{2}, \qquad z = \frac{-1 \pm \sqrt{17}}{4}.$$

验算

- 由 $z^2 - wz - 1 = 0$ 得出的每一对根，乘积都是 -1，可以用来快速检查计算有没有出错：例如 $(1 + \sqrt{2})(1 - \sqrt{2}) = -1$。
- 对于八次方程，各根之和应为 $\frac{3}{2}$（即 z^7 的系数除以 z^8 的系数，再取负号）。四对根的贡献是 $1 - 1 + 2 - \frac{1}{2} = \frac{3}{2}$。正确。
- 以数值计算，$z = \frac{1}{4}(-1 + \sqrt{17}) \approx 0.7808$ 代入八次多项式所得的值与零相差不超过 10^{-9}，其余七个根以及四次方程的四个根亦然。

常见错误

- 写成 $z^3 - z^{-3} = w^3$：交叉项会多出 $3w$。
- 得到 w 的四次方程后便停下来，尝试用试根的方法把它因式分解，而没有留意到它的系数具有相同的对称性。
- 只给出 w（或 v）的值，忘记回到 z。

STEP 之外

如果一个多项式的系数倒过来读仍然一样（或者只差交错的正负号），便称为*倒数多项式*（reciprocal polynomial）或*回文多项式*（palindromic polynomial）；它们的根成对出现，每对之间以 $z \mapsto 1/z$ 或 $z \mapsto -1/z$ 相联系，而本题的代换正是这种对称性所容许的化简。在伽罗瓦理论（Galois theory）中，这相当于观察到方程在根的某一组变换所构成的群（group）之下保持不变，因此可以经由一连串较小的方程分阶段求解，本题的情况正是

如此：先是八次，然后四次，再到二次。Allenby 的 *Rings, Fields and Groups*（环、域与群）是了解这些想法如何发展的好读物。

详细解答

STEP II 2002，第 3 题

纯数学 · 费马数与质数

> **STEP II 2002，第 3 题**
>
> 继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：
>
> step.maths.org/questions/02-s2-q3

本题在 *Beyond STEP* 第 1.1 节质数工厂中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch01/fermat-numbers.html

关键思路

写出 $F_k - 2 = 2^{2^k} - 1$。把它乘以 $F_k = 2^{2^k} + 1$，便得到一个平方差（difference of two squares），而这正是归纳法（induction）所需要的那一步。乘积公式一旦成立，任何两个费马数（Fermat number）的公因数（common factor）都必须整除 2。由于每个费马数都是奇数，这个因数只可能是 1。

解答

首四个费马数

由于 $2^0 = 1$、$2^1 = 2$、$2^2 = 4$ 及 $2^3 = 8$，

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

现在就 k 的首三个值验证 $(*)$：

$$\begin{aligned} k = 1: &\quad F_0 = 3 = 5 - 2 = F_1 - 2, \\ k = 2: &\quad F_0 F_1 = 15 = 17 - 2 = F_2 - 2, \\ k = 3: &\quad F_0 F_1 F_2 = 15 \times 17 = 255 = 257 - 2 = F_3 - 2. \end{aligned}$$

证明 $(*)$ 对所有 $k \geq 1$ 成立

我们对 k 使用归纳法。$k = 1$ 的情况已在上面完成。

假设 $(*)$ 对某个 $k \geq 1$ 成立，即 $F_0F_1 \cdots F_{k-1} = F_k - 2$。那么

$$\begin{aligned} F_0F_1 \cdots F_{k-1}F_k &= (F_k - 2)\, F_k \\ &= \left(2^{2^k} - 1\right)\left(2^{2^k} + 1\right) \\ &= \left(2^{2^k}\right)^2 - 1 \\ &= 2^{2\cdot 2^k} - 1 = 2^{2^{k+1}} - 1 \\ &= F_{k+1} - 2. \end{aligned}$$

所以 $(*)$ 对 $k+1$ 同样成立，根据归纳法，它对每个 $k \geq 1$ 都成立。

> **想一想** 为甚么 $\left(2^{2^k}\right)^2$ 等于 $2^{2^{k+1}}$，而不是 $2^{2^{2k}}$ 或其他式子?

把一个幂平方，指数（exponent）便会加倍：$\left(2^a\right)^2 = 2^{2a}$。这里 $a = 2^k$，而 $2a = 2 \cdot 2^k = 2^{k+1}$。归纳步骤的大部分工夫，就在于小心处理这类层层叠起的指数。

任何两个费马数都没有公因数

取两个不同的费马数 F_m 和 F_n，其中 $m < n$。那么 $n \geq 1$，而在 $k = n$ 时的 $(*)$ 中，F_m 是因数 $F_0, \ldots, F_{n-1}$ 的其中一个。所以

$$F_n - 2 = F_m \times (\text{某个整数}),$$

这表示 F_m 整除 $F_n - 2$。

现在假设正整数 d 同时整除 F_m 和 F_n。由于 $d \mid F_m$ 及 $F_m \mid F_n - 2$，我们有 $d \mid F_n - 2$。又由于 d 也整除 F_n，它便整除两者之差：

$$d \mid F_n - (F_n - 2) = 2,$$

所以 $d = 1$ 或 $d = 2$。但 $F_m = 2^{2^m} + 1$ 是一个偶数加 1，所以是奇数（注意 $2^m \geq 1$）。因此 $d \neq 2$，从而 $d = 1$。由此可见，任何两个费马数都没有大于 1 的公因数。

质数有无限多个

每个 F_n 都至少是 3，所以它至少有一个质因数（prime factor）。选取其中一个，称它为 p_n。这样，每个 $n = 0, 1, 2, \ldots$ 都对应一个质数。

> **想一想** 两个不同的费马数会不会给出同一个质数?

不会。假如 $p_m = p_n$ 而 $m \neq n$，这个质数便会同时整除 F_m 和 F_n。它就成了一个大于 1 的公因数，而上一部分已排除了这个可能。所以质数 $p_0, p_1, p_2, \ldots$ 全部互不相同。由于费马数有无限多个，互不相同的质数也有无限多个。

留意我们从来不需要 F_n 本身是质数，只需要每个费马数都有某个质因数，是其他费马数所没有的。

验算

- $k = 4$ 时的 $(*)$：$255 \times 257 = 65535 = 65537 - 2 = F_4 - 2$。我们亦用电脑以精确运算验证了 $(*)$ 对直至 12 的所有 k 都成立。
- $F_0, \ldots, F_5$ 的最小质因数分别是 $3, 5, 17, 257, 65537, 641$，全部不同，正如论证所预测。$F_5 = 4294967297 = 641 \times 6700417$ 是第一个不是质数的费马数。
- 我们以数值计算确认了对所有 $0 \leq m < n \leq 10$，都有 $\gcd(F_m, F_n) = 1$。

常见错误

- 把 2^{2^n} 读成 $(2^2)^n = 4^n$。指数是 2^n，所以 $F_3 = 2^8 + 1$，而不是 $4^3 + 1$。
- 得出 $d \mid 2$ 之后，忘记排除 $d = 2$。正是“费马数是奇数”这个事实，才令论证得以完成。
- 假设每个费马数都是质数。最后一部分不需要这个假设，而且它根本是错的（$641 \mid F_5$）。

STEP 之外

这个证明出自波利亚（Pólya）。它属于“质数有无限多个”的一系列证明之一。这类证明都是构造一个两两互质（pairwise coprime）的无穷整数数列，于是每一项都贡献一个属于自己的新质数。欧几里得（Euclid）的证明也可以这样理解。费马数在数论课程中还会再次出现：高斯（Gauss）证明了正 n 边形可以用尺规作图（ruler and compass construction），当且仅当 n 是 2 的幂乘以若干个互不相同的费马质数（Fermat prime）之积，而至今仍未发现 F_4 之后的费马质数。想了解更多，可阅读 Davenport 的 *The Higher Arithmetic*（一本经典的数论入门书）。

详细解答

STEP II 2002，第 4 题

纯数学 · 由积分推导不等式

STEP II 2002，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/02-s2-q4

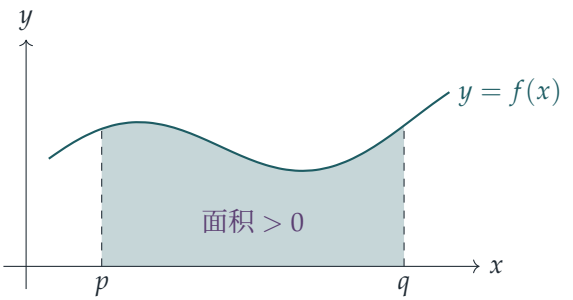
关键思路

二次式（quadratic）$as^2 - bs + c$ 若满足 $a > 0$ 和 $b^2 < 4ac$，便没有实根，而且首项系数为正，所以对每一个实数 s 都取正值。无论把 s 换成甚么（x、$\sin x$ 或 $1/x$），所得的函数都是正的，因此它在任何区间上的积分（integral）都是正的。于是每一部分要做的，只是选好代换和区间，使积分刚好给出所求不等式（inequality）中的那些数。

解答

开首的略图

积分 $\int_p^q f(x)\,\mathrm{d}x$ 是曲线 $y = f(x)$ 与 x 轴之间由 $x = p$ 至 $x = q$ 的有向面积（signed area）：在轴上方的面积算作正，在轴下方的面积算作负。若在整个 $p < x < q$ 上都有 $f(x) > 0$，整条曲线在该处便位于轴的上方，所以这个区域没有负的部分，并且有真正（非零）的面积。因此积分严格为正。（本题所有函数都是连续（continuous）的，所以面积有明确定义。）



每一部分都会用到的事实

设 $g(s) = as^2 - bs + c$，其中 $a > 0$ 及 $b^2 < 4ac$。配方（completing the square）后得

$$g(s) = a\left(s - \frac{b}{2a}\right)^2 + \frac{4ac - b^2}{4a}.$$

由于 $a > 0$，第一项至少为 0；又由于 $4ac - b^2 > 0$ 及 $a > 0$，第二项严格为正。所以对所有实数 s 都有 $g(s) > 0$。

第 (i) 部分

取 $f(x) = ax^2 - bx + c = g(x)$，它对所有 x 都是正的。我们可以在任何喜欢的区间上积分，而 $[0, 1]$ 给出的数最简单：

$$0 < \int_0^1 (ax^2 - bx + c)\,\mathrm{d}x = \frac{a}{3} - \frac{b}{2} + c.$$

乘以 6 得 $2a - 3b + 6c > 0$，即 $3b < 2a + 6c$，这正是所要证明的。

第 (ii) 部分

现在 $f(x) = a\sin^2 x - b\sin x + c = g(\sin x)$。既然 g 在每一个实数上都取正值，它在 $\sin x$ 上当然也取正值，所以对所有 x 都有 $f(x) > 0$。

> **想一想** 我们应该在哪一个区间上积分?

目标不等式中，π 乘在 a 和 c 上，却没有乘在 b 上。在 $[0, \frac{\pi}{2}]$ 上，$\int \sin x\,\mathrm{d}x = 1$ 不含 π，而 $\int \sin^2 x\,\mathrm{d}x$ 和 $\int 1\,\mathrm{d}x$ 则含有。利用 $\sin^2 x = \frac{1}{2}(1 - \cos 2x)$，

$$\int_0^{\pi/2} \sin^2 x\,\mathrm{d}x = \left[\frac{x}{2} - \frac{\sin 2x}{4}\right]_0^{\pi/2} = \frac{\pi}{4}, \qquad \int_0^{\pi/2} \sin x\,\mathrm{d}x = \Big[-\cos x\Big]_0^{\pi/2} = 1.$$

因此

$$0 < \int_0^{\pi/2} f(x)\,\mathrm{d}x = \frac{a\pi}{4} - b + \frac{c\pi}{2}.$$

乘以 4 得 $a\pi + 2c\pi - 4b > 0$，即 $4b < (a + 2c)\pi$，这正是所要证明的。

第 (iii) 部分

> **想一想** 哪一个函数的积分会给出 $\ln(q/p)$、$\frac{1}{p} - \frac{1}{q}$ 和 $q - p$?

它们分别是 $\int_p^q \frac{1}{x}\,\mathrm{d}x$、$\int_p^q \frac{1}{x^2}\,\mathrm{d}x$ 和 $\int_p^q 1\,\mathrm{d}x$。所以取

$$f(x) = \frac{a}{x^2} - \frac{b}{x} + c = g\left(\frac{1}{x}\right),$$

它在 $x > 0$ 时有定义而且连续, 并且由于 g 处处为正, 它在该处也是正的。因为 $q > p > 0$, 区间 $[p, q]$ 避开了 $x = 0$，所以

$$0 < \int_p^q \left(\frac{a}{x^2} - \frac{b}{x} + c\right) \mathrm{d}x = \left[-\frac{a}{x} - b\ln x + cx\right]_p^q$$
$$= a\left(\frac{1}{p} - \frac{1}{q}\right) - b\ln\frac{q}{p} + c(q-p).$$

移项后得 $b\ln(q/p) < a\left(\frac{1}{p} - \frac{1}{q}\right) + c(q-p)$，这正是所要证明的。（条件 $p > 0$ 也让我们可以写出 $\ln q - \ln p = \ln(q/p)$，其中两个对数都有定义。）

验算

取 $a = b = c = 1$，则 $b^2 = 1 < 4 = 4ac$。这时 (i) 说 $3 < 8$，(ii) 说 $4 < 3\pi \approx 9.42$，而 (iii) 取 $p = 1$、$q = 2$ 时说 $\ln 2 \approx 0.693 < \frac{1}{2} + 1 = 1.5$。全部成立。我们还用 20 000 组符合条件的随机三元组 (a, b, c) 测试了全部三个不等式，当中包括 b^2 仅略小于 $4ac$ 的情况，并以辛普森法则（Simpson's rule）核对了每个积分。留意若没有这些条件，不等式可以不成立：$a = 1$、$b = 10$、$c = 1$ 会令 (i) 失效，因为 $30 > 8$。

常见错误

- 忘记说明为何 $f > 0$：条件 $a > 0$ 和 $b^2 < 4ac$ 正是令二次式取正值的原因，这一点必须写出来。
- 在 (ii) 中，误取 $\int_0^{\pi/2} \sin^2 x \,\mathrm{d}x = \frac{\pi}{2}$，而正确的值是 $\frac{\pi}{4}$。
- 在 (iii) 中，没有提到 $p > 0$ 保证了 $x = 0$（f 在该处无定义）不在区间之内。

STEP 之外

“正的连续函数有正的积分”这个原理，是大学第一门分析课中为黎曼积分（Riemann integral）证明的基本性质之一。它与线性性质（linearity）结合起来，便是许多经典不等式的来源，其中包括柯西－施瓦兹不等式（Cauchy–Schwarz inequality）的积分形式：把平方 $(f + tg)^2 \geq 0$ 积分，再用上与本题相同的判别式（discriminant）论证，便可得出。Abbott 的 *Understanding Analysis*（一本分析学入门教科书）正是从这个角度仔细地建立积分理论。

详细解答

STEP II 2002，第 5 题

纯数学 · 逻辑斯谛映射的迭代

STEP II 2002，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/02-s2-q5

关键思路

一切都取决于抛物线（parabola）$y = kx(1-x)$，它在 $[0,1]$ 上的最大值是 $\frac{k}{4}$，在 $x = \frac{1}{2}$ 处取得。单凭这一点，就可以用归纳法（induction）解决第 (i) 部分。第 (ii) 和第 (iii) 部分先后要求找出不动点（fixed point）和周期（period）为二的点。第 (iii) 部分的窍门是把 x_1 用 b 表示：$x_1 = ka(1-a) = ab$。这样 $x_2 = a$ 就变成一条关于 b 的三次方程（cubic），而这条三次方程有一个明显的根（root），正好就是第 (ii) 部分的不动点。

解答

第 (i) 部分

我们对 n 使用归纳法。$n = 0$ 的情况是已知条件。假设 $0 < x_n < 1$，则 $x_n > 0$ 且 $1 - x_n > 0$，所以 $x_n(1-x_n) > 0$。利用配方法（completing the square），

$$x_n(1-x_n) = \tfrac{1}{4} - \left(x_n - \tfrac{1}{2}\right)^2 \leq \tfrac{1}{4}.$$

由于 $0 < k < 4$，各项乘以 k 后不等式（inequality）方向不变，于是

$$0 < x_{n+1} = kx_n(1-x_n) \leq \tfrac{k}{4} < 1.$$

所以 $0 < x_{n+1} < 1$，由归纳法可知对所有 n 都有 $0 < x_n < 1$。留意每个假设在哪里用到：$k > 0$ 用于下界，$k < 4$ 用于上界。

第 (ii) 部分

要得到常数数列（constant sequence），只需 $x_1 = x_0$，即 $a = ka(1-a)$。由于 $a \neq 0$，可以除以 a 得 $1 = k(1-a)$；又由于 $a \neq 1$，可以除以 $1-a$：

$$k = \frac{1}{1-a}.$$

反过来，取这个 k 时有 $x_1 = a$，因此其后每一项也都是 a。

第 (iii) 部分

设 $b = k(1-a)$，则

$$x_1 = ka(1-a) = ab.$$

若 $x_2 = a$，数列便以周期二重复，所以整个条件就只是 $x_2 = a$：

$$k\,(ab)(1-ab) = a.$$

除以 $a \neq 0$，得 $kb(1-ab) = 1$。由于 $a \neq 1$，可以写成 $k = \dfrac{b}{1-a}$，所以

$$\frac{b^2(1-ab)}{1-a} = 1 \quad \Longrightarrow \quad b^2 - ab^3 = 1-a \quad \Longrightarrow \quad ab^3 - b^2 + (1-a) = 0,$$

这就是所要求的结果。

> ? **想一想** 这条三次方程有哪一个根可以一眼看出？它代表甚么？

代入 $b = 1$ 得 $a - 1 + 1 - a = 0$，所以 $b = 1$ 是一个根。但 $b = 1$ 即是 $k(1-a) = 1$，这正是第 (ii) 部分的不动点：此时 $x_1 = ab = a$。所以额外条件 $x_1 \neq a$，即 $ab \neq a$，恰好就是 $b \neq 1$。除去因式 $b - 1$：

$$ab^3 - b^2 + (1-a) = (b-1)\big(ab^2 + (a-1)b + (a-1)\big),$$

你可以展开来验证。所以 b 必须是

$$ab^2 - (1-a)b - (1-a) = 0$$

的一个不等于 1 的根。由于 $a \neq 0$，这确实是一条二次方程（quadratic），其根为

$$b = \frac{(1-a) \pm \sqrt{(1-a)^2 + 4a(1-a)}}{2a} = \frac{(1-a) \pm \sqrt{(1-a)(1+3a)}}{2a}.$$

要有实根，需要 $(1-a)(1+3a) \geq 0$，即 $-\frac{1}{3} \leq a \leq 1$。连同 $a \neq 0, 1$，余下的范围是 $-\frac{1}{3} \leq a < 1$，$a \neq 0$。对于 a 的任何其他值，都没有实数 k 符合要求。

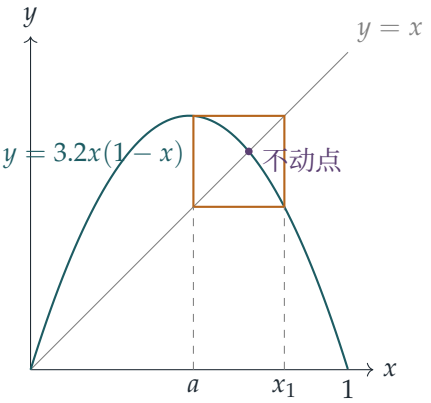
在这个范围内 $1-a>0$，所以除以 $1-a$ 便得到形式简洁的 $k=b/(1-a)$，当中用到 $\sqrt{(1-a)(1+3a)}/(1-a)=\sqrt{(1+3a)/(1-a)}$：

$$\boxed{k=\frac{1}{2a}\left(1\pm\sqrt{\frac{1+3a}{1-a}}\right),\qquad -\tfrac{1}{3}\le a<1,\ a\neq 0.}$$

最后还有一件事：我们必须确保这两个值都不是被排除的根 $b=1$。这条二次方程只有在 $a+(a-1)+(a-1)=0$，即 $a=\frac{2}{3}$ 时，才以 $b=1$ 为根。这时 $\sqrt{(1+3a)/(1-a)}=3$，两个值分别是 $k=3$（即 $b=1$，也就是不动点，故须排除）和 $k=-\frac{3}{2}$。所以：

- 当 $a=\frac{2}{3}$ 时，唯一的可能是 $k=-\frac{3}{2}$；
- 当 $a=-\frac{1}{3}$ 时，正负两个符号给出相同的值，$k=-\frac{3}{2}$；
- 对于其余所有容许的 a，两个符号都给出真正的 2-循环（cycle）。

下图显示了其中一个这样的循环，取 $k=3.2$。轨道（orbit）在 $a\approx 0.513$ 与 $x_1\approx 0.800$ 之间来回跳动，在抛物线与直线 $y=x$ 之间描出一个正方形。



验算

- 取 $a=0.7$。公式给出 $k\approx 3.0104$ 或 $k\approx -1.5818$。取 $k\approx 3.0104$ 时，得 $x_1\approx 0.6322$，然后 $x_2\approx 0.7000$，所以轨道的周期确实是二。
- 当 $a=\frac{2}{3}$ 时，$k=3$ 给出 $x_1=3\cdot\frac{2}{3}\cdot\frac{1}{3}=\frac{2}{3}$，即不动点，与预期相符。取 $k=-\frac{3}{2}$ 时，得 $x_1=-\frac{1}{3}$ 和 $x_2=-\frac{3}{2}\cdot(-\frac{1}{3})\cdot\frac{4}{3}=\frac{2}{3}$，是真正的 2-循环。
- 数值上，对数以千计随机选取的 $a\in[-\frac{1}{3},1)$，不论取哪个符号，$x_2=a$ 都成立。对于该区间以外的 a，在 k 的范围内逐一搜索，除了不动点 $k=1/(1-a)$ 之外找不到任何解。

常见错误

- 忘记三次方程必定包含不动点那个解。看出 $b = 1$ 这个根，正是令三次方程变得容易处理的关键。
- 遗漏实根条件 $-\frac{1}{3} \leq a < 1$，或没有留意到当 $a = \frac{2}{3}$ 时，k 的两个值之中有一个是不动点，必须舍去。
- 在第 (i) 部分只证明了 $x_{n+1} \leq \frac{1}{4}k$，却没有指出因为 $k < 4$，这个值严格小于 1。

STEP 之外

这就是逻辑斯谛映射 (logistic map)，是“规则简单而行为极其丰富”的经典例子。当 k 在通常的范围 $0 < k \leq 4$ 内，你找到的 2-循环恰好在 $k > 3$ 时存在：在 $k = 3$ 处，不动点失去稳定性，并分裂成两个点。这是第一次倍周期分岔（*period-doubling bifurcation*）。其后还会继续倍增，出现倍增的 k 值以几何方式愈聚愈密（其比率由费根鲍姆（Feigenbaum）常数决定），直至映射进入混沌（chaotic）状态。斯托加茨（Strogatz）的 *Nonlinear Dynamics and Chaos*（《非线性动力学与混沌》）以大学一年级的程度讲述了整个故事。

详细解答

STEP II 2002，第 6 题

纯数学 · 斜面上的直线

> **STEP II 2002，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/02-s2-q6

关键思路

整道题都建基于斜面（inclined plane）的一个事实：若斜面上的一条直线与最大斜率线（line of greatest slope）成角 ψ，则它与水平面所成的角 γ 满足

$$\sin\gamma = \cos\psi\,\sin\beta,$$

其中 β 是最大斜率线的倾角。有了这个结果，题目中每条直线都给出一条三角方程，而 l_1 与 l_3 互相垂直，则把一个余弦变成正弦。之后就只是代数运算了。

解答

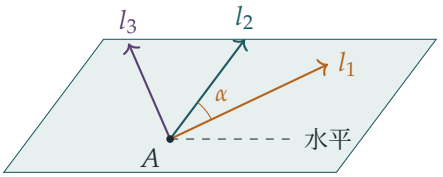
斜面上一条直线的斜度

以 A 为原点建立坐标轴，z 轴铅垂向上。设 $\mathbf{h}$ 为位于 P 内的水平单位向量（unit vector），并设 $\mathbf{u}$ 为沿最大斜率线 l_2 向上的单位向量，它在 P 内与 $\mathbf{h}$ 垂直。由于 l_2 与水平面成角 β，我们可取

$$\mathbf{h} = (1,0,0), \qquad \mathbf{u} = (0,\cos\beta,\sin\beta).$$

在 P 内与 $\mathbf{u}$ 成角 ψ 的单位向量是 $\mathbf{d} = \cos\psi\,\mathbf{u} + \sin\psi\,\mathbf{h}$，因为 $\mathbf{u}$ 与 $\mathbf{h}$ 标准正交（orthonormal），并且张成（span）P。它的铅垂分量是 $\cos\psi\sin\beta$。若一条直线的单位方向向量为 $\mathbf{d}$，则它与水平面所成的角 γ 满足 $\sin\gamma =$ |铅垂分量，取自$\mathbf{d}$|，所以

$$\sin\gamma = |\cos\psi|\,\sin\beta. \tag{$*$}$$



第一个位置

在 P 内，直线 l_1 与 l_2 成锐角（acute angle）α，而它与水平面所成的角是 $\pi/6$。由 $(*)$，

$$\cos\alpha\sin\beta = \sin\tfrac{\pi}{6} = \tfrac{1}{2},$$

正是所要证明的结果（这里 $\cos\alpha > 0$，因为 α 是锐角）。

想一想 l_3 与 l_2 所成的角是多少?

三条直线都位于同一平面 P 内，并且都通过 A。由于 $l_3 \perp l_1$，l_3 的方向就是 l_1 的方向在 P 内转过一个直角所得的方向，所以 l_3 所成的角是 $\alpha \pm \frac{\pi}{2}$，这是相对于 l_2 而言的（正负号取决于转动的方向，但并不影响结果）。无论哪一种情况，都有 $|\cos(\alpha \pm \frac{\pi}{2})| = \sin\alpha$，因为 α 是锐角。于是把 $(*)$ 用于 l_3，得

$$\sin\alpha\sin\beta = \sin\tfrac{\pi}{4} = \tfrac{1}{\sqrt{2}}.$$

把两个结果平方后相加，

$$\sin^2\beta\,(\cos^2\alpha + \sin^2\alpha) = \tfrac{1}{4} + \tfrac{1}{2}, \qquad \text{故} \qquad \sin^2\beta = \tfrac{3}{4}.$$

直线与水平面所成的角 β 位于 $[0, \frac{\pi}{2}]$ 之内，所以 $\sin\beta = \frac{\sqrt{3}}{2}$，即 $\beta = \pi/3$，正是所要证明的结果。（顺带一提，把两条方程相除可得 $\tan\alpha = \sqrt{2}$，这确实对应一个锐角，所以这个位置是存在的。）

转动后的位置

平面没有改变，因此 l_2 和 $\beta = \pi/3$ 也没有改变。现在 l_1 与 l_2 成锐角 θ，而 l_3 所成的角是 $\theta \pm \frac{\pi}{2}$，与之前完全一样。把 $(*)$ 分别用于两条直线，并代入 $\sin\beta = \frac{\sqrt{3}}{2}$：

$$\sin\phi = \tfrac{\sqrt{3}}{2}\cos\theta,$$
$$\sin 2\phi = \tfrac{\sqrt{3}}{2}\sin\theta.$$

由于 2ϕ 是与水平面所成的角，故 $0 \le 2\phi \le \frac{\pi}{2}$，而且 $\phi \neq 0$，因为 $\cos\theta > 0$。所以 $\sin\phi > 0$，我们可以把第二条方程除以第一条：

$$\tan\theta = \frac{\sin 2\phi}{\sin\phi} = \frac{2\sin\phi\cos\phi}{\sin\phi} = 2\cos\phi.$$

设 $t = \tan^2\theta$，则 $\cos^2\phi = t/4$ 及 $\sin^2\phi = 1 - t/4$。为了消去 θ，我们把 $\cos^2\theta = 1/(1+\tan^2\theta)$ 代入第一条方程的平方：

$$1 - \frac{t}{4} = \frac{3}{4}\cdot\frac{1}{1+t} \quad\Longrightarrow\quad (4-t)(1+t) = 3 \quad\Longrightarrow\quad t^2 - 3t - 1 = 0.$$

因此 $t = \frac{1}{2}\left(3 \pm \sqrt{13}\right)$。由于 $t = \tan^2\theta \geq 0$ 而 $\sqrt{13} > 3$，我们必须取正号：

$$\tan^2\theta = \frac{3+\sqrt{13}}{2},$$

正是所要证明的结果。

这个根确实可以达到：它给出 $\cos^2\phi = t/4 \approx 0.826$，所以 $\phi \approx 24.7^\circ$，$2\phi \approx 49.4^\circ < 90^\circ$，而 $\theta \approx 61.2^\circ$ 是锐角。

验算

- 在第一个位置，$\tan\alpha = \sqrt{2}$ 给出 $\cos\alpha = 1/\sqrt{3}$，$\sin\alpha = \sqrt{2/3}$，再配合 $\sin\beta = \sqrt{3}/2$，两个乘积分别是 $\frac{1}{2}$ 和 $1/\sqrt{2}$，与应有的结果相符。
- 以数值计算，$\tan^2\theta \approx 3.3028$，所以 $\tan\theta \approx 1.8174$，$\theta \approx 61.18^\circ$；于是 $\sin\phi = \frac{\sqrt{3}}{2}\cos\theta \approx 0.4175$ 给出 $\phi \approx 24.68^\circ$，而 $\frac{\sqrt{3}}{2}\sin\theta \approx 0.7589 \approx \sin(49.35^\circ) = \sin 2\phi$。
- $(*)$ 的极限情况：$\psi = 0$ 给出 l_2 本身（所成的角为 β），而 $\psi = \frac{\pi}{2}$ 给出一条水平线（所成的角为 0）。

常见错误

- 误用 $\tan\gamma = \cos\psi\tan\beta$ 而不是 $\sin\gamma = \cos\psi\sin\beta$。只有当 ψ 是在水平的平面图上量度时，正切的版本才正确，但这里 α 和 θ 都是在 P 内量度的角。
- 忘记舍去负根 $\frac{1}{2}(3-\sqrt{13})$：平方数不可能是负数。

STEP 之外

公式 $(*)$ 其实就是一个点积（dot product）：直线与水平面所成角的正弦，就是它的单位向量在铅垂方向上的分量，而把这个向量用平面的一组标准正交基（orthonormal basis）表示出来，便完成了全部工作。把平面看成标准正交向量所张成的空间，把角看成点积，正是大学一年级线性代数（linear algebra）以及旋转几何的起点。Strang 的 *Introduction to Linear Algebra*（线性代数入门）对这些想法有细致的阐述。

详细解答

STEP II 2002，第 7 题

纯数学 · 直线的向量几何

STEP II 2002，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/02-s2-q7

关键思路

这里的一切都归结为点积（dot product）。一条通过原点、单位方向向量为 $\mathbf{u}$ 的直线，与 m_1 的夹角为 $\pi/4$，当且仅当 $|\mathbf{u}\cdot\mathbf{d}_1| = |\mathbf{d}_1|\cos\frac{\pi}{4}$，于是第一部分就化为联立方程。至于后半部分，“垂直”即点积等于零；而空间中两条直线只有在共面（coplanar）且不平行时才会相交，共面则等价于纯量三重积（scalar triple product）等于零。出人意表的是，共面的那个情况原来正是平行的情况。

解答

直线 m_3 和 m_4

设 $\mathbf{d}_1 = \mathbf{i}+\mathbf{j}$ 及 $\mathbf{d}_2 = \mathbf{i}+\mathbf{k}$，两者的长度均为 $\sqrt{2}$。设 $\mathbf{u} = a\mathbf{i}+b\mathbf{j}+c\mathbf{k}$ 为沿着一条通过 O 的直线的单位向量（unit vector）。两条直线之间的夹角是指它们之间的锐角，所以条件是

$$|\mathbf{u}\cdot\mathbf{d}_1| = \sqrt{2}\cos\tfrac{\pi}{4} = 1, \qquad |\mathbf{u}\cdot\mathbf{d}_2| = 1,$$

即 $a+b = \pm1$ 及 $a+c = \pm1$。以 $-\mathbf{u}$ 取代 $\mathbf{u}$ 会得出同一条直线，并把两个正负号同时反转，所以只需考虑两种情况。

同号：$a+b=1$，$a+c=1$。这时 $b=c=1-a$，而由 $a^2+b^2+c^2=1$ 得

$$a^2+2(1-a)^2 = 1 \iff 3a^2-4a+1=0 \iff (3a-1)(a-1)=0.$$

所以 $a=1$，得 $\mathbf{u}=\mathbf{i}$；或 $a=\frac{1}{3}$，得 $\mathbf{u}=\frac{1}{3}(\mathbf{i}+2\mathbf{j}+2\mathbf{k})$。

异号：$a+b=1, a+c=-1$。这时 $b=1-a, c=-1-a$，而由 $a^2+(1-a)^2+(1+a)^2=1$ 得 $3a^2+2=1$，此方程没有实数解。

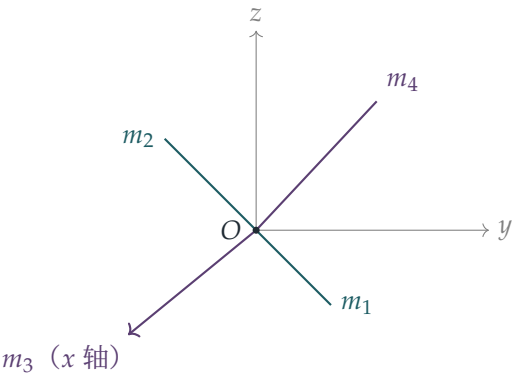
所以这两条直线的方向为

$$m_3:\ \mathbf{i}, \qquad m_4:\ \mathbf{i}+2\mathbf{j}+2\mathbf{k}$$

（两个标记对调亦无不可）。取单位向量 $\mathbf{e}_3=\mathbf{i}$ 及 $\mathbf{e}_4=\frac{1}{3}(\mathbf{i}+2\mathbf{j}+2\mathbf{k})$，则

$$\cos\theta=\mathbf{e}_3\cdot\mathbf{e}_4=\tfrac{1}{3},$$

此值为正数，所以这就是锐角：其余弦为 $\frac{1}{3}$。



设置各点

由于 A 位于 m_1 上，且与 O 的距离为 $\lambda\sqrt{2}$，它就是 $\pm\lambda(1,1,0)$；坐标须为非负数，故我们必须取 $A=\lambda(1,1,0)$，其中 $\lambda\geq 0$（由于 $\lambda\sqrt{2}$ 是距离，我们可以全程取 $\lambda\geq 0$）。同样地，$B=\lambda(1,0,1)$。在 m_3, m_4 上坐标为非负数的单位点是 $(1,0,0)$ 和 $\frac{1}{3}(1,2,2)$。

当中哪一点是 P、哪一点是 Q，取决于 m_3 和 m_4 的标记方式。但这并不要紧：反射 $(x,y,z)\mapsto(x,z,y)$ 把 A 和 B 对调，并保持 $(1,0,0)$ 和 $\frac{1}{3}(1,2,2)$ 两点不变，所以它把这个构形的一种标记方式变为另一种（直线 AQ 和 BP 的角色互换）。反射保持垂直关系和相交关系，所以只需取

$$A=(\lambda,\lambda,0),\quad B=(\lambda,0,\lambda),\quad P=(1,0,0),\quad Q=\left(\tfrac{1}{3},\tfrac{2}{3},\tfrac{2}{3}\right).$$

于是

$$\overrightarrow{AQ}=\left(\tfrac{1}{3}-\lambda,\ \tfrac{2}{3}-\lambda,\ \tfrac{2}{3}\right),\qquad \overrightarrow{BP}=(1-\lambda,\ 0,\ -\lambda).$$

第 (i) 部分

AQ 垂直于 BP，当且仅当 $\overrightarrow{AQ}\cdot\overrightarrow{BP}=0$：

$$\begin{aligned}\overrightarrow{AQ}\cdot\overrightarrow{BP}&=\left(\tfrac{1}{3}-\lambda\right)(1-\lambda)+0-\tfrac{2}{3}\lambda\\&=\lambda^2-\tfrac{4}{3}\lambda+\tfrac{1}{3}-\tfrac{2}{3}\lambda=\lambda^2-2\lambda+\tfrac{1}{3}.\end{aligned}$$

当 $(\lambda-1)^2=\frac{2}{3}$ 时，上式等于零，即

$$\lambda=1\pm\sqrt{\tfrac{2}{3}}=1\pm\tfrac{\sqrt{6}}{3}.$$

两个值都是正数（因为 $\sqrt{6}<3$），所以两者都符合非负条件，而这个二次方程再没有其他根。因此恰好有两个 λ 的值使得 $AQ\perp BP$，正如题目所求。

第 (ii) 部分

> **想一想** 怎样判断三维空间中的两条直线是否相交?

两条相异的直线相交，当且仅当它们位于同一平面上而且不平行。直线 AQ 和 BP 位于同一平面上，当且仅当 A、B、P、Q 四点共面，而这又当且仅当 $\overrightarrow{AB}$、$\overrightarrow{AP}$、$\overrightarrow{AQ}$ 的纯量三重积等于零。

这里 $\overrightarrow{AB}=(0,-\lambda,\lambda)$ 及 $\overrightarrow{AP}=(1-\lambda,-\lambda,0)$，所以

$$\begin{vmatrix} 0 & -\lambda & \lambda \\ 1-\lambda & -\lambda & 0 \\ \frac{1}{3}-\lambda & \frac{2}{3}-\lambda & \frac{2}{3} \end{vmatrix} = \lambda\cdot\tfrac{2}{3}(1-\lambda)+\lambda\Big[(1-\lambda)\big(\tfrac{2}{3}-\lambda\big)+\lambda\big(\tfrac{1}{3}-\lambda\big)\Big]$$

$$=\lambda\Big[\tfrac{2}{3}-\tfrac{2}{3}\lambda+\tfrac{2}{3}-\tfrac{4}{3}\lambda\Big]=2\lambda\big(\tfrac{2}{3}-\lambda\big),$$

这是沿第一行展开的结果。因此，对于 $\lambda\neq0$，两条直线只有在 $\lambda=\frac{2}{3}$ 时才有可能相交。

可是当 $\lambda=\frac{2}{3}$ 时，

$$\overrightarrow{AQ}=\big(-\tfrac{1}{3},\,0,\,\tfrac{2}{3}\big),\qquad \overrightarrow{BP}=\big(\tfrac{1}{3},\,0,\,-\tfrac{2}{3}\big)=-\overrightarrow{AQ},$$

所以两条直线平行。它们并不是同一条直线，因为 $\overrightarrow{BA}=(0,\frac{2}{3},-\frac{2}{3})$ 不是 $\overrightarrow{BP}$ 的倍数，所以 A 不在 BP 上。两条相异的平行线永不相交。

所以对于每一个非零的 λ，直线 AQ 和 BP 都不相交，正如题目所求。（当 $\lambda=0$ 时，$A=B=O$，两条直线确实相交于 O，这正是题目把它排除在外的原因。）

验算

- 一望而知，$\mathbf{i}$ 与 $\mathbf{i}+\mathbf{j}$ 的夹角为 $\pi/4$，这正好印证了二次方程的第一个根，令人安心。
- 当 $\lambda=1+\sqrt{2/3}\approx1.8165$ 时，点积 $\overrightarrow{AQ}\cdot\overrightarrow{BP}$ 的数值计算结果为 0（准确至 10^{-16}），当 $\lambda\approx0.1835$ 时亦然。
- 对 $\lambda=0.001,0.002,\ldots,5$ 计算直线 AQ 与 BP 之间的最短距离（并对 P 和 Q 的另一种标记方式作同样计算），结果从不为零；只有当 $\lambda\to0$ 时，距离才趋向零。

常见错误

- 遗漏 $a+b=1$，$a+c=-1$ 这个情况：它虽然得不出任何解，但仍须加以排除。
- 写到“三重积在 $\lambda=\frac{2}{3}$ 时等于零，所以两线相交”便停笔。共面的直线可以是平行的，而这里正是如此。

STEP 之外

纯量三重积是一个 3×3 行列式（determinant），它等于零即表示三个向量线性相关（linearly dependent）；两条歪斜线（skew lines）之间的最短距离 $|\overrightarrow{AB}\cdot(\mathbf{u}\times\mathbf{v})|/|\mathbf{u}\times\mathbf{v}|$，也是由同样的组件构成的。在大学一年级的线性代数（linear algebra）课程中，这些概念会通过行列式、秩（rank）和子空间（subspace）的几何统一起来。Strang 的 *Introduction to Linear Algebra*（线性代数入门）以大量三维图像阐述这些内容。

详细解答

STEP II 2002，第 8 题

纯数学 · 一阶微分方程

STEP II 2002，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/02-s2-q8

关键思路

每一条方程都是最简单的一阶线性方程（first-order linear equation）$y' = p(x)\,y$，所以在任何一个 p 由单一公式给出的区间上，解都是一个常数乘以 $\exp\big(\int p\,\mathrm{d}x\big)$。难处在于问题在 $x = 0$ 处被分成两段：两边各有一个任意常数（arbitrary constant），其中一边的常数由 y 的已知值确定，另一边的常数则只能靠要求两段在 $x = 0$ 处相接来确定。第一部分正是这个想法的预演，而第二部分则暗藏了一个绝对值（absolute value）。

解答

第一对方程

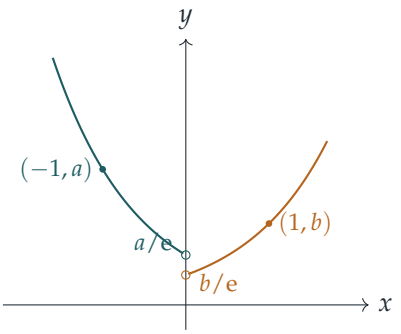
当 $x < 0$ 时，由 $\dfrac{\mathrm{d}y}{\mathrm{d}x} = -y$ 得 $y = A\mathrm{e}^{-x}$，其中 A 为某个常数。（要看出没有其他解，可留意 $\frac{\mathrm{d}}{\mathrm{d}x}\big(y\mathrm{e}^{x}\big) = \mathrm{e}^{x}(y' + y) = 0$，所以 $y\mathrm{e}^{x}$ 在区间 $x < 0$ 上是常数。）由条件“在 $x = -1$ 处 $y = a$”得 $A\mathrm{e} = a$，所以

$$y = a\,\mathrm{e}^{-x-1} \qquad (x < 0).$$

同样地，当 $x > 0$ 时有 $y = B\mathrm{e}^{x}$，而在 $x = 1$ 处 $y = b$ 给出 $B = b/\mathrm{e}$，所以

$$y = b\,\mathrm{e}^{x-1} \qquad (x > 0).$$

当 $x \to 0^{-}$ 时，$y \to a/\mathrm{e}$；当 $x \to 0^{+}$ 时，$y \to b/\mathrm{e}$。一般而言这两个值并不相同，曲线在 y 轴处会出现跳跃。下图所示为 $a, b > 0$ 且 $a > b$ 的典型情况：左边是一条衰减的指数曲线，愈接近该轴愈低；右边是一条增长的指数曲线；空心圆圈标示两个（不同的）极限（limit）值。



曲线没有跳跃，当且仅当两个极限相等，即 $a/\mathrm{e} = b/\mathrm{e}$，也就是

$$a = b.$$

在这情况下，曲线为 $y = a\,\mathrm{e}^{|x|-1}$，是一条对称、形状像“V”的曲线，在 $x = 0$ 处有一个尖角。

含 $|\mathrm{e}^x - 1|$ 的方程

由于 $\mathrm{e}^x > 1$ 当且仅当 $x > 0$，

$$|\mathrm{e}^x - 1| = \begin{cases} \mathrm{e}^x - 1, & x \geq 0, \\ 1 - \mathrm{e}^x, & x < 0. \end{cases}$$

所以这条方程和先前一样，分成两条普通的线性方程。

当 $x > 0$ 时： $y' = (\mathrm{e}^x - 1)y$。$\mathrm{e}^x - 1$ 的一个原函数（antiderivative）是 $\mathrm{e}^x - x$，所以 $\frac{\mathrm{d}}{\mathrm{d}x}\left(y\,\mathrm{e}^{-(\mathrm{e}^x - x)}\right) = 0$，于是

$$y = C\exp(\mathrm{e}^x - x).$$

在 $x = 1$ 处，$y = \mathrm{e}^{\mathrm{e}}$，所以 $C\,\mathrm{e}^{\mathrm{e}-1} = \mathrm{e}^{\mathrm{e}}$，得 $C = \mathrm{e}$。因此

$$y = \exp(\mathrm{e}^x - x + 1) \qquad (x \geq 0).$$

当 $x < 0$ 时： $y' = (1 - \mathrm{e}^x)y$。$1 - \mathrm{e}^x$ 的一个原函数是 $x - \mathrm{e}^x$，所以

$$y = D\exp(x - \mathrm{e}^x).$$

想一想 在 $x < 0$ 的范围内没有已知的数据点。那么 D 由甚么确定?

答案是在 $x = 0$ 处的连续性（continuity）。从右边趋近时，$y \to \exp(1 - 0 + 1) = \mathrm{e}^2$。从左边趋近时，$y \to D\exp(0 - 1) = D/\mathrm{e}$。令两者相等，得 $D = \mathrm{e}^3$，所以

$$y = \exp(x - \mathrm{e}^x + 3) \qquad (x < 0).$$

把两段合起来，

$$y=\begin{cases}\exp(\mathrm{e}^x-x+1), & x\geq 0,\\ \exp(x-\mathrm{e}^x+3), & x<0.\end{cases}$$

留意这一次在 $x=0$ 处连尖角也没有：两条公式都给出 $y'(0)=0$，与 $|\mathrm{e}^0-1|\,y=0$ 吻合。由于 $y>0$ 且 $|\mathrm{e}^x-1|\geq 0$，这个解处处递增，并在 $x=0$ 处有一个驻点拐点（stationary point of inflection），该处 $y=\mathrm{e}^2$。

极限

(i) 当 $x>0$ 时，

$$y\exp(-\mathrm{e}^x)=\exp(\mathrm{e}^x-x+1-\mathrm{e}^x)=\mathrm{e}^{1-x}\to 0\quad \text{当}x\to+\infty.$$

所以极限是 0：因子 $\exp(-\mathrm{e}^x)$ 恰好把双重指数增长完全抵消，只剩下衰减的 e^{1-x}。

(ii) 当 $x<0$ 时，

$$y\,\mathrm{e}^{-x}=\exp(x-\mathrm{e}^x+3-x)=\exp(3-\mathrm{e}^x)\to \mathrm{e}^3\quad \text{当}x\to-\infty,$$

因为 $\mathrm{e}^x\to 0$。所以极限是 e^3。换句话说，当 x 是绝对值很大的负数时，这个解的表现就像 $\mathrm{e}^3\mathrm{e}^x$，正如所料：在那里 $|\mathrm{e}^x-1|\approx 1$，方程近乎 $y'=y$。

验算

- 对答案求导：当 $x>0$ 时，$y'=(\mathrm{e}^x-1)y$；当 $x<0$ 时，$y'=(1-\mathrm{e}^x)y$。在 $x=1$ 处，$\exp(\mathrm{e}-1+1)=\mathrm{e}^{\mathrm{e}}$。
- 数值上，用龙格—库塔法（Runge–Kutta method）由 $y(1)=\mathrm{e}^{\mathrm{e}}\approx 15.154$ 逐步倒推至 $x=0$，得 $y(0)\approx 7.38906=\mathrm{e}^2$；继续推算至 $x=-5$，结果与 $\exp(x-\mathrm{e}^x+3)$ 吻合至六位有效数字（significant figures）。
- 就极限 (ii) 而言，在 $x=-10$ 处得 $y\mathrm{e}^{-x}\approx 20.0846$，接近 $\mathrm{e}^3\approx 20.0855$；就 (i) 而言，在 $x=10$ 处，$\mathrm{e}^{1-x}\approx 1.2\times 10^{-4}$。

常见错误

- 用 $x=1$ 处的条件去确定 $x=0$ 两边的常数。两个区间是各自独立的问题，左边的常数只能由连续性得出。
- 丢掉模（modulus）符号，对所有 x 求解 $y'=(\mathrm{e}^x-1)y$；这样在 $x<0$ 时会得出错误的曲线，(ii) 的极限也会出错。
- 写出 $\int(\mathrm{e}^x-1)\,\mathrm{d}x=\mathrm{e}^x-x$ 之后，却写成 $\exp(\mathrm{e}^x-x)=\mathrm{e}^{\mathrm{e}^x}/x$（正确的是 $\mathrm{e}^{\mathrm{e}^x}\mathrm{e}^{-x}$）。

STEP 之外

方程 $y' = |e^x - 1|\, y$ 的系数在 $x = 0$ 处连续但不可微（differentiable）。关于 $y' = f(x, y)$ 的存在唯一性定理（existence and uniqueness theorem，即 Picard–Lindelöf 定理）只要求 f 对 x 连续、对 y 满足利普希茨（Lipschitz）条件，所以在这里单凭一个条件就能确定整条实数线上的解，而在 0 处的连续性是自动成立的。在第一部分，方程右边本身在 $x = 0$ 处出现跳跃，所以才需要额外的条件把两段接起来。Boyce 与 DiPrima 的 *Elementary Differential Equations*（微分方程入门教科书）讨论了一阶方程的这些存在性与唯一性问题。

详细解答

STEP II 2002，第 9 题

力学 · 抛体运动

STEP II 2002，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：

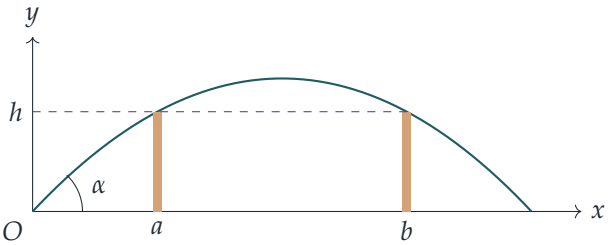
step.maths.org/questions/02-s2-q9

关键思路

先写出轨迹（path）的方程，把 y 表示成 x 的函数。“刚好越过两道墙”表示轨迹通过两道墙的墙顶，所以方程 $y = h$ 的根恰好是 $x = a$ 和 $x = b$。这是一条已知两根的二次方程（quadratic equation），把它与根之和及根之积（sum and product of roots）比较，便可一次过得出所有结果。至于第二部分，只需把两条公式对 h 求导（differentiate）。

解答

我们把质点（particle）看成只受重力作用而运动的一点（不计空气阻力（air resistance）），并取 O 为原点，x 轴水平，y 轴铅垂向上。按图中所示的情况，我们假设 $0 < a < b$、$h > 0$ 及 $0 < \alpha < \frac{\pi}{2}$。



求 $\tan\alpha$ 和 V

初速度的分量为 $V\cos\alpha$ 和 $V\sin\alpha$，经过时间 t 后

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

利用 $t = x/(V\cos\alpha)$ 消去时间，得轨迹

$$y = x\tan\alpha - \frac{gx^2}{2V^2\cos^2\alpha} = x\tan\alpha - kx^2, \qquad \text{其中} k = \frac{g\sec^2\alpha}{2V^2} = \frac{g(1+\tan^2\alpha)}{2V^2} > 0.$$

质点刚好越过每道墙，所以轨迹通过墙顶 (a,h) 和 (b,h)。因此 a 和 b 是以下方程的根：

$$x\tan\alpha - kx^2 = h, \qquad \text{即} \qquad kx^2 - x\tan\alpha + h = 0.$$

这是 x 的二次方程（因为 $k \neq 0$），所以它最多有两个根，故此它的根恰好就是 a 和 b。与根之和及根之积比较，

$$a + b = \frac{\tan\alpha}{k}, \qquad ab = \frac{h}{k}.$$

由第二式得 $k = h/(ab)$，再由第一式得

$$\boxed{\tan\alpha = \frac{h(a+b)}{ab}.}$$

最后，由 k 的定义，

$$\begin{aligned}\frac{2V^2}{g} = \frac{1+\tan^2\alpha}{k} &= \frac{ab}{h}\left(1 + \frac{h^2(a+b)^2}{a^2b^2}\right)\\ &= \frac{ab}{h} + \frac{(a+b)^2 h}{ab},\end{aligned}$$

这正是所求的结果。

（值得留意的是，代入这些值后，轨迹就是 $y = \dfrac{h}{ab}x(a+b-x)$，一眼可见它在 $x = a$ 和 $x = b$ 时等于 h。）

把墙加高

第二个质点同样刚好越过两道墙，墙与原点的距离仍是 a 和 b，但高度变为 $h + \delta h$。所以把 h, α, V 换成 $h + \delta h, \alpha + \delta\alpha, V + \delta V$ 后，上面的公式依然成立：

$$\tan\alpha = \frac{a+b}{ab}h, \qquad \frac{2V^2}{g} = \frac{ab}{h} + \frac{(a+b)^2}{ab}h.$$

两者都只是 h 的函数（a, b 固定），所以对于微小的变化，我们可以用 $\delta(\text{量}) \approx \dfrac{\mathrm{d}(\text{量})}{\mathrm{d}h}\delta h$。

角度。 把 $\tan\alpha = \frac{a+b}{ab}h$ 对 h 求导，得 $\sec^2\alpha\dfrac{\mathrm{d}\alpha}{\mathrm{d}h} = \dfrac{a+b}{ab}$，所以

$$\sec^2\alpha\,\delta\alpha \approx \frac{a+b}{ab}\delta h.$$

由于 $\sec^2\alpha > 0$、$a, b > 0$ 及 $\delta h > 0$，右边是正数，所以 $\delta\alpha > 0$：质点必须以更陡的角度抛出。

想一想 这个“≈”只是一阶（first-order）近似。我们真的可以凭它断定 $\delta\alpha$ 的正负号吗?

可以，而且其实完全不需要取近似。$\tan\alpha$ 的公式对 h 是线性的，所以以下等式准确成立：$\tan(\alpha+\delta\alpha)-\tan\alpha=\frac{a+b}{ab}\,\delta h>0$。由于 tan 在 $(0,\frac{\pi}{2})$ 上严格递增（strictly increasing），这就迫使 $\alpha+\delta\alpha>\alpha$，所以 $\delta\alpha>0$。

速率。 把 $\dfrac{2V^2}{g}$ 对 h 求导，

$$\frac{4V}{g}\frac{\mathrm{d}V}{\mathrm{d}h}=-\frac{ab}{h^2}+\frac{(a+b)^2}{ab}=\frac{(a+b)^2}{abh^2}\left(h^2-\frac{a^2b^2}{(a+b)^2}\right),$$

所以

$$\delta V\approx\frac{g(a+b)^2}{4Vabh^2}\left(h-\frac{ab}{a+b}\right)\left(h+\frac{ab}{a+b}\right)\delta h.$$

除了 $h-\frac{ab}{a+b}$ 之外，每个因子都是正数（因为 $V,a,b,h,\delta h>0$）。因此 δV 的正负号与 $h-\frac{ab}{a+b}$ 相同：若 $h>ab/(a+b)$ 则为正，若 $h<ab/(a+b)$ 则为负，这正是所求的结果。（由于这个导数在 $h=ab/(a+b)$ 以外都不等于零，只要 δh 足够小，一阶项确实决定了正负号。）

诠释。 与 $\tan\alpha=h(a+b)/(ab)$ 比较，条件 $h>ab/(a+b)$ 正好就是 $\tan\alpha>1$，即 $\alpha>45°$。作为 h 的函数，$2V^2/g$ 具有 $p/h+qh$ 的形式，它在 $h=\sqrt{p/q}=ab/(a+b)$ 时最小，此时 $\alpha=45°$ 而 $2V^2/g=2(a+b)$。所以：如果抛射角本来已经比 45° 陡，把墙加高便需要更大的速率；如果抛射角较平，把墙加高（同时令抛射角变陡，趋近 45°）所需的速率反而更小。

验算

- 取 $a=3\,\mathrm{m}$、$b=7\,\mathrm{m}$、$h=2.5\,\mathrm{m}$。那么 $\tan\alpha=25/21$，而 $2V^2/g=8.4+11.90\ldots=20.30\ldots\,\mathrm{m}$。用这些数值以数值方法逐步计算质点的运动，质点在 $x=3$ 和 $x=7$ 两处的高度都是 $2.5\,\mathrm{m}$（准确至 10^{-11}）。
- 对称性：答案对 a 和 b 是对称的，这是必然的，因为轨迹是一条以 $x=\frac{1}{2}(a+b)$ 为对称轴的抛物线（parabola）。
- 单位：ab/h 和 $(a+b)^2h/(ab)$ 都是长度，与 $2V^2/g$ 相符；而 $\tan\alpha$ 是无量纲（dimensionless）的。
- 当 $h=ab/(a+b)$ 时，我们得到 $\alpha=45°$ 及 $2V^2/g=2(a+b)$；对于 $a=3$、$b=7$，这个值是 $20\,\mathrm{m}$，比上面的 $20.30\,\mathrm{m}$ 略小，与它是最小值相符。

常见错误

- 为两道墙分别写出两条方程，然后硬解联立方程：这样做是可行的，但利用根之和与根之积，两行便可完成。
- 把 k 换回 V 时，忘记了 $\sec^2\alpha = 1 + \tan^2\alpha$。
- 没有把 $2V^2/g$ 的导数因式分解（factorise）便直接说出 δV 的正负号：那个简洁的条件是由 $h^2 > a^2b^2/(a+b)^2$ 和 $h > 0$ 得出的。

STEP 之外

第二部分是一个小型的灵敏度分析（sensitivity analysis）：我们把映射 $h \mapsto (\alpha, V)$ 在某个解的附近线性化，然后读出输出如何随之变化。线性化（linearisation），以及与之相关的轨迹族包络线（envelope）的概念（即“安全抛物线”（parabola of safety），它是以某一给定速率所能到达的所有位置的边界），都是大学一年级力学和微积分课程的标准内容。Kleppner 与 Kolenkow 的 *An Introduction to Mechanics*（力学入门教科书）对抛体运动（projectile motion）和这些近似技巧都有细致的论述。

详细解答

STEP II 2002，第 10 题

力学 · 运动学与最优化

> **STEP II 2002，第 10 题**
>
> 继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：
>
> step.maths.org/questions/02-s2-q10

关键思路

总路程是固定的，所以完成比赛的时间 T 与转换速度的时刻 t 由一条方程联系在一起。去分母后，它变成一条关于 t 的二次方程（quadratic），而某个时间 T 可以达到，当且仅当这条二次方程有实根。由判别式（discriminant）条件可直接得出 T 的最小值，完全不需要微积分。至于第二部分，一旦知道两名跑手的位置如何随时间变化，二人之间的距离便是一个分段二次函数（piecewise quadratic），我们用惯常的方法求其最大值即可。

解答

以下距离一律以 km 为单位，时间以小时为单位，并且 $42\frac{3}{8} = \frac{339}{8}$。

最短的比赛时间

第一名参赛者在最初 t 小时内跑了 $13t$，然后在余下的 $T-t$ 小时内跑了 $\left(14+\frac{2t}{T}\right)(T-t)$。因此

$$13t + \left(14 + \frac{2t}{T}\right)(T-t) = \frac{339}{8}.$$

展开后得 $13t + 14T - 14t + 2t - \dfrac{2t^2}{T} = \dfrac{339}{8}$，即

$$t + 14T - \frac{2t^2}{T} = \frac{339}{8}.$$

两边乘以 $T > 0$，并按 t 的幂次整理：

$$2t^2 - Tt + \left(\frac{339}{8}T - 14T^2\right) = 0. \qquad (*)$$

想一想 我们要求最小的 T，但 T 并不是以 t 的函数形式给出的。怎样才能判断哪些 T 值是有可能出现的呢?

某个 T 值能够出现，当且仅当 $(*)$ 有实数解 t，所以它的判别式必须非负：

$$T^2 - 8\Big(\frac{339}{8}T - 14T^2\Big) \geq 0 \iff 113T^2 - 339T \geq 0 \iff 113T(T-3) \geq 0.$$

由于 $T > 0$，这就迫使 $T \geq 3$。

这个下界是可以达到的。当 $T = 3$ 时，判别式为零，$(*)$ 变成

$$2t^2 - 3t + \Big(\frac{1017}{8} - 126\Big) = 2t^2 - 3t + \frac{9}{8} = 2\Big(t - \frac{3}{4}\Big)^2 = 0,$$

所以 $t = \frac{3}{4}$。这个值介乎 0 与 T 之间，因此确实是一个可行的比赛方案：先以速率 13 跑 45 分钟，然后在余下的 $2\frac{1}{4}$ 小时以速率 $14 + \frac{2(3/4)}{3} = 14.5$ 跑完。事实上 $13 \cdot \frac{3}{4} + 14.5 \cdot \frac{9}{4} = 9.75 + 32.625 = 42.375$。因此 T 的最小可能值是 3，正如题目所要求。

（等价地，对 $(*)$ 作隐函数求导（implicit differentiation）并令 $\frac{\mathrm{d}T}{\mathrm{d}t} = 0$，可得 $T = 4t$，代回去便得 $T = 3$。判别式的论证有一个好处：它证明了 3 是全局最小值（global minimum），而不仅仅是一个驻值（stationary value）。）

最大的相隔距离

设 τ 为起跑后所经过的时间。第二名参赛者的速率是 $16 - k\tau$，其中 k 为某个常数，所以其所跑的距离是

$$s_2(\tau) = 16\tau - \tfrac{1}{2}k\tau^2.$$

由 3 小时跑毕全程得 $48 - \frac{9}{2}k = \frac{339}{8}$，所以 $\frac{9}{2}k = \frac{45}{8}$，$k = \frac{5}{4}$。于是 $s_2 = 16\tau - \frac{5}{8}\tau^2$，而速率由 16 下降至 12.25，一直保持为正。

第一名参赛者同样用了 3 小时，根据第一部分，这表示 $t = \frac{3}{4}$。其所跑的距离是

$$s_1(\tau) = \begin{cases} 13\tau, & 0 \leq \tau \leq \frac{3}{4}, \\ \frac{39}{4} + \frac{29}{2}\big(\tau - \frac{3}{4}\big), & \frac{3}{4} \leq \tau \leq 3. \end{cases}$$

设 $d = s_2 - s_1$。第二名跑手起步较快，所以我们预期 $d \geq 0$。

当 $0 \leq \tau \leq \frac{3}{4}$：$d = 3\tau - \frac{5}{8}\tau^2$，而 $\frac{\mathrm{d}d}{\mathrm{d}\tau} = 3 - \frac{5}{4}\tau > 0$。所以 d 在这个区间上递增，升至 $d\big(\frac{3}{4}\big) = \frac{9}{4} - \frac{45}{128} = \frac{243}{128}$。

当 $\frac{3}{4} \leq \tau \leq 3$：

$$d = 16\tau - \tfrac{5}{8}\tau^2 - \tfrac{39}{4} - \tfrac{29}{2}\tau + \tfrac{87}{8} = \tfrac{1}{8}\big(9 + 12\tau - 5\tau^2\big) = \tfrac{1}{8}(3-\tau)(3+5\tau).$$

这是一条开口向下的抛物线，顶点位于 $\tau = \frac{12}{10} = \frac{6}{5}$，落在 $\left[\frac{3}{4}, 3\right]$ 之内。（等价地，当两人速率相等时，差距便停止扩大：$16 - \frac{5}{4}\tau = 14.5$。）在该处

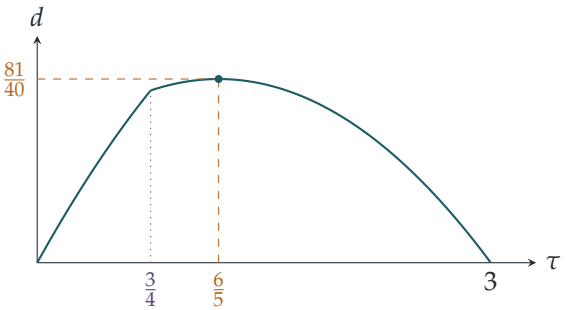
$$d\left(\tfrac{6}{5}\right) = \tfrac{1}{8} \cdot \tfrac{9}{5} \cdot 9 = \tfrac{81}{40}.$$

由因式分解后的形式可见，$d \geq 0$ 在整个区间上成立，并在终点降至 0，这是必然的。

比较两段的结果，$\frac{81}{40} = 2.025 > \frac{243}{128} \approx 1.898$。所以两名跑手之间的最大距离是

$$\boxed{\tfrac{81}{40}\ \text{km} = 2.025\ \text{km}},$$

在开赛后 1.2 小时（72 分钟）达到，当时第二名参赛者领先。



验算

- 两名跑手都在 $\tau = 3$ 时冲线，而公式给出 $d(3) = 0$，这是必然的。
- 第二段的最大值恰好出现在两人速率相等（同为 14.5）之时，这正是差距停止扩大的自然条件。
- 数值验证：在一组细密的 t 值上解这条关于 T 的二次方程，得出最小的 T 为 3.000，出现在 $t = 0.75$；在 $[0,3]$ 内的 300 000 个时刻计算 d，得出最大值为 2.025，出现在 $\tau = 1.2$，并且 $d \geq 0$ 始终成立。

常见错误

- 直接套用第一名参赛者的最短时间方案，却没有留意到正是“跑步时间为 3 小时”这个条件确定了 $t = \frac{3}{4}$。
- 只在转折点 $\tau = \frac{3}{4}$ 处求最大值：转换速度之后差距仍会继续扩大，因为直到 $\tau = \frac{6}{5}$ 之前，第二名跑手的速率仍然高于 14.5。
- 把第二名跑手所跑的距离写成 $16\tau - k\tau^2$，忘记了对线性递减的速率积分时所产生的 $\frac{1}{2}$。

STEP 之外

第一部分是在约束条件下求最优值（optimisation），而约束是由 $F(t,T)=0$ 以隐式给出的。提出“对哪些 T，解 t 才存在?”这个问题，正是隐函数定理（implicit function theorem）背后的想法，更一般地说，也是拉格朗日乘数（Lagrange multipliers）背后的想法：在最优点处，约束的等高线（level curve）满足 $\partial F/\partial t=0$，而这正是本题中的重根条件。Kleppner 与 Kolenkow 的 *An Introduction to Mechanics*（力学导论）全书以微积分处理运动学（kinematics），而 Siklos 的 *Advanced Problems in Mathematics*（高等数学难题）则有更多 STEP 风格的最优化题目。

详细解答

STEP II 2002，第 11 题

力学 · 摩擦力与极限平衡

STEP II 2002，第 11 题

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

step.maths.org/questions/02-s2-q11

关键思路

第一部分是直接的质心（centre of mass）积分；代换 $t = x/l$ 之后一行便可完成。至于梯子，“两端同时处于极限状态”令四个接触力全部化为同一个未知量的倍数，因此分解两次，再对 A 取力矩（moment），立即得出 $\tan\theta$。最后一部分还需要多一个想法：横梁竖得愈陡便愈安全，所以“对每个 $\theta < \pi/4$ 都会滑动”即是说极限角至少为 $\pi/4$。

解答

质心

距离 B 为 x 处、长度为 δx 的一小段，其重量为 $\alpha W l^{-1}(x/l)^{\alpha-1}\,\delta x$。设 $t = x/l$（故 $\mathrm{d}x = l\,\mathrm{d}t$），总重量为

$$\int_0^l \frac{\alpha W}{l}\left(\frac{x}{l}\right)^{\alpha-1}\mathrm{d}x = \alpha W\int_0^1 t^{\alpha-1}\,\mathrm{d}t = W,$$

正如所料（由于 $\alpha > 0$，此积分收敛）。重量对 B 的力矩为

$$\int_0^l x\,\frac{\alpha W}{l}\left(\frac{x}{l}\right)^{\alpha-1}\mathrm{d}x = \alpha W l\int_0^1 t^{\alpha}\,\mathrm{d}t = \frac{\alpha W l}{\alpha+1}.$$

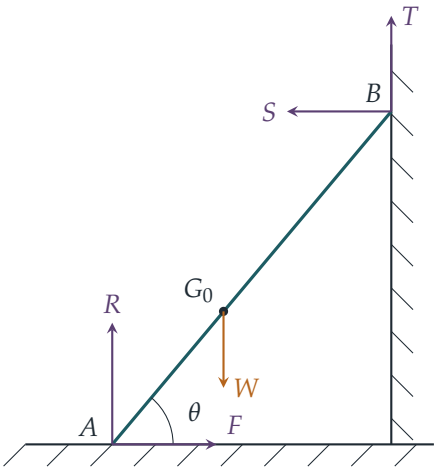
除以 W，可知质心 G_0 与 B 的距离为 $\alpha l/(\alpha+1)$，正是所求。因此它与 A 的距离为

$$d = l - \frac{\alpha l}{\alpha+1} = \frac{l}{\alpha+1}$$

稍后对 A 取力矩时，用的就是这个距离。

两端同处极限平衡

把横梁看成刚性杆（rigid rod），在 A 和 B 两处均为粗糙接触。若横梁滑动，A 会远离墙壁，B 会向下移动，所以摩擦力（friction）在 A 处指向墙壁，在 B 处向上。设地面在 A 处施加法向反作用力（normal reaction）R 和摩擦力 F，墙壁在 B 处施加法向反作用力 S 和摩擦力 T。重量作用于 G_0，与 A 相距 $l/(\alpha+1)$。



沿水平和垂直方向分解，并对 A 取力矩（重量的作用线与 A 的水平距离为 $d\cos\theta$；B 的高度为 $l\sin\theta$，水平距离为 $l\cos\theta$）：

$$F = S,$$
$$R + T = W,$$
$$S\,l\sin\theta + T\,l\cos\theta = W\,\frac{l}{\alpha+1}\cos\theta.$$

两端同处极限状态，即 $F=\mu R$ 及 $T=\mu S$。于是 $S=F=\mu R$，$T=\mu S=\mu^2 R$，而垂直方向的方程给出

$$R(1+\mu^2)=W, \qquad \text{所以} \qquad S=\frac{\mu W}{1+\mu^2},\quad T=\frac{\mu^2 W}{1+\mu^2}.$$

把力矩方程除以 $l\cos\theta$ 再代入，

$$\frac{\mu W}{1+\mu^2}\tan\theta+\frac{\mu^2 W}{1+\mu^2}=\frac{W}{1+\alpha} \quad\Longrightarrow\quad \mu\tan\theta=\frac{1+\mu^2}{1+\alpha}-\mu^2=\frac{1-\alpha\mu^2}{1+\alpha},$$

即

$$\tan\theta=\frac{1-\alpha\mu^2}{(1+\alpha)\mu},$$

正是所求。把这个角记作 θ_0，并设 $t_0=\tan\theta_0$。

μ 的最大值

想一想 为甚么“对每个 $\theta < \pi/4$ 都会滑动”可以化为关于 θ_0 的条件?

我们证明:平衡(equilibrium)可能成立,当且仅当 $\tan\theta \geq t_0$。设 $t = \tan\theta$ 及 $k = 1/(1+\alpha)$,则力矩方程可写成 $St + T = kW$。摩擦定律为 $F \leq \mu R$ 及 $|T| \leq \mu S$。

必要性(*necessity*)。由 $T \leq \mu S$ 及 $S = F \leq \mu R = \mu(W - T)$ 得 $T \leq \mu^2(W - T)$,所以 $T \leq \mu^2 W/(1+\mu^2)$。另外 $S \leq \mu(W - T)$,故

$$kW = St + T \leq \mu t(W - T) + T = \mu t W + (1 - \mu t)T.$$

若 $\mu t < 1$,利用 T 的上界可得 $kW \leq W(\mu t + \mu^2)/(1+\mu^2)$,整理后即 $t \geq t_0$。若 $\mu t \geq 1$,则无论如何都有 $t \geq 1/\mu > t_0$,因为 $\mu t_0 = (1 - \alpha\mu^2)/(1+\alpha) < 1$。两种情况下,平衡都迫使 $t \geq t_0$。

充分性(*sufficiency*)。若 $t \geq t_0$(且 $t > 0$),取墙壁的摩擦力为极限值 $T = \mu S$,由力矩方程得 $S = kW/(t+\mu)$,并取 $F = S$,$R = W - T$。余下唯一的条件是 $S \leq \mu R$,即 $S(1+\mu^2) \leq \mu W$,即 $k(1+\mu^2) \leq \mu(t+\mu)$,而这恰好就是 $t \geq t_0$。(此时亦有 $R > 0$。)

因此,横梁在角 θ 时滑动,当且仅当 $\tan\theta < t_0$。“对每个 $\theta < \pi/4$ 都会滑动”即 $t_0 \geq 1$:若 $t_0 \geq 1$,每个 $\theta < \pi/4$ 都满足 $\tan\theta < 1 \leq t_0$;若 $t_0 < 1$,任何满足 $\max(t_0, 0) < \tan\theta < 1$ 的角都小于 $\pi/4$,却处于平衡。

取 $\alpha = \frac{3}{2}$,

$$t_0 = \frac{1 - \frac{3}{2}\mu^2}{\frac{5}{2}\mu} = \frac{2 - 3\mu^2}{5\mu} \geq 1 \iff 3\mu^2 + 5\mu - 2 \leq 0 \iff (3\mu - 1)(\mu + 2) \leq 0,$$

其中两边乘以了 $5\mu > 0$。由于 $\mu + 2 > 0$,上式成立当且仅当 $\mu \leq \frac{1}{3}$。所以最大可能值为

$$\mu = \tfrac{1}{3}.$$

验算

- 当 $\alpha = 1$ 时横梁是均匀的,公式变成 $\tan\theta = (1 - \mu^2)/(2\mu)$,正是两端摩擦系数相同的均匀梯子的常见结果。
- 当 $\mu = \frac{1}{3}$、$\alpha = \frac{3}{2}$ 时:$t_0 = (2 - \frac{1}{3})/\frac{5}{3} = 1$,恰好等于 1,所以极限角为 $\pi/4$。数值上,$\mu = 0.332$ 给出 $t_0 \approx 1.004$,而 $\mu = 0.334$ 给出 $t_0 \approx 0.996$。
- 当 $\alpha \to \infty$ 时,质量集中于 B($d \to 0$),对任何 $\mu > 0$,$\tan\theta$ 都变为负数:横梁在任何角度都能竖立。这很合理,因为此时重量几乎作用在墙壁上。

常见错误

- 在力矩方程中把 $\alpha l/(\alpha+1)$ 当作与 A 的距离；它其实是由 B 量起的。
- 把墙壁的摩擦力画成向下：横梁滑动时 B 向下滑，所以横梁在 B 处所受的摩擦力向上。
- 把最后的不等式方向弄反：较陡的横梁较安全，所以“在 $\pi/4$ 以下会滑动”需要 $\tan\theta_0 \geq 1$，而最大的 μ 令 θ_0 恰好等于 $\pi/4$。

STEP 之外

上面的必要性与充分性论证，其实是一个小型的线性规划（linear programming）问题：摩擦定律界定了一个由可容许的力组成的凸区域（convex region），而当力矩直线与这个区域相交时，平衡便可能成立。像本题这类静不定（statically indeterminate）问题，即单凭方程不足以确定各力的问题，是大学一年级力学和工程学的常见课题。Kleppner and Kolenkow, *An Introduction to Mechanics*（力学入门教材）正是以这种严谨的方式处理摩擦、力矩和刚体平衡。

详细解答

STEP II 2002，第 12 题

概率与统计 · 正态近似与泊松近似

STEP II 2002，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/02-s2-q12

关键思路

这里有三层随机性，一层套一层：一枚硬币之内的正面次数、一天之内的硬币、一星期之内的日子。每一层都是一个二项计数，其“成功”概率来自下一层。公式的构造方法是：最底层用正态分布（normal distribution）近似（并作连续性校正（continuity correction）），中间层先用泊松分布（Poisson distribution）近似，再用正态分布近似，而最顶层则保留为精确的二项分布（binomial distribution）。一旦看出这三层结构，余下的工作就是辨认公式中每个式子出自哪一个近似，然后再问：对于题目给定的数值，每个近似是否可靠。

解答

某一天的一枚硬币

固定某一天和某一枚硬币。设 X 为该硬币 M 次投掷中正面的次数。各次投掷互相独立，每次掷出正面的概率为 p，所以

$$X \sim \mathrm{B}(M, p), \qquad \mathrm{E}(X) = Mp, \qquad \mathrm{Var}(X) = Mp(1-p).$$

当 M 很大时，我们以正态变量 $X' \sim \mathrm{N}\big(Mp,\, Mp(1-p)\big)$ 近似 X。由于 X 取整数值，“正面少于 m 次”即 $X \le m-1$，作连续性校正后就变成 $X' < m - \frac{1}{2}$。所以

$$r = \mathrm{P}(X < m) = \mathrm{P}(X \le m-1) \approx \Phi\left(\frac{m - \frac{1}{2} - Mp}{\sqrt{Mp(1-p)}}\right) = \Phi\left(\frac{2m - 1 - 2Mp}{2\sqrt{Mp(1-p)}}\right).$$

与题目比较，这说明 $h \approx Lr$。

某一天的所有硬币

设 Y 为 L 枚硬币中，当天正面少于 m 次的硬币数目。各枚硬币的投掷互相独立，而且每枚硬币的表现相同，所以每枚硬币各自独立地以概率 r “成功”，于是

$$Y \sim \mathrm{B}(L, r), \qquad \mathrm{E}(Y) = Lr = h, \qquad \mathrm{Var}(Y) = Lr(1-r) = h(1-r).$$

想一想 q 的公式中，分母是 $\sqrt{h}$。为甚么标准差（standard deviation）是 $\sqrt{h}$ 而不是 $\sqrt{h(1-r)}$?

这条公式是为 r 很小的情况而设的：对单独一枚硬币来说，正面少于 m 次是罕见事件，而 L 则很大。在这种情况下 $1-r \approx 1$，所以方差（variance）近似等于平均值（mean），这正是泊松近似 $Y \approx \mathrm{Po}(h)$ 适用的情形。如果 h 也相当大，$\mathrm{Po}(h)$ 又可以用 $\mathrm{N}(h,h)$ 近似。（等价的做法是：直接以正态分布近似 $\mathrm{B}(L,r)$，再把 $1-r$ 换成 1。）

因此取 $Y' \sim \mathrm{N}(h,h)$。“多于 l 枚硬币”即 $Y \geq l+1$，作连续性校正后就变成 $Y' > l + \frac{1}{2}$。于是

$$q = \mathrm{P}(Y > l) \approx 1 - \Phi\left(\frac{l + \frac{1}{2} - h}{\sqrt{h}}\right) = \Phi\left(\frac{h - l - \frac{1}{2}}{\sqrt{h}}\right) = \Phi\left(\frac{2h - 2l - 1}{2\sqrt{h}}\right),$$

当中用到标准正态分布的对称性 $1 - \Phi(z) = \Phi(-z)$。这就是题目给出的 q。

所有日子

如果某一天有多于 l 枚硬币的正面次数少于 m，就称这一天为好日子。每一天成为好日子的概率同样是 q，而不同日子涉及不同的投掷，所以各天互相独立。因此 K 天之中好日子的数目服从 $\mathrm{B}(K,q)$，而

$$\mathrm{P}(\text{恰好有 } k \text{ 个好日子}) = \binom{K}{k} q^k (1-q)^{K-k},$$

其中 q 按上文近似。这就是所要求的结果。

数值例子

这里 $Mp = 60$，$Mp(1-p) = 24$，所以

$$\frac{2m - 1 - 2Mp}{2\sqrt{Mp(1-p)}} = \frac{96 - 1 - 120}{2\sqrt{24}} = \frac{-25}{9.798} \approx -2.55,$$

由此得 $r \approx \Phi(-2.55) \approx 0.0054$ 及 $h = 500r \approx 2.68$。然后

$$\frac{2h - 2l - 1}{2\sqrt{h}} \approx \frac{5.36 - 9}{3.27} \approx -1.11, \qquad q \approx \Phi(-1.11) \approx 0.133,$$

而近似公式给出 $\binom{7}{2} q^2 (1-q)^5 \approx 21 \times 0.0178 \times 0.490 \approx 0.18$。

想一想 每一个近似都是标准做法。当中哪些是在其可靠条件以外使用的?

逐层检视这三层。

- **顶层**。$\mathrm{B}(7,q)$ 这一步是精确的（在 q 已知的前提下），所以它本身不会带来误差。但由于 $k=2$，它会把 q 的误差平方。
- **底层**。$M=100$，$Mp=60$，$M(1-p)=40$，完全处于通常使用正态分布近似二项分布的范围之内。但我们是在深入尾部 2.55 个标准差的位置使用它，用来求一个约为 0.005 的概率。就*绝对*误差而言，正态近似在分布的主体部分表现良好；但在尾部，它的*相对*误差可以很大，尤其因为 $p\neq\frac{1}{2}$ 令二项分布偏斜。r 的任何相对误差都会直接传到 $h=Lr$，而 q 对 h 十分敏感。
- **中间层**。以 $\mathrm{Po}(h)$ 取代 $\mathrm{B}(500,r)$ 效果极佳，因为 r 极小而 L 很大。但接着在 $h\approx 2.7$ 的情况下以正态分布近似 $\mathrm{Po}(h)$，却是没有根据的：平均值小于 3 的泊松分布明显偏斜，而常用的经验法则要求平均值至少约为 10（有人说 15 或 20）。

所以答案是否定的：我们不会预期这个近似准确。有疑问的步骤是对 $\mathrm{B}(100,0.6)$ 尾部的正态近似，以及对平均值很小的泊松分布的正态近似。

事实上，r 的精确值是 $\mathrm{P}(X\leq 47)\approx 0.00576$，比正态估计值高出约 7%，所以 h 应约为 2.88；而精确的 $\mathrm{P}(Y>4)$ 则约为 0.165，而不是 0.133。最终的概率约为 0.23，而不是 0.18。（就这组特定的数值而言，对 $\mathrm{Po}(2.68)$ 的正态近似所给出的尾部概率，碰巧与泊松分布本身几乎相同，即 0.133 对 0.134，所以大部分的偏差来自底层。但你事前不可能指望有这种运气。）

验算

- 连续性校正的方向正确：“少于 m”即 $X\leq m-1$，所以我们在 $m-\frac{1}{2}$ 处截止；“多于 l”即 $Y\geq l+1$，所以我们由 $l+\frac{1}{2}$ 开始。题目给出的公式中正好含有 $2m-1$ 和 $2l+1$。
- 数值上，$\Phi(-2.5516)=0.005362$，$h=2.681$，$q=\Phi(-1.1108)=0.1333$，由此得 0.1825。精确的二项求和给出 $r=0.005761$，$q=0.1645$（已由模拟 2×10^4 天加以确认，所得结果为 0.164），最终答案为 0.2314。

常见错误

- 用了 m 而不是 $m-\frac{1}{2}$（或用了 l 而不是 $l+\frac{1}{2}$）：这样公式就对不上。
- 用了 $\sqrt{h(1-r)}$，然后无法得出 $\sqrt{h}$。题目的公式略去了因子 $1-r$，因为它实际上是在使用泊松近似，而你应该指出这一点。

- 只凭“$Mp > 5$ 及 $M(1-p) > 5$”来判断准确度。这条法则针对的是分布的中心部分，而不是一个 0.005 的尾部概率。

STEP 之外

这道题目所关乎的，是误差如何在一连串近似之中传播，以及尾部的绝对误差与相对误差之间的分别。在大学里，中心极限定理（central limit theorem）会由 Berry–Esseen 定理进一步加强，该定理给出正态近似的绝对误差的上界。尾部概率则需要另一些工具，例如切尔诺夫界（Chernoff bounds）和大偏差理论（large-deviation theory），它们描述尾部概率缩小得有多快。Blitzstein 与 Hwang 的 *Introduction to Probability*（概率论入门）清楚地讲解了泊松近似、正态近似以及切尔诺夫型不等式；Grimmett 与 Welsh 的 *Probability: An Introduction*（概率论导引）则是一本很好的辅助读物。

详细解答

STEP II 2002，第 13 题

概率与统计 · 变换后的分布函数

STEP II 2002，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/02-s2-q13

关键思路

这一题虽然放在力学部分，其实是一道关于变换分布函数（distribution function）的概率题。一切都建基于一条公式：对 G 求导得 $G'(y) = f(y) \cdot \dfrac{2}{(2-F(y))^2}$，其中 $f = F'$ 是 X 的概率密度函数（density）。因此，Y 的密度就是 X 的密度乘以一个权重，而这个权重被夹在 $\frac{1}{2}$ 与 2 之间。有了这一点，Y 的每一个期望值（expectation）都是 X 的密度乘以一个在 $[\frac{1}{2}, 2]$ 内的权重的积分，只要比较被积函数（integrand），便可得出各个不等式。条件 $a > 0$ 的作用，是确保被积函数非负。

解答

G 的性质

我们按题目的原意，假设 X 在 $[a,b]$ 上有密度 $f(y) = F'(y) \geq 0$，因此 X 的全部概率都落在 $[a,b]$ 内。

由于 $F(a) = 0$ 及 $F(b) = 1$，

$$G(a) = \frac{0}{2-0} = 0, \qquad G(b) = \frac{1}{2-1} = 1.$$

由商法则（quotient rule），

$$G'(y) = \frac{F'(y)\big(2-F(y)\big) - F(y)\big(-F'(y)\big)}{(2-F(y))^2} = \frac{2F'(y)}{(2-F(y))^2}.$$

分子非负，因为分布函数是非递减的，所以 $F'(y) \geq 0$。分母是 $2 - F(y) \geq 1$ 的平方，故为正数。因此 $G'(y) \geq 0$。

至于那个界，由于 F 是分布函数，故 $0 \le F(y) \le 1$。于是

$$1 \le 2 - F(y) \le 2 \quad \Longrightarrow \quad 1 \le (2 - F(y))^2 \le 4 \quad \Longrightarrow \quad \tfrac{1}{2} \le \frac{2}{(2 - F(y))^2} \le 2,$$

其中各项皆为正数，所以取平方后不等号方向不变；取倒数会令不等号反向，再乘以 2 则只是把各项按比例放大。这就是所求的双重不等式。

综合而言，$G(a) = 0$、$G(b) = 1$ 以及 G 非递减，说明 G 确实是 $[a, b]$ 上的一个分布函数，所以随机变量（random variable）Y 是有意义的。它的密度是

$$g(y) = G'(y) = w(y)\, f(y), \qquad \text{其中} w(y) = \frac{2}{(2 - F(y))^2} \in \left[\tfrac{1}{2}, 2\right].$$

$\mathrm{E}(Y)$ 的界

> **想一想** 为甚么我们可以把不等式 $\frac{1}{2} \le w(y) \le 2$ 乘以 $yf(y)$，而不等号方向保持不变?

因为在 $[a, b]$ 上 $yf(y) \ge 0$：我们有 $f(y) \ge 0$，而且 $y \ge a > 0$。假设 $a > 0$ 正是用在这里。因此

$$\tfrac{1}{2}\, yf(y) \le y\, w(y) f(y) \le 2\, yf(y) \qquad \text{对于} a \le y \le b,$$

而在 $[a, b]$ 上积分后，不等式依然成立：

$$\tfrac{1}{2} \int_a^b yf(y)\, \mathrm{d}y \;\le\; \int_a^b y\, g(y)\, \mathrm{d}y \;\le\; 2 \int_a^b yf(y)\, \mathrm{d}y,$$

即

$$\tfrac{1}{2}\, \mathrm{E}(X) \le \mathrm{E}(Y) \le 2\, \mathrm{E}(X),$$

这正是所要证明的。

$\mathrm{Var}(Y)$ 的界

同样地，$y^2 f(y) \ge 0$（这只需要 $f \ge 0$），所以把 $w(y) \le 2$ 乘以它再积分，得

$$\mathrm{E}(Y^2) = \int_a^b y^2\, w(y) f(y)\, \mathrm{d}y \le 2 \int_a^b y^2 f(y)\, \mathrm{d}y = 2\, \mathrm{E}(X^2).$$

对于被减去的那一项，我们需要的是 $(\mathrm{E}(Y))^2$ 的下界。由上一部分，$\mathrm{E}(Y) \ge \frac{1}{2}\mathrm{E}(X)$，而且因为 $X \ge a > 0$，所以 $\mathrm{E}(X) > 0$。两边都是正数，故可以放心取平方：

$$(\mathrm{E}(Y))^2 \ge \tfrac{1}{4}(\mathrm{E}(X))^2.$$

因此

$$\begin{aligned}\operatorname{Var}(Y) &= \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2 \\ &\le 2\,\mathrm{E}(X^2) - \tfrac{1}{4}(\mathrm{E}(X))^2 \\ &= 2\big(\operatorname{Var}(X) + (\mathrm{E}(X))^2\big) - \tfrac{1}{4}(\mathrm{E}(X))^2 \\ &= 2\operatorname{Var}(X) + \tfrac{7}{4}(\mathrm{E}(X))^2,\end{aligned}$$

其中第三行用了 $\mathrm{E}(X^2) = \operatorname{Var}(X) + (\mathrm{E}(X))^2$。这就是所求的界。

验算

取 X 为 $[1,3]$ 上的均匀分布（uniform distribution），则 $F(y) = \frac{1}{2}(y-1)$，且 $\mathrm{E}(X) = 2$。由于 $Y \ge 1$，故 $\mathrm{E}(Y) = 1 + \int_1^3 (1 - G(y))\,\mathrm{d}y$；代入 $u = F(y)$ 得

$$\mathrm{E}(Y) = 1 + 2\int_0^1 \frac{2(1-u)}{2-u}\,\mathrm{d}u = 1 + 4(1 - \ln 2) \approx 2.227,$$

稳稳地落在 1 与 4 之间。直观地说，w 在 F 小的地方较小，在 F 接近 1 的地方较大，所以 Y 偏向区间的上端，在这个例子中 $\mathrm{E}(Y) > \mathrm{E}(X)$。我们亦以数值方法验证过：对于随机区间上的 300 个随机幂律分布（power-law distribution），$\frac{1}{2}\mathrm{E}(X) \le \mathrm{E}(Y) \le 2\mathrm{E}(X)$ 与方差（variance）的界都成立；观察所得的比值 $\mathrm{E}(Y)/\mathrm{E}(X)$ 大约介乎 0.92 与 1.31 之间。

常见错误

- 忘记了 $a > 0$ 为何重要：假如 X 可以取负值，$yf(y)$ 便可能是负数，乘以它会令不等号反向。
- 用 $\mathrm{E}(Y) \le 2\mathrm{E}(X)$ 来估计 $\operatorname{Var}(Y)$ 的界：$(\mathrm{E}(Y))^2$ 这一项是被减去的，所以你需要的是它的下界，而不是上界。
- 试图把 $(y-\mu)^2$ 对 g 积分，直接比较 $\operatorname{Var}(Y)$ 与 $\operatorname{Var}(X)$：两个方差所用的平均值并不相同，所以应改用 $\mathrm{E}(Y^2)$ 来处理。

STEP 之外

写成 $g = wf$，是测度变换（change of measure）的一个简单例子：新的分布相对于旧的分布有一个密度（称为似然比（likelihood ratio），或 Radon–Nikodym 导数），而这个比值的界可以直接转化为期望值的界。统计学中的重要性抽样（importance sampling），以及金融数学中所用的概率测度变换，都以这个想法为基础。Grimmett 与 Welsh 的 *Probability: An Introduction*（概率论入门）讲解分布函数、密度与期望值，程度正好适合由此再深入探究。

详细解答

STEP II 2002，第 14 题

概率与统计 · 混合分布与贝叶斯定理

STEP II 2002，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/02-s2-q14

关键思路

这是一个分两阶段进行的试验：先选出一个区域，然后该区域内发生意外。因此应以区域为条件来考虑。全期望定律（law of total expectation）给出平均值，全概率定律（law of total probability）给出分布，最后由贝叶斯定理（Bayes' theorem）把问题反过来问。题目只给出平均值，所以我们以泊松分布（Poisson distribution）作为某区域内意外宗数的模型；而要证明的公式带有 $\mathrm{e}^{-\lambda}\lambda^k/k!$ 的形式，也印证了这个模型。

解答

设置

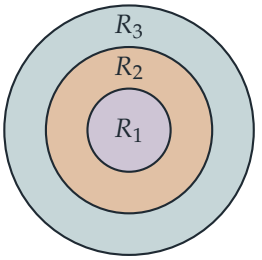
设 S 为所选区域的编号，$\lambda_n = 2 - \frac{n}{N}$ 为 R_n 内意外宗数的平均值。区域 R_n 是一个圆环（annulus），内外半径分别为 $n-1$ 和 n，所以它的面积是

$$\pi n^2 - \pi(n-1)^2 = \pi(2n-1).$$

整个岛的面积是 πN^2。（验算一下：$\sum_{n=1}^{N}(2n-1) = N^2$。）因此

$$\mathrm{P}(S=n) = \frac{2n-1}{N^2}, \qquad n = 1, \ldots, N.$$

在 $S = n$ 的条件下，我们把 X 看成平均值为 λ_n 的泊松变量。由于对每个 n 都有 $\lambda_n \geq 1 > 0$，这些泊松分布全都有意义。



区域面积：$R_n = \pi(2n-1)$

长期平均值

根据大数定律（law of large numbers），长远而言，每天记录到的平均宗数就是 $\mathrm{E}(X)$。以区域为条件，可得

$$\begin{aligned}\mathrm{E}(X) &= \sum_{n=1}^{N} \mathrm{P}(S=n)\,\mathrm{E}(X \mid S=n) = \sum_{n=1}^{N} \frac{2n-1}{N^2}\Big(2-\frac{n}{N}\Big)\\ &= 2\sum_{n=1}^{N}\frac{2n-1}{N^2} \;-\; \frac{1}{N^3}\sum_{n=1}^{N}(2n^2-n)\\ &= 2-\frac{1}{N^3}\Big(\frac{N(N+1)(2N+1)}{3}-\frac{N(N+1)}{2}\Big).\end{aligned}$$

第一个和等于 1，因为各区域被选中的概率加起来是 1。括号内我们用了题目给出的 $\sum n^2$ 公式，以及 $\sum n = \frac{1}{2}N(N+1)$。抽出 $\frac{1}{6}N(N+1)$ 后，余下 $2(2N+1)-3 = 4N-1$，所以

$$\mathrm{E}(X) = 2-\frac{(N+1)(4N-1)}{6N^2} = 2-\frac{1}{6}\Big(1+\frac{1}{N}\Big)\Big(4-\frac{1}{N}\Big),$$

其中把 N^2 拆成 $N\cdot N$，分给两个因子。这就是所要求的结果。

X 的分布

根据全概率定律，对 $k = 0, 1, 2, \ldots$，

$$\mathrm{P}(X=k) = \sum_{n=1}^{N}\mathrm{P}(S=n)\,\mathrm{P}(X=k \mid S=n) = \sum_{n=1}^{N}\frac{2n-1}{N^2}\cdot\frac{\mathrm{e}^{-\lambda_n}\lambda_n^k}{k!}.$$

现在把每一部分用 N 表示：

$$\lambda_n^k = \Big(\frac{2N-n}{N}\Big)^k = N^{-k}(2N-n)^k, \qquad \mathrm{e}^{-\lambda_n} = \mathrm{e}^{-2+n/N} = \mathrm{e}^{-2}\,\mathrm{e}^{n/N}.$$

因子 e^{-2}、N^{-2}、N^{-k} 和 $1/k!$ 都与 n 无关，所以可以提到求和符号之外：

$$\mathrm{P}(X=k) = \frac{\mathrm{e}^{-2}N^{-k-2}}{k!}\sum_{n=1}^{N}(2n-1)(2N-n)^k\,\mathrm{e}^{n/N},$$

正如所求。

$N = 3$、两宗意外的情况

我们要求 $\mathrm{P}(S = 2 \mid X = 2)$。根据贝叶斯定理，

$$\mathrm{P}(S = 2 \mid X = 2) = \frac{\mathrm{P}(S = 2)\,\mathrm{P}(X = 2 \mid S = 2)}{\mathrm{P}(X = 2)}.$$

> **想一想** 我们需要把 $\mathrm{P}(X = 2)$ 完整地计算出来吗？

不需要。分子正是 $\mathrm{P}(X = 2)$ 的求和式中 $n = 2$ 的那一项。因此公因子 $\mathrm{e}^{-2}N^{-k-2}/k!$ 可以约去，所求概率就是 $\sum(2n-1)(2N-n)^k\mathrm{e}^{n/N}$ 中 $n = 2$ 的一项除以整个和。

取 $N = 3$ 和 $k = 2$，各项 $(2n-1)(6-n)^2\mathrm{e}^{n/3}$ 分别是

$$n = 1:\ 1 \cdot 25\,\mathrm{e}^{1/3}, \qquad n = 2:\ 3 \cdot 16\,\mathrm{e}^{2/3} = 48\,\mathrm{e}^{2/3}, \qquad n = 3:\ 5 \cdot 9\,\mathrm{e} = 45\,\mathrm{e}.$$

于是，把分子和分母同除以 $\mathrm{e}^{2/3}$，得

$$\mathrm{P}(S = 2 \mid X = 2) = \frac{48\,\mathrm{e}^{2/3}}{25\,\mathrm{e}^{1/3} + 48\,\mathrm{e}^{2/3} + 45\,\mathrm{e}} = \frac{48}{48 + 45\,\mathrm{e}^{1/3} + 25\,\mathrm{e}^{-1/3}},$$

这就是所要求的结果。

验算

- 当 $N = 1$ 时只有一个区域，平均值为 1，而公式给出 $2 - \frac{1}{6} \cdot 2 \cdot 3 = 1$。当 $N \to \infty$ 时，平均值趋向 $2 - \frac{4}{6} = \frac{4}{3}$。这与 $\int_0^1 2x(2-x)\,\mathrm{d}x = \frac{4}{3}$ 吻合，其中均匀选取的点位于相对半径 x 处的密度为 $2x$。
- 当 $N = 3$ 时，平均值是 $\frac{32}{27} \approx 1.185$。模拟 4×10^5 天（在圆盘内均匀地选一点，然后抽取一个泊松计数），得出的平均值是 1.182。模拟同时得出 $\mathrm{P}(X = 2) \approx 0.209$（准确值为 0.2094），而在发生两宗意外的日子当中，落在 R_2 的比例为 0.372。最后答案的数值是 0.3729。
- 把分布公式对 k 求和，在我们试过的每个 N 之下，总概率都是 1。

常见错误

- 以相等的概率 $1/N$ 选取区域，而不是按面积 $2n-1$ 的比例选取。
- 把各个泊松分布的*平均值*取平均，然后声称 X 是以该值为平均值的泊松变量。泊松分布的混合并不是泊松分布。只有期望值才能这样简单地通过取平均而得出。

STEP 之外

变量 X 服从一个混合分布（mixture distribution）：区域充当隐变量（hidden variable），而最后一部分就是对它作贝叶斯推断（Bayesian inference）。上面第二个错误可以用全方差定律（law of total variance）来解释：$\mathrm{Var}(X) = \mathrm{E}(\mathrm{Var}(X \mid S)) + \mathrm{Var}(\mathrm{E}(X \mid S))$，它说明混合分布总比平均值相同的单一泊松分布更为分散。Grimmett 与 Welsh 所著的 *Probability: An Introduction*（概率论入门）对条件期望（conditional expectation）和这类混合分布有很好的讲解。

详细解答

STEP III 2002，第 1 题

纯数学 · 广义积分与体积

STEP III 2002，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：

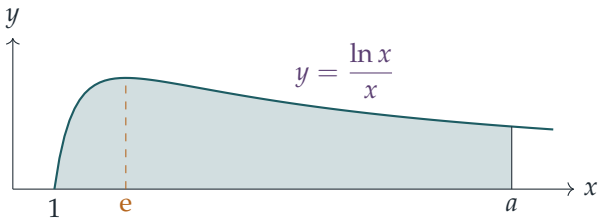
step.maths.org/questions/02-s3-q1

关键思路

只要留意到 $\frac{1}{x}$ 是 $\ln x$ 的导数（derivative），两个积分都变得容易。面积的被积函数（integrand）是“一个函数乘以它自己的导数”，所以一眼就能积出来。体积的被积函数 $(\ln x)^2/x^2$ 多出了一个 x 的幂，因此我们要用两次分部积分法（integration by parts），每次都对对数的幂求导，把它的次数降低；这里所说的对数就是 $\ln x$。有趣之处在于两个极限（limit）的对比：曲线衰减得刚好够慢，令面积无止境地增长；但把它平方之后，衰减就够快，令体积稳定下来。

解答

当 $x \geq 1$ 时，$\ln x \geq 0$，所以曲线位于 x 轴上或其上方，面积就是该积分本身。曲线由 $(1,0)$ 出发，上升至 $x = \mathrm{e}$ 处的极大值（该处 $y' = (1-\ln x)/x^2 = 0$），然后缓缓地向横轴衰减。



面积

由于 $\dfrac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{2}(\ln x)^2\right) = \dfrac{\ln x}{x}$，

$$A(a) = \int_1^a \frac{\ln x}{x}\,\mathrm{d}x = \left[\tfrac{1}{2}(\ln x)^2\right]_1^a = \tfrac{1}{2}(\ln a)^2,$$

因为 $\ln 1 = 0$。（等价的做法是用代换法（substitution）：设 $u = \ln x$，$\mathrm{d}u = \mathrm{d}x/x$，便得到 $\int_0^{\ln a} u\,\mathrm{d}u$。）

当 $a \to \infty$ 时，$\ln a \to \infty$，所以 $A(a) \to \infty$：面积是无界的，尽管它增长得非常缓慢。

体积

绕 x 轴旋转，所得的体积为 $V(a) = \pi\displaystyle\int_1^a y^2\,\mathrm{d}x = \pi\int_1^a \frac{(\ln x)^2}{x^2}\,\mathrm{d}x$。

用分部积分法，取 $u = (\ln x)^2$ 及 $\mathrm{d}v = x^{-2}\,\mathrm{d}x$，于是 $\mathrm{d}u = \frac{2\ln x}{x}\,\mathrm{d}x$ 及 $v = -\frac{1}{x}$：

$$\int \frac{(\ln x)^2}{x^2}\,\mathrm{d}x = -\frac{(\ln x)^2}{x} + 2\int \frac{\ln x}{x^2}\,\mathrm{d}x.$$

再做一次同样的步骤，取 $u = \ln x$，$\mathrm{d}v = x^{-2}\,\mathrm{d}x$：

$$\int \frac{\ln x}{x^2}\,\mathrm{d}x = -\frac{\ln x}{x} + \int \frac{1}{x^2}\,\mathrm{d}x = -\frac{\ln x}{x} - \frac{1}{x}.$$

把两者合并，

$$\int \frac{(\ln x)^2}{x^2}\,\mathrm{d}x = -\frac{(\ln x)^2 + 2\ln x + 2}{x} + C.$$

在 $x = 1$ 处，括号内的值是 $0 + 0 + 2$，所以下限的贡献是 $-(-2) = +2$，于是

$$V(a) = \pi\left(2 - \frac{(\ln a)^2 + 2\ln a + 2}{a}\right).$$

想一想 $(\ln a)^2/a$ 的分子和分母都趋向无穷大，我们怎样知道它趋向 0?

对数的增长比任何正数次幂都要慢。要把这句话说得严谨，可由以下事实出发：对所有 $t > 0$ 都有 $\ln t < t$（直线 $y = t$ 位于 $y = \ln t$ 的上方；两者的差距 $t - \ln t$ 在 $t = 1$ 处取得最小值 1）。代入 $t = a^{1/4}$：则 $\frac{1}{4}\ln a < a^{1/4}$，即 $\ln a < 4a^{1/4}$。所以当 $a > 1$ 时，

$$0 < \frac{(\ln a)^2}{a} < \frac{16\,a^{1/2}}{a} = \frac{16}{\sqrt{a}} \to 0, \qquad 0 < \frac{\ln a}{a} < \frac{4}{a^{3/4}} \to 0,$$

而 $2/a \to 0$ 自然不在话下。因此 $V(a)$ 中的分式趋向 0，而

$$V(a) \to 2\pi \quad \text{当} a \to \infty.$$

由于被积函数为正，$V(a)$ 随 a 递增，所以体积稳步上升并趋近 2π，但永远不会达到这个值：即使被旋转的区域面积无限，所得的立体却只有有限的体积。

验算

- 在 $a = 1$ 时，两条公式都给出 0，这是必然的：$A(1) = 0$ 及 $V(1) = \pi(2 - 2) = 0$。
- 对 $-\big((\ln x)^2 + 2\ln x + 2\big)/x$ 求导，可还原出 $(\ln x)^2/x^2$（$2\ln x$ 和 2 这两项与括号的导数互相抵消）。
- 数值上，对 $\pi \int_1^{10} (\ln x)^2/x^2 \,\mathrm{d}x$ 使用辛普森法则（Simpson's rule）得出 2.54247，与公式吻合；而设 $x = \mathrm{e}^t$ 后，无限延伸时的体积是 $\pi \int_0^{\infty} t^2 \mathrm{e}^{-t} \,\mathrm{d}t$，其数值计算结果为 $6.283185\ldots = 2\pi$。

常见错误

- 忘记下限：$\ln 1 = 0$ 令大部分的项在 $x = 1$ 处消失，但常数项 $-2/x$ 却不会消失，而整个极限值 2π 正是由它提供的。
- 毫无理由地断言 $(\ln a)^2/a \to 0$。只需写一行比较，例如 $\ln a < 4a^{1/4}$，便已足够。

STEP 之外

这是“加百列号角”（Gabriel's horn，即把 $y = 1/x$ 旋转而成）的近亲，后者是一个体积有限但表面积无限的立体：而在这里，连平面区域本身的面积也是无限的。代换 $x = \mathrm{e}^t$ 把体积化为 $\pi \int_0^{\infty} t^2 \mathrm{e}^{-t} \,\mathrm{d}t = \pi\,\Gamma(3) = 2!\,\pi$，让我们初次窥见伽玛函数（gamma function），$\Gamma(n+1) = n!$。怎样理解瑕积分（improper integral），以及“对数输给幂”这类比较，是大学一年级分析学（analysis）的核心课题；Abbott 的 *Understanding Analysis* 是阅读这方面内容的好去处。

详细解答

STEP III 2002，第 2 题

纯数学 · 反正切级数

STEP III 2002，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/02-s3-q2

关键思路

第一个结果其实是乔装了的正切加法公式（tangent addition formula）。真正要下工夫的，只是验证等式两边的角都落在同一个区间内，而 tan 在该区间上是一对一（one-to-one）的。有了这条公式，归纳法（induction）那一步只需套用它一次，再做一些因式分解（factorising）。至于最后一个和，窍门在于留意到 $r^2 - r + 1$ 不过是把 $r^2 + r + 1$ 中的 r 平移一位。因此新的级数（series）就是原来的级数，在前面多加一项而已。

解答

加法公式

设 $A = \arctan a$ 及 $B = \arctan b$。则 $\tan A = a$ 及 $\tan B = b$，并且

$$\tan(A + B) = \frac{\tan A + \tan B}{1 - \tan A \tan B} = \frac{a + b}{1 - ab},$$

由于 $ab < 1$，此式有意义。这说明 $A + B$ 与 $\arctan\left(\frac{a+b}{1-ab}\right)$ 的正切值相同。但单凭这一点并不足以令两者相等，因为 tan 每隔 π 便重复一次。所以我们需要两个角都落在同一个区间内，而 tan 在该区间上是一对一的。

? **想一想** 为甚么得到 $\tan(A + B) = \frac{a+b}{1-ab}$ 之后，证明还未完结?

证明还未完结，是因为 arctan 只会取 $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ 内的值。我们必须验证 $A + B$ 也落在这个区间内。

- 由于 $0 < a < 1$，而 arctan 是递增函数，故 $0 < A < \frac{\pi}{4}$。同理 $0 < B < \frac{\pi}{4}$，所以 $0 < A + B < \frac{\pi}{2}$。

► 由于 $a+b>0$ 及 $1-ab>0$，分数 $\frac{a+b}{1-ab}$ 为正数。所以它的 arctan 也落在 $(0,\frac{\pi}{2})$ 内。

函数 tan 在 $(0,\frac{\pi}{2})$ 上严格递增（strictly increasing），因而是一对一的。所以该区间内正切值相同的两个角必定相等，由此得

$$\arctan a+\arctan b=\arctan\left(\frac{a+b}{1-ab}\right). \qquad \checkmark$$

归纳法

记 $S_n=\sum_{r=1}^{n}\arctan\left(\frac{1}{r^2+r+1}\right)$。要证明的是 $S_n=\arctan\left(\frac{n}{n+2}\right)$。

基础情况（*base case*）。当 $n=1$ 时，两边均为 $\arctan\frac{1}{3}$。

归纳步骤（*inductive step*）。假设对某个 $n\geq 1$，有 $S_n=\arctan\left(\frac{n}{n+2}\right)$。下一项中有

$$(n+1)^2+(n+1)+1=n^2+3n+3,$$

所以

$$S_{n+1}=\arctan\left(\frac{n}{n+2}\right)+\arctan\left(\frac{1}{n^2+3n+3}\right).$$

两个自变量都严格介乎 0 与 1 之间，所以加法公式适用，其中 $a=\frac{n}{n+2}$ 及 $b=\frac{1}{n^2+3n+3}$。把 $\frac{a+b}{1-ab}$ 的分子和分母同乘以 $(n+2)(n^2+3n+3)$，得

$$\begin{aligned}\frac{a+b}{1-ab}&=\frac{n(n^2+3n+3)+(n+2)}{(n+2)(n^2+3n+3)-n}\\&=\frac{n^3+3n^2+4n+2}{n^3+5n^2+8n+6}.\end{aligned}$$

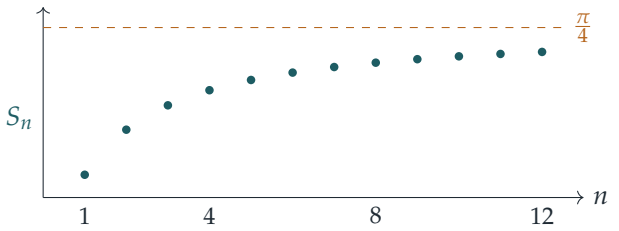
两个三次式都在某个负整数处取值为零：分子在 $n=-1$，分母在 $n=-3$。因式分解后，

$$\frac{a+b}{1-ab}=\frac{(n+1)(n^2+2n+2)}{(n+3)(n^2+2n+2)}=\frac{n+1}{n+3},$$

这里可以约去公因式，是因为 $n^2+2n+2=(n+1)^2+1>0$。因此 $S_{n+1}=\arctan\left(\frac{n+1}{(n+1)+2}\right)$，这正是 $n+1$ 时的命题。根据归纳法，结果对所有 $n\geq 1$ 成立。✓

无穷和。当 $n\to\infty$ 时，$\frac{n}{n+2}\to 1$。由于 arctan 是连续（continuous）的，

$$\sum_{r=1}^{\infty}\arctan\left(\frac{1}{r^2+r+1}\right)=\lim_{n\to\infty}\arctan\left(\frac{n}{n+2}\right)=\arctan 1=\frac{\pi}{4}.$$



最后的和

设 $s = r - 1$。则

$$r^2 - r + 1 = (s+1)^2 - (s+1) + 1 = s^2 + s + 1.$$

当 r 取遍 $1, 2, 3, \ldots$ 时，s 取遍 $0, 1, 2, \ldots$。$s = 0$ 的一项是 $\arctan 1 = \frac{\pi}{4}$，而其余各项（$s \geq 1$）恰好组成我们刚刚求和的级数。因此

$$\sum_{r=1}^{\infty} \arctan\left(\frac{1}{r^2 - r + 1}\right) = \arctan 1 + \sum_{s=1}^{\infty} \arctan\left(\frac{1}{s^2 + s + 1}\right) = \frac{\pi}{4} + \frac{\pi}{4} = \frac{\pi}{2}. \qquad \checkmark$$

第二个级数收敛（converge），是因为第一个级数收敛，所以把首项分拆出来是合理的。

验算

- 当 $n = 2$ 时：$\arctan\frac{1}{3} + \arctan\frac{1}{7}$ 的正切值为 $\frac{10/21}{20/21} = \frac{1}{2} = \frac{2}{4}$，与公式所预测的一致。
- 数值上，$S_{2000} \approx 0.784898$，而 $\arctan\frac{2000}{2002}$ 与之相差不超过 10^{-12}，另外 $\frac{\pi}{4} \approx 0.785398$。
- 把第二个级数求和至 $r = 10^6$，得 $1.5707953\ldots$，与 $\frac{\pi}{2}$ 相差不超过 10^{-6}。这与取 n 项后的差距 $\frac{\pi}{4} - \arctan\frac{n-1}{n+1} = \arctan\frac{1}{n} \approx \frac{1}{n}$ 吻合。

常见错误

- 写到“两边的 tan 相等”便停笔。缺少关于取值范围的论证，证明并不完整。（例如当 $a = b = 2$ 时公式不成立：左边是正数，但 $\arctan(-\frac{4}{3})$ 是负数。）
- 在归纳步骤中使用公式之前，忘记验证两个自变量都落在 $(0, 1)$ 内。
- 把最后的和又写成 $\frac{\pi}{4}$。平移 $r \mapsto r - 1$ 会在前面多加一项 $\arctan 1$。

STEP 之外

归纳法那部分还有一条更巧妙的路径。由于 $\frac{1}{r^2+r+1} = \frac{(r+1)-r}{1+r(r+1)}$，每一项都等于 $\arctan(r+1) - \arctan r$。于是这个和可以裂项相消（telescope），得 $\arctan(n+1) - \frac{\pi}{4}$，再由减法公式可知它等于 $\arctan\frac{n}{n+2}$。看出裂项相消的结构，以及利用连续性来论证部分和（partial sum）的极限，都是大学第一门分析（analysis）课程中的日常工具。Abbott 的 *Understanding Analysis*（分析学入门教材）细致地讨论了级数与连续性，而 Siklos 的 *Advanced Problems in Mathematics*（进阶数学难题集）则有更多这类 STEP 风格的练习。

详细解答

STEP III 2002，第 3 题

纯数学 · 曲线描绘

STEP III 2002，第 3 题

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

step.maths.org/questions/02-s3-q3

关键思路

一切都取决于图像的形状：图像由 $x = b$ 出发，起点处有一条铅垂切线（vertical tangent），曲线向下走，降至唯一的极小值（minimum），然后无止境地上升。只要知道起点的高度和极小值的高度，要数出 $\mathrm{f}(x) = c$ 有多少个解，只需在高度 c 处画一条水平线，看看它与曲线相交多少次。其后两条数值方程都是特殊情况，它们的 c^2 值分别落在这些范围的边界上和内部。

解答

图像

在左端，$\mathrm{f}(b) = a\sqrt{b}$。对 $x > b$ 求导，得

$$\mathrm{f}'(x) = \frac{a}{2\sqrt{x}} - \frac{1}{2\sqrt{x-b}}.$$

当 $x \to b^+$ 时，第二项趋向 $-\infty$，所以曲线离开点 $(b, a\sqrt{b})$ 时是笔直向下的（铅垂切线）。由于两个平方根均为正数，

$$\mathrm{f}'(x) < 0 \iff a\sqrt{x-b} < \sqrt{x} \iff a^2(x-b) < x \iff x < \frac{a^2 b}{a^2-1},$$

其中最后一步用到 $a^2 - 1 > 0$。因此 f 在 $[b, x_0]$ 上严格递减（strictly decreasing），在 $[x_0, \infty)$ 上严格递增（strictly increasing），其中

$$x_0 = \frac{a^2 b}{a^2 - 1}.$$

在极小点处，$x_0 - b = \dfrac{b}{a^2-1}$，所以

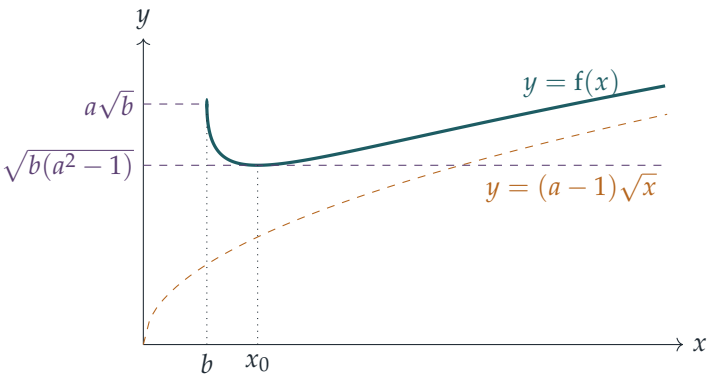
$$\mathrm{f}(x_0) = a\cdot\frac{a\sqrt{b}}{\sqrt{a^2-1}} - \frac{\sqrt{b}}{\sqrt{a^2-1}} = \frac{(a^2-1)\sqrt{b}}{\sqrt{a^2-1}} = \sqrt{b(a^2-1)}.$$

当 x 很大时，有理化（rationalising）后得

$$\mathrm{f}(x) = (a-1)\sqrt{x} + \left(\sqrt{x}-\sqrt{x-b}\right) = (a-1)\sqrt{x} + \frac{b}{\sqrt{x}+\sqrt{x-b}},$$

所以，因为 $a>1$，$\mathrm{f}(x)\to\infty$，而且曲线位于 $y=(a-1)\sqrt{x}$ 稍上方，并随 $x\to\infty$ 向它靠近。

最后，起点的高度高于极小值：$a^2b > (a^2-1)b$，所以 $a\sqrt{b} > \sqrt{b(a^2-1)}$。



数算解的个数

记极小值为 $m=\sqrt{b(a^2-1)}$，起点的值为 $M=a\sqrt{b}$，于是 $0<m<M$。在两段单调（monotonic）的部分上，f 各自连续（continuous）且严格单调，所以它对值域内的每个值都恰好取一次：

- 在 $[b,x_0]$ 上，f 对 $[m,M]$ 内的每个值恰好取一次；
- 在 $[x_0,\infty)$ 上，f 对 $[m,\infty)$ 内的每个值恰好取一次。

这两段只有 x_0 一个公共点，该处的函数值为 m。

由于对所有 $x\geq b$ 均有 $\mathrm{f}(x)\geq m$，当 $c<m$ 时方程 $\mathrm{f}(x)=c$ 无解。对 $c>0$ 而言，这等同于 $c^2<m^2$，即

$$c^2 < b(a^2-1) \quad\Longrightarrow\quad \text{无解},$$

这正是要证明的结果。至于其余情况：

想一想 为甚么 $c^2 = a^2 b$ 时有两个解，而不是一个?

当 $c = M$ 时，左边一段提供的解正是端点 $x = b$ 本身；由于定义域（domain）包含 $x = b$，它是一个名正言顺的解。右边一段则在更远处再提供一个解。只有当 $c > M$ 时，左边一段才不再提供解。因此，对 $c > 0$：

- 当 $c^2 = b(a^2 - 1)$（只有 $x = x_0$，即直线刚好碰到极小点），或当 $c^2 > a^2 b$（只有上升的一支能到达这个高度）时，**恰有一个解**；
- 当 $b(a^2 - 1) < c^2 \le a^2 b$（两支各有一个）时，**恰有两个解**。

第 (i) 部分

这里 $a = 3$，$b = 2$，$c = 4$，所以 $b(a^2 - 1) = 16 = c^2$：直线刚好碰到极小点，方程恰有一个解，

$$x = x_0 = \frac{9 \cdot 2}{8} = \frac{9}{4}.$$

确实有 $3 \cdot \frac{3}{2} - \sqrt{\frac{1}{4}} = \frac{9}{2} - \frac{1}{2} = 4$。直接计算亦可：把方程改写成 $3\sqrt{x} - 4 = \sqrt{x - 2}$ 再两边平方，得 $8x - 24\sqrt{x} + 18 = 0$，即 $(2\sqrt{x} - 3)^2 = 0$。重根（repeated root）正是相切在代数上的标记。

第 (ii) 部分

这里 $a = 3$，$b = 3$，$c = 5$，所以 $b(a^2 - 1) = 24 < 25 = c^2 < 27 = a^2 b$，方程恰有两个解。要求出这两个解，把方程写成 $3\sqrt{x} - 5 = \sqrt{x - 3}$ 再两边平方：

$$9x - 30\sqrt{x} + 25 = x - 3 \quad \Longrightarrow \quad 4x - 15\sqrt{x} + 14 = 0.$$

把它看成 $\sqrt{x}$ 的二次方程，有 $(\sqrt{x} - 2)(4\sqrt{x} - 7) = 0$，所以 $\sqrt{x} = 2$ 或 $\sqrt{x} = \frac{7}{4}$，由此得

$$x = 4 \qquad 或 \qquad x = \frac{49}{16}.$$

平方可能会引入增根（spurious roots），所以我们把两个值都代入原方程验算。两者都满足 $x \ge 3$（因为 $\frac{49}{16} = 3.0625$），而且

$$3 \cdot 2 - \sqrt{1} = 5, \qquad 3 \cdot \frac{7}{4} - \sqrt{\frac{1}{16}} = \frac{21}{4} - \frac{1}{4} = 5.$$

所以两个都是真正的解，与之前数出的两个解吻合。一如所料，它们分别位于 $x_0 = \frac{27}{8} = 3.375$ 的两侧。

验算

- 取 $a = \frac{3}{2}$，$b = 2$（即草图所用的数值），则 $x_0 = \frac{18}{5}$，而 $\mathrm{f}(\frac{18}{5}) = \frac{3}{2}\sqrt{3.6} - \sqrt{1.6} = \sqrt{2.5} \approx 1.581$，低于 $\mathrm{f}(2) = \frac{3}{2}\sqrt{2} \approx 2.121$。
- 随机选取 200 组 (a, b)，在细密的格点上对 f 作数值扫描，结果证实极小值 $\sqrt{b(a^2-1)}$ 出现在 x_0 处；而数算曲线与 $y = c$ 的交点，在 c^2 的三个范围内分别得到 0、2 和 1 个解。
- 在 (i) 中，二重根与 c^2 恰好等于 $b(a^2-1)$ 这一点吻合；在 (ii) 中，$c^2 = 25$ 严格介乎 24 与 27 之间，而我们确实找到了两个根。

常见错误

- 把 $c^2 = a^2 b$ 归入“一个解”的情况：$x = b$ 在定义域内，所以要计算在内。
- 平方之后没有验算。这里两个根都保留下来，但你必须逐一检验 $3\sqrt{x} - 5 \geq 0$ 和 $x \geq 3$。
- 忘记了正是 $a > 1$ 才令 $\mathrm{f} \to \infty$；假如 $a \leq 1$，便不会有极小值，整个图像会完全不同。

STEP 之外

这个数算解的个数的论证，建基于分析学（analysis）的两个事实：区间上的连续函数会取遍两端函数值之间的每一个值（介值定理，intermediate value theorem）；而严格单调函数对每个值最多取一次。两者合起来，便把一幅草图变成一个证明。Abbott 的 *Understanding Analysis* 对这两点都有细致的阐述，并说明为何连续性恰恰就是所需的假设。

详细解答

STEP III 2002，第 4 题

纯数学 · 丢番图不等式

> **STEP III 2002，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/02-s3-q4

关键思路

方程 $x^3+x^2=y^3-y^2$ 的一边增长得快，另一边却被 $-y^2$ 拖住，所以 y 必须“超越” x，但超出的幅度又不足 1。这两个不等式（inequality）把 y 严格夹在 x 与 $x+1$ 之间，而这个空隙容不下任何整数。至于最后一部分，额外的解只可能来自零或负数的取值，而一个巧妙的对称性（symmetry）——把 (m,n) 换成 $(-n,-m)$——可以把分类讨论的工夫减半。

解答

第一个结果

设 $x,y>0$ 且 $x^3+x^2=y^3-y^2$，并用反证法（proof by contradiction）假设 $x\geq y$。由于立方和平方在正实数上都是递增的，

$$x^3+x^2\geq y^3+y^2>y^3-y^2,$$

其中严格不等的一步用到 $y^2>0$。这与方程矛盾，所以 $x<y$。

第二个结果

现在设 $0<x\leq y-1$，于是 $y\geq x+1>1$。记 $f(t)=t^3-t^2$。那么当 $t>\frac{2}{3}$ 时，$f'(t)=3t^2-2t=t(3t-2)>0$，所以 f 在 $[1,\infty)$ 上是递增函数（increasing function）。

由于 $y \geq x+1$，而且两者都在这个区间内，

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

因为当 $x > 0$ 时 $x^2 + x > 0$。所以 $x^3 + x^2 < y^3 - y^2$，这正是所要证明的。

把两个结果合起来可知：对于正数 x，任何满足 $x^3 + x^2 = y^3 - y^2$ 的正数 y 都符合

$$x < y < x + 1.$$

没有正整数对

假设正整数 a 和 b 满足 $a^3 - b^3 = a^2 + b^2$。右边是正数，所以 $a^3 > b^3$，即 $a > b$。移项后得

$$b^3 + b^2 = a^3 - a^2,$$

这正是上面的方程，其中 $x = b$、$y = a$，两者皆为正数。所以 $b < a < b + 1$。

想一想 整数这个条件究竟用在哪里?

只用在这里：没有整数严格介乎 b 与 $b+1$ 之间。换个说法，$a > b$ 而 a, b 都是整数，就必有 $a \geq b + 1$，即 $0 < b \leq a - 1$，于是第二个结果给出 $b^3 + b^2 < a^3 - a^2$，与方程矛盾。所以这样的正整数对并不存在。（若在实数范围内，解多的是：对每个 $x > 0$ 都有一个 $y \in (x, x+1)$；只是整数刚好碰不上它们。）

所有整数对

现在要找出所有整数 m, n，使得

$$m^3 - n^3 = m^2 + n^2. \qquad (*)$$

以下两点观察可以省却不少工夫。

- 若 (m, n) 是解，则 $(-n, -m)$ 也是解：事实上 $(-n)^3 - (-m)^3 = m^3 - n^3$，而平方和保持不变。
- $(*)$ 的右边非负，所以 $m^3 \geq n^3$，即 $m \geq n$，而等号只在 $m^2 + n^2 = 0$ 时成立，即 $m = n = 0$。

其中一个是零。若 $n = 0$，则 $m^3 = m^2$，所以 $m = 0$ 或 $m = 1$。若 $m = 0$，则 $-n^3 = n^2$，所以 $n = 0$ 或 $n = -1$。这样得到 $(0, 0)$、$(1, 0)$ 和 $(0, -1)$（最后一个是 $(1, 0)$ 在上述对称下的像）。

两个都是正数。根据上一部分，这是不可能的。

两个都是负数。这时 $-n$ 和 $-m$ 都是正数，而 $(-n,-m)$ 就会是一个两项皆正的解，这是不可能的。

正负号相反。由于 $m>n$，我们有 $m>0>n$。记 $n=-p$，其中 $p\geq 1$。那么 $(*)$ 变成

$$m^3+p^3=m^2+p^2, \qquad \text{即} \qquad m^2(m-1)+p^2(p-1)=0.$$

对于正整数，两项都 ≥ 0，所以两项都等于零，于是必有 $m=p=1$。这样得到 $(1,-1)$，而确实 $1-(-1)=2=1+1$。

所以 $m^3-n^3=m^2+n^2$ 的整数解 (m,n) 恰好是

$$(m,n)=(0,0),\quad (1,0),\quad (0,-1),\quad (1,-1).$$

如果“整数对”的意思不包括平凡（trivial）的那一对，就剔除 $(0,0)$。

验算

每一对都可以用手验算：对于 $(1,0)$，$1-0=1=1+0$；对于 $(0,-1)$，$0-(-1)=1=0+1$；对于 $(1,-1)$，$2=2$。用电脑搜索所有 $|m|,|n|\leq 300$ 的情况，找到的正好是这四个解。至于头两个结果的验算：对随机选取的 $x>0$ 以数值方法解 $y^3-y^2=x^3+x^2$，每次都得到 $x<y<x+1$；例如 $x=1$ 时 $y\approx 1.696$。

常见错误

- 在最后一部分忘记了零和负整数，而所有的解偏偏都在那里。
- 想当然地认定两个整数中哪一个排在“前面”。这里立方差是正数，所以较大的数必须排在前面；这一点要写明。
- 证明第二个结果时只是“展开”了事，没有指出 t^3-t^2 在 $t\geq 1$ 时递增；$f(y)\geq f(x+1)$ 这一步正需要这个性质。

STEP 之外

“实数解被严格夹在两个相邻整数之间，所以没有整数解”这个论证，是研究丢番图方程（Diophantine equation）的标准工具，面对整数多项式方程时，往往也是第一样值得一试的方法。许多这类方程只有有限多个解，也是靠这样给解定出界限来证明的，然后就可以用有限的搜索把解全部找出来。Davenport 的 *The Higher Arithmetic*（一本数论入门书）是这类数论的绝佳入门读物。

详细解答

STEP III 2002，第 5 题

纯数学 · 四次方程的因式分解

STEP III 2002，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/02-s3-q5

关键思路

一个二次式是某个完全平方（perfect square）的常数倍，当且仅当它的判别式（discriminant）等于零。第一部分的全部内容就是这一点，而它也是其余部分的原动力：如果我们能选取 t，使得某个二次式的平方减去该四次式（quartic）后是一个完全平方的倍数，那么这个四次式就是*两个平方之差*（difference of two squares），可以立即因式分解（factorise）。这就是费拉里（Ferrari）解四次方程的经典方法，它把四次方程的问题换成一个关于 t 的三次方程（cubic）。

解答

二次式的条件

设 $p \neq 0$。配方（completing the square）得

$$px^2 + qx + r = p\left(x + \frac{q}{2p}\right)^2 + r - \frac{q^2}{4p}.$$

这个式子可写成 $p(x+h)^2$（h 为某个数）的形式，当且仅当常数 $r - \frac{q^2}{4p}$ 等于零，即

$$q^2 = 4pr,$$

此时 $h = \frac{q}{2p}$。（反过来，展开 $p(x+h)^2 = px^2 + 2phx + ph^2$ 可得 $q^2 = 4p^2h^2 = 4pr$，所以这个条件既是充分条件，也是必要条件。）若 $p = 0$，唯一的可能是 $q = r = 0$；条件 $q^2 = 4pr$ 仍然迫使 $q = 0$，但我们还需要 $r = 0$。

t 所满足的方程

展开得

$$\left(x^2+\tfrac{1}{2}bx+t\right)^2 = x^4+bx^3+\left(\tfrac{1}{4}b^2+2t\right)x^2+btx+t^2.$$

选用 $\frac{1}{2}b$，正是为了在减去四次式时令 x^3 项互相抵消（x^4 项也同样抵消）：

$$\left(x^2+\tfrac{1}{2}bx+t\right)^2-\left(x^4+bx^3+cx^2+dx+e\right)=\left(\tfrac{1}{4}b^2+2t-c\right)x^2+(bt-d)x+\left(t^2-e\right).$$

这是一个二次式，其中 $p=\frac{1}{4}b^2+2t-c$、$q=bt-d$ 及 $r=t^2-e$，而我们希望它等于 $k(x+h)^2$，即一个完全平方的倍数（倍数就是 $k=p$）。根据第一部分，t 必须满足

$$(bt-d)^2=4\left(\tfrac{1}{4}b^2+2t-c\right)\left(t^2-e\right).$$

这是一个关于 t 的三次方程（右边 t^3 项的系数是 8，而左边只是 t 的二次式），所以它必定有至少一个实根（real root）。

分解四次式

这里 $b=6$、$c=9$、$d=-2$、$e=-7$，所以

$$\tfrac{1}{4}b^2+2t-c=2t,\qquad bt-d=6t+2,\qquad t^2-e=t^2+7.$$

t 所满足的方程变成

$$\begin{aligned}(6t+2)^2&=8t\left(t^2+7\right)\\ 36t^2+24t+4&=8t^3+56t\\ 0&=2t^3-9t^2+8t-1.\end{aligned}$$

> **想一想** 怎样才能快速找出这个三次方程的一个根?

先试较小的数值：各系数之和是 $2-9+8-1=0$，所以 $t=1$ 是一个根。任何一个根都可以用，因为每个根都给出一个有效的因式分解；最简单的根给出最简单的分解。

取 $t=1$，所得的差是

$$2x^2+8x+8=2(x+2)^2,$$

所以 $k=2$ 及 $h=2$，而且确实有 $8^2=4\cdot 2\cdot 8$。因此

$$\left(x^2+3x+1\right)^2-\left(x^4+6x^3+9x^2-2x-7\right)=2(x+2)^2,$$

即

$$x^4+6x^3+9x^2-2x-7=\left(x^2+3x+1\right)^2-\left(\sqrt{2}\,(x+2)\right)^2.$$

这是两个平方之差，所以

$$\begin{aligned} x^4+6x^3+9x^2-2x-7 &= \left(x^2+3x+1-\sqrt{2}(x+2)\right)\left(x^2+3x+1+\sqrt{2}(x+2)\right) \\ &= \left(x^2+(3-\sqrt{2})x+1-2\sqrt{2}\right)\left(x^2+(3+\sqrt{2})x+1+2\sqrt{2}\right). \end{aligned}$$

由于 $k=2>0$，它的平方根是实数，所以这是两个实系数的二次因式，正如题目所要求。

三次方程的另外两个根是由 $2t^2-7t+1=0$ 得出的 $t=\frac{1}{4}\left(7\pm\sqrt{41}\right)$，它们同样满足 $k=2t>0$，并给出另外两种把四次式的四个（实）根配对成实二次因式的方法。这两种分解的系数繁复得多，所以应该选用 $t=1$。

验算

- 乘开：两个因式之积的 x^3 系数是 $(3-\sqrt{2})+(3+\sqrt{2})=6$，常数项是 $(1-2\sqrt{2})(1+2\sqrt{2})=1-8=-7$，两者皆正确。
- 当 $x=1$ 时，四次式的值是 $1+6+9-2-7=7$，而因式分解给出 $(5-\sqrt{2}-2\sqrt{2})(5+\sqrt{2}+2\sqrt{2})=25-18=7$。
- 两个因式的判别式分别是 $7+2\sqrt{2}\approx 9.83$ 和 $7-2\sqrt{2}\approx 4.17$，两者皆为正数，所以四次式有四个实根，这与三次方程的三个根全都给出实因式分解一致。我们亦在 200 个随机点上以数值方法核对过这个因式分解。

常见错误

- 把 $(x^2+\frac{1}{2}bx+t)^2$ 的 x^2 系数算错：它应是 $\frac{1}{4}b^2+2t$，因为除了 $(\frac{1}{2}bx)^2$ 之外，交叉项 $2\cdot x^2\cdot t$ 也有贡献。
- 把判别式条件写成 $q^2=pr$，或者忘记 k 就是首项系数（leading coefficient）$\frac{1}{4}b^2+2t-c$，而它必须是正数才能得到实因式分解。

STEP 之外

这个关于 t 的三次方程称为*预解三次式*（resolvent cubic），而这就是费拉里在十六世纪提出的方法：每个四次方程都可以化为一个三次方程，再用卡尔达诺（Cardano）公式求解。这个方法行得通的深层原因，以及一般五次方程（quintic）没有这类公式的原因，都属于伽罗瓦理论（Galois theory）的范畴，该理论研究的是多项式的根可以怎样置换（permute）。若想初步认识域（field）与多项式的代数，可以读 Allenby 的 *Rings, Fields and Groups*（环、域与群）。

详细解答

STEP III 2002，第 6 题

纯数学 · 非线性微分方程

STEP III 2002，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/02-s3-q6

关键思路

这条方程因为有四次方而显得吓人，但 $y\,\frac{\mathrm{d}y}{\mathrm{d}x}$ 正好是 y^2 的导数（derivative）的一半。因此作代换（substitution）$u = y^2 - 1$：方程随即化简为只含 u 和 u' 的方程，取平方根之后便可分离变量（separate variables）。唯一要小心的是正负号：u 的正负号把平面分成带状区域 $|y| < 1$ 和区域 $|y| > 1$，两者给出不同的曲线族（family of curves）。

解答

设 $u = y^2 - 1$，则 $u' = 2y\,\frac{\mathrm{d}y}{\mathrm{d}x}$，而 $y^4\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^4 = \left(\frac{1}{2}u'\right)^4$。方程变成

$$\tfrac{1}{16}(u')^4 = u^2 \quad\Longleftrightarrow\quad (u')^2 = 4|u|,$$

这里对两边取非负平方根（$(u')^2$ 和 $4|u|$ 都是非负的，所以没有遗漏任何解）。

首先留意，解曲线（solution curve）永远不会与 $y = 0$ 相遇：在那里原方程的左边是 0，右边却是 1。所以凡经过任何一个给定点的解，都处处满足 $y > 0$。

想一想 怎样解 $(u')^2 = 4u$，而不必为 u' 该取哪个正负号而烦恼?

在 $u > 0$ 之处，可以写 $u = v^2$，其中 $v = \sqrt{u} > 0$。于是 $u' = 2vv'$，所以 $(u')^2 = 4u$ 变成 $4v^2(v')^2 = 4v^2$，即 $(v')^2 = 1$。由于 v' 连续而且永不为零，它的正负号保持不变，所以 $v' = \pm 1$，并且 $v = \pm(x - a)$，其中 a 为某个常数。无论哪一种情况，都有

$$u = v^2 = (x - a)^2.$$

正负号的不确定性在平方之后自然消失。对 $u < 0$ 的情况，以 $u = -v^2$ 作同样的论证即可。

第 (i) 部分：经过 $(0, \frac{1}{2}\sqrt{3})$

这里 $y^2 = \frac{3}{4} < 1$，所以在这点附近 $u < 0$，即 $|y| < 1$。写 $1 - y^2 = -u = v^2$，其中 $v > 0$；那么 $(u')^2 = -4u$ 给出 $(v')^2 = 1$，与上面完全一样，所以 $1 - y^2 = (x - a)^2$，也就是

$$(x - a)^2 + y^2 = 1 :$$

圆心在 x 轴上、半径为 1 的圆。在 $x = 0$ 处需要 $a^2 = 1 - \frac{3}{4} = \frac{1}{4}$，所以 $a = \pm\frac{1}{2}$。由于 $y > 0$，经过 $(0, \frac{1}{2}\sqrt{3})$ 的解曲线是上半圆

$$y = \sqrt{1 - \left(x - \tfrac{1}{2}\right)^2} \qquad 及 \qquad y = \sqrt{1 - \left(x + \tfrac{1}{2}\right)^2}.$$

直接验证：在 $(x - a)^2 + y^2 = 1$ 上，隐函数微分（implicit differentiation）给出 $y\frac{\mathrm{d}y}{\mathrm{d}x} = -(x - a)$，所以 $y^4\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^4 = (x - a)^4 = (1 - y^2)^2 = (y^2 - 1)^2$。✓

第 (ii) 部分：经过 $(0, \frac{1}{2}\sqrt{5})$

现在 $y^2 = \frac{5}{4} > 1$，所以 $u > 0$，上面的论证给出 $y^2 - 1 = (x - a)^2$，也就是

$$y^2 - (x - a)^2 = 1 :$$

中心为 $(a, 0)$ 的等轴双曲线（rectangular hyperbola）。在 $x = 0$ 处，$a^2 = \frac{5}{4} - 1 = \frac{1}{4}$，所以同样有 $a = \pm\frac{1}{2}$，而由于 $y > 0$，解曲线是上支

$$y = \sqrt{1 + \left(x - \tfrac{1}{2}\right)^2} \qquad 及 \qquad y = \sqrt{1 + \left(x + \tfrac{1}{2}\right)^2},$$

对所有 x 都有定义。直接验证：$y\frac{\mathrm{d}y}{\mathrm{d}x} = x - a$，所以左边是 $(x - a)^4 = (y^2 - 1)^2$。✓

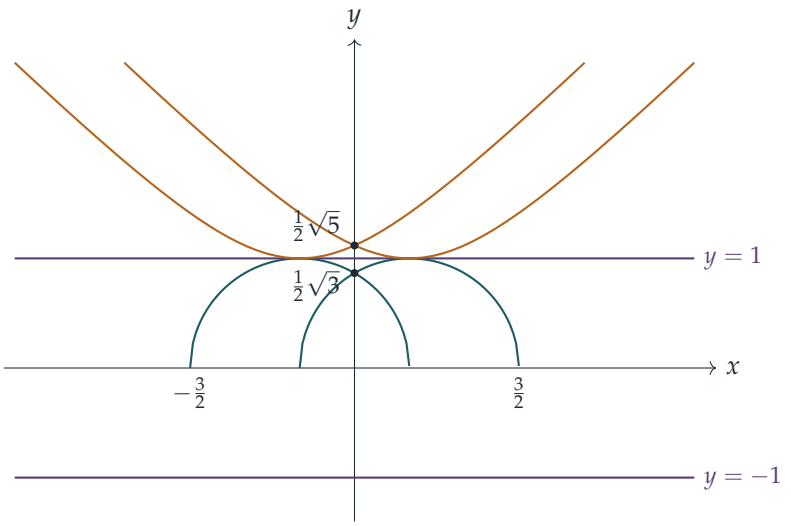
直线 $y = \pm 1$

若对所有 x 都有 $y = 1$ 或 $y = -1$，则 $\frac{\mathrm{d}y}{\mathrm{d}x} = 0$，所以左边是 0；而 $y^2 - 1 = 0$，所以右边也是 0。两条直线都是解。

想一想 这两条直线与那些圆和双曲线有甚么关系?

每个圆 $(x - a)^2 + y^2 = 1$ 和每条双曲线 $y^2 - (x - a)^2 = 1$ 都在 $(a, 1)$ 以零斜率与 $y = 1$ 相切（同样地，在 $(a, -1)$ 与 $y = -1$ 相切）。所以 $y = \pm 1$ 是这两个曲线族的包络线（envelope），而它们并不属于任何一族：它们来自 $u = 0$，而我们除以 v 时已把这情况排除在外。由此得出一个结果：经过 $y = 1$ 上的点的解并不唯一。例如，沿着 $a = \frac{1}{2}$ 的圆由 $(0, \frac{1}{2}\sqrt{3})$ 走到它的顶点 $(\frac{1}{2}, 1)$，然后沿 $y = 1$ 走一段，再在 $x \geq b$ 时沿双曲线 $y = \sqrt{1 + (x - b)^2}$ 向上走（任意 $b \geq \frac{1}{2}$ 皆可），便得到一条足够光滑（斜率连续）而且处处满足方程的曲线。上面四条曲线是经过给定点的“基本”解；经过这些点的其他每一个解，都是由这些曲线、$y = 1$，以及其他与 $y = 1$ 相切的圆和双曲线的片段拼合而成。

略图



半圆（蓝绿色）的圆心为 $(\pm\frac{1}{2}, 0)$，半径为 1，在 $x = \pm\frac{1}{2} \pm 1$ 处垂直地与 x 轴相遇，并在 $x = \pm\frac{1}{2}$ 处与 $y = 1$ 相切。双曲线的分支（琥珀色）在同样的两点与 $y = 1$ 相切，其渐近线（asymptote）为 $y = \pm(x \mp \frac{1}{2})$，即它们的增长与 $|x|$ 相若。每一对曲线都关于 y 轴对称，而每对的两条曲线在该轴上的给定点相交。

验算

- 在 $a = \frac{1}{2}$ 的圆上的点 $(0, \frac{1}{2}\sqrt{3})$：$y' = -(x-a)/y = 1/\sqrt{3}$，所以左边是 $\frac{9}{16} \cdot \frac{1}{9} = \frac{1}{16}$，而右边是 $(\frac{3}{4} - 1)^2 = \frac{1}{16}$。
- 由 $(0, \frac{1}{2}\sqrt{5})$ 出发，取正号对 $y' = \pm\sqrt{|y^2 - 1|}/y$ 作数值积分，得 $y(1.5) = 2.2360680$，与 $a = -\frac{1}{2}$ 的双曲线上的 $\sqrt{1 + 2^2} = \sqrt{5}$ 吻合。

常见错误

- 在两个区域都把平方根取成 $(u')^2 = 4u$；对 $|y| < 1$ 而言，这没有实数解，于是那些圆就遗漏了。
- 在每一步只保留平方根的其中一个正负号，结果经过每个点只找到一条曲线，而不是两条。
- 忘记了 $y = \pm 1$ 并不属于任何一族，尽管它们与每一个成员都相切。

STEP 之外

直线 $y = \pm 1$ 是奇解 (singular solution)：它们是通解 (general solution) 的包络线，无论常数 a 取甚么值都得不出来。它们恰恰出现在一阶方程的唯一性定理（皮卡–林德勒夫定理, Picard–Lindelöf) 失效之处，因为 $\sqrt{|y^2-1|}$ 在 $y = \pm 1$ 处不满足利普希茨 (Lipschitz) 条件，这正是解可以在那里分岔的原因。这类包络线常见于对 y' 而言并非线性的一阶方程（克莱罗（Clairaut）方程是经典例子）；Boyce 与 DiPrima 的 *Elementary Differential Equations*（初等微分方程教科书）讨论了存在性与唯一性，以及它们在甚么情况下失效。

详细解答

STEP III 2002，第 7 题

纯数学 · 坐标几何恒等式

STEP III 2002，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/02-s3-q7

关键思路

$s^2+b^2-a^2$ 这个式子应该似曾相识：在三角形 OAB 中（其中 $OA=s$、$OB=b$）使用余弦定理（cosine rule），得出的正是这个式子。因此，第一个恒等式（identity）其实就是开首关于 $\cos^2\alpha+\cos^2\beta$ 的结果，应用于 OB 与两条坐标轴所成的两个角。其后，第二个恒等式纯粹是建基于第一个恒等式的代数运算，而不等式（inequality）亦随之而来，因为两个平方之和必不小于其中任何一个平方。

解答

关于角的结果

首先假设 $\alpha+\beta=\frac{1}{2}\pi$。则 $\cos\beta=\cos(\frac{1}{2}\pi-\alpha)=\sin\alpha$，所以

$$\cos^2\alpha+\cos^2\beta=\cos^2\alpha+\sin^2\alpha=1.$$

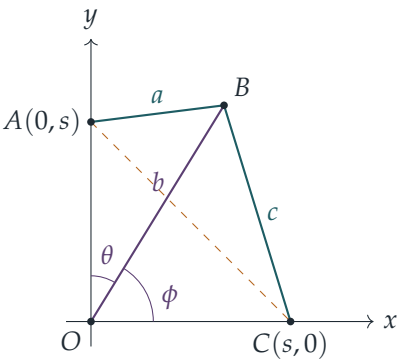
反过来，假设 $\cos^2\alpha+\cos^2\beta=1$。则 $\cos^2\beta=1-\cos^2\alpha=\sin^2\alpha$。$\alpha$ 和 β 都是锐角（acute angle），所以 $\cos\beta>0$ 且 $\sin\alpha>0$，取正平方根便得

$$\cos\beta=\sin\alpha=\cos(\tfrac{1}{2}\pi-\alpha).$$

现在 β 和 $\frac{1}{2}\pi-\alpha$ 都在 $(0,\frac{1}{2}\pi)$ 之内，而余弦在这个区间上严格递减（strictly decreasing），因此每个值只取一次。故此 $\beta=\frac{1}{2}\pi-\alpha$，即 $\alpha+\beta=\frac{1}{2}\pi$。这就证明了“当且仅当”（if and only if）。

第一个恒等式

设 $\theta=\angle AOB$ 及 $\phi=\angle BOC$。



在三角形 OAB（边长 $OA = s$、$OB = b$、$AB = a$）和三角形 OCB（边长 $OC = s$、$OB = b$、$CB = c$）中使用余弦定理，得

$$\cos\theta = \frac{s^2 + b^2 - a^2}{2sb}, \qquad \cos\phi = \frac{s^2 + b^2 - c^2}{2sb}.$$

由于 B 位于开的第一象限（open first quadrant）内，射线 OB 严格地夹在正 y 轴（沿 OA 方向）与正 x 轴（沿 OC 方向）之间。所以 θ 和 ϕ 都是锐角，且 $\theta + \phi = \frac{1}{2}\pi$。根据上面“仅当”的方向，$\cos^2\theta + \cos^2\phi = 1$，即

$$\frac{(s^2 + b^2 - a^2)^2}{4s^2b^2} + \frac{(s^2 + b^2 - c^2)^2}{4s^2b^2} = 1.$$

两边乘以 $4s^2b^2$，便得第一个恒等式：

$$(s^2 + b^2 - a^2)^2 + (s^2 + b^2 - c^2)^2 = 4s^2b^2.$$

（若要直接验证，可设 $B = (x, y)$。则 $a^2 = x^2 + (y - s)^2$、$b^2 = x^2 + y^2$ 及 $c^2 = (x - s)^2 + y^2$，所以 $s^2 + b^2 - a^2 = 2sy$ 及 $s^2 + b^2 - c^2 = 2sx$，而左方等于 $4s^2(x^2 + y^2) = 4s^2b^2$。）

第二个恒等式

为了令代数运算保持整洁，设

$$P = s^2 + b^2 - a^2, \qquad Q = s^2 + b^2 - c^2, \qquad S = s^2 + b^2, \qquad D = s^2 - b^2,$$

于是 $a^2 = S - P$ 及 $c^2 = S - Q$。第一个恒等式说的是 $P^2 + Q^2 = 4s^2b^2$。由于 $4s^2b^2 = (s^2 + b^2)^2 - (s^2 - b^2)^2 = S^2 - D^2$，它可以改写成

$$D^2 = S^2 - P^2 - Q^2. \tag{†}$$

现在 $a^2 + c^2 = 2S - P - Q$，而 $2s^2 = S + D$、$2b^2 = S - D$，所以

$$2s^2 - a^2 - c^2 = (P + Q - S) + D, \qquad 2b^2 - a^2 - c^2 = (P + Q - S) - D.$$

? **想一想** 为甚么值得把它们拆成 $(P+Q-S)\pm D$ 的形式?

因为 $(u+D)^2+(u-D)^2=2u^2+2D^2$：交叉项互相抵销，于是 D 只以 D^2 的形式出现，而 (†) 让我们可以把它换成只涉及 P、Q 和 S 的式子。取 $u=P+Q-S$，

$$\begin{aligned}&(2s^2-a^2-c^2)^2+(2b^2-a^2-c^2)^2\\&\quad=2(P+Q-S)^2+2D^2\\&\quad=2(P^2+Q^2+S^2+2PQ-2PS-2QS)+2(S^2-P^2-Q^2)\\&\quad=4S^2-4PS-4QS+4PQ\\&\quad=4(S-P)(S-Q)\\&\quad=4a^2c^2,\end{aligned}$$

其中第二行用了 (†)。这就是第二个恒等式。

不等式

第二个恒等式左方的两个平方都是非负的，所以

$$(2b^2-a^2-c^2)^2\le 4a^2c^2,\qquad \text{所以}\qquad |2b^2-a^2-c^2|\le 2ac,$$

这里取了非负平方根（注意 $a,c>0$）。这即是说

$$-2ac\le 2b^2-a^2-c^2\le 2ac,$$

亦即

$$a^2-2ac+c^2\le 2b^2\le a^2+2ac+c^2,$$

也就是 $(a-c)^2\le 2b^2\le(a+c)^2$，正是所求。

等号恰好在 $2s^2=a^2+c^2$ 时成立。由于 $AC^2=2s^2$，根据毕氏定理的逆定理（converse of Pythagoras），这表示 $\angle ABC$ 是直角：B 位于以 AC 为直径的圆上。

验算

- 取 $s=2$ 及 $B=(1,3)$。则 $a^2=2$、$b^2=c^2=10$。第一个恒等式成为 $12^2+4^2=160=4\cdot4\cdot10$；第二个恒等式成为 $(-4)^2+8^2=80=4\cdot2\cdot10$；而不等式则是 $3.06\le 20\le 20.94$。
- 点 $B=(s,s)$ 是正方形的第四个顶点，所以 $\angle ABC$ 是直角，且 $a=c=s$、$b=s\sqrt{2}$：于是 $2b^2=4s^2=(a+c)^2$，正是等号成立的情况。
- 在 2×10^5 个点 B 上对全部三个结果作随机测试，没有发现任何不成立的情况。

常见错误

- 在逆命题中漏掉锐角这个条件：没有这个条件，$\cos^2\beta = \sin^2\alpha$ 并不能推出 $\beta = \frac{1}{2}\pi - \alpha$（例如 $\beta = \frac{1}{2}\pi + \alpha$ 同样满足）。
- 不利用第一个恒等式，而以 a, b, c, s 硬生生地展开第二个恒等式。这样做是可行的，但很容易漏掉某一项；采用 S、D 的代换，可令运算不致失控。
- 对 $(2b^2 - a^2 - c^2)^2 \leq 4a^2c^2$ 取平方根时漏掉绝对值（modulus），结果只得到不等式的其中一半。

STEP 之外

$2b^2 - a^2 - c^2$ 这个量等于 $2s(x + y - s)$，即三角形 ABC 面积的 ± 4 倍，所以第二个恒等式其实是把余弦定理和面积公式 $\frac{1}{2}ac\sin B$ 合并成 $\cos^2 B + \sin^2 B = 1$。上界亦是托勒密不等式（Ptolemy's inequality）的一个特例：对任意四点，$OB \cdot AC \leq OA \cdot BC + OC \cdot AB$，而等号在圆内接四边形（cyclic quadrilateral）的情况下成立，在这里即是 $bs\sqrt{2} \leq s(a + c)$。这类距离之间的关系，可以引伸至度量空间（metric space）和距离几何（distance geometry）的研究；Siklos 的 *Advanced Problems in Mathematics*（高等数学难题集）中有更多同类风格的题目。

详细解答

STEP III 2002，第 8 题

纯数学 · 托勒密定理

> **STEP III 2002，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/02-s3-q8

关键思路

单位圆上两点之差，只要抽出“平均”角，便可漂亮地因式分解：$e^{i\alpha} - e^{i\beta} = e^{i(\alpha+\beta)/2}\left(e^{i(\alpha-\beta)/2} - e^{-i(\alpha-\beta)/2}\right)$，而括号内正是 $2i$ 乘以一个正弦。这样一来，每条弦（chord）的模（modulus）和辐角（argument）便同时求得。至于最后一部分，两个辐角相同的复数指向同一方向，所以它们的模直接相加；余下只需一条把三个乘积连系起来的代数恒等式。

解答

$u_1 - u_2$ 的辐角

设 $u_k = e^{i\theta_k}$。抽出因子 $e^{i(\theta_1+\theta_2)/2}$，

$$\begin{aligned} u_1 - u_2 &= e^{i(\theta_1+\theta_2)/2}\left(e^{i(\theta_1-\theta_2)/2} - e^{-i(\theta_1-\theta_2)/2}\right) \\ &= e^{i(\theta_1+\theta_2)/2} \cdot 2i \sin\tfrac{1}{2}(\theta_1 - \theta_2) \\ &= 2\sin\tfrac{1}{2}(\theta_2 - \theta_1) \cdot (-i)\, e^{i(\theta_1+\theta_2)/2} \\ &= 2\sin\tfrac{1}{2}(\theta_2 - \theta_1)\, e^{i(\theta_1+\theta_2-\pi)/2}, \end{aligned}$$

最后一行用了 $-i = e^{-i\pi/2}$。

> ? **想一想** 为甚么可以从这条式子直接读出辐角？又为甚么正弦要写成 $\sin\frac{1}{2}(\theta_2 - \theta_1)$ 而不是 $\sin\frac{1}{2}(\theta_1 - \theta_2)$？

形如 $re^{i\phi}$ 的数，只有在 $r > 0$ 时，其辐角才是 ϕ（可相差 2π 的倍数）。这里 $0 < \theta_2 - \theta_1 < 2\pi$，所以 $0 < \frac{1}{2}(\theta_2 - \theta_1) < \pi$，正弦严格为正。若反过来写，正弦便是负数，辐角就会相差 π。

因此 $|u_1-u_2| = 2\sin\frac{1}{2}(\theta_2-\theta_1)$，而 $\frac{1}{2}(\theta_1+\theta_2-\pi)$ 是 u_1-u_2 的一个辐角。取 arg 为 $(-\pi,\pi]$ 内的主值（principal value）：由于 $-\pi<\theta_1<\theta_2<\pi$，

$$-\tfrac{3\pi}{2} < \tfrac{1}{2}(\theta_1+\theta_2-\pi) < \tfrac{\pi}{2}.$$

若此值大于 $-\pi$，它本身已是主辐角（$n=0$）；否则它落在 $(-\frac{3\pi}{2},-\pi]$ 内，加上 2π 便回到范围之内（$n=1$）。故此

$$\arg(u_1-u_2) = \tfrac{1}{2}(\theta_1+\theta_2-\pi)+2n\pi, \qquad n=0 \text{ 或} 1,$$

正如题目所求。

两个乘积的辐角相同

以上推导除了用到 $\theta_1<\theta_2$ 且两者皆在 $(-\pi,\pi)$ 内之外，并没有用到 u_1,u_2 的任何其他性质。所以对任意 $j<k$，

$$\begin{aligned} u_j-u_k &= 2\sin\tfrac{1}{2}(\theta_k-\theta_j)\,\mathrm{e}^{\mathrm{i}(\theta_j+\theta_k-\pi)/2},\\ u_k-u_j &= -(u_j-u_k) = 2\sin\tfrac{1}{2}(\theta_k-\theta_j)\,\mathrm{e}^{\mathrm{i}(\theta_j+\theta_k+\pi)/2}, \end{aligned}$$

其中用了 $-1=\mathrm{e}^{\mathrm{i}\pi}$，而两式前面的实数因子都是正数。对 $j<k$，记 $s_{jk}=\sin\frac{1}{2}(\theta_k-\theta_j)>0$，并记 $\Sigma=\theta_1+\theta_2+\theta_3+\theta_4$。相乘（辐角相加），

$$\begin{aligned} (u_1-u_2)(u_4-u_3) &= 4s_{12}s_{34}\,\mathrm{e}^{\mathrm{i}[(\theta_1+\theta_2-\pi)+(\theta_3+\theta_4+\pi)]/2} = 4s_{12}s_{34}\,\mathrm{e}^{\mathrm{i}\Sigma/2},\\ (u_1-u_4)(u_3-u_2) &= 4s_{14}s_{23}\,\mathrm{e}^{\mathrm{i}[(\theta_1+\theta_4-\pi)+(\theta_2+\theta_3+\pi)]/2} = 4s_{14}s_{23}\,\mathrm{e}^{\mathrm{i}\Sigma/2}. \end{aligned}$$

两者都是 $\mathrm{e}^{\mathrm{i}\Sigma/2}$ 的正数倍，所以各自的辐角都是 $\frac{1}{2}\Sigma$（可相差 2π 的倍数）。因此它们的主辐角只可能相差 2π 的倍数：

$$\arg\big((u_1-u_2)(u_4-u_3)\big) = \arg\big((u_1-u_4)(u_3-u_2)\big)+2n\pi$$

其中 n 为某个整数。（事实上 $n=0$，因为两个主辐角若相差 2π 的倍数，就必定相等；但题目只要求较弱的这个结论。）

最后一部分

把三个乘积记为

$$A=(u_1-u_2)(u_4-u_3), \qquad B=(u_1-u_4)(u_3-u_2), \qquad C=(u_1-u_3)(u_4-u_2).$$

展开得

$$\begin{aligned} A+B &= (u_1u_4-u_1u_3-u_2u_4+u_2u_3)+(u_1u_3-u_1u_2-u_3u_4+u_2u_4)\\ &= u_1u_4+u_2u_3-u_1u_2-u_3u_4,\\ C &= u_1u_4-u_1u_2-u_3u_4+u_2u_3, \end{aligned}$$

所以 $A+B=C$。这纯粹是代数运算，对任意四个复数都成立。

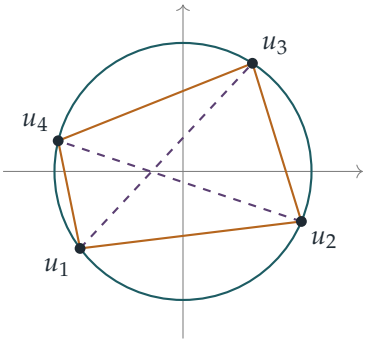
现在利用上一部分的结果：$A = a\,\mathrm{e}^{\mathrm{i}\Sigma/2}$ 及 $B = b\,\mathrm{e}^{\mathrm{i}\Sigma/2}$，其中 $a = 4s_{12}s_{34} > 0$，$b = 4s_{14}s_{23} > 0$。于是 $|A| = a$，$|B| = b$，并且

$$|C| = |A+B| = \left|(a+b)\,\mathrm{e}^{\mathrm{i}\Sigma/2}\right| = a+b = |A|+|B|,$$

因为 $a+b>0$ 而 $|\mathrm{e}^{\mathrm{i}\Sigma/2}| = 1$。即是说，

$$|(u_1-u_2)(u_4-u_3)| + |(u_1-u_4)(u_3-u_2)| = |(u_1-u_3)(u_4-u_2)|.$$

重点在于：三角不等式（triangle inequality）$|A+B| \le |A|+|B|$ 恰好在 A 与 B 指向同一方向时变成等式，而这正是中间部分所证明的。



从几何角度看，$|u_j-u_k|$ 是一条弦的长度；又因为辐角递增，$u_1u_2u_3u_4$ 是一个顶点依次排列的圆内接四边形（quadrilateral inscribed in the circle）。这条恒等式说明：两组对边之积的和，等于两条对角线（虚线）之积。这就是托勒密定理（Ptolemy's theorem）。

验算

- 正方形：$\theta_k = -\frac{3\pi}{4}, -\frac{\pi}{4}, \frac{\pi}{4}, \frac{3\pi}{4}$。每条边长 $\sqrt{2}$，每条对角线长 2，而确实 $2+2=4$。
- 取 $\theta = (-2.5, -0.4, 1.0, 2.9)$，得 $|A| \approx 2.8223$，$|B| \approx 1.1013$ 及 $|C| \approx 3.9236 = |A|+|B|$。
- 随机选取 2×10^5 组依次排列的 θ_k，$\arg(u_1-u_2)$ 的公式每次都以 $n=0$ 或 1 成立（两个值都有出现），而恒等式的误差不超过 2×10^{-15}。

常见错误

- 写出 $u_1-u_2 = 2\mathrm{i}\sin\frac{1}{2}(\theta_1-\theta_2)\,\mathrm{e}^{\mathrm{i}(\theta_1+\theta_2)/2}$ 后，把辐角读成 $\frac{1}{2}(\theta_1+\theta_2+\pi)$，忘记了正弦是负数。

- 把 $u_4 - u_3$ 当作与 $u_3 - u_4$ 有相同的辐角：把差的次序倒转，辐角会增加 π。
- 断言 $|A| + |B| = |A + B|$ 而不说明原因；这需要 A 与 B 的辐角相同。

STEP 之外

托勒密定理有逆定理（converse），也有更一般的形式：对平面上任意四点，$|A| + |B| \geq |C|$（只是把三角不等式应用于 $A + B = C$），等号恰好在四点依次位于同一个圆或同一条直线上时成立。这背后是交比（cross-ratio）$\frac{(u_1-u_2)(u_3-u_4)}{(u_1-u_4)(u_3-u_2)}$：它恰好在四点共圆（concyclic）时为实数，并且在莫比乌斯变换（Möbius transformations）下保持不变，是大学一年级复分析（complex analysis）与几何的核心对象。Needham 的 *Visual Complex Analysis*（《复分析：可视化方法》）以精彩的图画阐述了这些想法。

详细解答

STEP III 2002，第 9 题

力学 • 质心与倾倒

> **STEP III 2002，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/02-s3-q9

关键思路

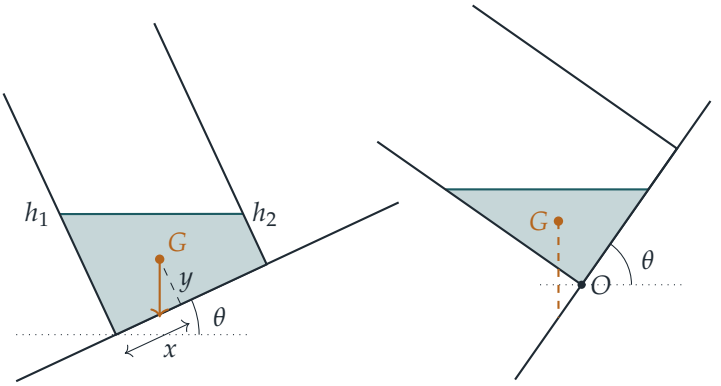
流体的自由表面（free surface）是水平的，所以在容器的参考系（容器倾斜了 θ 角）中，它是一条与底面成 θ 角的斜直线。因此，若液面与两边侧壁都相交，流体的横截面就是梯形（trapezium）；若液面未到达上方侧壁便先碰到底面，横截面就是直角三角形。这样，求质心（centre of mass）便成为一个二维的形心（centroid）计算。至于倾倒（toppling），由于容器是轻的，一切取决于通过流体质心的铅垂线与斜面相交于何处。

解答

建立模型

我们把容器看成一个轻的棱柱体，器壁厚度可忽略；把流体看成相对于容器静止的均匀连续体；并假设斜面足够粗糙，容器不会滑动。垂直于底面水平边的每一个铅垂横截面都相同，而流体是均匀的，所以流体的质心就是其中一个横截面的形心。该横截面的宽度为 a，面积为 $ka^3/a = ka^2$。

沿底面由下方侧壁起（沿斜面向上）量度 X，并垂直于底面量度 Y。随着 X 增加，底面以 θ 角相对于水平面上升，所以水平的液面相对于底面以 θ 角下降：X 每增加一个单位，液面离底面的高度便减少 $\tan\theta$。记 $d = a\tan\theta$ 为整个宽度上的总下降量。



左图显示 $k > \frac{1}{2}\tan\theta$ 的情况（流体接触两边侧壁）；右图显示 $k < \frac{1}{2}\tan\theta$ 且 $\theta > 45°$ 的情况，此时通过 G 的铅垂线落在底面之外。

$k > \frac{1}{2}\tan\theta$ 的情况

设下方侧壁和上方侧壁处的深度分别为 h_1 和 h_2。那么 $h_1 - h_2 = d$，而梯形的面积为 $\frac{1}{2}a(h_1 + h_2) = ka^2$，所以 $h_1 + h_2 = 2ka$。因此

$$h_2 = ka - \tfrac{1}{2}d = a\big(k - \tfrac{1}{2}\tan\theta\big),$$

它为正数当且仅当 $k > \frac{1}{2}\tan\theta$。所以这正是液面与两边侧壁都相交的情况。

把梯形分割成一个 $a \times h_2$ 的长方形，以及叠在其上的一个直角三角形，其两条直角边分别为 a（沿 $Y = h_2$）和 d（沿下方侧壁向上）。

- 长方形：面积 ah_2，形心 $\big(\frac{1}{2}a,\ \frac{1}{2}h_2\big)$。
- 三角形：顶点为 $(0, h_2)$、(a, h_2)、$(0, h_2 + d)$，所以面积为 $\frac{1}{2}ad$，形心（各顶点的平均）为 $\big(\frac{1}{3}a,\ h_2 + \frac{1}{3}d\big)$。

对下方侧壁取矩（taking moments），总面积为 ka^2，

$$ka^2\,x = ah_2 \cdot \tfrac{1}{2}a + \tfrac{1}{2}ad \cdot \tfrac{1}{3}a \quad\Longrightarrow\quad kx = \tfrac{1}{2}h_2 + \tfrac{1}{6}d = \tfrac{1}{2}ka - \tfrac{1}{4}d + \tfrac{1}{6}d = \tfrac{1}{2}ka - \tfrac{1}{12}d.$$

除以 ka 并利用 $d = a\tan\theta$，

$$\frac{x}{a} = \frac{1}{2} - \frac{\tan\theta}{12k}.$$

对底面取矩，

$$ka^2\,y = ah_2 \cdot \tfrac{1}{2}h_2 + \tfrac{1}{2}ad\big(h_2 + \tfrac{1}{3}d\big),$$
$$ka\,y = \tfrac{1}{2}h_2^2 + \tfrac{1}{2}dh_2 + \tfrac{1}{6}d^2 = \tfrac{1}{2}h_2(h_2 + d) + \tfrac{1}{6}d^2.$$

现在 $h_2 = ka - \frac{1}{2}d$ 而 $h_2 + d = ka + \frac{1}{2}d$，所以 $h_2(h_2 + d) = k^2a^2 - \frac{1}{4}d^2$，于是

$$ka\,y = \tfrac{1}{2}k^2a^2 - \tfrac{1}{8}d^2 + \tfrac{1}{6}d^2 = \tfrac{1}{2}k^2a^2 + \tfrac{1}{24}d^2.$$

除以 ka^2，

$$\frac{y}{a} = \frac{k}{2} + \frac{\tan^2\theta}{24k},$$

正如所求。

$k < \frac{1}{2}\tan\theta$ 的情况

这时 h_2 会是负数，所以液面在距离下方侧壁 $b < a$ 的某处与底面相交，而流体的横截面是一个直角三角形，其两条直角边分别为沿底面的 b 和沿下方侧壁向上的 h。由液面的斜率得 $h = b\tan\theta$，由面积得 $\frac{1}{2}bh = ka^2$。所以 $b^2\tan\theta = 2ka^2$，于是

$$b = a\sqrt{\frac{2k}{\tan\theta}}, \qquad h = a\sqrt{2k\tan\theta}.$$

（事实上，$b < a$ 正是因为 $2k < \tan\theta$；容器很高，所以 h 不会造成问题。）三角形的形心位于每条直角边的三分之一处，所以

$$\frac{x}{a} = \frac{1}{3}\sqrt{\frac{2k}{\tan\theta}} = \frac{1}{3}\sqrt{2k\cot\theta}, \qquad \frac{y}{a} = \frac{1}{3}\sqrt{2k\tan\theta}.$$

当 $k = \frac{1}{2}\tan\theta$ 时，两组公式都给出 $x/a = \frac{1}{3}$ 和 $y/a = \frac{1}{3}\tan\theta$，所以两种情况互相衔接。

倾倒

> **想一想** 容器在放手的一刻会否翻倒，取决于甚么？

容器是轻的，所以唯一的重量就是流体的重量，它铅垂地作用于流体的质心 G。斜面只能对底面施力，所以法向反作用力（normal reaction）作用于底面上某处，即下方边缘 O 与上方边缘之间。假如通过 G 的铅垂线与斜面相交于 O 的下坡一侧，重量对 O 便有一个力矩（moment），而这样的反作用力无法把它平衡，于是容器绕 O 转动：它会倾倒。

取 O 为原点，x 沿斜面向上，y 垂直于斜面。在这组坐标轴中，铅垂向下的方向的分量为 $(-\sin\theta, -\cos\theta)$，所以由 $G = (x, y)$ 沿铅垂线向下走一段距离 $y\sec\theta$，便到达斜面上的

$$x - y\tan\theta.$$

所以若 $x < y\tan\theta$，容器便会倾倒。

假设 $k < \frac{1}{2}$ 且 $\theta > 45°$。那么 $\tan\theta > 1 > 2k$，所以 $k < \frac{1}{2}\tan\theta$，我们处于三角形的情况。此时

$$y\tan\theta - x = \frac{a}{3}\left(\sqrt{2k\tan\theta}\,\tan\theta - \sqrt{\frac{2k}{\tan\theta}}\right) = \frac{a}{3}\sqrt{\frac{2k}{\tan\theta}}\left(\tan^2\theta - 1\right) > 0,$$

因为 $\tan\theta > 1$。因此 $x < y\tan\theta$：通过 G 的铅垂线落在下方边缘之下，无论 $k < \frac{1}{2}$ 取甚么值，容器都会倾倒。（在三角形的情况中，同样的计算显示若 $\theta < 45°$ 则容器是稳定

的，所以 45° 正是该情况下的临界值。）这就是所要的结果。条件 $k < \frac{1}{2}$ 的作用，是保证当 $\theta > 45^\circ$ 时必定处于三角形的情况。

验算

- 当 $\theta = 0$ 时，第一种情况给出 $x = \frac{1}{2}a$ 和 $y = \frac{1}{2}ka$，即一块深度为 ka 的流体的中心。
- 两种情况在 $k = \frac{1}{2}\tan\theta$ 时吻合，此时梯形退化成三角形。
- 数值验算：对 300 组随机的 (k, θ) 直接对流体区域进行积分，结果与两组公式的误差都在 1.4×10^{-6} 之内（取 $a = 1$）。例如，$\theta = 55^\circ$、$k = 0.3$ 给出 $x \approx 0.216a$、$y \approx 0.309a$，而 $y\tan\theta \approx 0.441a > x$，所以容器会倾倒。

常见错误

- 把液面画成与底面平行，或者斜向错误的一边：流体是堆向下方侧壁的。
- 忘记容器是轻的，而为容器另加一个重量：整体的质心就是流体的质心。
- 把通过 G 的铅垂线与底面的中心比较，而不是与下方边缘 O 比较。

STEP 之外

“通过质心的铅垂线是否与底面相交?”这个判别法，是稳定性分析（stability analysis）中最简单的情形。浮在水面的船只也是用同样的方法处理，不过在那里浮力（buoyancy）作用于被排开的水的质心，而这个点会随船身倾侧而移动；由此便引出定倾中心（metacentre）的概念。流动中的流体（放手后的晃荡）则是流体动力学（fluid dynamics）中困难得多的问题。Kleppner 与 Kolenkow 的 *An Introduction to Mechanics*（力学入门）对刚体平衡（rigid-body equilibrium）及绕轴转动有细致的论述。

详细解答

STEP III 2002，第 10 题

力学 • 涉及摩擦力的转动

STEP III 2002，第 10 题

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

step.maths.org/questions/02-s3-q10

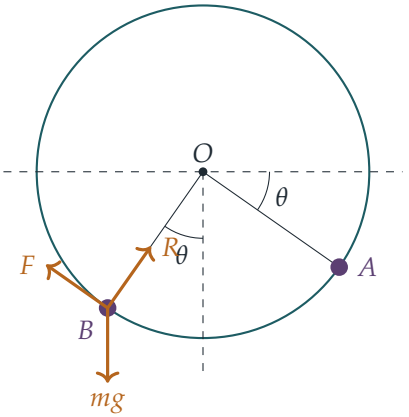
关键思路

只要第二个质点（particle）不滑动，它便随圆柱一起运动，于是整个系统是一个绕固定轴转动的刚体（rigid body）：两个相等的质量位于半径 a 处，彼此相隔四分之一圈。这样，用能量（或对转轴取力矩（moments））一行便可得出运动方程。要找出何时开始滑动，我们单独考虑那个没有固定的质点，沿表面方向和垂直于表面方向应用牛顿第二定律（Newton's second law），求出它所需的摩擦力（friction）F 和法向反作用力（normal reaction）R。当 F 首次达到 μR 时，它便开始滑动。

解答

建立模型

把两个质点都视为轻圆柱上的点质量，轮轴光滑。设 O 为转轴，A 为固定的质点，B 为没有固定的质点。系统由静止释放后，A 下落，圆柱因而转动（在图中为顺时针方向），并把 B 带上远离 A 的一侧。转过角度 θ 后，OA 在水平线以下成角 θ，而 OB 在另一侧与向下的铅垂线成角 θ。



运动方程

只要 B 不滑动，两个质点都在半径为 a 的圆上以角速率（angular speed）$\dot{\theta}$ 运动，而圆柱本身没有质量。取 O 所在的水平面为势能（potential energy）零点，则 A 的高度为 $-a\sin\theta$，B 的高度为 $-a\cos\theta$。系统在 $\theta = 0$ 时由静止开始运动，光滑轮轴不作功，而 B 与圆柱之间没有相对运动，所以摩擦力所作的净功为零。因此能量守恒：

$$2\cdot\tfrac{1}{2}ma^2\dot{\theta}^2 - mga\sin\theta - mga\cos\theta = -mga,$$

即

$$\dot{\theta}^2 = \frac{g}{a}(\sin\theta + \cos\theta - 1). \qquad (*)$$

对时间求导，再除以 $2\dot{\theta}$（运动一旦开始，它便不等于零），得

$$\ddot{\theta} = \frac{g}{2a}(\cos\theta - \sin\theta),$$

这正是所要求的结果。另一个等价的做法是对 O 取力矩：A 的重量沿转动方向的力矩为 $mga\cos\theta$，B 的重量逆转动方向的力矩为 $mga\sin\theta$，而转动惯量（moment of inertia）为 $2ma^2$。

B 何时滑动？

现在单独考虑 B。作用于它的力有：重量 mg、指向 O 的法向反作用力 R，以及沿表面的摩擦力 F；我们取摩擦力沿 B 的运动方向（即沿侧面向上）为正。B 在半径为 a 的圆上运动，所以它有指向 O 的加速度 $a\dot{\theta}^2$ 和沿运动方向的加速度 $a\ddot{\theta}$。重量有一个大小为 $mg\cos\theta$、方向正背离 O 的分量，以及一个大小为 $mg\sin\theta$、与运动方向相反的分量。因此

$$\text{指向}O:\quad R - mg\cos\theta = ma\dot{\theta}^2,$$
$$\text{沿运动方向:}\quad F - mg\sin\theta = ma\ddot{\theta}.$$

利用 $(*)$ 和运动方程，

$$R = mg\cos\theta + mg(\sin\theta + \cos\theta - 1) = mg\,(2\cos\theta + \sin\theta - 1),$$
$$F = mg\sin\theta + \tfrac{1}{2}mg(\cos\theta - \sin\theta) = \tfrac{1}{2}mg\,(\sin\theta + \cos\theta).$$

当 $0 \le \theta \le \frac{\pi}{3}$ 时，两者皆为正。（$R > 0$ 是因为当 $0 < \theta \le \frac{\pi}{3}$ 时，利用 $2\cos\theta \ge 1$ 可得 $2\cos\theta + \sin\theta - 1 \ge \sin\theta > 0$，而在 $\theta = 0$ 时 $R = mg$。）所以 B 保持接触，而摩擦力把它沿侧面向上推。只要 $F \le \mu R$，质点便不会滑动，也就是说，只要

$$\rho(\theta) = \frac{F}{R} = \frac{\sin\theta + \cos\theta}{2(2\cos\theta + \sin\theta - 1)} \le \mu.$$

在 $\theta = \frac{\pi}{3}$ 时，$\sin\theta = \frac{\sqrt{3}}{2}$ 而 $\cos\theta = \frac{1}{2}$，于是

$$\rho\left(\tfrac{\pi}{3}\right) = \frac{\frac{1}{2}(\sqrt{3}+1)}{2\cdot\frac{\sqrt{3}}{2}} = \frac{\sqrt{3}+1}{2\sqrt{3}} = \frac{3+\sqrt{3}}{6} = \mu.$$

想一想 为甚么质点不可能在此之前已经开始滑动？

我们需要对所有 $0 \le \theta < \frac{\pi}{3}$ 都有 $\rho(\theta) < \mu$。记 $N = \sin\theta + \cos\theta$ 和 $D = 2\cos\theta + \sin\theta - 1$。则

$$\begin{aligned} N'D - ND' &= (\cos\theta - \sin\theta)(2\cos\theta + \sin\theta - 1) - (\sin\theta + \cos\theta)(\cos\theta - 2\sin\theta) \\ &= \cos^2\theta + \sin^2\theta + \sin\theta - \cos\theta = 1 + \sin\theta - \cos\theta, \end{aligned}$$

它在 $0 < \theta < \pi$ 时为正。由于在 $[0, \frac{\pi}{3}]$ 上 $D > 0$，故在该区间上 $\rho' = (N'D - ND')/(2D^2) > 0$，所以 ρ 由 $\rho(0) = \frac{1}{2}$ 起严格递增。因此它一直小于 μ，直至在 $\theta = \frac{\pi}{3}$ 时恰好达到 μ。此外，由 $(*)$ 可知，当 $0 < \theta < \frac{\pi}{2}$ 时 $\dot\theta^2 > 0$，所以圆柱确实会一直转动并到达这个角度。

所以，当圆柱转过 60° 时，第二个质点开始滑动，这正是所要求的结果。

验算

- 在 $\theta = 0$ 时：B 静止于最低点，但 $F = \frac{1}{2}mg \ne 0$。这是对的：A 的重量正试图转动圆柱，而摩擦力必须令 B 随之加速。事实上，此时 $F = ma\ddot\theta = \frac{1}{2}mg$，而 $R = mg$，因为 B 还未有速率。
- 在 $\theta = \frac{\pi}{4}$ 时，$\ddot\theta = 0$，角速率达到最大值；两个重量对 O 的力矩互相平衡时，正应如此。
- 以数值方法验证：对运动方程作数值积分，并直接由 B 的加速度计算作用于它的接触力，所得结果与 R 和 F 的公式吻合，而 F/R 首次达到 $(3+\sqrt{3})/6 \approx 0.7887$ 是在 60.00°。

常见错误

- 把圆柱的转动惯量也计算在内：圆柱是轻的，所以只有两个质点贡献 $2ma^2$。
- 忘记了 B 的切向加速度（tangential acceleration）$a\ddot{\theta}$，这会得出 $F = mg\sin\theta$ 和错误的角度。
- 只验证在 $60°$ 时 $F = \mu R$，却没有证明在此之前 F/R 较小。

STEP 之外

这是一个单自由度（one-degree-of-freedom）系统：一旦加上约束（constraint）条件（不滑动），单一个坐标 θ 便足以描述一切，而能量可直接给出运动。约束力是事后才求出的，而当约束所需的力超过表面所能提供的力时，约束便会失效。这与质点离开球面的问题是同一套逻辑，而在拉格朗日力学（Lagrangian mechanics）中，这套方法藉拉格朗日乘数（Lagrange multipliers）而变得有系统。Kleppner 与 Kolenkow 的 *An Introduction to Mechanics*（力学入门教科书）正是以这种精神阐述刚体转动和约束力。

详细解答

STEP III 2002，第 11 题

力学 • 与墙壁的斜碰撞

STEP III 2002，第 11 题

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

step.maths.org/questions/02-s3-q11

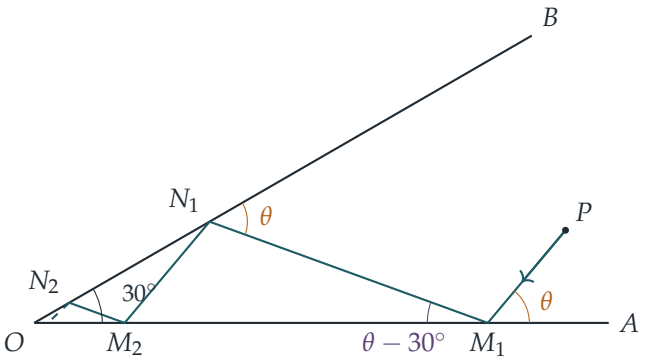
本题在 *Beyond STEP* 第 10.5 节有限时间内的无限次反弹中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch10/zeno-wedge.html

关键思路

在光滑（smooth）的墙上，冲量（impulse）垂直于墙。因此沿墙方向的速度分量保持不变，只有垂直方向的分量乘以 e。把这一点与三角形 OM_1N_1 的各个角结合起来，第一部分便化为一条三角方程。至于第二部分，每次反弹之后，运动都会以较小的规模重演一次。各段路程愈来愈短，但质点（particle）同时也在减慢，所以我们要相加的是时间，而不是距离。这个和是一个等比级数（geometric series）。

解答

我们把质点视为在光滑水平面上运动的一点，因此在两次碰撞之间，它以恒定速率沿直线运动。两面墙都是光滑的，而每次碰撞都遵从牛顿实验定律（Newton's experimental law），其系数为 e。



求 e 及 v_1/u_1

由于角 $PM_1A = \theta$，质点到达 M_1 时正朝着 O 运动，与墙成角 θ。碰撞前的一刻，它的速度分量为

$$u_1\cos\theta \text{ 沿墙方向（朝向 } O\text{）}, \qquad u_1\sin\theta \text{ 垂直于墙（指向墙内）}.$$

墙是光滑的，所以质点所受的冲量垂直于 OA。因此：

$$\begin{aligned}&\text{沿墙方向：}u_1\cos\theta\text{（不变）},\\ &\text{垂直方向：}e\,u_1\sin\theta\text{（方向逆转，并乘以 }e\text{）}.\end{aligned}$$

所以质点离开 M_1 时仍然朝着 O 运动，与墙成角 φ，其中

$$\tan\varphi = \frac{e\,u_1\sin\theta}{u_1\cos\theta} = e\tan\theta, \qquad v_1\cos\varphi = u_1\cos\theta. \tag{$*$}$$

接着我们利用三角形 OM_1N_1 的几何关系求 φ。O、N_1、B 三点共线，而角 $M_1N_1B = \theta$，所以角 $ON_1M_1 = 180° - \theta$。位于 O 的角是 $30°$，故

$$\varphi = \angle OM_1N_1 = 180° - 30° - (180° - \theta) = \theta - 30°.$$

要 φ 为正，必须有 $\theta > 30°$，这是本题的布局所隐含的假设。于是由 $(*)$ 得

$$\boxed{e = \frac{\tan(\theta - 30°)}{\tan\theta}}$$

以及

$$\boxed{\frac{v_1}{u_1} = \frac{\cos\theta}{\cos(\theta - 30°)} = \frac{2}{\sqrt{3} + \tan\theta}},$$

其中最后一个形式来自展开 $\cos(\theta - 30°) = \frac{\sqrt{3}}{2}\cos\theta + \frac{1}{2}\sin\theta$。你也可以写成 $e = \dfrac{\sqrt{3}\tan\theta - 1}{\tan\theta\,(\sqrt{3} + \tan\theta)}$。由于 $0 < \theta - 30° < \theta < 90°$，我们得到 $0 < e < 1$，符合题目的要求。

在有限时间内到达 O

质点到达 N_1 时正朝着 O 运动，与墙 OB 成角 θ（因为角 $M_1N_1B=\theta$）。这与在 M_1 的情况完全相同，只是两面墙的角色互换了。所以同样的计算仍然适用：它反弹后与 N_1O 成角 $\theta-30^\circ$，而速率则乘以

$$r=\frac{\cos\theta}{\cos(\theta-30^\circ)}.$$

想一想 质点到达 M_2 时是否又再成角 θ，使这个规律真的重复下去?

是的。在三角形 ON_1M_2 中，位于 O 的角是 30°，位于 N_1 的角是 $\theta-30^\circ$。所以角 $OM_2N_1=180^\circ-\theta$，即角 $N_1M_2A=\theta$。由归纳法（induction）可知，每次碰撞都是与墙成角 θ 发生的。每一段路程都以角 $\theta-30^\circ$ 离开墙。而每个三角形 OM_nN_n、ON_nM_{n+1} 的三个角都是 30°、$\theta-30^\circ$、$180^\circ-\theta$，所以它们全部相似（similar）。

长度。在三角形 OM_1N_1 中应用正弦定理（sine rule），

$$\frac{ON_1}{\sin(\theta-30^\circ)}=\frac{OM_1}{\sin(180^\circ-\theta)}=\frac{M_1N_1}{\sin 30^\circ}.$$

所以每个新三角形都是前一个三角形按以下比例缩放而成：

$$k=\frac{ON_1}{OM_1}=\frac{\sin(\theta-30^\circ)}{\sin\theta}.$$

把各段路程 M_1N_1，N_1M_2，$M_2N_2,\ldots$ 称为第 $1,2,3,\ldots$ 段。第 n 段的长度为 ℓ_1k^{n-1}，其中 $\ell_1=M_1N_1=\dfrac{OM_1}{2\sin\theta}$。

速率。质点以恒定速率 v_1r^{n-1} 走过第 n 段。

时间。因此第 n 段所需的时间为

$$\tau_n=\frac{\ell_1k^{n-1}}{v_1r^{n-1}}=\frac{\ell_1}{v_1}q^{n-1},\qquad q=\frac{k}{r}=\frac{\sin(\theta-30^\circ)\cos(\theta-30^\circ)}{\sin\theta\cos\theta}=\frac{\sin(2\theta-60^\circ)}{\sin 2\theta}.$$

想一想 各段路程按等比缩短，但质点同时也在减慢。为甚么单凭这一点还不足以下结论?

我们要相加的是时间，而时间之比 q 所比较的，正是路程的缩短（k）与质点的减慢（r）。我们需要 $q<1$。现在假设 $30^\circ<\theta<60^\circ$。那么

$$0^\circ<2\theta-60^\circ<60^\circ\ \Rightarrow\ \sin(2\theta-60^\circ)<\tfrac{\sqrt3}{2},\qquad 60^\circ<2\theta<120^\circ\ \Rightarrow\ \sin 2\theta>\tfrac{\sqrt3}{2},$$

所以 $0<q<1$。于是总时间是一个收敛（convergent）的等比级数：

$$T=\frac{PM_1}{u_1}+\sum_{n=1}^{\infty}\frac{\ell_1}{v_1}q^{n-1}=\frac{PM_1}{u_1}+\frac{\ell_1}{v_1(1-q)}.$$

这是有限的。要化简，可利用 $\sin 2\theta - \sin(2\theta - 60^\circ) = 2\cos(2\theta - 30^\circ)\sin 30^\circ$，由此得 $1 - q = \dfrac{\cos(2\theta - 30^\circ)}{\sin 2\theta}$。于是由 M_1 起计的时间为

$$\frac{\ell_1}{v_1(1-q)} = \frac{OM_1\cos(\theta - 30^\circ)}{u_1\cos(2\theta - 30^\circ)},$$

当 $30^\circ < \theta < 60^\circ$ 时，它是正数而且有限。

最后，质点确实会到达 O。第 n 次碰撞之后，它与 O 的距离是 $OM_1k^n \to 0$，因为 $0 < k < 1$。在走每一段路程期间，它与 O 的距离不会超过该段两个端点的距离中较大的一个。所以当 $t \to T$ 时，质点与 O 的距离趋向 0：它经过无限多次碰撞之后，在有限时间 T 到达 O。■

验算

- 当 $\theta = 45^\circ$：$e = \tan 15^\circ = 2 - \sqrt{3} \approx 0.268$，$v_1/u_1 = \sqrt{3} - 1 \approx 0.732$，而 $q = \sin 30^\circ / \sin 90^\circ = \frac{1}{2}$，所以每一段所需的时间都是前一段的一半。取 $OM_1 = u_1 = 1$，由 M_1 到 O 的时间是 $\cos 15^\circ / \cos 60^\circ = 2\cos 15^\circ \approx 1.932$。逐次模拟各次碰撞（六十次反弹）得到 1.931852，两者吻合。
- 当 $\theta = 60^\circ$：$q = 1$，所以每一段所需的时间都相同，总时间是无限的。当 $\theta > 60^\circ$ 时，$q > 1$。所以 60° 正好是分界线，而 T 的公式在该处因 $\cos(2\theta - 30^\circ) = \cos 90^\circ = 0$ 而趋向无限大。
- 当 $\theta \to 30^{\circ+}$ 时，$e \to 0$：质点几乎不反弹，差不多笔直地滑向 O，正如你所预期的。

常见错误

- 把三角形弄错：角 M_1N_1B 是位于 N_1 的外角（exterior angle），所以内角是 $180^\circ - \theta$，而在 M_1 的反弹角是 $\theta - 30^\circ$，不是 $\theta + 30^\circ$。
- 只因为距离 $OM_n \to 0$ 便断言质点到达 O。无限多段路程需要一个收敛的时间级数，而质点在每次碰撞时都会减慢。
- 在碰撞时把切向分量（tangential component）逆转。在光滑的墙上，只有法向分量（normal component）会改变。

STEP 之外

在有限时间内发生无限多次碰撞，称为芝诺现象（Zeno behaviour）。大家熟悉的例子是一个以 $e < 1$ 在地板上弹跳的球，而这里的楔形就扮演了地板的角色。研究“混合”系统

（hybrid systems，即把平滑运动与突然跳变混合起来的系统）的数学家，必须决定在芝诺时间之后会发生甚么事。在颗粒物料中，同样的效应称为非弹性坍缩（inelastic collapse），会在模拟中造成实际的麻烦。若想看以大学一年级程度仔细处理碰撞与冲量的论述，可参阅 Kleppner 与 Kolenkow 的 *An Introduction to Mechanics*（力学入门）。

详细解答

STEP III 2002，第 12 题

概率与统计 • 以条件化求期望值

> **STEP III 2002，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/02-s3-q12

关键思路

按游戏如何开始来取条件（conditioning）。每一局游戏的开头不外乎三种：H、TH 或 TT。在前两种情况下，并没有取得任何有用的进展（连续反面已被打断），所以游戏等于从头再来，而余下部分的期望长度仍然是 E。这种“更新”（renewal）给出一条关于 E 的线性方程，解之即可。同样的想法——按第一个正面何时出现来分类——也能处理连续 r 次反面的情况。

解答

全文中，p 是掷出正面的概率，$q = 1 - p$ 是掷出反面的概率，各次投掷互相独立（independent）。我们假设 $0 < p < 1$，这样游戏以概率 1 结束，而且下文所有期望值（expectation）都是有限的。

解释关于 E 的方程

考虑最初的一两次投掷。以下三种情况恰好有一种发生：

- **先掷出 H**（概率为 p）。已用去一次投掷，而且没有正在累积的连续反面。由于各次投掷互相独立，游戏余下的部分就是整个游戏的一个全新复本，期望长度为 E。期望总长度：$1 + E$。
- **先 T 后 H**（概率为 qp）。已用去两次投掷，而正面再次把连续反面清零，所以游戏重新开始。期望总长度：$2 + E$。
- **先 T 后 T**（概率为 q^2）。游戏恰好在 2 次投掷后结束。

这三种情况互斥（exclusive）而且穷尽（exhaustive）所有可能（它们的概率之和为 $p + qp + q^2 = p + q = 1$）。根据全期望定律（law of total expectation），

$$E = p(1 + E) + (1 - p)p(2 + E) + 2(1 - p)^2,$$

这正是题目给出的方程。

求 E

把含 E 的项合并：

$$E\,(1 - p - qp) = p + 2qp + 2q^2.$$

左边，$1 - p - qp = q - qp = q(1 - p) = q^2$。右边，$p + 2q(p + q) = p + 2q = 1 + q$。由于 $q \neq 0$，

$$E = \frac{1 + q}{q^2} = \frac{2 - p}{(1 - p)^2}.$$

期望奖金

设 W 为正面次数的期望值，亦即以英镑计的期望奖金。沿用同样的三种情况。掷出 H 之后，我们已赢得 £1，游戏重新开始；掷出 TH 之后，我们同样已赢得 £1，游戏重新开始；掷出 TT 之后，我们一无所获，游戏结束。所以

$$W = p(1 + W) + qp(1 + W) + q^2 \cdot 0.$$

? **想一想** 在 TH 的情况中，为甚么是 $1 + W$ 而不是 W?

在这种情况下，第二次投掷*确实*是正面，所以在游戏重新开始之前已经赚得 £1。这一题最容易出错的地方，就是漏掉了它。

合并各项，得 $W(1 - p - qp) = p(1 + q)$，所以 $Wq^2 = p(1 + q)$，而

$$W = \frac{p(1 + q)}{q^2} = \frac{p(2 - p)}{(1 - p)^2} \text{ 英镑}.$$

留意 $W = pE$：平均而言，所有投掷中有 p 这个比例是正面，即使游戏停止的时刻取决于投掷的结果。

连续 r 次反面的游戏

现在设 E 为游戏在连续出现 r 次反面时停止的情况下，投掷次数的期望值。按第一个正面来取条件。要么最初 r 次投掷全是反面，要么第一个正面出现在第 $j + 1$ 次投掷，其中 j 满足 $0 \leq j \leq r - 1$：

- j **次反面后一次正面**（概率为 $q^j p$）。这用去了 $j+1$ 次投掷，连续反面已被打断，游戏重新开始。期望总长度：$j+1+E$。
- r **次反面**（概率为 q^r）。游戏恰好在 r 次投掷后结束。

所以

$$E = \sum_{j=0}^{r-1} q^j p\,(j+1+E) + rq^r.$$

右边 E 的系数是 $p\sum_{j=0}^{r-1} q^j = p\cdot\dfrac{1-q^r}{1-q} = 1-q^r$，因为 $1-q=p$。把它移到左边，

$$q^r E = p\sum_{j=0}^{r-1}(j+1)q^j + rq^r. \qquad (*)$$

想一想 没有公式表在手，怎样求 $\sum(j+1)q^j$ 的和?

记 $S=\sum_{j=0}^{r-1}(j+1)q^j = 1+2q+3q^2+\cdots+rq^{r-1}$，然后减去 $qS = q+2q^2+\cdots+(r-1)q^{r-1}+rq^r$：

$$(1-q)S = 1+q+q^2+\cdots+q^{r-1}-rq^r = \frac{1-q^r}{p} - rq^r.$$

因此 $pS = \dfrac{1-q^r}{p} - rq^r$，代入 $(*)$ 后，rq^r 两项互相抵消：

$$q^r E = \frac{1-q^r}{p} - rq^r + rq^r = \frac{1-q^r}{p}.$$

于是

$$E = \frac{1-q^r}{pq^r},$$

这就是所要证明的结果。当 $r=2$ 时，上式为 $\dfrac{(1-q)(1+q)}{pq^2} = \dfrac{1+q}{q^2}$，与第一个游戏的结果一致。

验算

- 若硬币是公平的，$p=q=\frac{1}{2}$：第一个游戏平均持续 $E=\frac{3/2}{1/4}=6$ 次投掷，赢得 $W=£3$，正好是 E 的一半，合乎预期。
- $r=1$：游戏在第一次出现反面时停止，所以长度服从几何分布（geometric distribution），平均值为 $1/q$；公式给出 $\dfrac{1-q}{pq} = \dfrac{1}{q}$。
- 当 $q\to 0$（几乎每次投掷都是正面）时，E 和 W 都趋向无穷大，这是理所当然的。

- 我们以分数作精确计算，对 $r = 1, \ldots, 7$ 及若干个 p 值解出“目前已连续出现 k 次反面”这一串状态的方程，确认了这些公式。另外，以 $p = 0.4$ 模拟 2×10^5 局游戏，在 $r = 2$ 时得出平均 4.438 次投掷（理论值 4.444）及奖金 £1.772（理论值 £1.778），在 $r = 3$ 时得出 9.076 次投掷（理论值 9.074）。

常见错误

- 在奖金的方程中，把 $qp(1 + W)$ 写成 qpW，结果得出错误答案 p/q^2。
- 只按第一次投掷来取条件：掷出一次 T 之后，游戏*并不会*重新开始，因为连续反面中的一次已经到手。这正是题目所给方程要往前看两次投掷的原因。

STEP 之外

“一旦出现正面，未来的情形就与开局时完全一样”这个论证，是马可夫链（Markov chain）的*更新法*或*首步分析法*（first-step method），一般情况下的期望到达时间（expected hitting time）正是这样计算的。$W = pE$ 这个简洁的事实是沃尔德恒等式（Wald's identity）的一个特例：把随机个数 N 个独立同分布（independent, identically distributed）的项相加，其和的期望值等于 $\mathbb{E}[N]$ 乘以单独一项的平均值，前提是 N 为停时（stopping time）。这两个课题都可以在 Grimmett 与 Welsh 的 *Probability: An Introduction*（概率论入门），或 Blitzstein 与 Hwang 的 *Introduction to Probability*（概率论导论）中读到。

详细解答

STEP III 2002，第 13 题

概率与统计 · 指数分布变量的最小值

> **STEP III 2002，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/02-s3-q13

关键思路

“至少有一个小于 x”是“两个都不小于 x”的补事件（complement），而后者的概率可凭独立性（independence）分解成乘积。指数分布（exponential distribution）的尾概率非常简单，$\mathrm{P}(X_i \geq x) = \mathrm{e}^{-\lambda_i x}$，所以乘积仍然是一个指数尾概率，而关于最小值的证明就仅此而已。巴士一问则是直接应用：等候下一班巴士的时间，就是两个指数间隔之中较小的一个。

解答

其中一个或两个都小于 x 的概率

先求单一指数变量的尾概率。对于 $x \geq 0$，

$$\mathrm{P}(X_i \geq x) = \int_x^\infty \lambda_i \mathrm{e}^{-\lambda_i t}\,\mathrm{d}t = \left[-\mathrm{e}^{-\lambda_i t}\right]_x^\infty = \mathrm{e}^{-\lambda_i x}.$$

设 E 为 $X_1 < x$ 或 $X_2 < x$（或两者皆是）这个事件。它的补事件是 $X_1 \geq x$ 与 $X_2 \geq x$ 同时成立。由于 X_1 与 X_2 独立，

$$\mathrm{P}(X_1 \geq x \text{ 且 } X_2 \geq x) = \mathrm{P}(X_1 \geq x)\,\mathrm{P}(X_2 \geq x) = \mathrm{e}^{-\lambda_1 x}\mathrm{e}^{-\lambda_2 x},$$

所以，对于 $x \geq 0$，

$$\boxed{\mathrm{P}(E) = 1 - \mathrm{e}^{-(\lambda_1+\lambda_2)x}.}$$

（对于 $x < 0$，这个概率是 0，因为两个变量都不可能取负值。）

> **想一想** 为甚么不直接把 $\mathrm{P}(X_1 < x)$ 与 $\mathrm{P}(X_2 < x)$ 相加?

因为这两个事件有重叠：两个变量可以同时小于 x，相加便会把这种情况计算两次。你也可以用容斥原理（inclusion–exclusion）来做，$(1-\mathrm{e}^{-\lambda_1 x})+(1-\mathrm{e}^{-\lambda_2 x})-(1-\mathrm{e}^{-\lambda_1 x})(1-\mathrm{e}^{-\lambda_2 x})$，化简后得到相同的答案，但用补事件会快捷得多。

最小值服从指数分布

设 $X=\min(X_1,X_2)$。两个数中较小的一个小于 x，当且仅当其中至少有一个小于 x，所以 $\{X<x\}$ 就是上面的事件 E。因此 X 的分布函数（distribution function）是

$$F(x)=\mathrm{P}(X\le x)=1-\mathrm{e}^{-(\lambda_1+\lambda_2)x}\qquad (x\ge 0),$$

而当 $x<0$ 时 $F(x)=0$。（对于连续变量，$\mathrm{P}(X<x)=\mathrm{P}(X\le x)$。）求导后，$X$ 的密度函数（density）是

$$f(x)=F'(x)=(\lambda_1+\lambda_2)\,\mathrm{e}^{-(\lambda_1+\lambda_2)x}\qquad (0\le x<\infty),$$

这正是参数为 $\lambda_1+\lambda_2$ 的指数密度函数。所以 X 服从指数分布，这正是所要证明的。

巴士

参数为 λ 的指数分布，其平均值为 $\int_0^\infty t\lambda\mathrm{e}^{-\lambda t}\,\mathrm{d}t=1/\lambda$（用分部积分法）。以分钟为时间单位，平均间隔 15 分钟和 30 分钟所对应的参数是

$$\lambda_A=\tfrac{1}{15},\qquad \lambda_B=\tfrac{1}{30}.$$

两条路线各有一班巴士刚刚开出，所以距离下一班 A 线巴士的时间，就是 A 线巴士之间一个完整的间隔 T_A，B 线同理为 T_B。两者互相独立，分别服从参数为 λ_A 和 λ_B 的指数分布。两条路线都可以送我到学院，所以我的等候时间是 $W=\min(T_A,T_B)$，而根据刚才证明的结果，W 服从指数分布，参数为

$$\lambda_A+\lambda_B=\tfrac{1}{15}+\tfrac{1}{30}=\tfrac{1}{10}.$$

中位数（median）m 满足 $\mathrm{P}(W\le m)=\frac{1}{2}$：

$$1-\mathrm{e}^{-m/10}=\tfrac{1}{2}\iff \mathrm{e}^{-m/10}=\tfrac{1}{2}\iff m=10\ln 2.$$

由于 $\ln 2\approx 0.693$，等候时间的中位数是 $m\approx 6.93$ 分钟，即大约 7 分钟，与题目要求相符。

验算

- 若 $\lambda_2\to 0$（B 线巴士几乎永远不来），答案变成 $1-\mathrm{e}^{-\lambda_1 x}$，正好是 X_1 的分布，合乎预期。

- 平均等候时间是 10 分钟，比两条路线各自的平均值 15 或 30 分钟都要短；而中位数 $10\ln 2$ 低于平均值，这是右偏分布（right-skewed distribution）的常见现象。
- 模拟 2×10^5 对平均值分别为 15 和 30 的独立指数间隔，所得最小值的中位数约为 6.91 分钟，而 $10\ln 2 = 6.931\ldots$。

常见错误

- 把平均值当作参数：平均值为 15 的指数分布，其 $\lambda = \frac{1}{15}$，而不是 15。
- 把两个概率相加，即 $\mathrm{P}(X_1 < x) + \mathrm{P}(X_2 < x)$，这样会重复计算，而且结果可能超过 1。
- 取两条路线中位数的平均，或只取 A 线的中位数（$15\ln 2 \approx 10.4$）：有两条路线可选，等候时间会缩短。

STEP 之外

巴士之间的间隔服从指数分布，意味着巴士的到达构成一个泊松过程（Poisson process），而这里的结果说明：把两个独立的泊松过程合并，会得到另一个泊松过程，其速率是两个速率之和。这个结果连同无记忆性（memoryless property）$\mathrm{P}(X > s+t \mid X > s) = \mathrm{P}(X > t)$，是排队论（queueing theory）和连续时间马可夫链（continuous-time Markov chains）的基础，在大学一年级的概率课程中便会学到。Grimmett 与 Welsh 所著的 *Probability: An Introduction*（概率入门教材）涵盖了指数变量和泊松过程。

详细解答

STEP III 2002，第 14 题

概率与统计 · 方差与协方差

STEP III 2002，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/02-s3-q14

关键思路

这条方差恒等式其实就是方差（variance）的定义，再利用期望值的线性性质（linearity of expectation）展开而已。国际象棋那部分比表面看来更巧妙：我们从头到尾都没有直接计算 $\mathrm{E}(WL)$。相反，我们由二项分布（binomial distribution）得知 $W+L$ 的方差，同样由二项分布得知 $\mathrm{Var}(W)$ 和 $\mathrm{Var}(L)$，而刚证明的恒等式便把协方差（covariance）白白送到我们手上。

解答

方差恒等式

记 $\mu_X = \mathrm{E}(X)$ 及 $\mu_Y = \mathrm{E}(Y)$。我们采用以下定义

$$\mathrm{Var}(X) = \mathrm{E}\big[(X-\mu_X)^2\big], \qquad \mathrm{Cov}(X,Y) = \mathrm{E}\big[(X-\mu_X)(Y-\mu_Y)\big],$$

以及期望值的线性性质 $\mathrm{E}(aU+bV) = a\,\mathrm{E}(U) + b\,\mathrm{E}(V)$；对于离散随机变量（discrete random variable），只要把和式 $\sum (au+bv)\,\mathrm{P}(U=u, V=v)$ 拆成两个和式，并在每个和式中对另一个变量求和，便可得出这个性质。

根据线性性质，$\mathrm{E}(X+Y) = \mu_X + \mu_Y$，所以

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathrm{E}\big[\big((X-\mu_X)+(Y-\mu_Y)\big)^2\big] \\
&= \mathrm{E}\big[(X-\mu_X)^2 + (Y-\mu_Y)^2 + 2(X-\mu_X)(Y-\mu_Y)\big] \\
&= \mathrm{E}\big[(X-\mu_X)^2\big] + \mathrm{E}\big[(Y-\mu_Y)^2\big] + 2\,\mathrm{E}\big[(X-\mu_X)(Y-\mu_Y)\big] \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y),
\end{aligned}$$

其中第三行再次用到线性性质。这就是所需的结果。留意我们并不需要独立性（independence）：这条恒等式对任何 X 和 Y 都成立。

第 (i) 部分

每一局棋不是胜、就是负、或者和，所以无论结果如何，都有 $W+L+D=m$。一个永远取同一个值的随机变量等于它的平均值，所以它的方差是 $\mathrm{E}\big[(m-m)^2\big]=0$。

第 (ii) 部分

如果一局棋分出了胜负（即胜或负），我们称它为*分胜负*的一局。每一局分胜负的概率是 $\frac{1}{3}+\frac{1}{3}=\frac{2}{3}$，各局互相独立，而且局数固定为 m。因此 $W+L$，即分胜负的局数，就是在 m 次成功概率相同的独立试验中成功的次数：

$$W+L\sim \mathrm{B}\big(m,\tfrac{2}{3}\big).$$

相关系数

同样道理，W 是 m 次独立试验中胜出的次数，每次胜出的概率是 $\frac{1}{3}$，所以 $W\sim \mathrm{B}(m,\frac{1}{3})$，同理 $L\sim \mathrm{B}(m,\frac{1}{3})$。利用二项分布的公式 $\mathrm{Var}=np(1-p)$，

$$\mathrm{Var}(W)=\mathrm{Var}(L)=m\cdot\tfrac{1}{3}\cdot\tfrac{2}{3}=\frac{2m}{9},\qquad \mathrm{Var}(W+L)=m\cdot\tfrac{2}{3}\cdot\tfrac{1}{3}=\frac{2m}{9}.$$

> **想一想** 怎样可以不计算 $\mathrm{E}(WL)$ 而求出 $\mathrm{Cov}(W,L)$?

把恒等式倒过来用。取 $X=W$ 及 $Y=L$，

$$\frac{2m}{9}=\frac{2m}{9}+\frac{2m}{9}+2\,\mathrm{Cov}(W,L),\qquad \text{所以}\qquad \mathrm{Cov}(W,L)=-\frac{m}{9}.$$

因此，当 $m\geq 1$ 时（这样方差才不是零），

$$\frac{\mathrm{Cov}(W,L)}{\sqrt{\mathrm{Var}(W)\,\mathrm{Var}(L)}}=\frac{-m/9}{2m/9}=-\frac{1}{2}.$$

答案与 m 无关。它是负数，这亦合乎情理：局数既然固定，胜的局数越多，留给负局的空间便越少。

第 (i) 部分提供了另一条途径。把恒等式应用两次，可知 $\mathrm{Var}(W+L+D)$ 等于三个方差之和，再加上三个两两协方差之和的两倍。根据对称性，三个协方差全部相等，设为 c，于是 $0=3\cdot\frac{2m}{9}+6c$，再次得出 $c=-\frac{m}{9}$。

验算

- 取 $m = 1$。这时 W 和 L 都是指示变量（indicator），$\mathrm{E}(W) = \mathrm{E}(L) = \frac{1}{3}$，而且恒有 $WL = 0$，所以 $\mathrm{Cov}(W, L) = 0 - \frac{1}{9} = -\frac{1}{9}$，与 $-m/9$ 吻合。
- 两条途径（经第 (ii) 部分和经第 (i) 部分）的结果一致。
- 对 $m = 1, \ldots, 7$ 逐一穷举全部 3^m 个结果，每次都得出 $\mathrm{Var}(W) = 2m/9$、$\mathrm{Cov}(W, L) = -m/9$，而相关系数（correlation）都恰好是 $-\frac{1}{2}$。

常见错误

- 因为各局棋互相独立，便以为 W 和 L 也互相独立。其实不然：它们与 D 合起来的总数是固定的。
- 取 $\mathrm{Var}(W + L) = \mathrm{Var}(W) + \mathrm{Var}(L) = 4m/9$，这样做漏掉了协方差一项，而且与二项分布得出的答案 $2m/9$ 矛盾。

STEP 之外

(W, L, D) 服从多项分布（multinomial distribution），即把二项分布自然地推广至每次试验有多于两种结果的情况。一般而言，若各种结果的概率为 $p_1, \ldots, p_k$，则计数 N_i 与 N_j 的协方差为 $-mp_ip_j$，相关系数为 $-\sqrt{p_ip_j/\big((1-p_i)(1-p_j)\big)}$，在本题中就是 $-\frac{1}{2}$。协方差矩阵（covariance matrix）及协方差的双线性（bilinearity），可参阅 Grimmett 与 Welsh 的 *Probability: An Introduction*（概率论入门），以及 Blitzstein 与 Hwang 的 *Introduction to Probability*（概率论导论）。

Beyond STEP

STEP 是大多数中学生一生中应考过最艰深的数学考试，也是最出色的考试之一：题目简短、陌生而深刻，当中最精彩的，其实是乔装了的大学数学小品。

本书精选最有趣的 STEP 试题，由 1987 年首次开考之前的样本试卷（specimen papers），到 2025 年的试卷，并以每一道题目作为一扇门。每道题目都会仔细解答，沿途提出优秀学生会问的问题，然后带你穿过这扇门：这个想法从何而来，到了大学会演变成甚么，以及接下来该读些甚么。

本书为正在准备 STEP、有天分又有志投身数学的学生而写。他们想要的不止是模拟试卷：他们想开始像数学系本科生那样思考，并一探前方有甚么正等着自己。

- 共 13 章，涵盖纯数学（pure mathematics）、力学（mechanics）、概率（probability）与统计（statistics）
- 136 道 STEP 历届试题，每题均附引导式解答
- 每道题目背后的大学数学
- 附答案的延伸练习，以及延伸阅读

2002 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。

