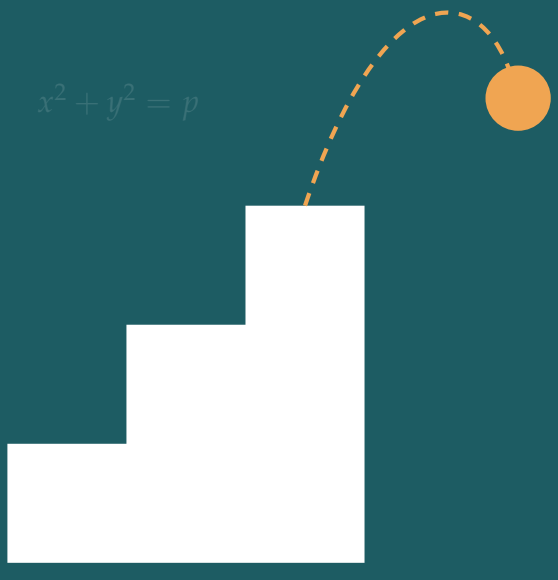
STEP 备试 · 大学数学

Beyond STEP

STEP I, II 及 III

2001 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

2001 年详细解答

Keith McNulty

Beyond STEP：以历届 STEP 试题为门，走进大学数学

2001 年详细解答

2001 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 几何不等式 纯数学 . . . 3

第 2 题 代数不等式 纯数学 . . . 6

第 3 题 三次方程与判别式 纯数学 . . . 9

第 4 题 反三角函数方程 纯数学 . . . 13

第 5 题 积分号下求导 纯数学 . . . 17

第 6 题 旋转曲面 纯数学 . . . 20

第 7 题 凹性与微分方程 纯数学 . . . 23

第 8 题 二阶线性方程 纯数学 . . . 27

第 9 题 拦截与相对速度 力学 . . . 30

第 10 题 抛体的飞行时间 力学 . . . 33

第 11 题 弹性绳与能量 力学 . . . 36

第 12 题 树形图 概率与统计 . . . 40

第 13 题 条件泊松分布 概率与统计 . . . 43

第 14 题 随机排序 概率与统计 . . . 46

STEP II

第 1 题 二项式近似 纯数学 . . . 49

第 2 题 整数部分的求和 纯数学 . . . 52

第 3 题 欧几里得几何 纯数学 . . . 55

第 4 题 倍角方程 纯数学 . . . 58

第 5 题 驻点与积分 纯数学 . . . 61

第 6 题 代换积分法 纯数学 . . . 65

第 7 题 复数与旋转 纯数学 . . . 68

第 8 题 周期函数与常微分方程 纯数学 . . . 71

第 9 题 二次空气阻力 力学 . . . 74
第 10 题 碰撞与恢复系数 力学 . . . 77
第 11 题 抛体运动 力学 . . . 81
第 12 题 泊松近似 概率与统计 . . . 85
第 13 题 连续分布与估计 概率与统计 . . . 88
第 14 题 最小方差估计量 概率与统计 . . . 92

STEP III

第 1 题 麦克劳林级数 纯数学 . . . 95
第 2 题 双曲代换与面积 纯数学 . . . 99
第 3 题 二次方程的根 纯数学 . . . 103
第 4 题 反三角函数 纯数学 . . . 107
第 5 题 抛物线的法线 纯数学 . . . 111
第 6 题 向量与三角形的心 纯数学 . . . 115
第 7 题 正交轨线 纯数学 . . . 119
第 8 题 复数轨迹 纯数学 . . . 123
第 9 题 摩擦力与简谐运动 力学 . . . 127
第 10 题 相对速度 力学 . . . 131
第 11 题 转动能量守恒 力学 . . . 135
第 12 题 随机指示变量求期望值 概率与统计 . . . 139
第 13 题 硬币序列游戏 概率与统计 . . . 142
第 14 题 方差的无偏估计量 概率与统计 . . . 146

详细解答

STEP I 2001，第 1 题

纯数学 · 几何不等式

STEP I 2001，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/01-s1-q1

关键思路

三点分别位于正方形的三条不同边上，其中必有两点落在一对对边上，而第三条边则连接这两条对边。为正方形设立坐标（coordinates），使每一点只有一个介乎 0 与 1 之间的自由参数（parameter）。然后用反证法（proof by contradiction）：假如三角形的三条边都比该上界长，每一个长度条件都会收紧其中一个参数的范围，而这些限制合起来，便会把其中一点推出正方形之外。

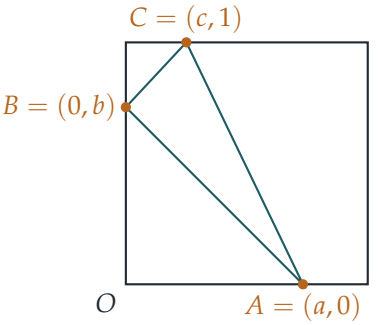
解答

以厘米为单位。正方形有四条边，而三点位于不同的边上，即这些点恰好不在正方形的其中一条边上，所以其中两点位于一对对边上，而第三点则位于与这两条边都相邻的一条边上。选取坐标轴，使正方形为 $0 \le x \le 1$，$0 \le y \le 1$，并且

$$A = (a, 0) \text{ 在底边}, \qquad B = (0, b) \text{ 在左边}, \qquad C = (c, 1) \text{ 在顶边},$$

其中 $a, b, c \in [0, 1]$。（把正方形旋转或反射并不会改变任何长度，所以这样做不失一般性（without loss of generality）。）于是

$$AB^2 = a^2 + b^2, \qquad BC^2 = c^2 + (1 - b)^2, \qquad CA^2 = (c - a)^2 + 1.$$



设 $k = \sqrt{6} - \sqrt{2}$ 及 $t = 2 - \sqrt{3}$。以下两个小结果能让数字变得容易处理：

$$k^2 = 8 - 2\sqrt{12} = 8 - 4\sqrt{3}, \qquad t^2 = 7 - 4\sqrt{3}, \qquad \text{所以} \qquad k^2 - 1 = t^2.$$

此外 $2t = 4 - 2\sqrt{3} = (\sqrt{3} - 1)^2$，所以 $\sqrt{2t} = \sqrt{3} - 1$，因此

$$\sqrt{2t} + t = (\sqrt{3} - 1) + (2 - \sqrt{3}) = 1.$$

假设三角形的三条边都比 k 长，以求得出矛盾。把正方形沿直线 $y = \frac{1}{2}$ 反射（这会把 A 与 C 的角色对调，亦把 a 与 c 对调），由这个对称性（symmetry），我们可以假设 $c \geq a$。

由 $BC > k$ 出发。由于 $c \leq 1$，

$$k^2 < BC^2 = c^2 + (1 - b)^2 \leq 1 + (1 - b)^2,$$

所以 $(1 - b)^2 > k^2 - 1 = t^2$。因为 $1 - b \geq 0$ 及 $t > 0$，由此得 $1 - b > t$，即 $0 \leq b < 1 - t$。

由 $AB > k$ 出发。由于 $0 \leq b < 1 - t$，我们有 $b^2 < (1 - t)^2$，所以

$$a^2 > k^2 - b^2 > k^2 - (1 - t)^2 = (1 + t^2) - (1 - 2t + t^2) = 2t.$$

因为 $a \geq 0$，由此得 $a > \sqrt{2t} = \sqrt{3} - 1$。

想一想 我们已经限定了 a 的范围；第三条边 CA 又会带来甚么？

边 CA 由底至顶横越整个正方形，所以它在垂直方向的跨度为 1，其余的长度必须靠水平方向补足。

由 $CA > k$ 出发。我们有 $(c - a)^2 + 1 > k^2$，所以 $(c - a)^2 > t^2$，而由于 $c \geq a$，这表示 $c - a > t$。因此

$$c > a + t > \sqrt{2t} + t = 1.$$

但 C 位于顶边上，所以 $c \leq 1$。这个矛盾说明三条边不可能全都比 k 长：三角形 ABC 至少有一条边的长度不超过 $(\sqrt{6} - \sqrt{2})$ cm，这正是所要证明的。

这个上界可以达到

当 $c=1$ 及 $a=b=\sqrt{3}-1$，即 $A=(\sqrt{3}-1,0)$、$B=(0,\sqrt{3}-1)$、$C=(1,1)$ 时，上面每一个不等式（inequality）都取到临界情况。这时

$$AB^2=2(\sqrt{3}-1)^2=8-4\sqrt{3},\qquad BC^2=CA^2=1+(2-\sqrt{3})^2=8-4\sqrt{3},$$

所以 ABC 是等边三角形（equilateral triangle），边长恰好为 k（C 位于角上，算作在顶边上）。因此 $\sqrt{6}-\sqrt{2}$ 是最佳的常数。用三角学的说法，$k=\sec 15^\circ$：这是能放进单位正方形内的最大等边三角形，与各边倾斜成 15°。

验算

- $k=\sqrt{6}-\sqrt{2}\approx 1.0353$ 比正方形的边长稍大，这是合理的：若三角形的三个顶点分别在三条边上，它的三条边可以全都比 1 长，但只能长一点点。
- 在 2×10^6 组随机选取的 $(a,b,c)\in[0,1]^3$ 中搜索，最短边从未超过 1.0337，低于 k；而上述等边三角形的例子，三条边在机器精度内均等于 1.03528。
- $\sec 15^\circ=4/(\sqrt{6}+\sqrt{2})=\sqrt{6}-\sqrt{2}$，与等边三角形的情况吻合。

常见错误

- 未经证明便假设三角形是等边的，或假设其中一个顶点在角上。极端情况能提示答案，却不能证明不等式对所有情况都成立。
- 忘记说明设置的理由：关键的结构性事实是，三条不同的边之中必定包含一对对边。
- 对 $(1-b)^2>t^2$ 或 $(c-a)^2>t^2$ 开平方时，没有说明所取的是哪一个正负号。

STEP 之外

这是一个小型的极大极小（max–min）问题：我们要在一个由各种情况组成的紧致集（compact set）上，使最短边尽量长，而极端情况会令所有约束条件同时取等号，这种模式在最优化（optimisation）和博弈论（game theory）中随处可见。由于 $[0,1]^3$ 是有界闭集，而最短边是一个连续函数（continuous function），最大值必定存在，这个事实会在大学第一门分析课程中证明（极值定理（extreme value theorem））。Abbott 的 *Understanding Analysis*（大学分析入门教材）对此有很好的讲解，而 Siklos 的 *Advanced Problems in Mathematics*（STEP 难题选讲）则有更多同类风格的 STEP 几何题。

详细解答

STEP I 2001，第 2 题

纯数学 · 代数不等式

STEP I 2001，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/01-s1-q2

关键思路

处理不等式（inequality）时，不能把两边同乘一个正负未知的量，所以 (i) 的稳妥做法是把所有项移到同一边，然后因式分解（factorise）：这样答案便可从符号图（sign diagram）直接读出。在 (ii) 中，平方根会引诱我们把两边平方，但这只有在已知两边皆非负时才是合理的。全题的关键就在于时刻留意这个条件。

解答

第 (i) 部分

我们不能乘以 x，因为 x 可能是负数。我们改为乘以 x^2，由于 $x \neq 0$，它必定严格为正，所以不等号方向不变：

$$\begin{aligned} 1+2x-x^2 > \frac{2}{x} &\iff x^2+2x^3-x^4 > 2x \\ &\iff 0 > x^4-2x^3-x^2+2x \\ &\iff x\left(x^3-2x^2-x+2\right) < 0. \end{aligned}$$

这个三次式（cubic）可用分组法因式分解：

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

所以我们需要

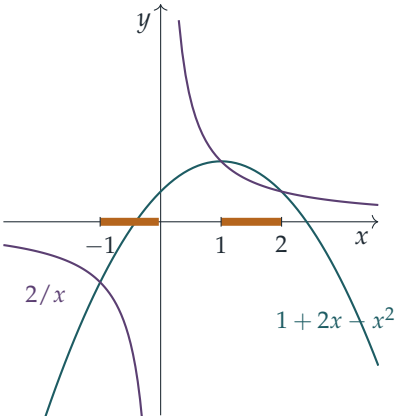
$$(x+1)\,x\,(x-1)(x-2) < 0.$$

这是一个首项系数（leading coefficient）为正的四次多项式（quartic），有四个单根（simple root）$-1, 0, 1, 2$，所以每当 x 越过一个根，它的正负号便交替一次：当 $x > 2$ 时为正，在

$(1,2)$ 上为负，在 $(0,1)$ 上为正，在 $(-1,0)$ 上为负，当 $x<-1$ 时为正。在各个根上，其值为 0，并不 <0（而且 $x=0$ 本来就已被排除）。因此

$$\boxed{-1<x<0 \quad \text{或} \quad 1<x<2.}$$

图像与此吻合：抛物线（parabola）$y=1+2x-x^2$ 位于双曲线（hyperbola）$y=2/x$ 上方的范围，恰好由它们在 $x=-1$、1、2 的交点划分，而在负的一侧，则只到双曲线在 $x=0$ 处急坠至 $-\infty$ 为止。



第 (ii) 部分

首先留意，$x\geq -\frac{10}{3}$ 除了给出 $3x+10\geq 0$ 之外，还给出 $x+4\geq \frac{2}{3}>0$，所以两个平方根在整个范围内都有定义。

两边皆非负，而对于 $a,b\geq 0$，我们有 $a>b \iff a^2>b^2$。所以我们可以把两边平方，既不会多出解，也不会丢失解：

$$\begin{aligned}\sqrt{3x+10}>2+\sqrt{x+4} &\iff 3x+10>4+4\sqrt{x+4}+(x+4)\\ &\iff 2x+2>4\sqrt{x+4}\\ &\iff x+1>2\sqrt{x+4}.\end{aligned}$$

? **想一想** 我们可以直接再平方一次吗?

不能马上这样做。右边的 $2\sqrt{x+4}$ 是正数，但 $x+1$ 可能是负数，这时平方便可能把一个不成立的不等式变成成立的不等式。所以我们要分情况讨论。

► 若 $x+1\leq 0$，左边至多为 0，而右边严格为正。不等式不成立。

- 若 $x+1>0$，两边皆为正，所以平方这一步同样是可逆的：

$$\begin{aligned} x+1>2\sqrt{x+4} &\iff x^2+2x+1>4x+16 \\ &\iff x^2-2x-15>0 \\ &\iff (x-5)(x+3)>0, \end{aligned}$$

  这在 $x>5$ 或 $x<-3$ 时成立。结合 $x>-1$，只余下 $x>5$。

因此

$$\boxed{x>5.}$$

验算

- 在 $x=5$：$\sqrt{25}=5$ 而 $2+\sqrt{9}=5$，所以该处等号成立，这正是边界上应有的情况。在 $x=12$：$\sqrt{46}\approx 6.78>2+4=6$。在 $x=0$：$\sqrt{10}\approx 3.16<4$。
- 在 (i) 中，$x=1.5$ 给出 $1.75>1.33$，$x=-0.5$ 给出 $-0.25>-4$，两者皆成立；$x=0.5$ 给出 $1.75<4$，不成立。
- 我们用电脑对两个不等式各自在 2×10^5 个随机点上作了测试，另加上与每个边界相距不超过 10^{-6} 的点，结果与上述答案完全一致。

常见错误

- 把 (i) 乘以 x，却没有分开 $x>0$ 和 $x<0$ 两种情况。这样会丢失区间 $(-1,0)$，或者把该处的答案弄反。
- 把 $x+1>2\sqrt{x+4}$ 平方，却没有要求 $x+1>0$。这样会得出多余的区间 $-\frac{10}{3}\le x<-3$；试把 $x=-3.2$ 代入原式，便可看到它并不成立（$\sqrt{0.4}\approx 0.63<2+\sqrt{0.8}$）。

STEP 之外

(ii) 中的原则是：一个运算只有在所涉及的数值上是递增函数（increasing function）时，才会保持不等式成立。这正是分析学家处理不等式的起点：$t\mapsto t^2$ 在 $[0,\infty)$ 上递增，但在 $\mathbb{R}$ 上则不然。乘积的符号图可推广为以下事实：连续函数（continuous function）只有经过零才能改变正负号，这就是介值定理（intermediate value theorem）。Abbott 的 *Understanding Analysis*（一本分析学入门书）对这两点都有细致的论述。

详细解答

STEP I 2001，第 3 题

纯数学 · 三次方程与判别式

STEP I 2001，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/01-s1-q3

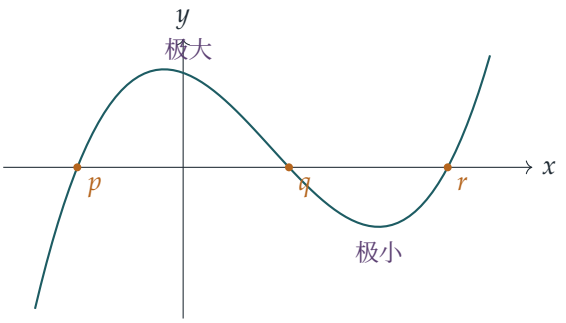
关键思路

一个有三个相异实根的三次多项式（cubic），在每两个相邻的根之间必定要转向，所以它的导数（derivative）——一个二次式——有两个相异实根。把“有两个相异实根”写成“判别式（discriminant）为正”，就得到第一个不等式。至于第二部分，题目给出的乘积是一个三次多项式，其系数正好就是目标不等式中的那些量，所以用第一部分的结果（或者配方（completing the square））便可完成。

解答

略图

展开得 $f(x) = x^3 - (p+q+r)x^2 + \ldots$，所以当 $x \to \infty$ 时 $f(x) \to +\infty$，而当 $x \to -\infty$ 时 $f(x) \to -\infty$。图像只在 p、q 和 r 三处与 x-轴相交，而且由于每个根都是单根（simple root），f 在每个根处都会变号：当 $x < p$ 时 $f < 0$，在 (p,q) 上 $f > 0$，在 (q,r) 上 $f < 0$，当 $x > r$ 时 $f > 0$。因此在 p 与 q 之间某处有一个局部极大值（local maximum），在 q 与 r 之间某处有一个局部极小值（local minimum）。y-截距为 $f(0) = -pqr$。



（图中是按 $p < 0 < q < r$ 的情况绘画的；无论各根位于何处，图像的形状都一样，只是 y-轴的位置有所不同。）

第一个不等式

完全展开，得

$$f(x) = x^3 - (p+q+r)x^2 + (qr+rp+pq)x - pqr,$$

所以

$$f'(x) = 3x^2 - 2(p+q+r)x + (qr+rp+pq).$$

由于 $f(p) = f(q) = 0$，根据罗尔定理（Rolle's theorem），f' 在 p 与 q 之间（不含端点）有一个根；同理，在 q 与 r 之间（不含端点）也有一个根。这两个区间互不重叠，所以二次方程 $f'(x) = 0$ 有两个相异实根。（不用罗尔定理的做法：$f'(p) = (p-q)(p-r) > 0$、$f'(q) = (q-p)(q-r) < 0$ 及 $f'(r) = (r-p)(r-q) > 0$，所以 f' 在每个区间上都会变号。）

一个二次式有两个相异实根，当且仅当它的判别式为正，所以

$$4(p+q+r)^2 - 12(qr+rp+pq) > 0, \qquad \text{即} \qquad (p+q+r)^2 > 3(qr+rp+pq),$$

这正是所要证明的。

第二个不等式

展开得

$$(x^2+gx+h)(x-k) = x^3 + (g-k)x^2 + (h-gk)x - hk.$$

把它与 f 的展开式比较：如果这个三次多项式的根是 p, q, r，则

$$p+q+r = -(g-k), \qquad qr+rp+pq = h-gk,$$

所以目标不等式 $(g-k)^2 > 3(h-gk)$ 正好就是把第一个不等式应用于这个三次多项式的根。

充分（*sufficient*）。假设 $g^2 > 4h$。那么 $x^2 + gx + h$ 有两个相异实根 $\alpha < \beta$，而这个三次多项式有实根 α、β、k。如果 k 与两者都不相同，三个根便互不相同，由第一部分立即得出 $(g-k)^2 > 3(h-gk)$。

想一想 如果 k 刚好等于 α 或 β，以致各根并非互不相同，那又如何?

第一部分假设了 $p < q < r$，所以我们还需要多做一点工夫。恒等式（identity）

$$(p+q+r)^2 - 3(qr+rp+pq) = \tfrac{1}{2}\left[(p-q)^2 + (q-r)^2 + (r-p)^2\right]$$

显示，除非 $p = q = r$，否则左边必为正。由于 $\alpha \neq \beta$，三个根绝不会全部相等，所以不等式在这个情况下同样成立。

另有一个直接的做法，可以完全避开根。乘开得

$$(g-k)^2 - 3(h-gk) = g^2 + gk + k^2 - 3h = \left(k + \tfrac{1}{2}g\right)^2 + \tfrac{3}{4}\left(g^2 - 4h\right),$$

当 $g^2 > 4h$ 时，这是一个平方加上一个正数，因此为正。无论用哪一个方法，$g^2 > 4h$ 都是充分条件。

非必要（*not necessary*）。我们需要找出满足 $g^2 \leq 4h$ 而不等式仍然成立的数值。取 $g = 0$、$h = 1$、$k = 2$。那么 $g^2 = 0 \leq 4 = 4h$，但是

$$(g-k)^2 = 4 > 3 = 3(h-gk).$$

所以即使没有 $g^2 > 4h$，不等式也可以成立，这个条件并非必要。

想一想 第一部分明明需要三个实根，为甚么会这样?

第一部分其实只需要 $f'(x) = 0$ 有两个相异实根，这个要求较弱。这里的三次多项式 $(x^2+1)(x-2) = x^3 - 2x^2 + x - 2$ 只有一个实根，但 $f'(x) = 3x^2 - 4x + 1 = (3x-1)(x-1)$，所以它仍然有两个驻点（stationary point）。不等式 $(g-k)^2 > 3(h-gk)$ 所说的，正正就是这个三次多项式有一个局部极大值和一个与之不同的局部极小值，不论它的图像是否与横轴相交三次。

验算

- 取 $p, q, r = 0, 1, 2$：$(p+q+r)^2 = 9 > 6 = 3(qr+rp+pq)$，而 $f'(x) = 3x^2 - 6x + 2$ 的根是 $1 \pm 1/\sqrt{3}$，一个在 $(0,1)$ 内，一个在 $(1,2)$ 内。
- 配方所得的公式显示，这个不等式的界限是最佳的，不能再改进：取 $g = 2$、$h = 1$（所以 $g^2 = 4h$）及 $k = -1$，得 $(g-k)^2 = 9 = 3(h-gk)$，是等式而不是严格不等式。这就是三次多项式 $(x+1)^3$，三个根全部相等。

- 我们用电脑以精确运算核对了两个恒等式和两个不等式，每个各用了 20 000 个随机有理数例子，另加例子 $g = 0$、$h = 1$、$k = 2$（得出 $4 > 3$）。

常见错误

- 把三次曲线上下颠倒地画：x^3 的系数是 $+1$，所以图像向右上升。
- 忘记了 k 是 $x^2 + gx + h$ 的根这个情况，这时第一部分不能照原样套用。
- 只用一个满足 $g^2 \leq 4h$ 的例子说明不等式成立，就称之为“充分”。一个例子证明的是*非必要*；要证明充分性，则须对所有 g、h、k 给出证明。

STEP 之外

$(p+q+r)^2 - 3(qr+rp+pq)$ 这个量，除了相差一个常数因子外，就是导数的判别式，而且它是各根的*对称函数*（*symmetric function*）：任何这类函数都可以用系数来表示，这就是为甚么这个不等式可以从根的形式转写成 g、h、k 的形式。三次多项式本身的完整判别式 $\prod(p-q)^2$，决定了三个根是否全为实数；本题用的是它较弱的近亲。对称函数和判别式再往前走，就通往伽罗瓦理论（Galois theory）；Allenby 的 *Rings, Fields and Groups*（环、域与群）是一个很好的起点。

详细解答

STEP I 2001，第 4 题

纯数学 · 反三角函数方程

> **STEP I 2001，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/01-s1-q4

关键思路

三倍角公式（triple-angle formula）可直接由 tan 的复角公式（addition formula）推出。之后，一切都取决于一个事实：tan 以 π 为周期（period）重复，所以若 $\tan 3\varphi = \tan 3\theta$，则 φ 与 θ 可以相差 $\pi/3$ 的任何整数倍。(i) 和 (ii) 真正要下工夫的地方，是留意 $\cos^{-1}$ 和 $\tan^{-1}$ 的值域（range），因为这些值域决定了那无穷多个角之中哪些才是容许的。

解答

三倍角公式

设 $t = \tan\theta$。由复角公式，$\tan 2\theta = \dfrac{2t}{1-t^2}$，于是

$$\tan 3\theta = \tan(2\theta+\theta) = \frac{\tan 2\theta + t}{1 - t\tan 2\theta} = \frac{\dfrac{2t}{1-t^2} + t}{1 - \dfrac{2t^2}{1-t^2}}$$

$$= \frac{2t + t(1-t^2)}{(1-t^2) - 2t^2} = \frac{3t - t^3}{1 - 3t^2},$$

其中我们把分子和分母同乘以 $1 - t^2$。这就是所求的公式。（只要等号两边都有定义，公式便成立。）

$\tan 3\theta$ 的值

由于 $\cos\theta = 2/\sqrt{5}$ 且 $0 < \theta < \pi/2$，我们有 $\sin\theta > 0$，所以

$$\sin\theta = \sqrt{1-\tfrac{4}{5}} = \frac{1}{\sqrt{5}}, \qquad \tan\theta = \frac{\sin\theta}{\cos\theta} = \frac{1}{2}.$$

（可以想像一个边长为 1、2、$\sqrt{5}$ 的直角三角形。）代入 $t = \frac{1}{2}$：

$$\tan 3\theta = \frac{\frac{3}{2} - \frac{1}{8}}{1 - \frac{3}{4}} = \frac{11/8}{1/4} = \frac{11}{2},$$

正如所求。

稍后会用到 θ 相当小这一点：$\tan\theta = \frac{1}{2} < \frac{1}{\sqrt{3}} = \tan\frac{\pi}{6}$，所以 $0 < \theta < \pi/6$，因而 $0 < 3\theta < \pi/2$。

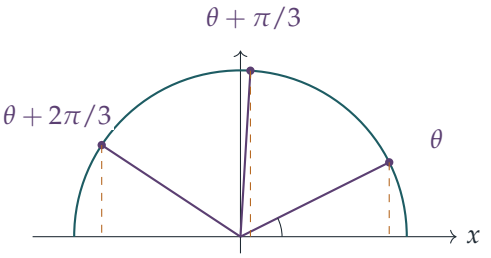
第 (i) 部分

设 $\varphi = \cos^{-1} x$，则 $0 \le \varphi \le \pi$ 且 $x = \cos\varphi$。方程即是说 $\tan 3\varphi = \frac{11}{2} = \tan 3\theta$。因为 tan 的周期为 π，而且在每个长度为 π 的区间上每个值都恰好取一次，

$$3\varphi = 3\theta + k\pi \quad\Longleftrightarrow\quad \varphi = \theta + \frac{k\pi}{3}, \qquad k \in \mathbb{Z}.$$

想一想 这些角之中，实际上有多少个是容许的?

我们需要 $0 \le \varphi \le \pi$。由于 $0 < \theta < \pi/6$，取 $k = 0, 1, 2$ 得 $\varphi = \theta,\ \theta + \frac{\pi}{3},\ \theta + \frac{2\pi}{3}$，全部都严格落在 $(0, \pi)$ 之内；$k = 3$ 给出 $\varphi > \pi$，而 $k = -1$ 给出 $\varphi < 0$。所以恰好有三个解。



x 的三个值就是这些角的余弦（即各条虚线的垂足）。利用 $\cos\theta = 2/\sqrt{5}$、$\sin\theta = 1/\sqrt{5}$ 和复角公式：

$$x_1 = \cos\theta = \frac{2}{\sqrt{5}},$$

$$x_2 = \cos\theta\cos\tfrac{\pi}{3} - \sin\theta\sin\tfrac{\pi}{3} = \frac{2}{\sqrt{5}}\cdot\frac{1}{2} - \frac{1}{\sqrt{5}}\cdot\frac{\sqrt{3}}{2} = \frac{2-\sqrt{3}}{2\sqrt{5}},$$

$$x_3 = \cos\theta\cos\tfrac{2\pi}{3} - \sin\theta\sin\tfrac{2\pi}{3} = -\frac{2}{\sqrt{5}}\cdot\frac{1}{2} - \frac{1}{\sqrt{5}}\cdot\frac{\sqrt{3}}{2} = -\frac{2+\sqrt{3}}{2\sqrt{5}}.$$

这些角都不会令 3φ 成为 $\pi/2$ 的奇数倍（每个 3φ 都是 $3\theta + k\pi$，其中 $0 < 3\theta < \pi/2$），所以 $\tan 3\varphi$ 在每个情况下都有定义。(i) 的解是

$$x = \frac{2}{\sqrt{5}}, \qquad x = \frac{2-\sqrt{3}}{2\sqrt{5}}, \qquad x = -\frac{2+\sqrt{3}}{2\sqrt{5}}.$$

（有理化（rationalising）后，它们是 $\frac{2\sqrt{5}}{5}$、$\frac{(2-\sqrt{3})\sqrt{5}}{10}$ 和 $-\frac{(2+\sqrt{3})\sqrt{5}}{10}$。）

另一途径。把 $T = \tan\varphi$ 代入三倍角公式，得 $2(3T - T^3) = 11(1 - 3T^2)$，即

$$2T^3 - 33T^2 - 6T + 11 = 0.$$

我们知道 $T = \frac{1}{2}$ 是一个根（root），因式分解（factorising）得 $(2T-1)(T^2 - 16T - 11) = 0$，所以 $T = \frac{1}{2}$ 或 $T = 8 \pm 5\sqrt{3}$。于是 $x = \cos\varphi = \pm 1/\sqrt{1+T^2}$，其正负号与 T 相同（T 为正表示 φ 是锐角；为负则表示是钝角）。这样得出同样的三个答案，而且 $1 + (8+5\sqrt{3})^2 = 140 + 80\sqrt{3} = 1/x_2^2$ 也是一个不错的验算。

第 (ii) 部分

设 $\psi = \frac{1}{3}\tan^{-1} y$。由于 $-\frac{\pi}{2} < \tan^{-1} y < \frac{\pi}{2}$，我们有 $-\frac{\pi}{6} < \psi < \frac{\pi}{6}$。

> ? **想一想** 条件 $\cos\psi = 2/\sqrt{5}$ 给出一个角还是两个角?

在 $(-\pi/6, \pi/6)$ 上，cos 是偶函数（even function），所以 $\cos\psi = 2/\sqrt{5} = \cos\theta$ 给出 $\psi = \theta$ 或 $\psi = -\theta$，而且因为 $0 < \theta < \pi/6$，两者都在容许的范围内。（再没有其他解：在 $[0, \pi/6)$ 上，函数 cos 严格递减（strictly decreasing），所以它在该处只取 $2/\sqrt{5}$ 这个值一次，在 $(-\pi/6, 0]$ 上亦然。）

所以 $\tan^{-1} y = \pm 3\theta$。由于 $0 < 3\theta < \pi/2$，$\pm 3\theta$ 两者确实都是 $\tan^{-1}$ 的值，而

$$y = \tan(\pm 3\theta) = \pm\frac{11}{2}.$$

(ii) 的解是 $y = \frac{11}{2}$ 和 $y = -\frac{11}{2}$。

验算

- $\theta = \cos^{-1}(2/\sqrt{5}) \approx 0.4636$，所以 $3\theta \approx 1.3909$，而 $\tan 3\theta \approx 5.5000$。
- $x_2 = \frac{2-\sqrt{3}}{2\sqrt{5}} \approx 0.05992$ 而 $x_3 \approx -0.83451$；用计算机算得两者的 $\tan(3\cos^{-1} x) \approx 5.5$，而在 $[-1, 1]$ 上作细密的数值扫描，亦恰好找到三个解。
- $\cos\left(\frac{1}{3}\tan^{-1}(\pm 5.5)\right) = \cos(\pm 0.4636) \approx 0.8944 = 2/\sqrt{5}$。

常见错误

- 在 (i) 中只给出 $x = 2/\sqrt{5}$，忘记了 tan 每隔 π 便重复一次，所以 3φ 在 $[0, 3\pi]$ 内有三个可能的值。
- 在 (ii) 中只给出 $y = 11/2$：cos 是偶函数，所以 $\psi = -\theta$ 同样成立。
- 没有检查范围：答案取决于 $\theta < \pi/6$，正因如此，$\theta + 2\pi/3$ 在 (i) 中才是容许的，而 3θ 在 (ii) 中才落在 $\tan^{-1}$ 的值域之内。

STEP 之外

反三角函数（inverse trigonometric functions）让我们初次体会到，要为一个并非一对一（one-to-one）的函数求反函数是怎样的难题：我们选定一个*主支*（*principal branch*），而这类方程其实是在问我们身处哪一个分支。在复分析（complex analysis）中，$\log z$ 和 $z^{1/3}$ 也会出现同样的问题：绕原点走一圈，便会由一个分支走到下一个分支，正如这里把 θ 加上 $2\pi/3$ 一样。Needham 的 *Visual Complex Analysis* 以图像把分支解释得十分精彩。

详细解答

STEP I 2001，第 5 题

纯数学 · 积分号下求导

STEP I 2001，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/01-s1-q5

关键思路

(i) 和 (ii) 的两个积分都可以直接积出，第二个要先作代换（substitution）$u = 1 + tx$。真正的重点在于题末的那句话：(ii) 的被积函数（integrand）是 (i) 的被积函数对 t 的导数（derivative），而两个答案之间也有同样的关系。由此看来，在积分号内对参数（parameter）t 求导，似乎可以与对 x 积分交换次序。再做一次，然后代入 $t = 1$，便得到所要的猜想（conjecture）。

解答

全题中 $t > 0$ 且 $0 \le x \le 1$，所以 $1 + tx \ge 1 > 0$，各个被积函数都不会趋向无穷大。

第 (i) 部分

由于 $\dfrac{\mathrm{d}}{\mathrm{d}x}(1+tx)^{-1} = -t(1+tx)^{-2}$，$(1+tx)^{-2}$ 的一个原函数（antiderivative）是 $-\dfrac{1}{t(1+tx)}$（这里用到 $t \neq 0$）。因此

$$\int_0^1 \frac{1}{(1+tx)^2}\,\mathrm{d}x = \left[-\frac{1}{t(1+tx)}\right]_0^1 = \frac{1}{t}\left(1 - \frac{1}{1+t}\right) = \frac{1}{t}\cdot\frac{t}{1+t} = \frac{1}{1+t},$$

这正是所要证明的结果。

第 (ii) 部分

分子中的 x 令直接积分变得不便，所以我们设 $u = 1 + tx$。于是 $x = \dfrac{u-1}{t}$，$\mathrm{d}x = \dfrac{\mathrm{d}u}{t}$，而积分上下限 $x = 0, 1$ 变成 $u = 1, 1+t$：

$$\begin{aligned}\int_0^1 \frac{-2x}{(1+tx)^3}\,\mathrm{d}x &= -\frac{2}{t^2}\int_1^{1+t}\frac{u-1}{u^3}\,\mathrm{d}u = -\frac{2}{t^2}\int_1^{1+t}\left(u^{-2}-u^{-3}\right)\mathrm{d}u \\ &= -\frac{2}{t^2}\left[-\frac{1}{u}+\frac{1}{2u^2}\right]_1^{1+t} = -\frac{2}{t^2}\left(\frac{1}{2}-\frac{1}{1+t}+\frac{1}{2(1+t)^2}\right).\end{aligned}$$

想一想 怎样才能不必把所有项通分展开，就把括号内的式子化简?

设 $w = \dfrac{1}{1+t}$，括号内的式子就是 $\frac{1}{2} - w + \frac{1}{2}w^2 = \frac{1}{2}(1-w)^2$，是一个完全平方（perfect square）。又因为 $1 - w = \dfrac{t}{1+t}$，所以

$$\int_0^1 \frac{-2x}{(1+tx)^3}\,\mathrm{d}x = -\frac{2}{t^2}\cdot\frac{1}{2}\cdot\frac{t^2}{(1+t)^2} = -\frac{1}{(1+t)^2},$$

这就是所求的结果。

猜想

把 x 固定，将被积函数看成 t 的函数：

$$\frac{\partial}{\partial t}(1+tx)^{-2} = -2x(1+tx)^{-3}.$$

所以 (ii) 的被积函数是 (i) 的被积函数对 t 的导数，正如 (ii) 的答案是 (i) 的答案的导数一样。这提示了以下法则：

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_0^1 f(x,t)\,\mathrm{d}x = \int_0^1 \frac{\partial f}{\partial t}\,\mathrm{d}x.$$

再应用一次，

$$\frac{\partial}{\partial t}\left(-2x(1+tx)^{-3}\right) = 6x^2(1+tx)^{-4}, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\left(-(1+t)^{-2}\right) = 2(1+t)^{-3},$$

于是我们猜想，对于 $t > 0$，

$$\int_0^1 \frac{6x^2}{(1+tx)^4}\,\mathrm{d}x = \frac{2}{(1+t)^3}.$$

代入 $t = 1$，猜想的值为

$$\int_0^1 \frac{6x^2}{(1+x)^4}\,\mathrm{d}x = \frac{2}{8} = \frac{1}{4}.$$

验算

- 当 $t \to 0$ 时，(i) 的被积函数趋向 1，而答案 $\frac{1}{1+t}$ 也趋向 1；在 (ii) 中，被积函数趋向 $-2x$，其积分为 -1，与 $-\frac{1}{(1+t)^2} \to -1$ 吻合。
- 题目没有要求我们验证这个猜想，但验证起来很快：设 $u = 1 + x$，积分变成 $6\int_1^2 (u^{-2} - 2u^{-3} + u^{-4})\,\mathrm{d}u = 6\left[-\frac{1}{u} + \frac{1}{u^2} - \frac{1}{3u^3}\right]_1^2 = 6\left(-\frac{7}{24} + \frac{1}{3}\right) = \frac{1}{4}$。
- 用辛普森法则（Simpson's rule）计算最后一个积分，得 0.250000；另外取 25 个不同的 t 值，结果都与 $\frac{2}{(1+t)^3}$ 吻合至小数点后九位。

常见错误

- 在 (ii) 中漏掉由 $\mathrm{d}x = \mathrm{d}u/t$ 而来的因子 $\frac{1}{t}$：这里共有*两个* $\frac{1}{t}$ 因子，一个来自 x，一个来自 $\mathrm{d}x$。
- 代换后没有把积分上下限改为 $u = 1$ 和 $u = 1 + t$。
- 对 $-(1+t)^{-2}$ 求导时得出 $-2(1+t)^{-3}$：两个负号互相抵消，所以猜想的值是正数（这也是必然的，因为被积函数是正的）。

STEP 之外

我们猜想出来的法则，就是*积分号下求导*（differentiation under the integral sign），又称莱布尼兹法则（Leibniz's rule）。实分析（real analysis）中有一条定理指出：只要 f 和 $\partial f/\partial t$ 在一个闭矩形上连续（continuous），这条法则便成立；本题中，对于 $0 \le x \le 1$ 以及任何区间 $[a, b]$（其中 $a > 0$）内的 t，这个条件都成立。重复同样的论证，可得对每一个 n 都有 $\int_0^1 \frac{x^n}{(1+tx)^{n+2}}\,\mathrm{d}x = \frac{1}{(n+1)(1+t)^{n+1}}$。理查德·费曼（Richard Feynman）曾用这一招计算其他方法都束手无策的积分，并因此为人津津乐道。至于交换极限（limit）、导数与积分次序背后的理论，可参阅 Bartle 与 Sherbert 的 *Introduction to Real Analysis*（实分析入门教材）。

详细解答

STEP I 2001，第 6 题

纯数学 · 旋转曲面

> **STEP I 2001，第 6 题**
>
> 继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：
>
> step.maths.org/questions/01-s1-q6

本题在 *Beyond STEP* 第 4.8 节*切球形面包*中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch04/spherical-loaf.html

关键思路

把半圆 $y=\sqrt{a^2-x^2}$ 绕 x 轴旋转，便得到球体。在表面积（surface area）公式中，被积函数（integrand）$y\sqrt{1+(\mathrm{d}y/\mathrm{d}x)^2}$ 原来是*常数*，所以任何一片面包上的面包皮只取决于该片的厚度，而与它所在的位置无关。至于第二部分，我们从一端开始逐片吃掉（正如提示中的积分上下限 $-a$ 至 $-a+t$ 所暗示），因此剩下的是一个球冠（spherical cap）；比值 V/A 成为球冠厚度的二次函数，我们要求出它的最大值。

解答

面包皮与片数成正比

设面包的半径为 a，中心位于原点，则其表面可由

$$y=\sqrt{a^2-x^2},\qquad -a\le x\le a,$$

绕 x 轴旋转而得。求导后，当 $|x|<a$ 时 $\dfrac{\mathrm{d}y}{\mathrm{d}x}=-\dfrac{x}{\sqrt{a^2-x^2}}=-\dfrac{x}{y}$，所以

$$y\left(1+\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2\right)^{\frac{1}{2}}=y\left(\frac{y^2+x^2}{y^2}\right)^{\frac{1}{2}}=y\cdot\frac{a}{y}=a,$$

其中用到 $x^2+y^2=a^2$ 及 $y>0$。因此，球体在平面 $x=x_1$ 与 $x=x_2$ 之间（其中 $-a\le x_1<x_2\le a$）的曲面面积为

$$2\pi\int_{x_1}^{x_2} a\,\mathrm{d}x = 2\pi a\,(x_2-x_1).$$

（被积函数只在端点 $x=\pm a$ 处没有定义，但在该处仍然有界，所以这个积分没有问题。）

现在假设面包被切成 N 片，每片厚度为 $h=2a/N$。每一片无论位于何处，所带的面包皮面积都恰好是 $2\pi ah$。所以，如果剩下 n 片，不论它们在甚么位置，剩余的面包皮面积都是

$$A = n\cdot 2\pi ah,$$

它与剩余的片数 n 成正比。这正是我们要证明的结果。（这是阿基米德（Archimedes）的古老结果：球体与其外切圆柱（circumscribing cylinder）在任意两个平行平面之间的面积相等。）

“脆皮”条件

面包是从一端开始逐片吃掉的，所以吃掉若干片之后，剩下的面包是由 $x=-a$ 至 $x=-a+t$ 的球冠，其中 t 满足 $0<t\le 2a$。由第一部分可知 $A=2\pi at$。至于体积，作代换 $u=x+a$，使得 $y^2=a^2-x^2=(a-x)(a+x)=(2a-u)u$：

$$V=\pi\int_{-a}^{-a+t}(a^2-x^2)\,\mathrm{d}x=\pi\int_0^t(2au-u^2)\,\mathrm{d}u=\pi\Big(at^2-\frac{t^3}{3}\Big).$$

因此

$$\frac{V}{A}=\frac{\pi t^2\left(a-\frac{t}{3}\right)}{2\pi at}=\frac{t}{2}-\frac{t^2}{6a}=:f(t).$$

想一想 这个比值取决于已经吃掉多少。我们需要留意的是哪一个 t 值?

无论吃掉了多少片，条件都必须成立，所以我们需要 f 的最大值。利用配方法（completing the square），

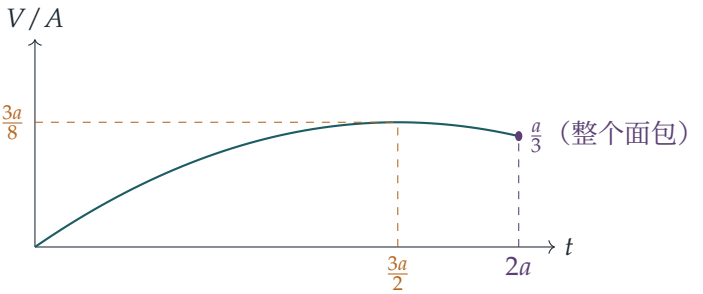
$$f(t)=\frac{3a}{8}-\frac{1}{6a}\Big(t-\frac{3a}{2}\Big)^2,$$

所以 f 在 $t=\frac{3a}{2}$ 时最大（这个值位于 $(0,2a]$ 之内），最大值为 $\frac{3a}{8}$。

$t=\frac{3}{2}a$ 是否真的会出现？会的，例如把面包切成 4 片并吃掉一片（或切成 $4m$ 片并吃掉 m 片）。一个面包若要作为平行切片面包而称得上“脆皮”，就必须在这样的切法下符合规定，所以需要

$$\frac{3a}{8}<1,\qquad \text{即}\qquad a<\frac{8}{3}=2\tfrac{2}{3}\text{ 米},$$

正如题目所求。反过来说，若 $a<\frac{8}{3}$，则对每一个 t 都有 $f(t)\le\frac{3a}{8}<1$，所以无论怎样切片，这个面包都是“脆皮”的。



上图显示 $a = 2$ 时的 $f(t)$：当我们向面包中央推进时，比值上升，其后回落，因为接近另一端时，我们加入了大量面包皮，却只加入少量面包。

验算

- 对整个面包而言，$t = 2a$ 给出 $f(2a) = a - \frac{2a}{3} = \frac{a}{3}$，与 $\frac{4}{3}\pi a^3/(4\pi a^2)$ 吻合。
- 取 $a = \frac{8}{3}$，切成 4 片。剩下 $1, 2, 3, 4$ 片时，比值分别为 $\frac{5}{9}, \frac{8}{9}, 1, \frac{8}{9}$，所以吃掉一片之后比值恰好是 1，刚好不满足严格不等式（strict inequality）。
- 数值上，对面积公式作数值积分（quadrature），所得结果在随机选取的切片上都与 $2\pi a(x_2 - x_1)$ 相符；而当 $a = 2.66$ 时，切成不多于 200 片的每一种切法，比值都保持在 1 以下。

常见错误

- 只考虑整个面包：$a/3 < 1$ 给出 $a < 3$，这是必要条件，但弱得多。最坏的一刻是面包宽度的四分之一被吃掉的时候。
- 把平坦的切面也算作面包皮。面包皮只是原来的曲面，这也正是题目所给公式所量度的。

STEP 之外

被积函数为常数，这就是阿基米德的帽盒定理（hat-box theorem）：把球面横向投影到其外切圆柱上，面积保持不变。正因如此，在球面上均匀地选取一点，其高度服从均匀分布（uniform distribution），这个事实在概率和物理中经常用到。像这样的曲面积分（surface integral），以及映射在甚么情况下保持面积，都会在向量微积分（vector calculus）的入门课程中详细讨论；Needham 的 *Visual Complex Analysis*（以几何直观讲解复分析的著作）对球面的相关映射有一段优美的几何讨论。

详细解答

STEP I 2001，第 7 题

纯数学 · 凹性与微分方程

STEP I 2001，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/01-s1-q7

关键思路

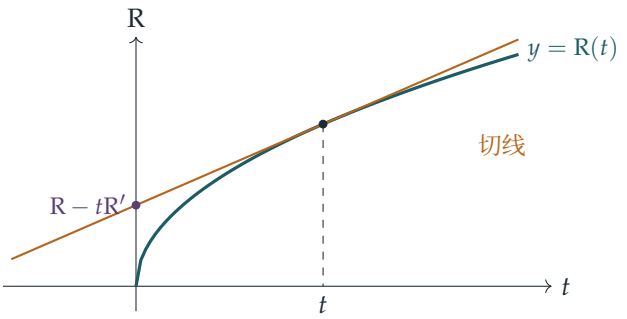
三个条件说明 R 由原点出发、一直上升，并且向下弯曲：它的图像是凹（concave）的。凹的图像位于它每一条切线（tangent）的下方，单凭这个几何事实就可得出 (i) 的不等式。第 (ii) 和 (iii) 部分则是可分离变量的微分方程（separable differential equations），因为 $\mathrm{H} = \mathrm{R}'/\mathrm{R}$ 正是 $\ln \mathrm{R}$ 的导数。一旦明确求出 R，我们便把它逐一对照各个条件检验，而 (i) 的不等式提供了一个快捷的方法来排除某个模型。

解答

首先，有一个事实会贯穿全题：由于 $\mathrm{R}(0) = 0$，而且 R 在 $t > 0$ 时严格递增（strictly increasing），所以当 $t > 0$ 时 $\mathrm{R}(t) > 0$。因此当 $t > 0$ 时，$\mathrm{H}(t)$ 有定义，且 $\mathrm{H}(t) > 0$。

第 (i) 部分

图像通过原点，斜率（gradient）为正，而且斜率随 t 增加而减少。



固定 $t > 0$。图像在时间 t 处的切线斜率为 $\mathrm{R}'(t)$，所以它与纵轴相交于高度

$$\mathrm{R}(t) - t\,\mathrm{R}'(t).$$

从图中可见这个截距（intercept）是正数，因为通过原点的曲线位于切线下方。要把这一点说得严谨，可对 R 在 $[0, t]$ 上应用中值定理（mean value theorem）：存在某个 c 满足 $0 < c < t$，使得

$$\mathrm{R}(t) - \mathrm{R}(0) = t\,\mathrm{R}'(c), \qquad \text{所以} \qquad \mathrm{R}(t) = t\,\mathrm{R}'(c).$$

由于 $\mathrm{R}'' < 0$，函数 R' 严格递减，而 $c < t$ 给出 $\mathrm{R}'(c) > \mathrm{R}'(t)$。因此

$$\mathrm{R}(t) > t\,\mathrm{R}'(t),$$

这正正就是“切线的截距在原点之上”这句话。两边除以 $\mathrm{R}'(t) > 0$，得

$$t < \frac{\mathrm{R}(t)}{\mathrm{R}'(t)} = \frac{1}{\mathrm{H}(t)},$$

这就是所要证明的结果。

（等价地说，切线与 t 轴相交于 $t - 1/\mathrm{H}(t)$，而这是负数。在宇宙学（cosmology）中，$1/\mathrm{H}$ 称为“哈勃时间”（Hubble time），这个结果说明：在任何减速膨胀的模型中，宇宙的年龄都比它的哈勃时间短。）

第 (ii) 部分

当 $t > 0$ 时，$\mathrm{R} > 0$，所以可以写成

$$\frac{\mathrm{d}}{\mathrm{d}t} \ln \mathrm{R}(t) = \frac{\mathrm{R}'(t)}{\mathrm{R}(t)} = \frac{a}{t}.$$

积分后得 $\ln \mathrm{R} = a \ln t + C$，所以

$$\mathrm{R}(t) = k\,t^a \qquad (t > 0),$$

其中 $k = \mathrm{e}^C > 0$。

现在利用 $k > 0$ 和 $t^a > 0$ 来检验各个条件：

- $\mathrm{R}'(t) = ka\,t^{a-1}$，它在 $t > 0$ 时为正，当且仅当 $a > 0$。
- $\mathrm{R}''(t) = ka(a-1)\,t^{a-2}$。已知 $a > 0$，它为负当且仅当 $a < 1$。
- 若 $a > 0$，则当 $t \to 0^+$ 时 $kt^a \to 0$，所以与 $\mathrm{R}(0) = 0$ 相容。（若 $a \le 0$，这个条件无论如何都不成立：kt^a 会趋向 $k \neq 0$，或者趋向无限大。）

所以条件 $(*)$ 成立，当且仅当

$$0 < a < 1.$$

这与第 (i) 部分吻合：在那里 $t < 1/\mathrm{H} = t/a$，由此必有 $a < 1$。

第 (iii) 部分

现在 $\mathrm{H}(t) = b/t^2$。首先，当 $t > 0$ 时 $\mathrm{H}(t) > 0$（见开首的说明），所以必须有 $b > 0$。若 $b \le 0$，结论已经成立。

因此假设 $b > 0$。根据第 (i) 部分，每个 $t > 0$ 都必须满足

$$t < \frac{1}{\mathrm{H}(t)} = \frac{t^2}{b}, \qquad \text{即} \qquad t > b$$

（两边除以 $t > 0$ 并乘以 $b > 0$）。但对于每个满足 $0 < t \le b$ 的 t，这个不等式都不成立。所以没有任何 b 的值能与 $(*)$ 相容。

> **想一想** 这样使用第 (i) 部分的结果是否合理?

合理。第 (i) 部分证明了*任何*满足 $(*)$ 的函数，对所有 $t > 0$ 都有 $t < 1/\mathrm{H}(t)$。所以，假如某个建议的 H 哪怕只在一个 t 值上违反这个不等式，就没有任何具有这个 H 的函数能满足 $(*)$。这是一个反证法（proof by contradiction）。

若想直接验证，可仿照 (ii) 解这个方程：$\ln \mathrm{R} = -b/t + C$，所以 $\mathrm{R}(t) = k\,\mathrm{e}^{-b/t}$，其中 $k > 0$。于是

$$\mathrm{R}'(t) = \frac{kb}{t^2}\,\mathrm{e}^{-b/t},$$

$$\mathrm{R}''(t) = kb\,\mathrm{e}^{-b/t}\left(\frac{b}{t^4} - \frac{2}{t^3}\right) = \frac{kb\,\mathrm{e}^{-b/t}}{t^4}\,(b - 2t).$$

当 $b > 0$ 时，$\mathrm{R}' > 0$，而且当 $t \to 0^+$ 时 $\mathrm{R}(t) \to 0$，所以头两个条件成立。但当 $0 < t < b/2$ 时 $\mathrm{R}''(t) > 0$，所以图像在原点附近弯向了错误的方向，$(*)$ 不成立。

验算

- 在 (ii) 中，取 $a = \frac{1}{2}$ 得 $\mathrm{R} = k\sqrt{t}$，正是草图中的形状：递增而且是凹的，并有 $t < 1/\mathrm{H} = 2t$。取 $a = 1$ 则得一条通过原点的直线，此时恰好有 $t = 1/\mathrm{H}$：这是 (i) 的边界情况。
- 在 (iii) 中取 $b = 1$、$k = 1$：数值上 $\mathrm{R}''(0.25) \approx 2.34 > 0$，而 $1/\mathrm{H}(0.5) = 0.25 < 0.5$，两者都违反 $(*)$。

我们用电脑以几个凹的测试函数检验了 (i)，以有限差分（finite differences）检验了 (ii) 和 (iii) 中的解及导数公式，并检验了 R'' 在 $t = b/2$ 处变号。

常见错误

- 在 (i) 中，只凭图像作论证。“曲线位于切线下方”需要理由：中值定理，加上 R' 递减。

- 在 (ii) 中，忘记了要有 $a > 0$ 才能保证 $\mathrm{R}' > 0$（以及 $\mathrm{R}(0) = 0$），结果只给出 $a < 1$。
- 在 (iii) 中，声称凹性在 t 很大时不成立。$b - 2t$ 的正负号显示，出问题的地方是在 t 很小的时候。

STEP 之外

第 (i) 部分是关于凸（convex）函数和凹函数的一个核心事实的特殊情况：一个可微（differentiable）函数是凹的，当且仅当它的图像位于所有切线的下方。这个事实是詹森不等式（Jensen's inequality）以及最优化（optimisation）中许多理论的基础。这里用中值定理的论证，是大学一年级分析学（analysis）的标准内容；可参阅 Abbott 的 *Understanding Analysis*。在真实的宇宙学中，$a = \frac{2}{3}$ 对应由物质主导的爱因斯坦–德西特（Einstein–de Sitter）宇宙，它的年龄是哈勃时间的三分之二。

详细解答

STEP I 2001，第 8 题

纯数学 · 二阶线性方程

STEP I 2001，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/01-s1-q8

关键思路

把每个已知解代入方程，都会在每个 x 值处得到一条联系两个未知函数 $p(x)$ 和 $q(x)$ 的方程。两个解便给出两条联立线性方程（simultaneous linear equations），我们可以像解两个未知数那样把 p 和 q 解出来。在最后一部分，中间部分所证明的线性（linearity）让我们可以把题目给出的两个解换成它们较简单的组合，而三角恒等式（trigonometric identities）令这些组合变得非常简单。

解答

求 p 和 q

对于 $y = x$，有 $y' = 1$ 及 $y'' = 0$，所以代入 $(*)$ 得

$$p(x) + x\,q(x) = 0. \tag{1}$$

对于 $y = 1 - x^2$，有 $y' = -2x$ 及 $y'' = -2$，所以

$$-2 - 2x\,p(x) + (1 - x^2)\,q(x) = 0. \tag{2}$$

由 (1) 得 $p(x) = -x\,q(x)$。把它代入 (2)，

$$-2 + 2x^2 q(x) + (1 - x^2)q(x) = 0 \quad \Longrightarrow \quad (1 + x^2)\,q(x) = 2.$$

由于 $1 + x^2 > 0$，我们可以相除，所以 $q(x) = 2(1 + x^2)^{-1}$，然后由 (1) 得 $p(x) = -x\,q(x) = -2x(1 + x^2)^{-1}$，正如题目所求。（我们从没有除以 x，所以结果对所有实数 x 都成立，包括 $x = 0$。）

任意组合 $ax+b(1-x^2)$ 都是解

记 $L[y]=y''+p\,y'+q\,y$。求导是线性的，所以若 $y=ay_1+by_2$，其中 a,b 为常数，则

$$\begin{aligned}L[ay_1+by_2] &= a\big(y_1''+p\,y_1'+q\,y_1\big)+b\big(y_2''+p\,y_2'+q\,y_2\big)\\ &= a\,L[y_1]+b\,L[y_2].\end{aligned}$$

取 $y_1=x$ 及 $y_2=1-x^2$，则 $L[y_1]$ 和 $L[y_2]$ 都等于零（这正是它们满足 $(*)$ 的意思），所以对所有常数 a 和 b，都有 $L[ax+b(1-x^2)]=0$。如果你想直接验证：$y=ax+b(1-x^2)$ 有 $y'=a-2bx$ 及 $y''=-2b$，而

$$\begin{aligned}y''+p\,y'+q\,y &= -2b-\frac{2x(a-2bx)}{1+x^2}+\frac{2\big(ax+b-bx^2\big)}{1+x^2}\\ &= \frac{-2b(1+x^2)-2ax+4bx^2+2ax+2b-2bx^2}{1+x^2}=0.\end{aligned}$$

最后一部分

刚才的论证完全没有用到 p 和 q 的具体形式：对于任何形如 $(*)$ 的方程，两个解的组合 ay_1+by_2 仍然是解。所以，如果 $\cos^2(\frac12x^2)$ 和 $\sin^2(\frac12x^2)$ 都满足 $(*)$，它们的和与差也同样满足，

$$\cos^2(\tfrac12x^2)+\sin^2(\tfrac12x^2)=1,\qquad \cos^2(\tfrac12x^2)-\sin^2(\tfrac12x^2)=\cos(x^2),$$

这里用了 $\cos^2\theta+\sin^2\theta=1$ 和 $\cos^2\theta-\sin^2\theta=\cos2\theta$，其中 $\theta=\frac12x^2$。

? **想一想** 为甚么可以用 1 和 $\cos(x^2)$ 来代替题目给出的函数?

因为这一步是可逆的：题目给出的函数是 $\frac12\big(1+\cos x^2\big)$ 和 $\frac12\big(1-\cos x^2\big)$，所以 p 和 q 使这两个函数都成为解，当且仅当它们使 1 和 $\cos(x^2)$ 都成为解。我们没有损失任何信息，而导数却容易计算得多。

解 $y=1$。这时 $y'=y''=0$，所以 $(*)$ 简化为 $q(x)\cdot1=0$，得

$$q(x)=0.$$

解 $y=\cos(x^2)$。由链式法则（chain rule）得 $y'=-2x\sin(x^2)$，再由积法则（product rule）得

$$y''=-2\sin(x^2)-4x^2\cos(x^2).$$

当 $q=0$ 时，$(*)$ 变成

$$-2\sin(x^2)-4x^2\cos(x^2)-2x\sin(x^2)\,p(x)=0.$$

凡是 $x\sin(x^2)\neq 0$ 之处，我们都可以相除，于是

$$p(x)=-\frac{\sin(x^2)+2x^2\cos(x^2)}{x\sin(x^2)}=-\frac{1}{x}-2x\cot(x^2).$$

所以 $p(x)=-\dfrac{1}{x}-2x\cot(x^2)$ 及 $q(x)=0$。

在 $x\sin(x^2)=0$ 的各点（即 $x=0$ 和 $x=\pm\sqrt{n\pi}$），p 的系数为零，所以方程对该处的 p 没有任何限制；事实上，我们求得的 p 在这些点趋向无穷。这并不是出错：在这些点上，题目给出的两个函数的导数都是零，所以它们不可能是一条性质良好、形如 $(*)$ 的方程的两个“独立”解，而 p 在该处必定是奇异（singular）的。

验算

- 作为第一部分的独立验算，两个解的朗斯基行列式（*Wronskian*）$W=y_1y_2'-y_2y_1'$ 满足 $W'=-pW$。取 $y_1=x$，$y_2=1-x^2$，得 $W=-2x^2-(1-x^2)=-(1+x^2)$，所以 $-p=W'/W=2x/(1+x^2)$，与我们求得的 p 吻合。至于最后一部分，解 1 和 $\cos(x^2)$ 给出 $W=-2x\sin(x^2)$，而同样有 $W'/W=1/x+2x\cot(x^2)=-p$。
- 数值上，在 $x=0.9$ 处，把题目给出的两个函数代入 $(*)$，并以有限差分（finite differences）计算导数，可迫使 $p(0.9)\approx -2.82465$ 及 $q(0.9)\approx 0$，与 $-1/0.9-1.8\cot(0.81)\approx -2.82465$ 相符。

常见错误

- 直接对 $\cos^2(\frac{1}{2}x^2)$ 求导两次，结果迷失在繁复的代数运算之中。应先用倍角恒等式（double-angle identities），然后才求导。
- 把 (1) 除以 x 而得到 $q=-p/x$；这样做是可行的，但之后必须另外处理 $x=0$ 的情况。改为消去 p 便可避开这个问题。
- 忘记说明最后一部分中 p 的公式在哪些地方成立。

STEP 之外

这道题让我们一窥二阶线性微分方程（second-order linear differential equations）的理论：解构成一个二维向量空间（vector space），任意两个独立的解便可决定方程，而朗斯基行列式（连同阿贝尔（Abel）公式 $W'=-pW$）则用来衡量独立性。朗斯基行列式为零的点，例如本题的 $x=\pm\sqrt{n\pi}$，正是系数必定变得奇异之处。Boyce 与 DiPrima 的 *Elementary Differential Equations*（微分方程入门教材）对这一切都有细致的阐述。

详细解答

STEP I 2001，第 9 题

力学 · 拦截与相对速度

> **STEP I 2001，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/01-s1-q9

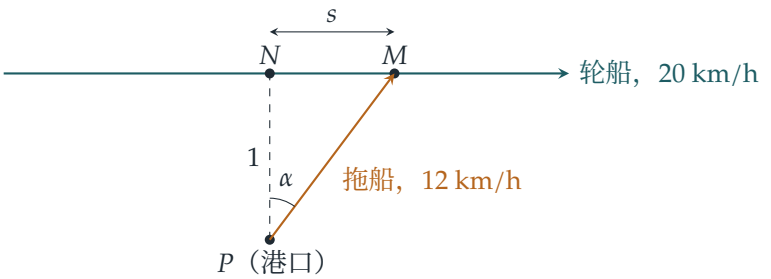
关键思路

要尽量迟出发，拖船就必须全速沿直线前进，所以真正要选择的，只是在轮船航线上的哪一点与它会合。对每一个可能的会合点，先求出拖船最迟可以在甚么时间出发而仍然及时赶到，然后对会合点求这个出发时间的最大值。只要把问题设置妥当，这就是一个单变量的最优化（optimisation）问题。

解答

建立模型

把轮船和拖船都视为质点（particle）。把港口 P 放在原点，并设轮船沿直线 $y = 1$ 向正 x 方向航行（距离以公里为单位，时间以小时为单位）。由 P 到航线的垂足（foot of the perpendicular）是 $N = (0, 1)$，即轮船最接近港口的一点。



假设两船在点 $M = (s, 1)$ 会合。轮船沿航线的位置是时间的线性函数（linear function），

所以它到达 M 的时间是

$$t_M = \frac{s-c}{20},$$

其中 c 是一个常数，由轮船的起始位置决定。（轮船仍在驶近，所以 c 小于我们会用到的任何会合位置。）

想一想 为甚么拖船必须全速沿直线前进?

如果拖船在时间 T 出发，并在时间 t_M 到达 M，它就以不超过 12 km/h 的速率走了至少等于直线距离 $PM = \sqrt{s^2+1}$ 的路程，所以 $t_M - T \geq \sqrt{s^2+1}/12$。以最高速率直驶往 M，等号便成立。任何较慢或较迂回的走法，都会迫使拖船提早出发。

因此，能够在 M 会合的最迟出发时间是

$$T(s) = \frac{s-c}{20} - \frac{\sqrt{s^2+1}}{12},$$

而我们要求的是 $T(s)$ 在所有 s 之中的最大值。

求出发时间的最大值

求导数（derivative），得

$$T'(s) = \frac{1}{20} - \frac{s}{12\sqrt{s^2+1}}.$$

令 $T'(s) = 0$，得

$$\frac{s}{\sqrt{s^2+1}} = \frac{12}{20} = \frac{3}{5}.$$

这要求 $s > 0$。两边平方，$25s^2 = 9(s^2+1)$，所以 $s^2 = \frac{9}{16}$，$s = \frac{3}{4}$。

这是极大值，而且是全局最大值（global maximum）：$\sqrt{s^2+1}$ 是凸函数（convex function），所以 T（一个线性函数减去一个凸函数）是凹函数（concave function）。换一个说法，$\dfrac{s}{\sqrt{s^2+1}}$ 随 s 递增，所以当 $s < \frac{3}{4}$ 时 $T'(s) > 0$，当 $s > \frac{3}{4}$ 时 $T'(s) < 0$。留意常数 c 已经消失：会合点与轮船从哪里出发无关。

所以拖船在越过 N 之后 $\frac{3}{4}$ 公里处与轮船会合，而它所走的距离是

$$PM = \sqrt{\tfrac{9}{16}+1} = \sqrt{\tfrac{25}{16}} \quad \Longrightarrow \quad \boxed{PM = \tfrac{5}{4} \text{ 公里} = 1.25 \text{ 公里}.}$$

拖船在海上航行 $\frac{5/4}{12} = \frac{5}{48}$ 小时，即 $6\frac{1}{4}$ 分钟。

几何解读

拖船的航向与 PN 成角 α，其中 $\sin\alpha = \frac{s}{\sqrt{s^2+1}} = \frac{3}{5}$，所以 $\sin\alpha$ 正是两船速率之比 $12/20$。拖船的速度（velocity）是 $(7.2,\ 9.6)$，而它相对于轮船的速度是 $(7.2-20,\ 9.6) = (-12.8,\ 9.6)$。两者的点积（dot product）是 $-92.16 + 92.16 = 0$。在临界出发时间，拖船的速度与它相对于轮船的速度互相垂直。这是较慢的追赶者仅仅来得及拦截时的标准条件。

验算

- 以几个不同的轮船起始位置，在一个细密的会合点网格上作穷举搜索，$T(s)$ 的最大值总是出现在 $s = 0.75$，所以拖船所走的距离是 1.25 公里。
- 由最迟出发时间开始，以航向 $(\frac{3}{5}, \frac{4}{5})$ 用数值方法逐步模拟运动，拖船会到达轮船所在之处。只要迟 0.002 小时出发，无论取甚么航向，都不能驶到轮船 20 米以内。
- 合理性检查：假如拖船与轮船一样快，$\sin\alpha$ 就会是 1，会合点会跑到无穷远处。这是合理的，因为速率相同的拖船可以无止境地追着轮船跑。

常见错误

- 在轮船最接近港口的一点 N 与它会合（拖船所走的距离为 1 公里）。这样做是把拖船的距离减至最小，而不是把出发时间推到最迟：拖船向前多走一点，就能赚到时间，因为轮船还要航行一段才到那里。
- 只把拖船的航行时间 $\sqrt{s^2+1}/12$ 减至最小，而漏掉 $s/20$ 一项。要求最大值的量是出发时间，它同时牵涉两艘船。

STEP 之外

求到达时间之差的最大值，是最优控制（optimal control）或追逐问题（pursuit problem）的一个小例子。速率之比的条件 $\sin\alpha = 12/20$，与超音速飞机后方的马赫锥（Mach cone）是同一种几何：圆锥半顶角的正弦等于音速与飞机速率之比，而这里那艘慢的拖船，就扮演着声音追赶快速声源的角色。Kleppner 与 Kolenkow 的 *An Introduction to Mechanics*（力学入门）仔细讨论了相对速度（relative velocity）和参考系（frame of reference），而 Siklos 的 *Advanced Problems in Mathematics*（高等数学难题）则有更多 STEP 风格的力学最优化题目。

详细解答

STEP I 2001，第 10 题

力学 • 抛体的飞行时间

STEP I 2001，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/01-s1-q10

关键思路

把选择分成两部分：仰角决定了飞行时间和水平射程，而方向只决定炮弹落在以炮为圆心的圆上的哪一点。因此，炮弹能击中直线，当且仅当其射程至少为 d。飞行时间随仰角增大而增加，但仰角超过 $45°$ 后，射程会缩短，所以我们把仰角一直调高，直到射程刚好降至 d 为止。

解答

给定仰角时的飞行时间与射程

把炮弹当作只受重力作用的质点（particle）（不计空气阻力），由地面发射，并落在平原上。设发射的仰角（elevation）为 α，其中 $0 < \alpha \leq \frac{\pi}{2}$。它的垂直速度分量是 $v\sin\alpha$，水平分量是 $v\cos\alpha$。

在垂直方向，经过时间 t 后它的高度是 $vt\sin\alpha - \frac{1}{2}gt^2$，而高度再次为零的时刻是

$$T = \frac{2v\sin\alpha}{g}.$$

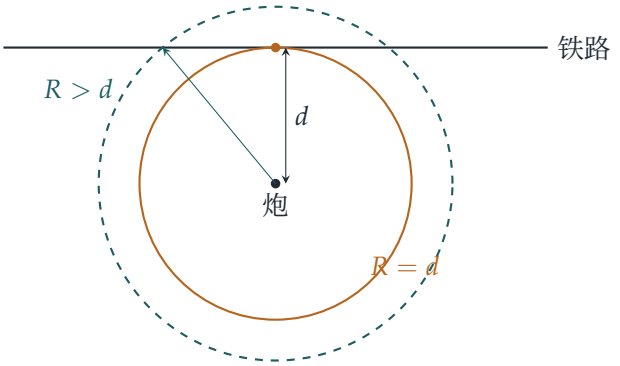
在这段时间内，它走过的水平距离是

$$R = v\cos\alpha \cdot \frac{2v\sin\alpha}{g} = \frac{v^2\sin 2\alpha}{g}.$$

炮弹何时能击中直线？

仰角固定时，转动炮身会使落点沿着一个以炮为圆心、半径为 R 的圆移动。直线上最近的一点与炮的距离为 d，而直线上其他各点都更远。所以这个圆与直线相交，当且仅当

$R \geq d$。当 $R = d$ 时，圆刚好与直线相切，炮必须正对着最近的那一点，即与直线垂直。



（俯视图。虚线圆对应较低的仰角，它在两点到达直线。实线圆对应最优的仰角，它刚好到达直线。）

所以炮手以仰角 α 发射时能击中直线，当且仅当

$$\sin 2\alpha \geq \frac{gd}{v^2}.$$

记 $k = gd/v^2$。题目告诉我们 $v^2 > gd$，所以 $0 < k < 1$，确实有一些仰角可行。（最大射程是 v^2/g，在 $45°$ 时取得，所以这个条件只是说直线在射程之内。）

使飞行时间最大

由于 $T = 2v\sin\alpha/g$，而 $\sin\alpha$ 在 $(0, \frac{\pi}{2}]$ 上是递增的，要使 T 最大，就要选取容许范围内最大的 α。

当 $0 < 2\alpha \leq \pi$ 时，条件 $\sin 2\alpha \geq k$ 成立，当且仅当

$$\arcsin k \leq 2\alpha \leq \pi - \arcsin k.$$

所以容许的最大仰角是 $\alpha = \frac{\pi}{2} - \frac{1}{2}\arcsin k$，它大于 $45°$。在这个仰角下 $\sin 2\alpha = k$，而由于 2α 位于 $(\frac{\pi}{2}, \pi)$ 之内，$\cos 2\alpha$ 是负数：

$$\cos 2\alpha = -\sqrt{1 - k^2}.$$

? **想一想** 沿着直线瞄准远处某一点，让炮弹飞得更远，不是更好吗？

不是。飞得更远需要更大的射程，而在 $45°$ 以上，这意味着更低的仰角，飞行时间因而更短。飞行时间只取决于仰角，所以我们能做到的最好结果，就是仍能到达直线的最陡一击。它的射程刚好是 d，所以炮弹落在由炮到直线的垂足上。

现在把 T 平方，并利用 $2\sin^2\alpha = 1 - \cos 2\alpha$：

$$\begin{aligned} g^2T^2 &= 4v^2\sin^2\alpha \\ &= 2v^2\left(1-\cos 2\alpha\right) \\ &= 2v^2 + 2v^2\sqrt{1-\frac{g^2d^2}{v^4}} \\ &= 2v^2 + 2\left(v^4 - g^2d^2\right)^{1/2}, \end{aligned}$$

其中把 v^2 移入了平方根之内（因为 $v^2 > 0$，这是容许的）。这就是所要求的结果。

验算

- 当 $d \to 0$ 时，公式给出 $g^2T^2 \to 4v^2$，所以 $T \to 2v/g$。这相当于垂直向上发射，是最长的飞行时间；当直线紧贴着炮经过时，这正是应有的结果。
- 当 $v^2 = gd$ 时，直线仅仅在射程之内：唯一可行的是以 45° 发射，而公式给出 $g^2T^2 = 2v^2$，与 $T = 2v\sin 45^\circ/g = \sqrt{2}\,v/g$ 吻合。
- 以数值计算，取 $g = 9.8\,\mathrm{m\,s^{-2}}$、$d = 100\,\mathrm{m}$ 和 $v = 40\,\mathrm{m\,s^{-1}}$，最佳仰角约为 71.1°，公式给出 $T \approx 7.72\,\mathrm{s}$。以数值方法逐步模拟运动，得到相同的时间，炮弹亦落在直线上。对所有仰角作穷举搜索，在数百个随机例子中都与公式一致。

常见错误

- 取了较小的根 $\alpha = \frac{1}{2}\arcsin k$。这样发射同样能击中直线，但得到的是最短而不是最长的飞行时间：实际上就是取错了 $\cos 2\alpha$ 的正负号。
- 一开始就假设炮弹必定落在直线上最近的一点。这个结论最终是对的，但需要上面的论证：其他任何一点都需要更大的射程，因而需要更低的仰角。

STEP 之外

这是一个小型的约束最优化（constrained optimisation）问题：在某个约束条件（到达直线）之下，使某个量（飞行时间）最大，而最优解位于容许范围的边界上。在大学一年级的力学和最优化课程中，你会遇到处理这类问题的一般工具，例如拉格朗日乘数（Lagrange multipliers），以及与之相关的概念：所有轨迹的包络线（envelope）（即“安全抛物线”（parabola of safety）），它围住了炮所能到达的每一点。Kleppner 与 Kolenkow 的 *An Introduction to Mechanics*（力学入门）是深入学习抛体与质点力学的好去处。

详细解答

STEP I 2001，第 11 题

力学 · 弹性绳与能量

STEP I 2001，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：

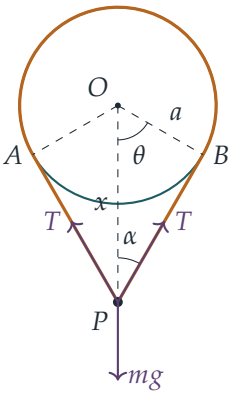
step.maths.org/questions/01-s1-q11

关键思路

质点把弹性带向下拉之后，弹性带大部分长度仍然贴在圆柱上，并在两侧沿切线方向离开圆柱，再笔直地连到质点。因此，整个问题归结为由圆外一点向圆所作两条切线（tangent）的几何。由于圆柱是光滑的，弹性带上各处的张力（tension）都相同，所以用总伸长量（extension）代入胡克定律（Hooke's law）便可求得张力。之后，第 (i) 部分是能量问题（质点在最低点瞬时静止），第 (ii) 部分是力的问题（速率最大即加速度为零）。

解答

我们把质点视为点质量，把弹性带视为轻而有弹性（遵从胡克定律），并把圆柱视为光滑。设 O 为横截面的圆心，并设质点 P 位于 O 下方距离 x 处，其中 $x \geq a$。弹性带在 A 和 B 两点离开圆，其中 PA 和 PB 为切线。



几何关系与张力

三角形 OAP 在 A 处为直角（切线垂直于半径），所以

$$PA = PB = \sqrt{x^2 - a^2}, \qquad \cos\theta = \frac{a}{x}, \qquad \sin\alpha = \frac{a}{x},$$

其中 $\theta = \angle AOP$，而 $\alpha = \angle APO$，即每段笔直的弹性带与铅垂线所成的角。除了最接近 P 的一段圆心角为 2θ 的弧 AB 之外，弹性带处处与圆接触，所以它的长度是

$$L = (2\pi a - 2a\theta) + 2\sqrt{x^2 - a^2}.$$

自然长度（natural length）为 $2\pi a$，所以伸长量是

$$e(x) = 2\left(\sqrt{x^2 - a^2} - a\theta\right), \qquad \theta = \arccos\frac{a}{x}.$$

由于圆柱是光滑的，沿弹性带方向没有摩擦力，所以张力 T 处处相同，而根据胡克定律，$T = \dfrac{\lambda e}{2\pi a}$。

当 $x = 2a$ 时，我们有 $\cos\theta = \frac{1}{2}$，所以 $\theta = \frac{\pi}{3}$，$\alpha = \frac{\pi}{6}$，而 $\sqrt{x^2 - a^2} = \sqrt{3}\,a$。因此

$$e(2a) = 2\left(\sqrt{3}\,a - \frac{\pi a}{3}\right) = \frac{2a}{3}\left(3\sqrt{3} - \pi\right).$$

第 (i) 部分

释放时，质点位于圆柱的底部，即 $x = a$，弹性带未被拉长，质点处于静止。到了 $x = 2a$，质点再次静止（它不再往下掉），所以在这两个时刻它都没有动能（kinetic energy）。圆柱是光滑的，所以没有能量损失，重力势能（gravitational potential energy）的减少等于弹性势能（elastic energy）的增加。质点下降了距离 $2a - a = a$，而自然长度为 $2\pi a$、伸长量为 e 的弹性带，其弹性势能为 $\dfrac{\lambda e^2}{2 \cdot 2\pi a}$。所以

$$mga = \frac{\lambda}{4\pi a} \cdot \frac{4a^2}{9}\left(3\sqrt{3} - \pi\right)^2 = \frac{\lambda a\left(3\sqrt{3} - \pi\right)^2}{9\pi},$$

由此得

$$\lambda = \frac{9\pi mg}{\left(3\sqrt{3} - \pi\right)^2},$$

正如所求。

> **想一想** 弹性带有一部分绕在圆柱上，对整条弹性带使用弹性势能公式，这样做合理吗？

合理。由于圆柱是光滑的，无论是否绕在圆柱上，整条弹性带的张力都相同，因而每单位长度的伸长也相同。弹性带的表现，与一条自然长度为 $2\pi a$、总伸长量为 e 的弹性绳完全一样，而这正是公式 $\lambda e^2/(2l)$ 所描述的情况。

第 (ii) 部分

加速度为零时，速率最大，也就是说，作用于质点的合力为零。在水平方向，两个张力因对称而互相抵消。在 P 处沿铅垂方向分解，

$$2T\cos\alpha = mg.$$

当 $x = 2a$ 时，$\cos\alpha = \cos\frac{\pi}{6} = \frac{\sqrt{3}}{2}$，所以 $\sqrt{3}\,T = mg$，得 $T = \dfrac{mg}{\sqrt{3}}$。利用胡克定律和上面求得的伸长量，

$$T = \frac{\lambda}{2\pi a}\cdot\frac{2a}{3}\left(3\sqrt{3}-\pi\right) = \frac{\lambda\left(3\sqrt{3}-\pi\right)}{3\pi}.$$

令 T 的两个表达式相等，

$$\lambda = \frac{3\pi mg}{\sqrt{3}\left(3\sqrt{3}-\pi\right)} = \frac{\sqrt{3}\,\pi mg}{3\sqrt{3}-\pi}.$$

? **想一想** 我们怎样知道这是速率的极大值，而不仅仅是一个驻点？

当 x 增加时，e 会增加（每段笔直的弹性带变长的速度，比弧长 $a\theta$ 缩短的速度快），而 $\cos\alpha = \sqrt{1-a^2/x^2}$ 也会增加，所以向上的力 $2T\cos\alpha$ 由 0 开始持续增加。质点在 $x = 2a$ 上方时，向下的合力为正，在其下方时则为负，所以质点一直加速至 $x = 2a$，之后便减速。

验算

- 一个巧妙的一致性检验：对 L 求导，得 $\dfrac{\mathrm{d}L}{\mathrm{d}x} = \dfrac{2x}{\sqrt{x^2-a^2}} - \dfrac{2a^2}{x\sqrt{x^2-a^2}} = \dfrac{2\sqrt{x^2-a^2}}{x} = 2\cos\alpha$。所以弹性势能的变化率 $T\dfrac{\mathrm{d}L}{\mathrm{d}x}$ 等于 $2T\cos\alpha$，即向上的力。(i) 的能量方法与 (ii) 的力学方法互相吻合。
- 以数值计算，(i) 中 $\lambda \approx 6.70\,mg$，(ii) 中 $\lambda \approx 2.65\,mg$。(ii) 中的弹性带较弱，这是合理的：在那里质点到达 $x = 2a$ 时仍在运动，并会继续越过该点，而在 (i) 中，较硬的弹性带已经令质点停下来。
- 由 $x = a$ 处的静止状态出发，以数值方法逐步求解运动方程 $m\ddot{x} = mg - 2T\cos\alpha$，取 λ 的第一个值时，最低点在 $x = 2.0000a$；取第二个值时，速率最大的位置在 $x = 2.0000a$。

常见错误

- 忘记了有一部分弹性带离开了圆柱：伸长量是 $2\sqrt{x^2-a^2}$ 减去失去的弧长 $2a\theta$，而不仅仅是两段笔直部分的长度。

- 由轴线而非释放点起量度下降的距离：质点下降的距离是 a，而不是 $2a$。
- 沿铅垂方向分解时，用了 $\sin\alpha$ 而不是 $\cos\alpha$；α 是由铅垂线起量度的，而 $\sin\alpha = a/x$。

STEP 之外

验算 d(弹性势能)/dx 等于回复力，其实是虚功原理（principle of virtual work）的一个例子：在保守系统（conservative system）中，任何方向上的力，等于势能沿该方向的变化率的负值。这种以一个坐标及其势能来描述系统的观点，正是拉格朗日力学（Lagrangian mechanics）的起点，你会在大学力学课程的第一或第二年遇到它。若想看对能量、力和势的严谨论述，可参阅 Kleppner and Kolenkow, *An Introduction to Mechanics*（力学入门教科书）。

详细解答

STEP I 2001，第 12 题

概率与统计 • 树形图

STEP I 2001，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/01-s1-q12

关键思路

一切取决于仔细记录谁知道甚么。一名学生要知道某人已听过谣言，只有两种情况：那人把谣言告诉了他，或者他把谣言告诉了那人；没有人会透露自己是从谁那里听来的。只要记下每名学生在每个阶段所知道的事，树形图（tree diagram）其实很小：每次分支都是在两个人之间选择，所以只有八条路径，而且结果它们的可能性全部相等。

解答

设置

以 A、B、C、D 表示四名学生。在任何阶段，学生 X 知道已听过谣言的人包括：X 自己、曾把谣言告诉 X 的人，以及 X 曾把谣言告诉的人。当 X 听到谣言时，X 会在这个集合以外的人当中均匀随机（uniformly at random）地选一人。如果这个集合已包含全部四名学生，谣言便在 X 处停止。

开始时，A 已告诉 C，所以 A 知道 $\{A, C\}$，而 C 知道 $\{A, C\}$。

树形图

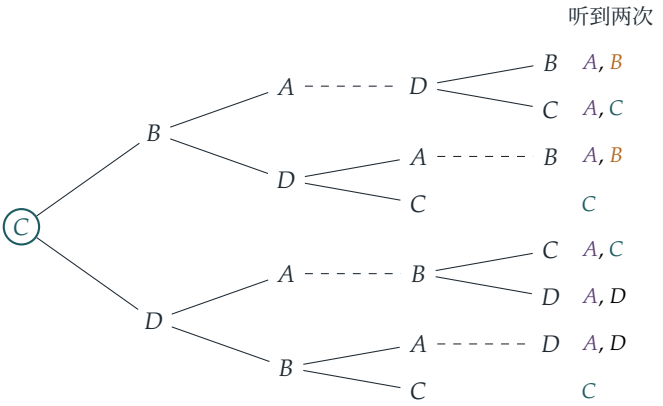
C 会略过 A，告诉 B 或 D，概率（probability）各为 $\frac{1}{2}$。先看分支 $C \to B$。这时 C 知道 $\{A, B, C\}$，而 B 知道 $\{B, C\}$（不包括 A，因为 C 没有说谣言从何而来）。所以 B 会告诉 A 或 D，概率各为 $\frac{1}{2}$。

- $B \to A$。Arthur 现在知道 $\{A, B, C\}$，所以他必须告诉 D，之后 Arthur 便知道人人都已听过。D 只知道 $\{A, D\}$，所以 D 会告诉 B 或 C，概率各为 $\frac{1}{2}$。无论是哪一个，听

者的集合都会齐全：B 本来已知道 $\{A,B,C\}$，C 亦然，而两人此时都得知 D 已听过。于是谣言停止。

- $B \to D$。这时 B 知道 $\{B,C,D\}$，而 D 知道 $\{B,D\}$，所以 D 会告诉 A 或 C。
  - $D \to C$：C 本来已知道 $\{A,B,C\}$，现在又得知 D 已听过，所以谣言停止。Arthur 始终没有再次听到谣言。
  - $D \to A$：Arthur 知道 $\{A,C,D\}$，所以他告诉 B，后者随即知道四人都已听过。谣言停止。

分支 $C \to D$ 的情况相同，只是 B 与 D 的角色互换。以下是整个树形图。除了以虚线表示的必然步骤（概率为 1）外，每条分支的概率都是 $\frac{1}{2}$。



每条通往树叶的路径都恰好包含三次概率为 $\frac{1}{2}$ 的选择（必然步骤只贡献因子 1），所以八个结果的概率各为 $\frac{1}{8}$。右方的标记显示在该路径上谁听到谣言两次。（在每条路径上，一名学生最多听到两次。）

> **想一想** 为甚么 B 刚从 C 那里听到谣言，却不知道 Arthur 已听过？

因为 C 不会说是谁告诉自己的。B 所能推断的，只是 C 已听过谣言。正是这一点令谣言得以继续流传：在 $C \to B \to A$ 这条分支上，Arthur 被告知一件他早已知道的事。

Arthur 再次听到谣言

除了两条短路径 $C \to B \to D \to C$ 和 $C \to D \to B \to C$ 之外，Arthur 在每条路径上都会再次听到谣言。所以

$$\mathrm{P}(\text{Arthur 再次听到谣言}) = 1 - 2 \times \tfrac{1}{8} = \tfrac{3}{4},$$

正如所求。

Bertha 与 Chandra

Bertha 只在路径 $C \to B \to A \to D \to B$ 和 $C \to B \to D \to A \to B$ 上听到两次，所以

$$\mathrm{P}(\text{Bertha 听到两次}) = \tfrac{2}{8} = \tfrac{1}{4}.$$

Chandra 在四条路径上听到两次：树形图的上下两半各有两条（$C \to B \to A \to D \to C$、$C \to B \to D \to C$、$C \to D \to A \to B \to C$ 和 $C \to D \to B \to C$），所以

$$\mathrm{P}(\text{Chandra 听到两次}) = \tfrac{4}{8} = \tfrac{1}{2}.$$

Chandra 比 Bertha 更有可能再次听到谣言，是因为 Chandra 很早便听到：在传话链中较后才被告知的人并不知道 Chandra 已听过，所以直到最后 Chandra 仍然是可能被告知的对象。

验算

根据 B 与 D 之间的对称性（symmetry），Delilah 听到两次的概率也应是 $\frac{1}{4}$，而树形图中确有两条这样的路径。各结果的概率之和正确：八片等可能的树叶，总概率为 1。以电脑对整个过程作精确枚举，得出 $\frac{3}{4}$、$\frac{1}{4}$ 和 $\frac{1}{2}$；而模拟（simulation）200 000 次谣言传播，则得出 0.750、0.250 和 0.501。

常见错误

- 让听者知道传话链中更早的每一个人。其实传递下去的只有告知者的身分；假如人人都掌握全部信息，谣言便永远不会回到 Arthur 那里。
- 忘记告知者同样会得知一些事情：告诉 Y 之后，X 便知道 Y 已听过谣言，而 Arthur 的步骤之所以是必然的，正是这个缘故。
- Arthur 一再次听到谣言便把树形图停下。其后的分支对 Bertha 和 Chandra 的计算是必需的。

STEP 之外

这个谣言是四名学生之间的一个随机漫步（random walk），其规则取决于过往历史，而这里以每名学生所知道的事来记录历史。只要把状态设置得足够丰富，使下一步只取决于当前状态，它便成为一条马可夫链（Markov chain），亦即随机过程（random processes）入门课程的核心对象；谣言模型和传染病模型是它的经典应用。Grimmett and Welsh, *Probability: An Introduction*（概率论入门）对马可夫链的讲解，程度恰到好处。

详细解答

STEP I 2001，第 13 题

概率与统计 · 条件泊松分布

STEP I 2001，第 13 题

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

step.maths.org/questions/01-s1-q13

关键思路

以打破碟子的总数作为条件，便可把一道泊松（Poisson）问题化为二项（binomial）问题。假如四名学生各自独立地以相同的比率打破碟子，那么一旦知道总共打破了多少只碟子，每一只破碟由四人中任何一人打破的可能性都相等，而且与其他破碟互相独立。因此，数学家在破碟中所占的数目服从二项分布，而未知的 λ 亦随之消失。

解答

设置

设 M 为数学家打破的碟子数目，R 为其余三名学生打破的碟子总数。题目指出每名学生打破的碟子数目均服从 $\mathrm{Po}(\lambda)$，且彼此独立。所以 $M \sim \mathrm{Po}(\lambda)$，而 R 是三个独立的 $\mathrm{Po}(\lambda)$ 变量之和，故服从 $\mathrm{Po}(3\lambda)$。此外，M 与 R 独立（independent），因为 R 只取决于其余三名学生。总数为 $T = M + R \sim \mathrm{Po}(4\lambda)$。

? **想一想** 为甚么独立泊松变量之和仍然服从泊松分布?

对于独立的 $X \sim \mathrm{Po}(\alpha)$ 和 $Y \sim \mathrm{Po}(\beta)$，就 n 的各种拆分方式求和：

$$\mathrm{P}(X+Y=n) = \sum_{k=0}^{n} \frac{\mathrm{e}^{-\alpha}\alpha^k}{k!} \cdot \frac{\mathrm{e}^{-\beta}\beta^{n-k}}{(n-k)!} = \frac{\mathrm{e}^{-(\alpha+\beta)}}{n!} \sum_{k=0}^{n} \binom{n}{k} \alpha^k \beta^{n-k} = \frac{\mathrm{e}^{-(\alpha+\beta)}(\alpha+\beta)^n}{n!},$$

最后一步用了二项式定理（binomial theorem）。所以 $X + Y \sim \mathrm{Po}(\alpha + \beta)$。应用这个结果两次，得 $R \sim \mathrm{Po}(3\lambda)$；再应用一次，得 $T \sim \mathrm{Po}(4\lambda)$。

M 的条件分布

对于 $m = 0, 1, \ldots, 5$，事件 $\{M = m,\ T = 5\}$ 与 $\{M = m,\ R = 5 - m\}$ 相同。由于 M 与 R 独立，

$$\begin{aligned}\mathrm{P}(M = m \mid T = 5) &= \frac{\mathrm{P}(M = m)\,\mathrm{P}(R = 5 - m)}{\mathrm{P}(T = 5)} \\ &= \frac{\dfrac{\mathrm{e}^{-\lambda}\lambda^m}{m!} \cdot \dfrac{\mathrm{e}^{-3\lambda}(3\lambda)^{5-m}}{(5-m)!}}{\dfrac{\mathrm{e}^{-4\lambda}(4\lambda)^5}{5!}} \\ &= \frac{5!}{m!\,(5-m)!} \cdot \frac{3^{5-m}}{4^5} = \binom{5}{m}\left(\tfrac{1}{4}\right)^m \left(\tfrac{3}{4}\right)^{5-m}.\end{aligned}$$

指数项互相约去，λ 的所有幂次也一样（分子和分母各有 λ^5）。所以，在 $T = 5$ 的条件下，不论 λ 取甚么值，都有 $M \sim \mathrm{B}(5, \frac{1}{4})$。

答案

把 $m = 3, 4, 5$ 的二项概率相加：

$$\begin{aligned}\mathrm{P}(M \geq 3 \mid T = 5) &= \binom{5}{3}\frac{3^2}{4^5} + \binom{5}{4}\frac{3}{4^5} + \binom{5}{5}\frac{1}{4^5} \\ &= \frac{10 \times 9 + 5 \times 3 + 1}{1024} = \frac{106}{1024} = \frac{53}{512}.\end{aligned}$$

所以，五只破碟之中有三只或以上由数学家打破的概率是 $\dfrac{53}{512} \approx 0.104$。

验算

- 六个条件概率（conditional probability）$\binom{5}{m}3^{5-m}/4^5$ 之和为 $(1+3)^5/4^5 = 1$，正如所料。
- 对称性：按同样的论证，每名学生在五只破碟中所占数目的期望值（expected value）为 $5 \times \frac{1}{4} = \frac{5}{4}$，而四人所占数目之和为 5。
- 我们直接由四个泊松分布计算 $\mathrm{P}(M \geq 3 \mid T = 5)$（就五只破碟在四名学生之间的每一种分配方式求和），取 $\lambda = 0.3, 1, 2.7$ 和 5，每次都得到 $0.103515625 = 53/512$。另外模拟了 4×10^5 轮洗碗，在恰好打破五只碟子的各轮之中，所得比例约为 0.105。

常见错误

- 答案中仍留有 λ。它必定会约去；如果约不去，就表示取条件的步骤出了错。

- 除以 $\mathrm{P}(M+R=5)$ 时，把 R 当作 $\mathrm{Po}(\lambda)$ 而不是 $\mathrm{Po}(3\lambda)$，这等于只有两名学生。
- 漏掉二项式系数，或把事件当作“恰好三只”而不是“三只或以上”。

STEP 之外

这是一个一般事实的小例子：若 $X_1,\ldots,X_k$ 是独立的泊松变量，平均值分别为 $\lambda_1,\ldots,\lambda_k$，则在它们的和为 n 的条件下，向量 $(X_1,\ldots,X_k)$ 服从多项分布（multinomial distribution），其概率为 $\lambda_i/\sum\lambda_j$。反过来看，这个事实说明：若事件的数目服从 $\mathrm{Po}(\mu)$，而每个事件各自独立地标记为 k 种类型之一，则各类型的数目是互相独立的泊松变量。这种“泊松分拆”（Poisson splitting）是研究泊松过程（Poisson process）的基本工具。Blitzstein 与 Hwang 合著的 *Introduction to Probability*（概率论入门教材）以大量例子讨论了这两个方向。

详细解答

STEP I 2001，第 14 题

概率与统计 • 随机排序

STEP I 2001，第 14 题

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

step.maths.org/questions/01-s1-q14

本题在 *Beyond STEP* 第 12.12 节何时停止寻找中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch12/secretary.html

关键思路

不要考虑整个面试次序，只需考虑个别候选人排在哪个位置。首 n 人中的最佳者就是全体最佳者，当且仅当最优秀的候选人在首 n 名接受面试的人之中；而根据对称性（symmetry），最优秀的候选人落在 N 个位置中任何一个的可能性均等。第 (ii) 部分只需多考虑一种情况，或者巧妙地利用补事件（complement）。

解答

按名次为候选人编号：候选人 1 是全体最佳，候选人 2 是第二佳，如此类推。由于全部 $N!$ 种次序的可能性均等，每位候选人占据任何指定位置的可能性均等，而首 n 场面试所见的候选人组成的集合，是 $\binom{N}{n}$ 个大小为 n 的子集（subset）中任何一个的可能性亦均等。以下一律假设 $1 \leq n \leq N$。

第 (i) 部分

设 A 为候选人 1 在首 n 名接受面试的人之中这一事件（event）。若 A 发生，候选人 1 必定是首 n 人中的最佳者；若不发生，首 n 人中的最佳者就是比候选人 1 逊色的人。因此我们所求的事件正正就是 A。

候选人 1 落在 N 个位置中每一个的可能性均等，而其中 n 个位置属于首 n 个，所以

$$\mathrm{P}(\text{首}n\text{ 人中的最佳者是全体最佳}) = \frac{n}{N}.$$

第 (ii) 部分

现在取 $N \geq 2$，使第二佳的候选人存在。首 n 人中的最佳者是全体最佳或第二佳，当且仅当

- 候选人 1 在首 n 人之中（事件 A），或
- 候选人 1 不在首 n 人之中，但候选人 2 在。

> **想一想** 在第二种情况中，我们怎样知道首 n 人中的最佳者是候选人 2，而不是更优秀的人?

比候选人 2 优秀的只有候选人 1，而在这种情况下他并不在较早接受面试的人之中。反过来说，若 1 和 2 都不在首 n 人之中，则首 n 人中的最佳者名次为 3 或以下。所以上述两种情况合起来正正就是我们所求的事件，而且两者互斥（mutually exclusive）。

由第 (i) 部分，第一种情况的概率为 $\frac{n}{N}$。至于第二种情况，$\mathrm{P}(A') = \frac{N-n}{N}$，而已知候选人 1 在最后 $N-n$ 个位置的其中之一时，候选人 2 落在余下 $N-1$ 个位置中任何一个的可能性均等，其中 n 个属于首 n 个位置。因此

$$\begin{aligned}\mathrm{P}(\text{首}n\text{ 人中的最佳者是全体第一或第二}) &= \frac{n}{N} + \frac{N-n}{N}\cdot\frac{n}{N-1}\\ &= \frac{n\big[(N-1)+(N-n)\big]}{N(N-1)}\\ &= \frac{n(2N-n-1)}{N(N-1)}.\end{aligned}$$

*另一方法*是利用补事件：所求事件不发生，当且仅当较早接受面试的 n 名候选人全部来自名次为 3 至 N 的 $N-2$ 名候选人，其概率为

$$\frac{\binom{N-2}{n}}{\binom{N}{n}} = \frac{(N-n)(N-n-1)}{N(N-1)},$$

而 $N(N-1)-(N-n)(N-n-1) = 2Nn-n^2-n = n(2N-n-1)$，得出相同的答案。

$N=4$、$n=2$ 的情况

按公式，(i) 的答案是 $\frac{2}{4}=\frac{1}{2}$，(ii) 的答案是 $\frac{2\cdot 5}{4\cdot 3}=\frac{5}{6}$。

要列出各种可能性，留意到重要的只是哪两名候选人最先接受面试。$\binom{4}{2}=6$ 个组合中的

每一个，都恰好由 $2! \times 2! = 4$ 种次序产生（全部共 24 种次序；先排好首两人，再排好后两人），所以这六个组合的可能性均等：

最先接受面试的两人	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{2,3\}$	$\{2,4\}$	$\{3,4\}$
其中的最佳者	1	1	1	2	2	3

首两人中的最佳者是全体最佳的情况占 3 种（共 6 种），概率为 $\frac{1}{2}$；是全体最佳或第二佳的情况占 5 种（共 6 种），概率为 $\frac{5}{6}$。两者都与公式吻合。

验算

- $n = N$：我们见过所有人，所以两个概率都应该是 1。果然 $\frac{N}{N} = 1$ 而 $\frac{N(N-1)}{N(N-1)} = 1$。
- $n = 1$：第一名候选人是均匀随机选出的一人，所以 (ii) 的答案应该是 $\frac{2}{N}$；公式给出 $\frac{2N-2}{N(N-1)} = \frac{2}{N}$。
- (ii) 的答案不小于 (i) 的答案，因为 $2N - n - 1 \geq N - 1$。
- 把全部 $N!$ 种次序逐一列举，对 $N \leq 7$ 及每个 n 的结果都与两条公式完全吻合（例如 $N = 5$、$n = 2$ 给出 $\frac{2}{5}$ 和 $\frac{7}{10}$）。

常见错误

- 在 (ii) 中忘记了“全体第二佳”只在全体最佳者*不*在首 n 人之中时才算数；直接把 $\frac{n}{N} + \frac{n}{N}$ 相加会重复计算。
- 以候选人 1 的位置为条件（conditioning）之后，仍对候选人 2 使用 $\frac{n}{N}$：候选人 1 的位置一旦确定，就只剩下 $N - 1$ 个位置。

STEP 之外

这正是著名的*秘书问题*（secretary problem）的设置：假如你必须即场决定取录或拒绝每名候选人，最佳策略是先观察首 n 人，然后取录其后第一名比他们全部都优秀的候选人；把第 (i) 部分的论证加以推广，可以证明最优的 n 大约是 N/e，成功的概率约为 $1/e$。这里所用的对称性论证（“每名候选人落在任何位置的可能性均等”）是研究随机排列（random permutation）的起点。Blitzstein and Hwang 的 *Introduction to Probability*（概率论入门教科书）对这类对称性论证有十分精彩的阐述。

详细解答

STEP II 2001，第 1 题

纯数学 · 二项式近似

> **STEP II 2001，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/01-s2-q1

关键思路

当 x 很小时，$(1-x)^{1/2}$ 的二项级数（binomial series）收敛（converge）得很快，所以只取首几项，便可得到极佳的有理近似（rational approximation）。诀窍在于如何选取 x：我们希望 $1-x$ 等于某个平方数乘以所求平方根的那个数，再除以十的某次幂。这样，$1-x$ 的平方根就是所求平方根的简单倍数。由于 $99 = 9 \times 11$ 和 $9999 = 9 \times 1111$，该怎样选几乎是不言而喻的。

解答

二次近似

由二项展开（binomial expansion），当 $|x| < 1$ 时，

$$\begin{aligned}(1-x)^{1/2} &= 1 + \tfrac{1}{2}(-x) + \frac{\frac{1}{2}\cdot\left(-\frac{1}{2}\right)}{2!}(-x)^2 + \frac{\frac{1}{2}\cdot\left(-\frac{1}{2}\right)\cdot\left(-\frac{3}{2}\right)}{3!}(-x)^3 + \cdots \\ &= 1 - \tfrac{1}{2}x - \tfrac{1}{8}x^2 - \tfrac{1}{16}x^3 - \cdots .\end{aligned}$$

所以当 x 很小时，

$$\sqrt{1-x} \approx 1 - \frac{x}{2} - \frac{x^2}{8}.$$

我们同时求得第一个被略去的项 $-\frac{1}{16}x^3$，稍后估计误差（error）时会用到。事实上，首项之后的每一项都是负数（当 $n \geq 1$ 时，乘积 $\frac{1}{2}\cdot(-\frac{1}{2})\cdots(\frac{1}{2}-n+1)$ 有 $n-1$ 个负因子，而 $(-x)^n$ 则带来 $(-1)^n$，合起来刚好剩下一个负号），所以当 $0 < x < 1$ 时，这个二次式是一个高估值（overestimate）。

第 (i) 部分

取 $x = \frac{1}{100}$，则

$$\sqrt{1-x} = \sqrt{\frac{99}{100}} = \frac{\sqrt{9\times 11}}{10} = \frac{3\sqrt{11}}{10}, \qquad \text{所以} \qquad \sqrt{11} = \frac{10}{3}\sqrt{1-\tfrac{1}{100}}.$$

利用二次近似，

$$\begin{aligned}\sqrt{11} &\approx \frac{10}{3}\left(1 - \frac{1}{200} - \frac{1}{80\,000}\right) = \frac{10}{3}\cdot\frac{80\,000 - 400 - 1}{80\,000}\\ &= \frac{10}{3}\cdot\frac{79\,599}{80\,000} = \frac{79\,599}{24\,000},\end{aligned}$$

这正是题目所要求的结果。

> **想一想** 第一个被略去的项是 $-\frac{1}{16}x^3$。这就是 $\sqrt{11}$ 的误差吗?

不完全是：那是 $\sqrt{1-x}$ 的近似值的误差，而我们之后还乘上了 $\frac{10}{3}$。因此 $\sqrt{11}$ 的误差约为

$$\frac{10}{3}\times\frac{1}{16}\left(\frac{1}{100}\right)^3 = \frac{10}{48\times 10^6} = \frac{1}{4\,800\,000} \approx 2.08\times 10^{-7}.$$

所以，准确至 1 位有效数字（significant figure），误差为 2×10^{-7}，而 $\frac{79\,599}{24\,000}$ 略大于 $\sqrt{11}$。

第 (ii) 部分

我们照搬第 (i) 部分的技巧。由于 $1111\times 9 = 9999 = 10^4 - 1$，取 $x = \frac{1}{10\,000}$：

$$\sqrt{1-x} = \sqrt{\frac{9999}{10\,000}} = \frac{3\sqrt{1111}}{100}, \qquad \text{所以} \qquad \sqrt{1111} = \frac{100}{3}\sqrt{1-\tfrac{1}{10\,000}}.$$

于是

$$\begin{aligned}\sqrt{1111} &\approx \frac{100}{3}\left(1 - \frac{1}{20\,000} - \frac{1}{8\times 10^8}\right) = \frac{100}{3}\cdot\frac{800\,000\,000 - 40\,000 - 1}{8\times 10^8}\\ &= \frac{100}{3}\cdot\frac{799\,959\,999}{8\times 10^8} = \frac{799\,959\,999}{24\,000\,000} = \frac{266\,653\,333}{8\,000\,000},\end{aligned}$$

其中最后一步约去了因子 3（799 959 999 的数字之和是 75，可被 3 整除）。

误差约为第一个被略去的项的 $\frac{100}{3}$ 倍：

$$\frac{100}{3}\times\frac{1}{16}\left(\frac{1}{10^4}\right)^3 = \frac{100}{48\times 10^{12}} \approx 2.08\times 10^{-12},$$

即大约 2×10^{-12}，符合题目的要求。所以一个合适的有理数是

$$\sqrt{1111} \approx \frac{266\,653\,333}{8\,000\,000}\left(= \frac{799\,959\,999}{24\,000\,000}\right).$$

验算

- $79\,599/24\,000 = 3.316625$，而 $\sqrt{11} = 3.31662479\ldots$；两者相差 2.096×10^{-7}，与估计值 2.08×10^{-7} 吻合，亦印证了这个近似值偏大。
- $266\,653\,333/8\,000\,000 = 33.331666625$，而 $\sqrt{1111} = 33.33166662291\ldots$；我们以高精度准确计算两者之差，得出 2.0835×10^{-12}。
- 第 (ii) 部分的估计更为准确，因为当 x 较小时，下一个被略去的项 $-\frac{5}{128}x^4$ 相对于 x^3 项也较小。

常见错误

- 忘记把被略去的项乘以 $\frac{10}{3}$（或 $\frac{100}{3}$），以致得出的误差小了十倍或一百倍。
- 弄错 x^2 系数的正负号：由于是 $(-x)^2$，该项为 $\frac{\frac{1}{2}\cdot(-\frac{1}{2})}{2}x^2 = -\frac{1}{8}x^2$，而不是 $+\frac{1}{8}x^2$。
- 尝试选取 x 的值，使 $1-x$ 等于 1111 除以十的某次幂。这样的值并不小，级数收敛得太慢，根本帮不上忙。

STEP 之外

“误差大约等于第一个被略去的项”这个说法，可以用泰勒定理（Taylor's theorem）和拉格朗日余项（Lagrange form of the remainder）严格地表述：它给出误差的准确公式，当中含有一个位于 $(0, x)$ 内的未知点。由此我们可以得到确切的上界和下界（upper and lower bounds），而不仅仅是估计。把这些误差界严格地建立起来，是大学第一门分析（analysis）课程的标准内容；参阅 Abbott 的 *Understanding Analysis*。

详细解答

STEP II 2001，第 2 题

纯数学 · 整数部分的求和

STEP II 2001，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/01-s2-q2

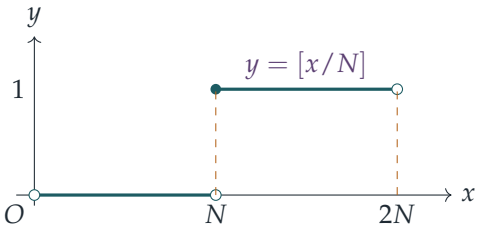
关键思路

指数 $[k/N]$ 看似棘手，但当 k 由 1 取到 $2N$ 时，它其实只有三个可能的值，而草图正好显示它在哪里跳升。弄清这一点之后，正负号 $(-1)^{[k/N]}$ 在 k 的三个区段上各自保持不变，于是每个和式都可以按区段拆开：第 (i) 部分拆成等差级数（arithmetic series），第 (ii) 部分拆成等比级数（geometric series）。唯一要小心的，是各区段的端点。

解答

草图

当 $0 < x < N$ 时，$0 < x/N < 1$，所以 $[x/N] = 0$。当 $N \le x < 2N$ 时，$1 \le x/N < 2$，所以 $[x/N] = 1$。图像是一级阶梯：跳升发生在 $x = N$，函数在该点已经取值 1；右端点 $x = 2N$ 并不包括在内（函数在该处的值会是 2）。



实心点表示包括该点，空心点表示不包括该点。

计算和式时，我们需要的是整数值 $k = 1, 2, \ldots, 2N$。从图像读出（并补上端点 $k = 2N$）：

$$[k/N] = \begin{cases} 0, & 1 \le k \le N-1, \\ 1, & N \le k \le 2N-1, \\ 2, & k = 2N. \end{cases}$$

因此，$(-1)^{[k/N]}$ 在 $k \le N-1$ 时是 $+1$，在 $N \le k \le 2N-1$ 时是 -1，而在 $k = 2N$ 这单独一项又变回 $+1$。（当 $N = 1$ 时，第一个区段是空的；由于空和等于 0，以下所有公式仍然成立。）

> **想一想** 为甚么不能直接说首 N 项取 $+$ 号、其后 N 项取 $-$ 号?

因为跳升恰好发生在 N 的倍数上：$k = N$ 这一项已经有 $[k/N] = 1$，而最后一项 $k = 2N$ 有 $[k/N] = 2$，所以它是正的。把这两个端点处理正确，就是整道题的关键。

第 (i) 部分

按三个区段拆开，

$$\sum_{k=1}^{2N} (-1)^{[k/N]} k = \sum_{k=1}^{N-1} k \;-\; \sum_{k=N}^{2N-1} k \;+\; 2N.$$

第一个和是 $\frac{1}{2}(N-1)N$。第二个和共有 N 项，由 N 到 $2N-1$，所以它等于 N 乘以首项与末项的平均值：$\frac{1}{2}N(3N-1)$。因此

$$\begin{aligned} \sum_{k=1}^{2N} (-1)^{[k/N]} k &= \tfrac{1}{2}\left(N^2 - N\right) - \tfrac{1}{2}\left(3N^2 - N\right) + 2N \\ &= -N^2 + 2N = 2N - N^2, \end{aligned}$$

这正是所要证明的结果。

第 (ii) 部分

用同样的拆法，得

$$S_N = \sum_{k=1}^{N-1} 2^{-k} \;-\; \sum_{k=N}^{2N-1} 2^{-k} \;+\; 2^{-2N}.$$

两个和都是公比为 $\frac{1}{2}$ 的等比级数。利用 $\sum_{k=a}^{b} 2^{-k} = 2^{-a}\left(1 - 2^{-(b-a+1)}\right) \big/ \frac{1}{2} = 2^{1-a} - 2^{-b}$，

$$\sum_{k=1}^{N-1} 2^{-k} = 1 - 2^{1-N},$$

$$\sum_{k=N}^{2N-1} 2^{-k} = 2^{1-N} - 2^{1-2N}.$$

所以

$$\begin{aligned} S_N &= 1 - 2^{1-N} - 2^{1-N} + 2^{1-2N} + 2^{-2N} \\ &= 1 - 4\cdot 2^{-N} + 3\cdot 2^{-2N}. \end{aligned}$$

设 $t = 2^{-N}$，上式就是 $1 - 4t + 3t^2$，可以漂亮地因式分解：

$$\boxed{S_N = 1 - \frac{4}{2^N} + \frac{3}{4^N} = \left(1 - \frac{1}{2^N}\right)\left(1 - \frac{3}{2^N}\right).}$$

当 $N \to \infty$ 时，$2^{-N} \to 0$ 且 $4^{-N} \to 0$，所以

$$\lim_{N\to\infty} S_N = 1.$$

这个结果合情合理：随着 N 增大，正号区段 $\frac{1}{2} + \frac{1}{4} + \cdots$ 愈来愈长，其和趋近 1；而负号区段的起点已经排得很后，它的总和大约是 2^{1-N}，可以忽略不计。

验算

- $N = 1$：各项为 $k = 1$（此时 $[1] = 1$）和 $k = 2$（此时 $[2] = 2$）。第 (i) 部分给出 $-1 + 2 = 1 = 2 - 1$，第 (ii) 部分给出 $-\frac{1}{2} + \frac{1}{4} = -\frac{1}{4}$，与 $(1 - \frac{1}{2})(1 - \frac{3}{2}) = -\frac{1}{4}$ 相符。
- $N = 2$：对应 $k = 1, 2, 3, 4$ 的正负号依次为 $+, -, -, +$。于是 $1 - 2 - 3 + 4 = 0 = 4 - 4$，而 $\frac{1}{2} - \frac{1}{4} - \frac{1}{8} + \frac{1}{16} = \frac{3}{16} = \frac{3}{4}\cdot\frac{1}{4}$。
- 我们以分数作精确计算，对 N 由 1 到 59 的每一个值验证了这两条公式。例如 $S_{50} \approx 0.999\,999\,999\,999\,996$。

常见错误

- 把 $k = N$ 归入正号区段，或者忘记最后一项 $k = 2N$ 的指数是 2，因而是正的。
- 数错 $\sum_{k=N}^{2N-1}$ 的项数：它共有 N 项，而不是 $N - 1$ 项。

STEP 之外

整数部分函数（integer-part function，又称下取整函数 floor function）在数论（number theory）中随处可见：经典的例子是勒让德公式（Legendre's formula），它指出整除 $n!$ 的质数 p 的最高次幂是 $\sum_{j\ge 1}[n/p^j]$。像第 (ii) 部分那样正负号按区段变化的和式，也是一个浅易的入门例子，让我们初步接触级数各项的重排（rearranging）与分组（grouping），而分析（analysis）课程会把这些做法严格化。想看看下取整函数如何大派用场，Davenport 的 *The Higher Arithmetic*（一本数论入门名著）是个绝佳的起点。

详细解答

STEP II 2001，第 3 题

纯数学 · 欧几里得几何

STEP II 2001，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/01-s2-q3

关键思路

把长方体（cuboid）放进坐标系，令其中一个顶点位于原点，各棱沿坐标轴方向。这样每个顶点的位置向量（position vector）都一目了然，对角线只是这些向量之差，而两条对角线之间的夹角可直接由纯量积（scalar product）求得。体积与表面积之比是 λ 的一个简单有理函数（rational function），最后一部分则把关于 R 的条件转化为关于 λ 的条件，从而把前两个结果连系起来。

解答

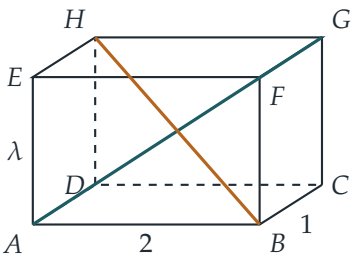
对角线之间的夹角

由于棱 AE、BF、CG、DH 垂直于各个面，各顶点处的角全是直角，因此可取坐标轴沿 AB、AD 和 AE 方向，并以 A 为原点：

$$A=(0,0,0),\quad B=(2,0,0),\quad C=(2,1,0),\quad D=(0,1,0),$$

而顶面就是把底面升高 λ：

$$E=(0,0,\lambda),\quad F=(2,0,\lambda),\quad G=(2,1,\lambda),\quad H=(0,1,\lambda).$$



两条对角线的方向向量（direction vector）为

$$\overrightarrow{AG} = \begin{pmatrix} 2 \\ 1 \\ \lambda \end{pmatrix}, \qquad \overrightarrow{BH} = \begin{pmatrix} -2 \\ 1 \\ \lambda \end{pmatrix},$$

长度均为 $\sqrt{5+\lambda^2}$，而且

$$\overrightarrow{AG} \cdot \overrightarrow{BH} = -4 + 1 + \lambda^2 = \lambda^2 - 3.$$

设 θ 为这两个向量之间的夹角，则

$$\cos\theta = \frac{\lambda^2 - 3}{5 + \lambda^2}.$$

想一想 绝对值符号（modulus sign）是从哪里来的?

两条相交的直线形成两个角 θ 和 $\pi - \theta$，它们的余弦互为相反数。其中的锐角（当 $\lambda^2 = 3$ 时则为直角）就是余弦非负的那一个，所以

$$\cos\alpha = |\cos\theta| = \left|\frac{3 - \lambda^2}{5 + \lambda^2}\right|,$$

正是所求。

比值 R

体积为 $2 \times 1 \times \lambda = 2\lambda$。表面由面积分别为 2×1、$2 \times \lambda$ 和 $1 \times \lambda$ 的面各两个组成，所以表面积为 $2(2 + 3\lambda)$。因此

$$R = \frac{2\lambda}{2(2+3\lambda)} = \frac{\lambda}{2+3\lambda}.$$

由于 $\lambda > 0$，分母为正数，于是

$$R < \tfrac{1}{3} \iff 3\lambda < 2 + 3\lambda \iff 0 < 2,$$

而这显然成立。所以对任何可能的 λ，均有 $R < \frac{1}{3}$。（当 $\lambda \to \infty$ 时，$R \to \frac{1}{3}$，所以这个界不能再改进。）

最后一部分

同样因为 $2 + 3\lambda > 0$，可以把不等式两边同乘分母：

$$R \geq \tfrac{1}{4} \iff 4\lambda \geq 2 + 3\lambda \iff \lambda \geq 2.$$

现设 $\lambda \geq 2$。则 $\lambda^2 \geq 4 > 3$，所以 $\lambda^2 - 3 > 0$，绝对值符号可以去掉：

$$\cos\alpha = \frac{\lambda^2 - 3}{\lambda^2 + 5} = 1 - \frac{8}{\lambda^2 + 5}.$$

写成这个形式，便可看出 $\cos\alpha$ 随 λ^2 增大而增大。由 $\lambda^2 \geq 4$ 得 $\lambda^2 + 5 \geq 9$，所以

$$\cos\alpha \geq 1 - \frac{8}{9} = \frac{1}{9}.$$

α 和 $\arccos\frac{1}{9}$ 都在 $[0, \frac{\pi}{2}]$ 内，而余弦在这个区间上严格递减（strictly decreasing），所以由 $\cos\alpha \geq \frac{1}{9}$ 可得

$$\alpha \leq \arccos\tfrac{1}{9},$$

这正是要证明的结果。等号成立当且仅当 $\lambda = 2$，即 $R = \frac{1}{4}$。

验算

- 当 $\lambda^2 = 3$ 时，公式给出 $\cos\alpha = 0$。验算：$\overrightarrow{AG} \cdot \overrightarrow{BH} = -4 + 1 + 3 = 0$，所以两条对角线确实互相垂直。
- 当 $\lambda \to \infty$ 时，长方体变成又高又窄的棱柱（prism），两条对角线都变得接近垂直，所以 $\alpha \to 0$；公式给出 $\cos\alpha \to 1$，与此吻合。
- 当 $\lambda = 2$ 时：体积为 4，表面积为 $2(2 + 4 + 2) = 16$，所以 $R = \frac{1}{4}$，而 $\cos\alpha = \frac{1}{9}$，所以 $\alpha \approx 83.6°$。以 10^5 个随机的 λ 值作数值检验，结果证实了夹角公式、$R < \frac{1}{3}$，以及每当 $R \geq \frac{1}{4}$ 时 $\alpha \leq \arccos\frac{1}{9}$。

常见错误

- 忘记加绝对值符号：当 $\lambda^2 < 3$ 时，向量 $\overrightarrow{AG}$ 与 $\overrightarrow{BH}$ 所成的是钝角，而所求的锐角是它的补角。
- 把最后一个不等式的方向弄反：余弦较大，角就较小，因为余弦在 $[0, \frac{\pi}{2}]$ 上递减。

STEP 之外

把物体放进坐标系，再用纯量积求角度，是向量几何的基本工具；在大学一年级的线性代数（linear algebra）课程中，它会推广为任意维数的内积（inner product）：$\mathbb{R}^n$ 中两个向量之间的夹角是用 $\cos\theta = \mathbf{u}\cdot\mathbf{v}/(|\mathbf{u}||\mathbf{v}|)$ 来定义的，而柯西－施瓦茨不等式（Cauchy–Schwarz inequality）保证了这个值落在 $[-1, 1]$ 之内。有一道不错的练习：证明 n 维单位立方体的长对角线与每条棱所成的角的余弦为 $1/\sqrt{n}$，所以当 n 增大时，这个角趋向直角。Strang 的 *Introduction to Linear Algebra*（线性代数入门）把这些概念讲解得十分清楚。

详细解答

STEP II 2001，第 4 题

纯数学 · 倍角方程

STEP II 2001，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/01-s2-q4

关键思路

f 的每一项都含有因子 $\sin x$：$\sin 2x = 2\sin x\cos x$，而题目的提示对 $\sin 3x$ 也有同样的作用。抽出 $\sin x$ 之后，剩下的是一个关于 $\cos x$ 的二次式（quadratic），而 P、Q、R 所满足的条件，正好就是“这个二次式的判别式（discriminant）为负”。第二部分是同一个函数，只是取 $P = 1$、$Q = 1$、$R = -1$，并把 x 换成 $2nx$。这时二次式确实有根，功夫就在于找出哪一个根最接近容许范围的上端。

解答

第一个结果

设 $c = \cos x$。利用 $\sin 2x = 2\sin x\cos x$ 及题目给出的 $\sin 3x$ 恒等式（identity），

$$\begin{aligned} \mathrm{f}(x) &= P\sin x + 2Q\sin x\cos x + R\sin x\,(4\cos^2 x - 1) \\ &= \sin x\left(4Rc^2 + 2Qc + (P - R)\right). \end{aligned}$$

所以 $\mathrm{f}(x) = 0$ 当且仅当 $\sin x = 0$ 或 $4Rc^2 + 2Qc + (P - R) = 0$。

首先，$R \neq 0$：若 $R = 0$，条件便会变成 $Q^2 < 0$，这是不可能的。因此括号内确实是一个关于 c 的二次式，其判别式为

$$(2Q)^2 - 4\cdot 4R\,(P - R) = 4\big(Q^2 - 4R(P - R)\big) < 0$$

这由假设可得。判别式为负的实系数二次式没有实根，所以对任何实数 c，括号内的式子都不会等于零，当 $c = \cos x$ 时自然也不例外。故此 $\mathrm{f}(x) = 0$ 必然导致 $\sin x = 0$，即 $x = m\pi$，其中 m 为整数。反过来，每一个 $x = m\pi$ 都是 f 的零点（zero），所以这些正好就是全部零点，结果得证。

g 的最大根

设 $y = 2nx$。那么 $\mathrm{g}(x) = \sin y + \sin 2y - \sin 3y$，即是取 $P = 1$、$Q = 1$、$R = -1$ 时的 $\mathrm{f}(y)$。这里 $Q^2 = 1$ 并不小于 $4R(P - R) = -8$，所以第一个结果并不适用，而二次式有实根。设 $c = \cos y$，由上面的因式分解（factorisation）可得

$$\mathrm{g}(x) = \sin y\,(-4c^2 + 2c + 2) = -2\sin y\,(2c + 1)(c - 1).$$

所以 $\mathrm{g}(x) = 0$ 当且仅当

$$\sin y = 0, \qquad \cos y = 1, \qquad \text{或} \qquad \cos y = -\tfrac{1}{2}.$$

$\cos y = 1$ 的情况已经包含在 $\sin y = 0$ 之内。因此以 y 表示的根为

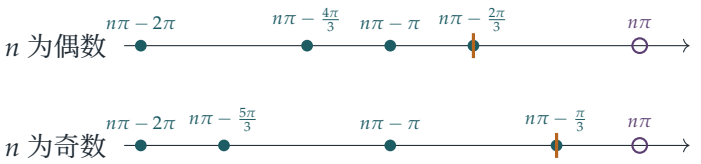
$$y = k\pi \quad \text{或} \quad y = 2k\pi \pm \tfrac{2\pi}{3} \qquad (k \in \mathbb{Z}).$$

在每一个周期（period）$[2k\pi, 2k\pi + 2\pi)$ 之内，根位于 $2k\pi$ 加上 0、$\frac{2\pi}{3}$、π、$\frac{4\pi}{3}$ 之处。

范围 $0 < x < \frac{1}{2}\pi$ 变成 $0 < y < n\pi$，而右端点 $y = n\pi$ 本身是一个根，却被排除在外。所以我们要找的是严格小于 $n\pi$ 的最大根，而关键只在于 $n\pi$ 在其所属周期内的位置。

> **想一想** 为甚么 n 的奇偶性（parity）有影响?

根的分布规律每隔 2π 重复一次，而不是每隔 π。若 n 为偶数，$n\pi$ 是 2π 的倍数；若 n 为奇数，$n\pi$ 是 π 的奇数倍。在这两种情况下，紧接在 $n\pi$ 之下的根，其间距并不相同。



（g 以 y 表示的根，范围为 $n\pi$ 之下的最后一段 2π，按比例绘画。空心圆圈标示被排除的端点，琥珀色的短线标示最大根。）

n 为偶数。这时 $n\pi = 2k\pi$，其中 $k = \frac{n}{2} \geq 1$，而在它之下的最后一个周期内，以 $n\pi$ 为基准量度，根为

$$n\pi - 2\pi, \quad n\pi - \tfrac{4\pi}{3}, \quad n\pi - \pi, \quad n\pi - \tfrac{2\pi}{3}.$$

其中最后一个是 $2(k-1)\pi + \frac{4\pi}{3}$，它是根，因为其余弦值为 $-\frac{1}{2}$。在 $(n\pi - \frac{2\pi}{3}, n\pi)$ 内没有根：这个区间内没有 π 的倍数，而 $\cos y = -\frac{1}{2}$ 则需要 $y \equiv \frac{2\pi}{3}$ 或 $\frac{4\pi}{3} \pmod{2\pi}$，但这里 y 在模 2π 之下严格介乎 $\frac{4\pi}{3}$ 与 2π 之间。此外，由于 $n \geq 2$，有 $n\pi - \frac{2\pi}{3} > 0$。所以最大根对应 $y = n\pi - \frac{2\pi}{3}$，而

$$x = \frac{y}{2n} = \frac{(3n-2)\pi}{6n} \qquad (n \text{ 为偶数}).$$

n **为奇数**。这时 $n\pi = 2k\pi + \pi$，其中 $k = \frac{n-1}{2} \geq 0$。在由 $2k\pi = n\pi - \pi$ 开始的周期内量度，$n\pi$ 之下的根位于偏移量 0 和 $\frac{2\pi}{3}$ 之处，所以候选的根是 $n\pi - \pi$ 和 $n\pi - \frac{\pi}{3}$。区间 $\left(n\pi - \frac{\pi}{3}, n\pi\right)$ 对应于 y 在模 2π 之下严格介乎 $\frac{2\pi}{3}$ 与 π 之间，当中没有根。而且 $n\pi - \frac{\pi}{3} \geq \frac{2\pi}{3} > 0$。所以最大根对应 $y = n\pi - \frac{\pi}{3}$，而

$$x = \frac{(3n-1)\pi}{6n} \qquad (n \text{ 为奇数}).$$

验算

当 $n = 1$ 时，在 $(0, \frac{1}{2}\pi)$ 上 $\mathrm{g}(x) = \sin 2x + \sin 4x - \sin 6x$：公式给出 $x = \frac{\pi}{3}$，而确实有 $\sin\frac{2\pi}{3} + \sin\frac{4\pi}{3} - \sin 2\pi = 0$。当 $n = 2$ 时，公式同样给出 $\frac{\pi}{3}$（$y = \frac{4\pi}{3}$）。当 $n \to \infty$ 时，两条公式都趋向 $\frac{1}{2}\pi$，这是理所当然的，因为根会愈来愈密集。我们以细密的网格对 $n = 1, \ldots, 12$ 的 g 作了数值扫描。每一个预测值（例如 $n = 3$ 时的 $x = \frac{4\pi}{9} \approx 1.3963$）都是根，而 g 在该值与 $\frac{1}{2}\pi$ 之间保持同一符号。至于第一部分，随机抽取 2000 组满足 $Q^2 < 4R(P-R)$ 的三元组，所得的括号内式子在 $[-1, 1]$ 上全部没有零点。

常见错误

- 答 $x = \frac{1}{2}\pi$ 本身：对每一个 n，它都是 g 的根，但它在开区间（open interval）之外。
- 忘记了由 $\sin y = 0$ 得出的根，只保留 $\cos y = -\frac{1}{2}$。当 n 为偶数时，$y = n\pi - \pi$ 是一个根，虽然并非最大的一个。
- 只写 $\cos y = -\frac{1}{2} \Rightarrow y = \frac{2\pi}{3} + 2k\pi$，漏掉了 $\frac{4\pi}{3}$ 那一族解。这样在 n 为偶数时会得出错误答案。

STEP 之外

这个因式分解之所以行得通，是因为 $\sin mx / \sin x$ 永远是 $\cos x$ 的多项式（polynomial）：这些就是第二类切比雪夫多项式（Chebyshev polynomials of the second kind）$U_{m-1}(\cos x)$。所以任何有限正弦级数（sine series）$\sum a_m \sin mx$ 都等于 $\sin x$ 乘以一个 $\cos x$ 的多项式，而数算它的零点，就变成数算多项式在 $[-1, 1]$ 内的根。这个想法是逼近理论（approximation theory）和傅里叶分析（Fourier analysis）中许多内容的基础。Needham 的 *Visual Complex Analysis*（《复分析：可视化方法》）通过 $\mathrm{e}^{\mathrm{i}x}$，为倍角公式（multiple-angle formulae）提供了一个优美的视角。

详细解答

STEP II 2001，第 5 题

纯数学 · 驻点与积分

STEP II 2001，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/01-s2-q5

关键思路

驻点（stationary point）可直接由斜率（gradient）求得：它们位于 $\frac{\mathrm{d}y}{\mathrm{d}x} = 0$ 之处，而其性质则取决于斜率的正负号如何变化。由于指数因子恒为正，斜率的正负号只取决于多项式因子 $x(1-x^2)$，而这个因子对两条曲线都相同。至于高度，第一个斜率可以准确地积分（留意 $u = x^2$）；第二个则不能，但我们并不需要它的值，只需比较两个积分。

解答

曲线 C_1

由于对所有 x 都有 $\mathrm{e}^{-x^2} > 0$，斜率为零当且仅当 $x(1-x^2) = x(1-x)(1+x) = 0$，即 $x = -1$、0 及 1。此外，斜率的正负号与 $x(1-x)(1+x)$ 相同：

	$x < -1$	$-1 < x < 0$	$0 < x < 1$	$x > 1$
$\frac{\mathrm{d}y}{\mathrm{d}x}$ 的正负号	$+$	$-$	$+$	$-$

在 $x = 0$ 处，斜率由负变正，所以原点（它在曲线上）是极小点（minimum）。在 $x = 1$ 处，斜率由正变负，所以这是极大点（maximum）；同理，$x = -1$ 也是极大点。

要求出高度，我们进行积分。由于曲线通过原点，

$$y = \int_0^x t(1-t^2)\,\mathrm{e}^{-t^2}\,\mathrm{d}t.$$

想一想 怎样对 $t(1-t^2)\mathrm{e}^{-t^2}$ 积分？

因子 t 旁边是一个 t^2 的函数，这提示我们用代换（substitution）$u = t^2$，$\mathrm{d}u = 2t\,\mathrm{d}t$：

$$y = \tfrac{1}{2}\int_0^{x^2} (1-u)\,\mathrm{e}^{-u}\,\mathrm{d}u.$$

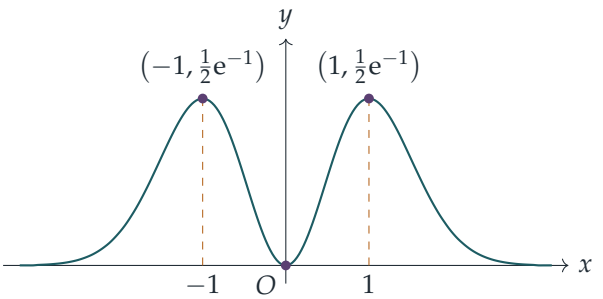
现在留意 $\frac{\mathrm{d}}{\mathrm{d}u}\left(u\mathrm{e}^{-u}\right) = \mathrm{e}^{-u} - u\mathrm{e}^{-u} = (1-u)\mathrm{e}^{-u}$（用分部积分法（integration by parts）也会得到同样的结果）。因此

$$y = \tfrac{1}{2}\left[u\,\mathrm{e}^{-u}\right]_0^{x^2} = \tfrac{1}{2}\,x^2\mathrm{e}^{-x^2}.$$

在 $x = 1$ 处，这给出 $y = \frac{1}{2}\mathrm{e}^{-1}$，所以极大点位于 $\left(1, \frac{1}{2}\mathrm{e}^{-1}\right)$，正如题目所求。由于 y 是 x 的偶函数（even function），另一个驻点是极大点

$$\left(-1, \tfrac{1}{2}\mathrm{e}^{-1}\right).$$

草图。 由 $y = \frac{1}{2}x^2\mathrm{e}^{-x^2}$ 可知：曲线关于 y 轴对称；$y \geq 0$，且只在原点取等号；又因为指数函数衰减的速度比 x^2 增长的速度快，当 $x \to \pm\infty$ 时 $y \to 0$。所以这是一条“双峰”曲线：它在原点与 x 轴相切，上升至 $x = \pm 1$ 处两个等高的极大点（高度为 $\frac{1}{2}\mathrm{e}^{-1} \approx 0.18$），然后在两边逐渐下降，趋向 x 轴。（下图的纵向比例经过放大。）



曲线 C_2

同样，对所有 x 都有 $\mathrm{e}^{-x^3} > 0$，所以斜率恰好在 $x = -1, 0, 1$ 处为零，而且正负号与 $x(1-x)(1+x)$ 相同，得出的正负号表与之前完全一样。所以原点（按题设在曲线上）是极小点，而 $x = 1$ 处的驻点是极大点。（$x = -1$ 处的驻点也是极大点，不过题目没有问及。）

它的高度是

$$k = \int_0^1 x(1-x^2)\,\mathrm{e}^{-x^3}\,\mathrm{d}x.$$

这个积分没有初等原函数（elementary antiderivative），所以我们改为把它与 C_1 的积分比较，而我们已知后者等于 $\frac{1}{2}\mathrm{e}^{-1}$。

想一想 怎样可以不计算 k 的值，而把 k 与 $\frac{1}{2}\mathrm{e}^{-1}$ 比较?

比较被积函数（integrand）。当 $0 < x < 1$ 时，我们有 $x^3 < x^2$（把 $x < 1$ 两边乘以 $x^2 > 0$），所以 $-x^3 > -x^2$；又因为 e^t 是递增函数，故 $\mathrm{e}^{-x^3} > \mathrm{e}^{-x^2}$。此外，在这个范围内 $x(1-x^2) > 0$，所以乘上它之后不等式的方向不变：

$$x(1-x^2)\,\mathrm{e}^{-x^3} > x(1-x^2)\,\mathrm{e}^{-x^2} \qquad \text{当} 0 < x < 1.$$

两个被积函数在 $[0,1]$ 上都连续，而不等式在开区间上严格成立，所以在 $[0,1]$ 上积分会得到严格不等式（strict inequality）：

$$k = \int_0^1 x(1-x^2)\,\mathrm{e}^{-x^3}\,\mathrm{d}x > \int_0^1 x(1-x^2)\,\mathrm{e}^{-x^2}\,\mathrm{d}x = \tfrac{1}{2}\mathrm{e}^{-1},$$

最后一个等式就是上面求得的 C_1 在 $x = 1$ 处的高度。所以极大点位于 $(1,k)$，其中 $k > \frac{1}{2}\mathrm{e}^{-1}$，正如题目所求。

验算

用积法则（product rule）对 $\frac{1}{2}x^2\mathrm{e}^{-x^2}$ 求导，得 $x\mathrm{e}^{-x^2} - x^3\mathrm{e}^{-x^2} = x(1-x^2)\mathrm{e}^{-x^2}$，而它在 $x = 0$ 处的值为 0，所以 C_1 的结果正确。二阶导数（second derivative）印证了驻点的分类：对两条曲线都有 $y''(0) = 1 > 0$ 及 $y''(1) = -2\mathrm{e}^{-1} < 0$。以数值方法（森逊法则（Simpson's rule））计算，$k \approx 0.2034$，而 $\frac{1}{2}\mathrm{e}^{-1} \approx 0.1839$。

常见错误

- 忘记积分常数（constant of integration）：正是“通过原点”这个条件确定了它的值，而它对高度有影响。
- 只凭 $\mathrm{e}^{-x^3} > \mathrm{e}^{-x^2}$ 就声称 $k > \frac{1}{2}\mathrm{e}^{-1}$，却没有指出 $x(1-x^2)$ 在 $(0,1)$ 上为正；若乘上的是负因子，不等式的方向便会倒转。
- 说 $x^3 < x^2$ “对所有 x 成立”：它只在 $x < 1$、$x \neq 0$ 时成立，这正是为甚么区间 $[0,1]$ 恰好符合我们所需。

STEP 之外

第二部分是一个比较论证（comparison argument）：对于一个写不出来的解，我们把它与一个写得出来的解比较，从而了解它的一些性质。这个想法贯穿整个微分方程（differential equations）理论，当中的比较定理（comparison theorem）以较简单方程的解来界定棘

手方程的解的范围；它也贯穿分析学（analysis），当中积分的单调性（monotonicity）是一件基本工具。Boyce 与 DiPrima 的 *Elementary Differential Equations*（微分方程入门教材）是个好去处，可以让你看到即使没有公式，也能从定性的角度对解作出多少论述。

详细解答

STEP II 2001，第 6 题

纯数学 · 代换积分法

STEP II 2001，第 6 题

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

step.maths.org/questions/01-s2-q6

关键思路

看似棘手的 $\tan^{-1}\left(\frac{1-x}{1+x}\right)$ 其实只是 $\frac{\pi}{4}-\tan^{-1}x$ 的另一种写法：它就是把正切的差角公式（tangent subtraction formula）倒过来读。一旦看穿这一点，第 (i) 部分便化为一次分部积分（integration by parts），而且刚好落在开首的积分上。第 (ii) 部分看来毫不相干，但映射 $y=\frac{1-x}{1+x}$ 把 0 和 1 互换，而且它是自身的逆映射，所以用它作代换（substitution），便可把 (ii) 变回 (i)。

解答

开首的积分

以 $1+x^2$ 除 x^4（或写成 $x^4=(x^4-1)+1=(x^2-1)(x^2+1)+1$），得

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

因此

$$\int_0^1\frac{x^4}{1+x^2}\,\mathrm{d}x=\left[\tfrac{1}{3}x^3-x+\tan^{-1}x\right]_0^1=\frac{1}{3}-1+\frac{\pi}{4}=\frac{\pi}{4}-\frac{2}{3},$$

正是所求。

第 (i) 部分

设 $\theta=\tan^{-1}x$，则当 $0\le x\le 1$ 时，有 $0\le\theta\le\frac{\pi}{4}$。由差角公式，

$$\tan\left(\frac{\pi}{4}-\theta\right)=\frac{\tan\frac{\pi}{4}-\tan\theta}{1+\tan\frac{\pi}{4}\tan\theta}=\frac{1-x}{1+x}.$$

想一想 有了 $\tan\left(\frac{\pi}{4}-\theta\right)=\frac{1-x}{1+x}$，是否真的就可以写 $\tan^{-1}\left(\frac{1-x}{1+x}\right)=\frac{\pi}{4}-\theta$?

只有当 $\frac{\pi}{4}-\theta$ 落在 $\tan^{-1}$ 的值域（range）$\left(-\frac{\pi}{2},\frac{\pi}{2}\right)$ 之内才可以。这里 $0\le\frac{\pi}{4}-\theta\le\frac{\pi}{4}$，所以条件成立，于是在整个积分区间上

$$\tan^{-1}\left(\frac{1-x}{1+x}\right)=\frac{\pi}{4}-\tan^{-1}x.$$

因此

$$\int_0^1 x^3\tan^{-1}\left(\frac{1-x}{1+x}\right)\mathrm{d}x=\frac{\pi}{4}\int_0^1 x^3\,\mathrm{d}x-\int_0^1 x^3\tan^{-1}x\,\mathrm{d}x=\frac{\pi}{16}-J,$$

其中 $J=\int_0^1 x^3\tan^{-1}x\,\mathrm{d}x$。取 $u=\tan^{-1}x$ 及 $\mathrm{d}v=x^3\,\mathrm{d}x$（于是 $v=\frac{1}{4}x^4$ 而 $\mathrm{d}u=\frac{\mathrm{d}x}{1+x^2}$）作分部积分：

$$\begin{aligned}J&=\left[\tfrac{1}{4}x^4\tan^{-1}x\right]_0^1-\frac{1}{4}\int_0^1\frac{x^4}{1+x^2}\,\mathrm{d}x\\&=\frac{\pi}{16}-\frac{1}{4}\left(\frac{\pi}{4}-\frac{2}{3}\right)=\frac{1}{6},\end{aligned}$$

当中用了开首的结果。因此

$$\int_0^1 x^3\tan^{-1}\left(\frac{1-x}{1+x}\right)\mathrm{d}x=\frac{\pi}{16}-\frac{1}{6}.$$

第 (ii) 部分

想一想 那个古怪的权重 $\frac{(1-y)^3}{(1+y)^5}$ 是从哪里来的?

它正是 $x^3\,\mathrm{d}x$ 在同一个变换下所变成的样子。代入 $y=\frac{1-x}{1+x}$，解出 $x=\frac{1-y}{1+y}$：这个映射是自身的逆映射，并把 $x=0$ 送到 $y=1$，把 $x=1$ 送到 $y=0$。于是

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

所以

$$\frac{(1-y)^3}{(1+y)^5}\,\mathrm{d}y=\frac{8x^3}{(1+x)^3}\cdot\frac{(1+x)^5}{32}\cdot\left(-\frac{2}{(1+x)^2}\right)\mathrm{d}x=-\frac{x^3}{2}\,\mathrm{d}x,$$

而 $\tan^{-1}y=\tan^{-1}\left(\frac{1-x}{1+x}\right)$。把积分上下限对调，负号便被抵销（$y$ 由 0 到 1 即 x 由 1 到 0）：

$$\begin{aligned}\int_0^1\frac{(1-y)^3}{(1+y)^5}\tan^{-1}y\,\mathrm{d}y&=\int_1^0\left(-\frac{x^3}{2}\right)\tan^{-1}\left(\frac{1-x}{1+x}\right)\mathrm{d}x\\&=\frac{1}{2}\int_0^1 x^3\tan^{-1}\left(\frac{1-x}{1+x}\right)\mathrm{d}x=\frac{1}{2}\left(\frac{\pi}{16}-\frac{1}{6}\right)=\frac{\pi}{32}-\frac{1}{12}\end{aligned}$$

这里用了第 (i) 部分的结果。

验算

- 两个答案都是正数，这是必然的（每个被积函数（integrand）在 $[0,1]$ 上均非负）：$\frac{\pi}{16}-\frac{1}{6}\approx 0.02968$ 及 $\frac{\pi}{32}-\frac{1}{12}\approx 0.01484$。对原来的被积函数使用辛普森法则(Simpson's rule)，结果吻合至小数点后 12 位。
- (ii) 的答案恰好是 (i) 的一半，你可以在动手做任何代数运算之前，先用数值方法粗略核对一下。
- 在恒等式中代入 $x=0$：$\tan^{-1}1=\frac{\pi}{4}$。代入 $x=1$：$\tan^{-1}0=0$。两者都与 $\frac{\pi}{4}-\tan^{-1}x$ 相符。

常见错误

- 在 (ii) 中漏掉 $\mathrm{d}y$，或处理错正负号：因子 $-\frac{2}{(1+x)^2}$ 与对调后的上下限，两者的负号互相抵销。
- 未检查值域便写下 $\tan^{-1}\left(\frac{1-x}{1+x}\right)=\frac{\pi}{4}-\tan^{-1}x$；当 $x<-1$ 时此式并不成立（两边相差 π）。

STEP 之外

代换 $y=\frac{1-x}{1+x}$ 是一个莫比乌斯变换（Möbius transformation），即形如 $\frac{ax+b}{cx+d}$ 的映射。这些映射在复合（composition）运算下构成一个群（group），而这一个的阶（order）为二。同样的技巧，配合对称性，可得出著名的结果 $\int_0^1\frac{\ln(1+x)}{1+x^2}\,\mathrm{d}x=\frac{\pi}{8}\ln 2$，值得接着一试。莫比乌斯映射在复分析（complex analysis）中占有核心地位；Needham 的 *Visual Complex Analysis*（《复分析：可视化方法》）对它们有一番优美的几何阐述。

详细解答

STEP II 2001，第 7 题

纯数学 · 复数与旋转

> **STEP II 2001，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/01-s2-q7

关键思路

沿着正多边形（regular polygon）绕一圈，每条边的长度都相同，而在每个顶点转过的都是同一个外角（exterior angle）。在阿根图（Argand diagram）上，转过角 θ 相当于乘以 $\mathrm{e}^{\mathrm{i}\theta}$，所以每条边都是前一条边乘以一个固定的复数（complex number）。一旦求出各顶点，圆心便可由垂直平分线（perpendicular bisector）求得，而旋转（rotation）只不过是再乘一次 $\mathrm{e}^{\mathrm{i}\theta}$。

解答

求 Q 和 S

每个顶点的内角（interior angle）是 $5\pi/6$，所以在每个顶点，前进方向转过的是外角 $\pi - 5\pi/6 = \pi/6$。由于多边形位于上半平面，而我们由 O 出发，沿正实轴走到 P，所以每次都必须向逆时针（anticlockwise）方向转（若在 P 顺时针转，Q 便会落在实轴之下）。设 $\omega = \mathrm{e}^{\mathrm{i}\pi/6} = \frac{\sqrt{3}}{2} + \frac{1}{2}\mathrm{i}$。边 $\overrightarrow{OP}$ 就是复数 2，而其后每条边都是前一条边乘以 ω（长度不变，转过 $\pi/6$）。因此

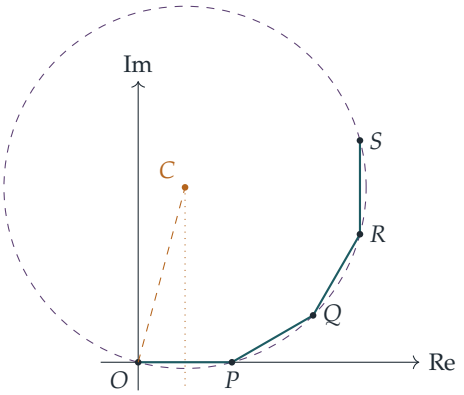
$$\begin{aligned}\overrightarrow{PQ} &= 2\omega = \sqrt{3} + \mathrm{i},\\ \overrightarrow{QR} &= 2\omega^2 = 2\mathrm{e}^{\mathrm{i}\pi/3} = 1 + \sqrt{3}\,\mathrm{i},\\ \overrightarrow{RS} &= 2\omega^3 = 2\mathrm{e}^{\mathrm{i}\pi/2} = 2\mathrm{i}.\end{aligned}$$

逐一相加，

$$\begin{aligned} Q &= 2 + (\sqrt{3} + \mathrm{i}) = 2 + \sqrt{3} + \mathrm{i}, \\ R &= Q + 1 + \sqrt{3}\,\mathrm{i} = (3 + \sqrt{3}) + (1 + \sqrt{3})\mathrm{i}, \\ S &= R + 2\mathrm{i} = (3 + \sqrt{3}) + (3 + \sqrt{3})\mathrm{i}. \end{aligned}$$

第一行就是题目要求的 Q 值，而

$$S = (3 + \sqrt{3})(1 + \mathrm{i}).$$



圆心 C

通过 O、P、Q 的圆就是整个正十二边形的外接圆（circumcircle），所以 C 是这个多边形的中心。由于 C 与 O 和 $P = 2$ 等距，它位于垂直平分线 $\operatorname{Re} z = 1$ 上；设 $C = 1 + h\mathrm{i}$。它与 O 和 Q 亦等距：

$$\begin{aligned} |C|^2 = |C - Q|^2 &\iff 1 + h^2 = (1 + \sqrt{3})^2 + (h - 1)^2 \\ &\iff 1 + h^2 = 4 + 2\sqrt{3} + h^2 - 2h + 1 \\ &\iff 2h = 4 + 2\sqrt{3}, \end{aligned}$$

所以 $h = 2 + \sqrt{3}$，而 $C = 1 + (2 + \sqrt{3})\mathrm{i}$，位于第一象限。

> ? **想一想** C 的辐角（argument）的准确值是甚么？

我们需要 $\tan(\arg C) = 2 + \sqrt{3}$。现在

$$\tan \tfrac{5\pi}{12} = \tan\left(\tfrac{\pi}{4} + \tfrac{\pi}{6}\right) = \frac{1 + 1/\sqrt{3}}{1 - 1/\sqrt{3}} = \frac{\sqrt{3} + 1}{\sqrt{3} - 1} = \frac{(\sqrt{3} + 1)^2}{2} = 2 + \sqrt{3},$$

而 C 在第一象限，所以 $\arg C = 5\pi/12$。（从几何角度看：三角形 OCP 是等腰三角形，在 C 的顶角为 $2\pi/12$，所以两个底角都是 $\frac{1}{2}(\pi - \pi/6) = 5\pi/12$。）

旋转

绕 O 逆时针旋转角 θ，相当于把每一点乘以 $\mathrm{e}^{\mathrm{i}\theta}$，所以 C 会移到辐角为 $5\pi/12+\theta$ 的位置。要 C 落在实轴上，$5\pi/12+\theta$ 必须是 π 的倍数；最小的正数 θ 是

$$\theta = \pi - \tfrac{5\pi}{12} = \tfrac{7\pi}{12},$$

这时 C 落在负实轴上。

现在把 S 写成模—辐角形式（modulus–argument form）：$1+\mathrm{i} = \sqrt{2}\,\mathrm{e}^{\mathrm{i}\pi/4}$，所以 $S = (3+\sqrt{3})\sqrt{2}\,\mathrm{e}^{\mathrm{i}\pi/4} = (3\sqrt{2}+\sqrt{6})\,\mathrm{e}^{\mathrm{i}\pi/4}$。$S$ 的新位置是

$$\begin{aligned} S' = S\,\mathrm{e}^{7\mathrm{i}\pi/12} &= (3\sqrt{2}+\sqrt{6})\,\mathrm{e}^{\mathrm{i}(3\pi/12+7\pi/12)} = (3\sqrt{2}+\sqrt{6})\,\mathrm{e}^{5\mathrm{i}\pi/6} \\ &= (3\sqrt{2}+\sqrt{6})\left(-\tfrac{\sqrt{3}}{2}+\tfrac{1}{2}\mathrm{i}\right) = -\tfrac{1}{2}(3\sqrt{2}+\sqrt{6})(\sqrt{3}-\mathrm{i}), \end{aligned}$$

正是所求。

验算

- $|C|^2 = 1+(2+\sqrt{3})^2 = 8+4\sqrt{3} = (\sqrt{6}+\sqrt{2})^2$，而 $|C-S|^2 = (2+\sqrt{3})^2+1$ 亦相同，所以 S 确实在同一个圆上。
- $|S'| = 3\sqrt{2}+\sqrt{6} = |S|$，旋转理应保持模不变。
- 数值上 $S \approx 4.732+4.732\mathrm{i}$，而 $S' \approx -5.7956+3.3461\mathrm{i}$；电脑验证了所有边长、各个 $5\pi/6$ 的角，以及 $7\pi/12$ 确是逆时针旋转时第一个令 C 落在实轴上的角。

常见错误

- 转了内角 $5\pi/6$，而不是外角 $\pi/6$。
- 旋转了 $-5\pi/12$（顺时针），或旋转了 $\pi/2-5\pi/12$：旋转是逆时针的，而 C 首次碰到实轴是在负半轴上。

STEP 之外

把乘以 $\mathrm{e}^{\mathrm{i}\theta}$ 看成旋转，是把复函数视为几何变换（geometric transformation）的起点：$z \mapsto az+b$ 可给出平面上所有旋转、放大（enlargement）和平移（translation），而这些变换构成一个群（group）。正多边形的顶点是等比级数的和（geometric sum）$2(1+\omega+\cdots+\omega^{k-1})$，与单位根（roots of unity）息息相关。Needham 的 *Visual Complex Analysis*（《复分析：可视化方法》）把这种几何观点发挥得淋漓尽致。

详细解答

STEP II 2001，第 8 题

纯数学 · 周期函数与常微分方程

STEP II 2001，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/01-s2-q8

关键思路

周期（period）为 1 的函数，在每一个长度为 1 的区间上的形状都相同，所以它在任何这样的区间上的积分都是同一个数。因此 F 每次增加的量都相同，都是 $F(1)$，只要 x 增加 1。最后一部分的微分方程（differential equation）是一阶线性的，所以利用积分因子（integrating factor）便可把它的解化为 $-F(x)$ 的指数函数，其余的结论则由 $F(1)$ 为正数这一事实推出。

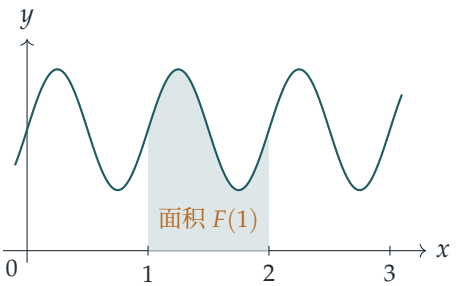
解答

第 (i) 部分

取

$$f(x) = 2 + \sin(2\pi x).$$

由于 sin 的周期为 2π，$\sin(2\pi(x+1)) = \sin(2\pi x + 2\pi) = \sin(2\pi x)$，所以 $f(x+1) = f(x)$。另外 $\sin(2\pi x) \geq -1$，故对每一个 x 都有 $f(x) \geq 1 > 0$。（常数函数如 $f(x) = 1$ 也符合要求，但没那么有趣。）



图像下方每一条单位阔度的长条，面积都相同；第 (ii) 部分说的就是这回事。

第 (ii) 部分

定义 $G(x) = F(x+1) - F(x)$。求导，并对第一项使用链式法则（chain rule），得

$$G'(x) = f(x+1) - f(x) = 0$$

对所有 x 成立，这是由 f 的周期性得出的。所以 G 是常数，其值为 $G(0) = F(1) - F(0) = F(1)$，因为 $F(0) = 0$。也就是说，

$$F(x+1) = F(x) + F(1) \quad \text{对所有} x.$$

现在对 n 使用归纳法（induction）。当 $n = 1$ 时，命题 $F(1) = 1 \cdot F(1)$ 显然成立。若对某个正整数 k 有 $F(k) = kF(1)$，则在上式中代入 $x = k$，得

$$F(k+1) = F(k) + F(1) = kF(1) + F(1) = (k+1)F(1).$$

因此对每一个正整数 n 都有 $F(n) = nF(1)$，命题得证。

（等价的做法是：若 f 连续，则 $F(x) = \int_0^x f(t)\,\mathrm{d}t$，而 $F(n)$ 是 n 个在 $[k, k+1]$ 上的积分之和，每一个积分经代换（substitution）$t \mapsto t + k$ 后都等于 $\int_0^1 f(t)\,\mathrm{d}t$。上面的论证则完全不必把 F 写成积分。）

第 (iii) 部分

这个方程对 y 而言是线性的，积分因子为 $\mathrm{e}^{\int f\,\mathrm{d}x} = \mathrm{e}^{F(x)}$，$F$ 就是在这里出场的。事实上，由乘积法则（product rule）及 $F' = f$，

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(y\,\mathrm{e}^{F(x)}\right) = \mathrm{e}^{F(x)}\left(\frac{\mathrm{d}y}{\mathrm{d}x} + f(x)\,y\right) = 0.$$

所以 $y\,\mathrm{e}^{F(x)}$ 是常数。在 $x = 0$ 时它等于 $1 \cdot \mathrm{e}^{F(0)} = 1$，故

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

由第 (ii) 部分，对每一个正整数 n，

$$y(n) = \mathrm{e}^{-F(n)} = \mathrm{e}^{-nF(1)} = \left(\mathrm{e}^{-F(1)}\right)^n.$$

想一想 我们怎样知道 $F(1) > 0$？题目并没有直接告诉我们任何关于 $F(1)$ 的资料。

由于对所有 x 都有 $F'(x) = f(x) > 0$，F 是严格递增（strictly increasing）的，所以 $F(1) > F(0) = 0$。（如果你较喜欢用积分来想：$F(1)$ 是一条正值曲线在 0 与 1 之间下方的面积。）

因此 $0 < \mathrm{e}^{-F(1)} < 1$，而一个满足 $0 < r < 1$ 的固定数 r 必有 $r^n \to 0$（当 $n \to \infty$ 时）。于是 $y(n) \to 0$（当 $n \to \infty$ 时），这正是我们要证明的。

留意我们只声称这个结论沿整数成立。在整数与整数之间，y 被夹住了：由于 F 递增，y 便递减，所以事实上 $y(x) \to 0$ 在 $x \to \infty$ 取遍所有实数值时同样成立。

验算

取 $f(x) = 2 + \sin(2\pi x)$，则 $F(1) = \int_0^1 (2 + \sin 2\pi t)\,\mathrm{d}t = 2$，因为正弦函数在一个完整周期上的积分为零。用辛普森法则（Simpson's rule）计算，$F(2), F(3), F(5), F(7)$ 分别等于 $4, 6, 10, 14$，误差小于 10^{-12}。以数值方法（四阶龙格—库塔法，Runge–Kutta）逐步求解这个微分方程，由 $y(0) = 1$ 出发，得 $y(1) \approx 0.135335 = \mathrm{e}^{-2}$ 及 $y(4) \approx 3.3546 \times 10^{-4} = \mathrm{e}^{-8}$，与 $\mathrm{e}^{-nF(1)}$ 吻合。我们亦在非整数的 x 处验证了 $F(x+1) - F(x) = F(1)$，所用的是对称性较低的例子 $f(x) = \mathrm{e}^{\cos 2\pi x} + 0.1$。

常见错误

- 把 $F(n) = nF(1)$ 当作由周期性“显然”得出的结论，却没有说明为何 $[k, k+1]$ 上的面积等于 $[0,1]$ 上的面积。这里必须用到函数 $G(x) = F(x+1) - F(x)$ 或代换。
- 忘记说明为何 $F(1) > 0$。没有这一点，$\mathrm{e}^{-nF(1)}$ 可能一直停留在 1，甚至趋向无穷大。
- 弄错积分因子的正负号，最后得出 $y = \mathrm{e}^{F(x)}$，它会增长而不是衰减。

STEP 之外

这是弗洛凯理论（Floquet theory）最简单的情形：对于系数具周期性的线性微分方程，经过一个周期之后的解，等于初始值乘以一个固定的因子（这里是 $\mathrm{e}^{-F(1)}$），所以长远的变化取决于这个因子是小于还是大于 1。同样的想法，把数换成矩阵，便可解释支点被上下摇动的单摆的稳定性。Boyce 与 DiPrima 合著的 *Elementary Differential Equations*（初等微分方程教科书）对线性方程和积分因子有详尽的讲解，很适合作为下一步的读物。

详细解答

STEP II 2001，第 9 题

力学 · 二次空气阻力

STEP II 2001，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：

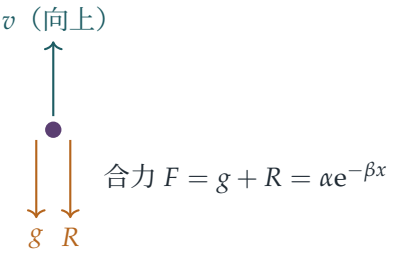
step.maths.org/questions/01-s2-q9

关键思路

题目给出的力是高度的函数，而不是时间的函数，所以最自然的工具是功能原理（work–energy principle）：质点在上升过程中克服 F 所作的功，恰好等于它损失的动能（kinetic energy）。这样就把 $\mathrm{e}^{-\beta x}$ 与 v^2 联系起来，而所得结果显示空气阻力（air resistance）只取决于速率。一旦知道阻力与速率的关系，下落的过程就是一条关于 v^2 的标准微分方程（differential equation）。

解答

模型。 质点的质量为一个单位，沿一条铅垂线运动。上升时，重力和空气阻力都向下作用，与运动方向相反，所以合力 F（大小为 $\alpha\mathrm{e}^{-\beta x}$）向下作用。取 x 为由抛射点起向上量度的高度。



所耗的能量

到达高度 x 时所耗的能量，就是克服 F 所作的功：

$$\int_0^x \alpha \mathrm{e}^{-\beta s}\,\mathrm{d}s = \frac{\alpha}{\beta}\Big(1-\mathrm{e}^{-\beta x}\Big).$$

这些能量来自质点的动能，所以根据功能原理（单位质量）

$$\tfrac{1}{2}u^2 - \tfrac{1}{2}v^2 = \frac{\alpha}{\beta}\Big(1-\mathrm{e}^{-\beta x}\Big). \qquad (*)$$

（等价地，把牛顿第二定律（Newton's second law）写成 $v\frac{\mathrm{d}v}{\mathrm{d}x} = -\alpha\mathrm{e}^{-\beta x}$ 的形式，再由 $x=0$、$v=u$ 起积分，也会得到同一条方程。）

以速率表示的力

把 $(*)$ 乘以 β 并整理：

$$\alpha - \alpha\mathrm{e}^{-\beta x} = \tfrac{1}{2}\beta u^2 - \tfrac{1}{2}\beta v^2 \quad\Longrightarrow\quad F = \alpha\mathrm{e}^{-\beta x} = \tfrac{1}{2}\beta v^2 + \alpha - \tfrac{1}{2}\beta u^2,$$

这正是所要求的结果。

> ? **想一想** 为甚么必定有 $\alpha = \frac{1}{2}\beta u^2 + g$？题目中似乎没有任何条件迫使它成立。

把 F 分成两部分：$F = g + R$，其中 R 是空气阻力。于是

$$R = F - g = \tfrac{1}{2}\beta v^2 + \Big(\alpha - \tfrac{1}{2}\beta u^2 - g\Big).$$

所以阻力只是速率的函数：一个 v^2 项加上一个常数。空气阻力是由物体在空气中运动而产生的，所以当质点在最高点瞬间静止时，阻力必定为零。在该处代入 $v=0$，得 $R=0$，所以 $F=g$（只有重力作用），而且

$$\alpha - \tfrac{1}{2}\beta u^2 = g, \qquad \text{即} \qquad \alpha = \tfrac{1}{2}\beta u^2 + g.$$

因此常数项为零，阻力定律就简单地是

$$R = \tfrac{1}{2}\beta v^2.$$

（顺带一提，最高点就是 $\alpha\mathrm{e}^{-\beta x} = g$ 之处，高度为 $\frac{1}{\beta}\ln(\alpha/g)$。）

下落的过程

假设无论运动方向如何，同一条阻力定律 $R = \frac{1}{2}\beta v^2$ 都成立；下落时阻力向上作用，与运动方向相反，而重力仍然向下作用。设 y 为质点在最高点以下的距离，v 为向下的速率，并且在 $y=0$ 时 $v=0$。取向下为正，由牛顿第二定律（单位质量）得：

$$v\frac{\mathrm{d}v}{\mathrm{d}y} = g - \tfrac{1}{2}\beta v^2.$$

由于 $v\frac{\mathrm{d}v}{\mathrm{d}y} = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}y}(v^2)$，设 $w = v^2$：

$$\frac{\mathrm{d}w}{\mathrm{d}y} = 2g - \beta w, \qquad w = 0, \text{ 当} y = 0.$$

这是一条常系数线性方程（也可以用分离变量法）：其解为

$$w = \frac{2g}{\beta}\left(1 - \mathrm{e}^{-\beta y}\right),$$

它确实满足 $w(0) = 0$ 和 $w' = 2g\mathrm{e}^{-\beta y} = 2g - \beta w$。因此，在最高点以下距离 y 处的空气阻力为

$$R = \tfrac{1}{2}\beta v^2 = g\left(1 - \mathrm{e}^{-\beta y}\right).$$

验算

当 $y \to \infty$ 时，$R \to g$：质点趋近终端速率（terminal speed）$\sqrt{2g/\beta}$，此时阻力与重力平衡。在 $y = 0$ 处，$R = 0$，与质点静止时应有的情况相符。上升时，在 $x = 0$ 处公式给出 $F = \alpha = g + \frac{1}{2}\beta u^2$，即重力加上速率为 u 时的阻力。数值上，取 $g = 9.8$、$\beta = 0.1$、$u = 20$（所以 $\alpha = 29.8$），用龙格－库塔法（Runge–Kutta method）逐步计算运动方程，所得结果与上升时的 $F = \alpha\mathrm{e}^{-\beta x}$ 和下落时的 $R = g(1 - \mathrm{e}^{-\beta y})$ 相符，误差约为 10^{-12}，而质点到达的最高高度 ≈ 11.12，与 $\frac{1}{\beta}\ln(\alpha/g)$ 吻合。

常见错误

- 把 F 当作只是空气阻力：它是合力，所以必须减去重力才得到 R。
- 下落时仍然沿用相同的阻力符号：阻力总是与运动方向相反，所以此时它向上作用，方程变为 $v\frac{\mathrm{d}v}{\mathrm{d}y} = g - \frac{1}{2}\beta v^2$。
- 以 v 表示答案，或以 u 和 α 表示答案；题目要求只以 y、g 和 β 表示 R。

STEP 之外

二次阻力（quadratic drag）$R \propto v^2$ 是描述物体在空气中高速运动的切合实际的模型，而这道题展示了一个巧妙的技巧：以距离为自变量时，v^2 所满足的方程是线性的，尽管以时间为自变量时 v 所满足的方程并非线性。大学本科的力学课程还会探讨先上升后下落这个问题的不对称性：质点返回起点时的速率小于 u，而下落所需的时间比上升长。Kleppner 与 Kolenkow 所著的 *An Introduction to Mechanics*（力学导论）详细讨论了取决于速度的力和终端速度。

详细解答

STEP II 2001，第 10 题

力学 · 碰撞与恢复系数

STEP II 2001，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/01-s2-q10

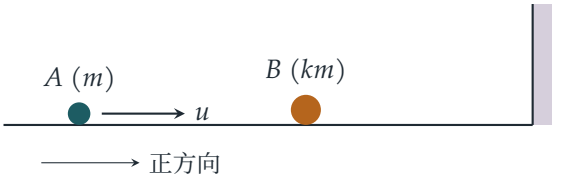
关键思路

这题是两次碰撞加一场追逐。第一次碰撞用的是标准的两条方程：动量守恒（conservation of momentum）和牛顿实验定律（Newton's experimental law）。B 从墙上反弹之后，唯一的问题是：正在折返的 B 能否与 A 相遇。这恰好在 A 朝墙方向的速度大于 B 的速度时发生，由此便得出关于 k 的不等式。至于能量部分，我们把动能（kinetic energy）写成 k 的函数，再与 $mu^2/27$ 比较。比较的结果原来是一个完全平方（perfect square）。

解答

第二次碰撞

把 A 和 B 当作光滑平面上的质点（particle），因此水平方向的冲量（impulse）只来自碰撞。取由 A 指向墙的方向为正方向。设 v_A 和 v_B 分别为 A 和 B 在第一次碰撞刚结束时的速度。



第一次碰撞。 动量守恒，而分离速率等于接近速率的 e_1 倍：

$$mv_A + km\,v_B = mu,$$
$$v_B - v_A = e_1 u.$$

把第二条方程乘以 k 再加到第一条方程，得 $(1+k)v_B = (1+e_1)u$。所以

$$v_B = \frac{(1+e_1)u}{1+k}, \qquad v_A = v_B - e_1 u = \frac{(1-ke_1)u}{1+k}.$$

由于 $v_B > 0$，B 继续向墙移动。A 则可能向前移动、静止或向后移动，视乎 $1-ke_1$ 的正负号而定。

墙。 B 以速率 v_B 撞墙，并以速率 $e_2 v_B$ 反弹，所以它的速度变为 $-e_2 v_B$。与此同时，A 继续以恒定速度 v_A 移动。注意 A 不可能先到达墙，因为 $v_A < v_B$（两者之差为 $e_1 u > 0$），而且 A 起初在 B 的后面。

> ? **想一想** 反弹之后，A 和 B 可能朝同一方向移动。我们怎样判断它们会否相遇？

反弹之后，B 位于 A 与墙之间，而且再没有其他作用施加于任何一个质点。所以两者之间的距离会缩短，当且仅当 A 的速度大于 B 的速度：

$$v_A > -e_2 v_B \iff v_A + e_2 v_B > 0.$$

若 $v_A > 0$，两者相向而行。若 $v_A \le 0$，则是 B 在追赶 A，而且只有在较快时才追得上。这一条不等式已涵盖两种情况。代入后，再乘以正数 $(1+k)/u$，得

$$1 - ke_1 + e_2(1+e_1) > 0 \iff k < \frac{1+e_2(1+e_1)}{e_1},$$

其中最后一步除以了 $e_1 > 0$。所以只要这条不等式成立，便会发生第二次碰撞，正如题目所求。（若等号成立，A 和 B 会以相同速度离开，永不相遇。）

动能

当 $e_1 = \frac{1}{3}$ 及 $e_2 = \frac{1}{2}$ 时，

$$v_A = \frac{(3-k)u}{3(1+k)}, \qquad e_2 v_B = \frac{1}{2}\cdot\frac{\frac{4}{3}u}{1+k} = \frac{2u}{3(1+k)}.$$

在 B 刚反弹之后，A 的速率仍是 $|v_A|$，而 B 的速率是 $e_2 v_B$。所以总动能为

$$\begin{aligned} T &= \tfrac{1}{2}mv_A^2 + \tfrac{1}{2}km\,(e_2 v_B)^2 \\ &= \frac{mu^2}{18(1+k)^2}\left[(3-k)^2 + 4k\right] = \frac{mu^2\,(k^2-2k+9)}{18(1+k)^2}. \end{aligned}$$

现在与 $mu^2/27$ 比较，取公分母 $54(1+k)^2$：

$$\begin{aligned} T - \frac{mu^2}{27} &= \frac{mu^2}{54(1+k)^2}\left[3(k^2-2k+9) - 2(1+k)^2\right] \\ &= \frac{mu^2}{54(1+k)^2}\left(k^2-10k+25\right) = \frac{mu^2\,(k-5)^2}{54(1+k)^2}. \end{aligned}$$

想一想 既然平方永不为负，条件 $k < 5$ 又有甚么用?

右边永不为负，而且只有在 $k = 5$ 时才等于零。所以当 $k < 5$ 时，我们有 $(k-5)^2 > 0$，因此

$$T > \frac{mu^2}{27},$$

正如题目所求。5 这个数值并非巧合。代入这些系数后，第一部分的上界是 $\left(1 + \frac{1}{2} \cdot \frac{4}{3}\right) \Big/ \frac{1}{3} = 5$。所以 $k < 5$ 正是发生第二次碰撞的条件。当 $k = 5$ 时，两个质点在反弹后都以 $-u/9$ 移动，而能量恰好是 $mu^2/27$。

验算

- 取 $e_1 = 1$ 及 $k = 1$（质量相等，完全弹性）。这时 $v_A = 0$ 及 $v_B = u$，即大家熟悉的速度互换。上界变成 $k < 1 + 2e_2$，而只要 $e_2 > 0$，$k = 1$ 便满足这条件：折返的 B 必定会撞上静止的 A。
- 当 $k = 5$ 时，公式给出 $v_A = -\frac{u}{9}$ 及 $-e_2 v_B = -\frac{1}{2} \cdot \frac{2u}{9} = -\frac{u}{9}$，两个速度相等，与上文所说一致。
- 数值方面，以精确分数计算 k 由 0.001 至 4.999 的情况，比值 $T/(mu^2/27)$ 的最小值为 1.00000001，出现在 $k = 5$ 附近，而且始终大于 1。另外随机测试了 20 000 组 (k, e_1, e_2)，没有发现不等式与直接比较反弹后速度所得结果不一致的情况。

常见错误

- 要求 $v_A > 0$。A 即使正在向后移动，仍然可以被 B 追上。正确的条件是比较反弹后两者的速度。
- 忘记墙会把 B 的方向反转。它的速度变为 $-e_2 v_B$，而不是 $e_2 v_B$。
- 计算能量时用了 B 撞墙前的速率。题目要求的是刚反弹后的能量，所以 B 的速率是 $e_2 v_B$。

STEP 之外

牛顿实验定律是一个模型，用来描述撞击期间那短暂而剧烈的形变过程。恢复系数（coefficient of restitution）概括了相对动能损失了多少。在大学一年级的力学课程中，碰撞通常会在质心参考系（centre-of-mass frame）中分析。在这个参考系中，碰撞只是把相对速度反转并乘以 e，而损失的能量为 $\frac{1}{2}\mu(1-e^2)w^2$，其中 μ 是约化质量（reduced mass），

w 是接近速率。Kleppner 与 Kolenkow 所著的 *An Introduction to Mechanics* 对这些概念有细致的阐述。

详细解答

STEP II 2001，第 11 题

力学 · 抛体运动

STEP II 2001，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：

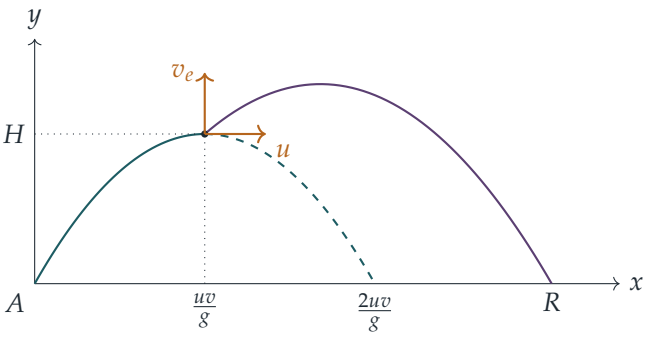
step.maths.org/questions/01-s2-q11

关键思路

以爆炸一刻为界，把整段飞行分成两段。爆炸之前，导弹是普通的抛体（projectile），最高点就是垂直速度为零之处。爆炸之后，碎片 P 是一个新的抛体，从该高度以速度 (u, v_e) 抛出，因此落地所需时间可由一条二次方程（quadratic）求得。至于第二部分，射程（range）是 v_e 的递增函数（increasing function），所以“$R > D$”就变成“v_e 大于恰好令 $R = D$ 的那个值”。

解答

我们把导弹和碎片 P 视为只受重力作用的质点（particle），并设地面为水平。取 A 为原点，x 轴沿水平方向，y 轴垂直向上。



虚线显示假如没有爆炸，导弹本来会落在哪里；紫色曲线则是 P 的路径。

P 的射程

爆炸之前。垂直速度为 $v - gt$，在 $t_1 = v/g$ 时变为零。此刻导弹的水平位移为

$$x_1 = ut_1 = \frac{uv}{g},$$

而它的高度，可由 $v^2 = 2gH$（末垂直速率为零）得出：

$$H = \frac{v^2}{2g}.$$

爆炸之后。碎片 P 从高度 H 出发，水平速度为 u，垂直速度为向上的 v_e。以爆炸一刻起计的时间为 t，它的高度是

$$y = H + v_e t - \tfrac{1}{2}gt^2.$$

当 $y = 0$ 时它落地：

$$\tfrac{1}{2}gt^2 - v_e t - \frac{v^2}{2g} = 0 \quad \Longleftrightarrow \quad g^2t^2 - 2gv_e t - v^2 = 0.$$

把它看成 gt 的二次方程，

$$gt = v_e \pm \sqrt{v_e^2 + v^2}.$$

两根之积为 $-v^2 < 0$，所以其中一个根是负数。这个根代表在爆炸之前抛物线（parabola）与地面相交的时刻，与本题无关。因此爆炸之后的飞行时间是

$$t_2 = \frac{v_e + \sqrt{v_e^2 + v^2}}{g}.$$

水平速度自始至终都是 u，所以总水平距离是

$$R = \frac{uv}{g} + ut_2 = \frac{u}{g}\left(v + v_e + \sqrt{v_e^2 + v^2}\right),$$

亦即

$$\frac{gR}{u} = v + v_e + \sqrt{v_e^2 + v^2},$$

这正是所需的结果。

$R > D$ 的条件

设

$$f(v_e) = v_e + \sqrt{v_e^2 + v^2}, \qquad k = \frac{gD}{u} - v = \frac{gD - uv}{u}.$$

那么 $R > D$ 等价于 $f(v_e) > k$。由于 $D > 2uv/g$，我们有 $gD > 2uv$，所以 $k > v > 0$。

? **想一想** 为甚么只需找出令 $R = D$ 的那一个 v_e 值就足够?

因为 f 是严格递增（strictly increasing）的：

$$f'(v_e) = 1 + \frac{v_e}{\sqrt{v_e^2 + v^2}} > 0,$$

这是由于当 $v \neq 0$ 时 $|v_e| < \sqrt{v_e^2 + v^2}$。（从物理上看，向上的冲击愈猛，$P$ 留在空中的时间就愈长。）所以若 $f(v_e^*) = k$，则 $f(v_e) > k$ 当且仅当 $v_e > v_e^*$。

要求出 v_e^*，解 $\sqrt{v_e^2 + v^2} = k - v_e$。两边平方得

$$v_e^2 + v^2 = k^2 - 2kv_e + v_e^2 \quad \Longrightarrow \quad v_e^* = \frac{k^2 - v^2}{2k} = \frac{(k-v)(k+v)}{2k}.$$

平方可能会引入增根（false root），所以我们要验证 $k - v_e^* \geq 0$：事实上 $k - v_e^* = \dfrac{k^2 + v^2}{2k} > 0$，所以 v_e^* 确实满足 $f(v_e^*) = k$。

现在把 k 代入：

$$k - v = \frac{gD - 2uv}{u}, \qquad k + v = \frac{gD}{u}, \qquad 2k = \frac{2(gD - uv)}{u},$$

所以

$$v_e^* = \frac{(gD - 2uv)\, gD/u^2}{2(gD - uv)/u} = \frac{gD}{2u}\left(\frac{gD - 2uv}{gD - uv}\right).$$

因此，由于 f 是递增的，只要

$$v_e > \frac{gD}{2u}\left(\frac{gD - 2uv}{gD - uv}\right),$$

便有 $R > D$，这就是所需的条件。事实上，以上论证说明这个条件也是必要的。注意因为 $k > v$，所以 $v_e^* > 0$，即确实需要一个向上的冲击。

验算

- 若 $v_e = 0$，公式给出 $gR/u = 2v$，所以 $R = 2uv/g$：这正是没有爆炸的抛体的一般射程，合乎预期。这也是题目假设 $D > 2uv/g$ 的原因，否则根本不需要任何冲击。
- 取 $u = v = 10$、$g = 10$ 及 $D = 30$，临界值为 $v_e^* = \frac{300}{20} \cdot \frac{100}{200} = 7.5$，于是 $gR/u = 10 + 7.5 + 12.5 = 30$，所以 $R = 30 = D$，恰好相等。
- 以数值方法逐步模拟这两段飞行，所得结果与射程公式相差约 10^{-9}；另外 10^5 个随机个案亦确认 $R > D$ 当且仅当 $v_e > v_e^*$。

常见错误

- 忘记爆炸之前已经走过的水平距离 uv/g，把 R 从爆炸点正下方起计，而不是从 A 起计。

- 把由最高点落地所需的时间又当作 v/g，忽略了新增的向上速度 v_e。
- 把 $\sqrt{v_e^2+v^2}=k-v_e$ 两边平方时，没有验证右边为非负，或者没有解释为何一个使等式成立的值可以化为一个不等式。

STEP 之外

爆炸过程中整个系统的动量守恒（conservation of momentum），所以即使 P 偏离了原来的抛物线，两块碎片的质心（centre of mass）仍然沿着原来的抛物线继续前进。内力不能改变质心的运动，这个原理正是大学一年级力学课程中火箭运动和碰撞的起点。Kleppner 与 Kolenkow 合著的 *An Introduction to Mechanics*（力学入门教科书）对这两个课题都有细致的论述。

详细解答

STEP II 2001，第 12 题

概率与统计 · 泊松近似

> **STEP II 2001，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/01-s2-q12

关键思路

合伙集团（syndicate）必定恰好持有一张中奖彩票，所以一切取决于有多少其他投注者选中了中奖号码。这个数目服从二项分布（binomial distribution），试验次数极多而成功概率极小，正是泊松分布（Poisson distribution）接手的情况。于是所得奖金的期望值（expectation）就是 W 乘以 $1/(1+X)$ 的期望值，而这个和可以把 $k+1$ 并入阶乘（factorial），从而求得闭合形式（closed form）。

解答

建立模型

合伙集团买了 N 张彩票，所以其他投注者买了 $3.8N - N = 2.8N$ 张。设 X 为这些非集团彩票中印有中奖号码的张数。

每名非集团投注者都是均匀随机地选一个号码，与其他人的选择及开奖结果独立，所以无论中奖号码是甚么，他们每张彩票与之相符的概率都是 $1/N$，而且各张彩票互相独立。因此

$$X \sim \mathrm{B}\left(2.8N,\ \tfrac{1}{N}\right).$$

由于 N 非常大，$2.8N$ 很大而 $1/N$ 很小，同时平均值 $2.8N \cdot \frac{1}{N} = 2.8$ 保持不变，所以 X 近似服从平均值为 $\lambda = 2.8$ 的泊松分布：

$$\mathrm{P}(X = k) \approx \frac{\mathrm{e}^{-2.8}\,(2.8)^k}{k!}, \qquad k = 0, 1, 2, \ldots$$

合伙集团每个号码都恰好持有一张彩票，所以它总是恰好持有一张中奖彩票，而中奖彩票的总数是 $T = 1 + X$。

最可能出现的中奖彩票数目

记 $p_k = \mathrm{P}(X = k)$。比较相邻的概率，

$$\frac{p_k}{p_{k-1}} = \frac{\lambda}{k} = \frac{2.8}{k},$$

当 $k \le 2$ 时此比值大于 1，当 $k \ge 3$ 时小于 1。所以 $p_0 < p_1 < p_2$ 而 $p_2 > p_3 > p_4 > \cdots$，X 最可能取的值是 2。（对精确的二项分布，结论相同：此时 $p_k/p_{k-1} = \frac{(2.8N-k+1)}{k(N-1)}$，它大于 1 当且仅当 $k < 2.8 + \frac{1}{N}$。）

因此最可能出现的中奖彩票数目是

$$T = 1 + 2 = 3.$$

净损失的期望值

头奖奖金由 $T = 1 + X$ 张中奖彩票平分，而合伙集团持有其中一张，所以它得到 $\dfrac{W}{1+X}$。它付出了 $£N$，所以净损失是

$$L = N - \frac{W}{1+X}, \qquad \text{而} \qquad \mathrm{E}(L) = N - W\,\mathrm{E}\left(\frac{1}{1+X}\right).$$

> **想一想** 怎样求 $\sum \frac{1}{k+1} \cdot \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}$ 这个和?

窍门在于 $(k+1)\cdot k! = (k+1)!$，所以那个碍事的因子 $\frac{1}{k+1}$ 可以并入阶乘。同时乘以和除以 λ，这个和就变成缺了一项的泊松级数：

$$\begin{aligned}\mathrm{E}\left(\frac{1}{1+X}\right) &\approx \sum_{k=0}^{\infty} \frac{1}{k+1} \cdot \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!} = \frac{\mathrm{e}^{-\lambda}}{\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k+1}}{(k+1)!} \\ &= \frac{\mathrm{e}^{-\lambda}}{\lambda} \sum_{j=1}^{\infty} \frac{\lambda^j}{j!} = \frac{\mathrm{e}^{-\lambda}}{\lambda}\left(\mathrm{e}^{\lambda} - 1\right) = \frac{1 - \mathrm{e}^{-\lambda}}{\lambda},\end{aligned}$$

其中我们令 $j = k+1$，并利用 $\sum_{j\ge 0} \lambda^j/j! = \mathrm{e}^{\lambda}$，再减去 $j = 0$ 的一项，即 1。

代入 $\lambda = 2.8 = \frac{14}{5}$，上式等于 $\frac{5}{14}\left(1 - \mathrm{e}^{-2.8}\right)$，于是

$$\mathrm{E}(L) \approx N - \frac{5\left(1 - \mathrm{e}^{-2.8}\right)}{14} W,$$

正是所求的结果。

验算

- 同一窍门也适用于精确的二项分布 $X \sim \mathrm{B}(n,p)$：$\mathrm{E}\left(\frac{1}{1+X}\right) = \dfrac{1-(1-p)^{n+1}}{(n+1)p}$，当 $n \to \infty$ 而 $np = \lambda$ 时，它趋向 $(1 - \mathrm{e}^{-\lambda})/\lambda$。取 $N = 1000$ 时，精确值是 0.33536，而泊松近

似值是 $\frac{5}{14}(1-\mathrm{e}^{-2.8}) \approx 0.33542$。

- 以 $N = 200$ 模拟 10^5 次开奖，所得奖金的平均值为 $0.335W$，而最常出现的结果是 3 张中奖彩票。
- 极限情况：如果没有其他人投注（$\lambda \to 0$），则 $(1-\mathrm{e}^{-\lambda})/\lambda \to 1$，合伙集团独得全部头奖奖金，与预期相符。

常见错误

- 把其他投注者的彩票数目当作 $3.8N$ 而不是 $2.8N$：合伙集团自己的彩票并非随机，它们恰好贡献一张中奖彩票。
- 写成 $\mathrm{E}\left(\frac{W}{1+X}\right) = \frac{W}{1+\mathrm{E}(X)} = \frac{W}{3.8}$。期望值不能直接穿过非线性函数；事实上 $\frac{1}{3.8} \approx 0.263$ 明显小于 0.335（延森不等式（Jensen's inequality）指出它必定不会较大，因为 $1/(1+x)$ 是凸函数（convex function））。

STEP 之外

以 $\mathrm{Po}(\lambda)$ 逼近 $\mathrm{B}(n, \lambda/n)$，就是所谓的“稀有事件定律”（law of rare events）；大学一年级的课程会证明各个概率收敛，或者证明概率生成函数（probability generating function）收敛，从而证明这个定律。计算 $\mathrm{E}\big(1/(1+X)\big)$ 其实就是对生成函数求积分 $\int_0^1 G_X(s)\,\mathrm{d}s$，这是生成函数在计算平均值和方差（variance）以外的一个巧妙用途。Grimmett 与 Welsh 合著的 *Probability: An Introduction*（概率论入门）涵盖了这两个概念。

详细解答

STEP II 2001，第 13 题

概率与统计 · 连续分布与估计

STEP II 2001，第 13 题

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

step.maths.org/questions/01-s2-q13

关键思路

题目直接给出了生存函数（survival function）$\mathrm{P}(T > t)$，所以令它等于 $\frac{1}{2}$ 便可求出中位数（median）；至于平均值（mean），则先求导得出密度函数，再积分。实验提供了样本中位数（九个由小至大排列的寿命中的第五个）和样本平均值，把它们与公式对应起来，便得到关于 α 和 k 的两条方程。两式相除可消去 k，余下一条只含 α 的方程。最后要求的是一个条件概率（conditional probability），而那里正有一个令人惊喜的结果等着我们。

解答

中位数与平均值

设 $S(t) = \mathrm{P}(T > t) = (1+kt)^{-\alpha}$，其中 $t \geq 0$。要使它由 1 下降至 0，必须有 $k > 0$。

*中位数。*中位数 M 满足 $\mathrm{P}(T > M) = \frac{1}{2}$，所以

$$(1+kM)^{\alpha} = 2 \quad \Longrightarrow \quad 1 + kM = 2^{1/\alpha} \quad \Longrightarrow \quad M = \frac{2^{1/\alpha} - 1}{k}.$$

*平均值。*累积分布函数（cumulative distribution function）是 $F(t) = 1 - (1+kt)^{-\alpha}$，所以概率密度函数（probability density function）是

$$f(t) = F'(t) = \alpha k\,(1+kt)^{-\alpha-1}, \qquad t \geq 0.$$

然后用分部积分法（integration by parts），取 $u = t$ 和 $\mathrm{d}v/\mathrm{d}t = \alpha k(1+kt)^{-\alpha-1}$，于是

$v = -(1+kt)^{-\alpha}$，

$$m = \int_0^\infty t\,\alpha k(1+kt)^{-\alpha-1}\,\mathrm{d}t = \Big[-t(1+kt)^{-\alpha}\Big]_0^\infty + \int_0^\infty (1+kt)^{-\alpha}\,\mathrm{d}t$$
$$= 0 + \left[-\frac{(1+kt)^{1-\alpha}}{k(\alpha-1)}\right]_0^\infty = \frac{1}{k(\alpha-1)}.$$

两个在无穷远处的极限都用到 $\alpha > 1$：$t(1+kt)^{-\alpha}$ 的表现类似 $t^{1-\alpha}/k^\alpha \to 0$，而 $(1+kt)^{1-\alpha} \to 0$。（若 $\alpha \le 1$，平均值便会是无限大，这正是题目坚持要求 $\alpha > 1$ 的原因。）所以

$$M = \frac{2^{1/\alpha}-1}{k}, \qquad m = \frac{1}{k(\alpha-1)}.$$

估计 α 和 k

共有九个灯泡，第五个失效的灯泡正是九个顺序排列的寿命中居中的一个：它就是样本中位数（sample median）。所以我们估计 $M \approx 1000$，并由样本平均值得 $m \approx 2400$。

想一想 我们有两条方程和两个未知数，但 k 在两条方程中都出现。怎样才能单独求出 α?

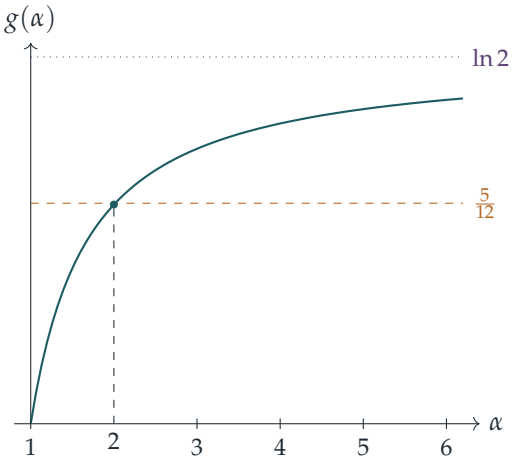
相除：由于 M 和 m 都与 $1/k$ 成正比，k 会被约去。由此得

$$\frac{M}{m} = \left(2^{1/\alpha}-1\right)(\alpha-1) \approx \frac{1000}{2400} = \frac{5}{12} \approx 0.4167.$$

把左方记作 $g(\alpha)$。当 $\alpha = 2$ 时，

$$g(2) = \left(\sqrt{2}-1\right)(2-1) = \sqrt{2}-1 \approx 0.4142,$$

这与 5/12 相差不到 0.003。要看出没有其他数值同样吻合，可留意 $g(1.9) \approx 0.396$ 及 $g(2.1) \approx 0.430$，所以 5/12 这个值介乎两者之间。事实上，g 由 $g(1) = 0$ 开始稳步递增，当 $\alpha \to \infty$ 时趋向 $\ln 2 \approx 0.693$（见图），所以它只会取 5/12 这个值一次，而且是在略大于 2 的某个 α 处。因此 $\alpha \approx 2$。



现在把 $\alpha = 2$ 代回任何一条公式。由平均值得

$$k = \frac{1}{m(\alpha - 1)} \approx \frac{1}{2400},$$

而由中位数得 $k = (\sqrt{2} - 1)/1000 \approx 4.14 \times 10^{-4}$。两者十分接近，于是

$$k \approx 4.2 \times 10^{-4} \text{ 每小时}.$$

条件概率

一个灯泡运作了 M 小时后再多运作 $m - M$ 小时，合共便运作了 m 小时。由于 $m > M$，事件 $\{T > m\}$ 包含于 $\{T > M\}$ 之内，所以

$$\mathrm{P}(T > m \mid T > M) = \frac{\mathrm{P}(T > m)}{\mathrm{P}(T > M)} = \frac{(1 + km)^{-\alpha}}{1/2} = \frac{2}{(1 + km)^{\alpha}}.$$

（严格来说，是在 M 小时那一刻发现灯泡仍在运作；对于连续的寿命，$T \geq M$ 与 $T > M$ 的概率相同。）惊喜来了：不论 k 是多少，都恰好有 $km = 1/(\alpha - 1)$，所以 $1 + km = \alpha/(\alpha - 1)$，而

$$\mathrm{P}(T > m \mid T > M) = 2\left(\frac{\alpha - 1}{\alpha}\right)^{\alpha}.$$

当 $\alpha = 2$ 时，这等于 $2 \cdot \left(\frac{1}{2}\right)^2 = \frac{1}{2}$。所以概率的估计值是

$$\mathrm{P}(\text{再多运作} m - M \text{ 小时}) \approx \tfrac{1}{2}.$$

最后这一步完全用不着 k 的估计值。

验算

- 在中位数处，$S(M) = (2^{1/\alpha})^{-\alpha} = \frac{1}{2}$，与预期相符。随机选取 $\alpha > 1$ 和 $k > 0$，对 $S(t)$ 在 $[0, \infty)$ 上作数值积分，结果与 $m = 1/(k(\alpha - 1))$ 吻合（回想对于非负的变量，有 $\mathrm{E}(T) = \int_0^\infty \mathrm{P}(T > t)\,\mathrm{d}t$，这正是我们计算过程中的第二个积分）。
- 准确地解 $g(\alpha) = 5/12$ 可得 $\alpha \approx 2.015$，这时条件概率是 $2(1 - 1/\alpha)^\alpha \approx 0.502$，所以把它约为 $\alpha = 2$ 几乎没有任何损失。
- 以 $\alpha = 2$、$k = 1/2400$ 模拟 400 000 个寿命（取 $T = (U^{-1/2} - 1)/k$，其中 U 在 $(0, 1)$ 上均匀分布），得出 $\mathrm{P}(T > m \mid T > M) \approx 0.501$。

常见错误

- 把 $\mathrm{P}(T > t)$ 当作累积分布函数，求导时漏掉负号，结果得出负的“密度”。
- 以 $\mathrm{P}(T > m - M)$ 作答，仿佛灯泡在时刻 M 重新开始计算寿命。这个分布并不具有无记忆性（memoryless；只有指数分布才有），所以条件 $T > M$ 是有影响的。
- 把第五个失效的时间当作平均值的估计，而不是中位数的估计。

STEP 之外

这是一种帕累托（Pareto）型分布，称为 Lomax 分布，是描述寿命和收入的标准模型，带有重尾（heavy tail）：$\mathrm{E}(T^r)$ 只在 $r < \alpha$ 时才是有限的。我们在这里所用的是*矩估计法*（method of moments；严格来说，是配对一个分位数和一个矩）：选取参数，使模型的概括统计量等于样本的概括统计量。大学一年级的统计学会把这个方法与最大似然估计（maximum likelihood）作比较，并探讨只凭九个灯泡得出的估计究竟有多准确。比值 $\mathrm{P}(T > s + t \mid T > s)$ 是可靠性理论（reliability theory）中所用的生存函数：在这里它随 s *递增*，所以旧灯泡比新灯泡更有可能继续运作下去。Grimmett 与 Welsh 合著的 *Probability: An Introduction*（概率入门）是很好的下一步读物。

详细解答

STEP II 2001，第 14 题

概率与统计 · 最小方差估计量

STEP II 2001，第 14 题

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

step.maths.org/questions/01-s2-q14

关键思路

T 是两个独立（independent）伯努利变量（Bernoulli variable）的加权组合，因此它的平均值（mean）和方差（variance）可直接由期望值的线性性质（linearity of expectation），以及适用于独立的 X、Y 的法则 $\operatorname{Var}(aX+cY)=a^2\operatorname{Var}(X)+c^2\operatorname{Var}(Y)$ 得出。权重经过特意选取，使得无论 λ 取何值，平均值都是 p：T 是 p 的*无偏估计量*（unbiased estimator），而题目要找的正是其中最好的一个。至于最后一部分，中心极限定理（central limit theorem）令 $\overline{T}$ 近似服从正态分布（normal distribution），而平均值固定的正态变量，方差最小时最为集中。

解答

T 的平均值与方差

这里 X 和 Y 都是伯努利变量，且

$$\mathrm{E}(X)=p,\quad \operatorname{Var}(X)=p(1-p),\qquad \mathrm{E}(Y)=2p,\quad \operatorname{Var}(Y)=2p(1-2p).$$

（若一个变量以概率 q 取值 1，否则取值 0，则 $\mathrm{E}=q$ 且 $\mathrm{E}(\cdot^2)=q$，所以方差是 $q-q^2=q(1-q)$。）由期望值的线性性质，

$$\mathrm{E}(T)=\lambda p+\tfrac{1}{2}(1-\lambda)\cdot 2p=\lambda p+(1-\lambda)p=p,$$

正如所求。由于两枚硬币互相独立，X 和 Y 亦独立，所以没有协方差（covariance）项，而

$$\begin{aligned}\operatorname{Var}(T)&=\lambda^2\operatorname{Var}(X)+\tfrac{1}{4}(1-\lambda)^2\operatorname{Var}(Y)\\&=\lambda^2 p(1-p)+\tfrac{1}{2}(1-\lambda)^2 p(1-2p).\end{aligned}$$

求方差的最小值

记 $V(\lambda) = \mathrm{Var}(T)$。这是 λ 的二次函数，首项系数 $p(1-p) + \frac{1}{2}p(1-2p)$ 为正（因为 $0 < p < \frac{1}{2}$，两项皆为正），所以它只有一个最小值，出现在 $V'(\lambda) = 0$ 之处：

$$V'(\lambda) = 2\lambda p(1-p) - (1-\lambda)p(1-2p) = 0.$$

两边除以 $p > 0$，再把含 λ 的项合并，

$$\lambda\big[2(1-p) + (1-2p)\big] = 1 - 2p \quad \Longrightarrow \quad \lambda = \frac{1-2p}{3-4p},$$

这正是题目给出的值。（这里 $3 - 4p > 1 > 0$，所以不会出现除以零的情况。）

我们还需要最小值本身。设 $a = p(1-p)$ 及 $c = \frac{1}{2}p(1-2p)$，则 $V = a\lambda^2 + c(1-\lambda)^2$，在 $\lambda = c/(a+c)$（即上面的值）处取得最小值，由此得

$$V_{\min} = a\Big(\frac{c}{a+c}\Big)^2 + c\Big(\frac{a}{a+c}\Big)^2 = \frac{ac(a+c)}{(a+c)^2} = \frac{ac}{a+c}.$$

现在 $a + c = \frac{1}{2}p\big[2(1-p) + (1-2p)\big] = \frac{1}{2}p(3-4p)$ 而 $ac = \frac{1}{2}p^2(1-p)(1-2p)$，所以

$$V_{\min} = \frac{p(1-p)(1-2p)}{3-4p}.$$

n 次掷币的平均值

设 $T_1, \ldots, T_n$ 为 n 次掷币中 T 的取值；它们互相独立，各自的平均值为 p，方差为 $V(\lambda)$。于是

$$\mathrm{E}(\overline{T}) = p, \qquad \mathrm{Var}(\overline{T}) = \frac{V(\lambda)}{n}.$$

> **想一想** T 只取四个值。为甚么我们可以把 $\overline{T}$ 当作正态变量来处理?

因为 $\overline{T}$ 是 n 个独立同分布（independent, identically distributed）且方差有限的变量的平均值，而 $n > 30$ 够大，所以根据中心极限定理，不论单独一个 T 的分布形状如何，$\overline{T}$ 都近似服从 $\mathrm{N}\big(p, V(\lambda)/n\big)$。记 $\sigma_\lambda = \sqrt{V(\lambda)/n}$ 及 $Z = (\overline{T} - p)/\sigma_\lambda \approx \mathrm{N}(0,1)$，则

$$\mathrm{P}\big(|\overline{T} - p| < b\big) = \mathrm{P}\Big(|Z| < \frac{b}{\sigma_\lambda}\Big) \approx \Phi\Big(\frac{b}{\sigma_\lambda}\Big) - \Phi\Big(-\frac{b}{\sigma_\lambda}\Big) = 2\Phi\Big(\frac{b}{\sigma_\lambda}\Big) - 1,$$

其中用到对称性 $\Phi(-z) = 1 - \Phi(z)$。

由于 $b > 0$ 固定，而 Φ 严格递增（strictly increasing），上式在 b/σ_λ 最大时最大，亦即在 σ_λ 最小时最大，也就是 $V(\lambda)$ 取最小值之时。这发生在 $\lambda = (1-2p)/(3-4p)$，此时

$$\sigma_\lambda^2 = \frac{V_{\min}}{n} = \frac{p(1-p)(1-2p)}{(3-4p)n} = s^2.$$

所以当 λ 变动时，$\mathrm{P}\big(|\overline{T} - p| < b\big)$ 的最大值约为 $2\Phi(b/s) - 1$，正如所求。

验算

- 当 $\lambda = 1$ 时，$T = X$，公式给出 $\operatorname{Var}(T) = p(1-p)$；当 $\lambda = 0$ 时，$T = \frac{1}{2}Y$，公式给出 $\frac{1}{4} \cdot 2p(1-2p)$。两者都正确。
- $V_{\min}$ 比这两个值都小：它等于 $p(1-p)$ 乘以 $\frac{1-2p}{3-4p} < 1$。把两枚硬币结合起来，胜过单独使用其中任何一枚。
- 数值上，取 $p = 0.3$、$n = 100$、$b = 0.03$：最佳的 λ 是 $\frac{2}{9}$，$s \approx 0.0216$，而 $2\Phi(b/s) - 1 \approx 0.835$。模拟 40 000 轮、每轮掷币 100 次，得出 0.836；而 $\lambda = 1$（只用硬币 A）则只得约 0.487。

常见错误

- 写成 $\operatorname{Var}(\frac{1}{2}(1-\lambda)Y) = \frac{1}{2}(1-\lambda)\operatorname{Var}(Y)$：常数从方差中提出来时要平方。
- 忘记说明为甚么把方差减至最小就能使概率最大：这需要 Φ 递增，以及对每个 λ 而言 $\overline{T}$ 的平均值都是 p。

STEP 之外

这是估计理论（estimation theory）的初次体验：在 p 的所有无偏估计量 $\lambda X + \frac{1}{2}(1-\lambda)Y$ 之中，我们找出了方差最小的一个。答案是按各个观测值的方差成反比来加权，这与统计学中用来结合精确度不同的量度结果的法则相同（逆方差加权，inverse-variance weighting），而最小值 $ac/(a+c)$ 正是“并联电阻”公式 $1/V_{\min} = 1/a + 1/c$。Grimmett 与 Welsh 的 *Probability: An Introduction*（概率论入门教材）涵盖了最后一步背后的中心极限定理。

详细解答

STEP III 2001，第 1 题

纯数学 · 麦克劳林级数

> **STEP III 2001，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/01-s3-q1

关键思路

函数 $y = \ln\left(x + \sqrt{x^2+1}\right)$ 就是 $\sinh^{-1} x$，而它的一阶导数（derivative）远比函数本身简单。把这个导数平方，便得到一个不含平方根的关系式 $(x^2+1)(y')^2 = 1$，再对它求导，就得到递推关系（recurrence）中 $n = 0$ 的情况。每多求导一次，递推关系便上移一步，这正是归纳法（induction）所需要的。在 $x = 0$ 处，含 x 的一项消失，递推关系把 $y^{(n+2)}(0)$ 直接与 $y^{(n)}(0)$ 连系起来，于是所有麦克劳林（Maclaurin）系数几乎唾手可得。

解答

导数

设 $s = \sqrt{x^2+1}$，则 $\dfrac{\mathrm{d}s}{\mathrm{d}x} = \dfrac{x}{s}$。留意对任意实数 x 都有 $x + s > 0$，因为 $s > \sqrt{x^2} = |x|$，所以这个对数处处有定义。由链式法则（chain rule），

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{x+s}\left(1 + \frac{x}{s}\right) = \frac{1}{x+s} \cdot \frac{s+x}{s} = \frac{1}{\sqrt{x^2+1}},$$

正如所求。

归纳法

设 $P(n)$ 为以下命题：

$$(x^2+1)\,y^{(n+2)} + (2n+1)\,x\,y^{(n+1)} + n^2\,y^{(n)} = 0 \quad \text{对所有实数} x.$$

基础情况，$n = 0$。我们需要证明 $(x^2+1)y'' + xy' = 0$。由第一部分，$(x^2+1)(y')^2 = 1$。

对 x 求导，

$$2x(y')^2 + (x^2+1)\cdot 2y'y'' = 0.$$

由于 $y' = 1/\sqrt{x^2+1}$ 永不为零，我们可以除以 $2y'$，得到 $(x^2+1)y'' + xy' = 0$。这就是 $P(0)$。（另一做法是直接对 $y' = (x^2+1)^{-1/2}$ 求导：$y'' = -x(x^2+1)^{-3/2}$，所以 $(x^2+1)y'' = -x(x^2+1)^{-1/2} = -xy'$。）

*归纳步骤。*假设 $P(k)$ 对某个 $k \geq 0$ 成立：

$$(x^2+1)\,y^{(k+2)} + (2k+1)\,x\,y^{(k+1)} + k^2\,y^{(k)} = 0.$$

此式对所有 x 都成立，所以我们可以对它求导。对首两项运用积法则（product rule），

$$2x\,y^{(k+2)} + (x^2+1)\,y^{(k+3)} + (2k+1)\,y^{(k+1)} + (2k+1)\,x\,y^{(k+2)} + k^2\,y^{(k+1)} = 0.$$

合并同类项：

$$(x^2+1)\,y^{(k+3)} + (2k+3)\,x\,y^{(k+2)} + (k^2+2k+1)\,y^{(k+1)} = 0,$$

而由于 $2k+3 = 2(k+1)+1$ 及 $k^2+2k+1 = (k+1)^2$，这正正就是 $P(k+1)$。

> **想一想** 为甚么 $y^{(k+2)}$ 的系数刚好增加 2?

对第一项中的 (x^2+1) 求导，会多出一个 $2x\,y^{(k+2)}$，它与第二项的 $(2k+1)x\,y^{(k+2)}$ 合并。同样地，对第二项中的 x 求导所得的 $(2k+1)$ 与 k^2 合并，而 k^2+2k+1 是一个完全平方（perfect square）。这个递推关系天生就会复制自己。

由于 $P(0)$ 成立，且对每个 $k \geq 0$ 都有 $P(k) \Rightarrow P(k+1)$，根据归纳法，$P(n)$ 对所有 $n \geq 0$ 成立。

麦克劳林级数

在 $P(n)$ 中代入 $x = 0$。中间一项消失，剩下

$$y^{(n+2)}(0) = -n^2\,y^{(n)}(0) \qquad (n \geq 0).$$

我们需要两个初始值：$y(0) = \ln(0+1) = 0$ 和 $y'(0) = 1/\sqrt{1} = 1$。

*偶数阶导数。*由 $y(0) = 0$，递推关系给出 $y''(0) = 0$，继而 $y^{(4)}(0) = 0$，如此类推：每个偶数阶导数在 0 处都等于零。（这亦可由 y 是奇函数（odd function）推出：$y(-x) = \ln\bigl(\sqrt{x^2+1} - x\bigr) = -\ln\bigl(\sqrt{x^2+1} + x\bigr)$，因为 $\bigl(\sqrt{x^2+1} - x\bigr)\bigl(\sqrt{x^2+1} + x\bigr) = 1$。）

*奇数阶导数。*由 $y'(0) = 1$ 出发，

$$\begin{aligned} y'''(0) &= -1^2 \cdot y'(0) = -1, \\ y^{(5)}(0) &= -3^2 \cdot y'''(0) = 9, \\ y^{(7)}(0) &= -5^2 \cdot y^{(5)}(0) = -225. \end{aligned}$$

想一想 要由 $y'''(0)$ 得到 $y^{(5)}(0)$，n 应取甚么值?

要得到 $n+2=5$ 的 $y^{(n+2)}$，我们需要 $n=3$，所以因子是 $-3^2=-9$，而不是 -5^2。这个因子用的是较低阶导数的阶数。

因此麦克劳林级数（Maclaurin series）$y=\sum y^{(k)}(0)\,x^k/k!$ 为

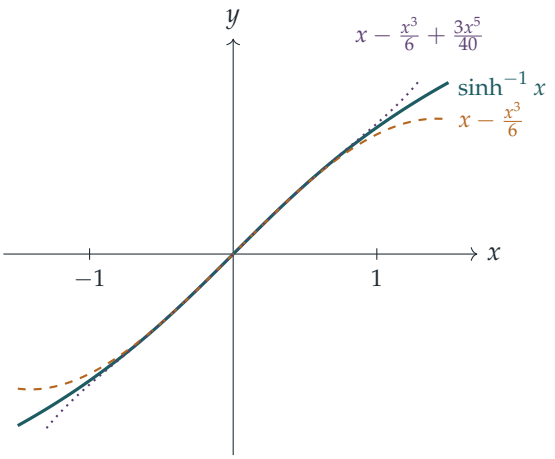
$$\begin{aligned} y &= x+\frac{-1}{3!}x^3+\frac{9}{5!}x^5+\frac{-225}{7!}x^7+\cdots \\ &= x-\frac{x^3}{6}+\frac{9}{120}x^5-\frac{225}{5040}x^7+\cdots \\ &= x-\frac{x^3}{6}+\frac{3x^5}{40}-\frac{5x^7}{112}+\cdots, \end{aligned}$$

其开首各项与题目所求相符，而下一个非零项是

$$-\frac{5x^7}{112}.$$

另一途径。把二项式级数（binomial series）$(1+x^2)^{-1/2}=1-\frac{1}{2}x^2+\frac{3}{8}x^4-\frac{5}{16}x^6+\cdots$ 逐项积分（对 $|x|<1$ 成立），并利用 $y(0)=0$，便得到 $x-\frac{1}{6}x^3+\frac{3}{40}x^5-\frac{5}{112}x^7+\cdots$，结果一致。

下图显示部分和（partial sums）在原点附近如何紧贴 $y=\sinh^{-1}x$，而在较远处则交替地偏高和偏低。



验算

- 当 $x=0.1$ 时，$\ln\left(0.1+\sqrt{1.01}\right)=0.0998340789$，而 $x-\frac{x^3}{6}+\frac{3x^5}{40}-\frac{5x^7}{112}=0.0998340789$ 亦然；两者之差约为 3×10^{-11}，即下一项（x^9 项）的大小。

- $n = 0$ 的情况很容易在 $x = 1$ 处检验：$y' = 1/\sqrt{2}$ 而 $y'' = -1/(2\sqrt{2})$，所以 $2y'' + y' = 0$。
- 在 0 处的奇数阶导数为 $1, -1, 9, -225$，即 $(-1)^m\big(1 \cdot 3 \cdots (2m-1)\big)^2$：这是一个赏心悦目、值得留意的规律。

常见错误

- 在 $y^{(n+2)}(0) = -n^2 y^{(n)}(0)$ 中用错了 n，得出因子 $-9, -25, -49$，而不是 $-1, -9, -25$。
- 忘记除以 $k!$：x^7 的系数是 $-225/5040$，而不是 -225。
- 在归纳步骤中，漏掉了对 $x^2 + 1$ 求导所产生的 $2x\,y^{(k+2)}$。

STEP 之外

泰勒级数（Taylor series）只在它收敛（converge）的地方才有用，而这个级数恰好在 $|x| \leq 1$ 时收敛，尽管 $\sinh^{-1} x$ 在整条实数线上都十分光滑。个中原因藏在复平面（complex plane）上：$1/\sqrt{1+z^2}$ 在 $z = \pm\mathrm{i}$ 处趋于无穷，这两点与原点的距离为 1，而这个距离就是收敛半径（radius of convergence）。Needham 的 *Visual Complex Analysis*（视觉化复分析）对此有精彩的解释；至于泰勒定理（Taylor's theorem）及其余项在实变量方面的论述，可参阅 Abbott 的 *Understanding Analysis*（理解分析学）。

详细解答

STEP III 2001，第 2 题

纯数学 · 双曲代换与面积

STEP III 2001，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/01-s3-q2

关键思路

两个不等式说的都是“留在某条等轴型双曲线（rectangular-type hyperbola）的其中一侧”，所以第一步是画出这两条双曲线的草图，并找出它们的交点。结果发现，所求区域是一块有界、具四重对称性的范围，而它的面积可写成两个形如 $\int \sqrt{x^2 \pm a^2}\,\mathrm{d}x$ 的积分。这类积分正好可用双曲代换（hyperbolic substitution）来处理，而题目的第一部分，就是为了让我们把所得的 $\cosh^{-1}$（以及 $\sinh^{-1}$）的值化为对数。

解答

$\cosh^{-1} x$ 的对数形式

设 $x \geq 1$，并设 $y = \cosh^{-1} x$；按定义，它是满足 $\cosh y = x$ 的 非负数。于是

$$x = \frac{\mathrm{e}^{y} + \mathrm{e}^{-y}}{2} \quad \Longrightarrow \quad \mathrm{e}^{2y} - 2x\,\mathrm{e}^{y} + 1 = 0,$$

这是一条以 e^{y} 为未知数的二次方程（quadratic），所以

$$\mathrm{e}^{y} = x \pm \sqrt{x^2 - 1}.$$

? **想一想** 应该取哪一个符号？为甚么？

两个根的积是 1（即二次方程的常数项），而取正号的根至少是 1。因此取负号的根至多是 1，这会得出 $y \leq 0$。由于 $y \geq 0$，我们必须取正号（当 $x = 1$ 时，两个根本来就相等），所以

$$\cosh^{-1} x = \ln\bigl(x + \sqrt{x^2 - 1}\bigr),$$

这正是所要证明的结果。以 $x = \sinh y$ 作完全相同的论证，可得 $\mathrm{e}^{2y} - 2x\,\mathrm{e}^{y} - 1 = 0$；这次两个根的积是 -1，所以只有取正号的根是正数，于是 $\sinh^{-1} x = \ln\left(x + \sqrt{x^2+1}\right)$。这两条公式稍后都会用到。

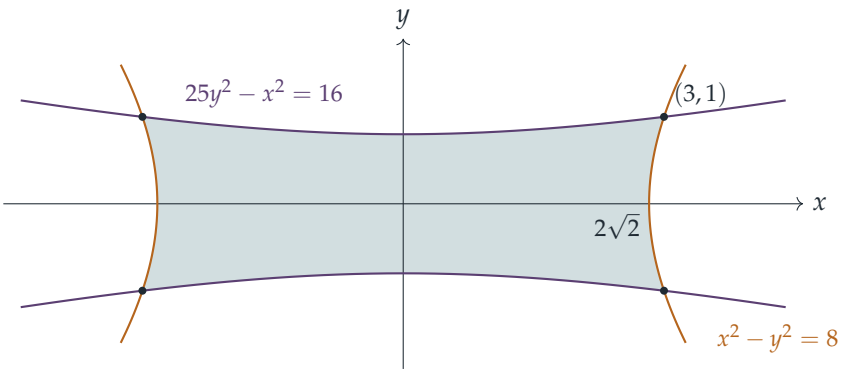
区域

把两个不等式改写为

$$x^2 - y^2 \le 8 \qquad 及 \qquad 25y^2 - x^2 \le 16.$$

曲线 $x^2 - y^2 = 8$ 是一条双曲线，两支分别向左、向右开口，并通过 $(\pm 2\sqrt{2}, 0)$；第一个不等式要求我们留在它的两支之间，即 $|x| \le \sqrt{y^2+8}$。曲线 $25y^2 - x^2 = 16$ 是一条双曲线，两支分别向上、向下开口，并通过 $(0, \pm\frac{4}{5})$；第二个不等式要求我们留在这两支之间，即 $|y| \le \frac{1}{5}\sqrt{x^2+16}$。

两条曲线相交之处满足 $x^2 = y^2 + 8$ 及 $25y^2 = x^2 + 16 = y^2 + 24$，所以 $y^2 = 1$ 及 $x^2 = 9$：交点就是 $(\pm 3, \pm 1)$ 这四点。



这个区域（图中阴影部分）是有界的：若 (x, y) 在区域内，则 $x^2 \le y^2 + 8 \le \frac{1}{25}(x^2+16) + 8$，所以 $x^2 \le 9$，继而 $y^2 \le \frac{25}{25} = 1$。两个不等式在 $x \mapsto -x$ 及 $y \mapsto -y$ 之下都保持不变，所以区域对两条坐标轴都对称，其面积是第一象限（first quadrant）内那部分的面积 A 的四倍。

在第一象限内，我们沿垂直方向切割。

- 当 $0 \le x \le 2\sqrt{2}$ 时，我们有 $x^2 - 8 \le 0 \le y^2$，所以第一个不等式自动成立，而 y 由 0 上升至 $\frac{1}{5}\sqrt{x^2+16}$。
- 当 $2\sqrt{2} \le x \le 3$ 时，y 由 $\sqrt{x^2-8}$ 上升至 $\frac{1}{5}\sqrt{x^2+16}$（而根据交点的计算，下方的曲线位于上方的曲线之下，当且仅当 $x^2 \le 9$）。

因此

$$A = \frac{1}{5}\int_0^3 \sqrt{x^2+16}\,\mathrm{d}x \;-\; \int_{2\sqrt{2}}^3 \sqrt{x^2-8}\,\mathrm{d}x = \tfrac{1}{5}I_1 - I_2.$$

第一个积分

代入 $x = 4\sinh s$，于是 $\mathrm{d}x = 4\cosh s\,\mathrm{d}s$ 及 $\sqrt{x^2+16} = 4\cosh s$（是正数，因为 $\cosh s > 0$）。积分上下限是 $s = 0$ 及 $s = S = \sinh^{-1}\frac{3}{4} = \ln\left(\frac{3}{4}+\frac{5}{4}\right) = \ln 2$。利用 $\cosh^2 s = \frac{1}{2}(\cosh 2s + 1)$，

$$I_1 = \int_0^S 16\cosh^2 s\,\mathrm{d}s = \int_0^S 8(\cosh 2s + 1)\,\mathrm{d}s = \Big[8\sinh s\cosh s + 8s\Big]_0^S$$
$$= 8\cdot\tfrac{3}{4}\cdot\tfrac{5}{4} + 8\ln 2 = \tfrac{15}{2} + 8\ln 2,$$

因为 $\sinh S = \frac{3}{4}$ 及 $\cosh S = \sqrt{1+\frac{9}{16}} = \frac{5}{4}$。

第二个积分

代入 $x = 2\sqrt{2}\cosh t$，其中 $t \geq 0$，于是 $\mathrm{d}x = 2\sqrt{2}\sinh t\,\mathrm{d}t$ 及 $\sqrt{x^2-8} = 2\sqrt{2}\sinh t$（是正数，因为 $t \geq 0$）。积分上下限是 $t = 0$ 及

$$t = T = \cosh^{-1}\frac{3}{2\sqrt{2}} = \ln\left(\frac{3}{2\sqrt{2}} + \sqrt{\tfrac{9}{8}-1}\right) = \ln\frac{4}{2\sqrt{2}} = \ln\sqrt{2} = \tfrac{1}{2}\ln 2,$$

这是由第一部分得出的。利用 $\sinh^2 t = \frac{1}{2}(\cosh 2t - 1)$，

$$I_2 = \int_0^T 8\sinh^2 t\,\mathrm{d}t = \int_0^T 4(\cosh 2t - 1)\,\mathrm{d}t = \Big[4\sinh t\cosh t - 4t\Big]_0^T$$
$$= 4\cdot\frac{1}{2\sqrt{2}}\cdot\frac{3}{2\sqrt{2}} - 2\ln 2 = \tfrac{3}{2} - 2\ln 2,$$

因为 $\cosh T = \frac{3}{2\sqrt{2}}$ 及 $\sinh T = \sqrt{\frac{9}{8}-1} = \frac{1}{2\sqrt{2}}$。

综合结果

$$A = \tfrac{1}{5}\left(\tfrac{15}{2} + 8\ln 2\right) - \left(\tfrac{3}{2} - 2\ln 2\right) = \tfrac{3}{2} + \tfrac{8}{5}\ln 2 - \tfrac{3}{2} + 2\ln 2 = \tfrac{18}{5}\ln 2.$$

代数部分刚好完全抵销，而总面积是

$$4A = \frac{72}{5}\ln 2,$$

这正是所要证明的结果。

验算

- $\frac{72}{5}\ln 2 \approx 9.981$。对两个积分使用森逊法则（Simpson's rule），得出 $I_1 \approx 13.0452$ 及 $I_2 \approx 0.11371$，与 $\frac{15}{2} + 8\ln 2$ 及 $\frac{3}{2} - 2\ln 2$ 吻合；另外，直接在一个细密的网格上数算同时满足两个不等式的点，得出的面积约为 9.98。

- 检查大小是否合理：这个区域可放进矩形 $[-3,3]\times[-1,1]$ 之内，该矩形的面积是 12，而区域只是欠缺了矩形四角几块细长的小片，所以面积略小于 10 是相当合理的。
- $\cosh^{-1}1=\ln 1=0$，与应有的结果一致。

常见错误

- 把不等式理解为双曲线“之外”：$y^2\geq x^2-8$ 是 $x^2-y^2=8$ 两支之间的区域，当中包含 y-轴。
- 忘记第一个不等式只在 $x>2\sqrt{2}$ 时才起作用，因而减去 $\int_0^3\sqrt{x^2-8}\,\mathrm{d}x$，而这个积分甚至不是实数。
- 在 $\mathrm{e}^y=x\pm\sqrt{x^2-1}$ 中取负号，这样得出的是 $-\cosh^{-1}x$。

STEP 之外

$\int\sqrt{x^2+a^2}\,\mathrm{d}x$ 和 $\int\sqrt{x^2-a^2}\,\mathrm{d}x$ 这两类积分可用双曲代换处理，道理与 $\int\sqrt{a^2-x^2}\,\mathrm{d}x$ 可用 $x=a\sin\theta$ 处理一样：恒等式 $\cosh^2-\sinh^2=1$ 能消去平方根。事实上，双曲函数（hyperbolic functions）量度的是面积：$(\cosh t,\sinh t)$ 中的参数 t 是单位双曲线上扇形面积的两倍，正如 θ 是圆扇形面积的两倍。Siklos 的 *Advanced Problems in Mathematics*（进阶数学难题集）还有很多这类风格的题目，而 Bartle 与 Sherbert 的 *Introduction to Real Analysis*（实分析导论）则说明指数函数、对数函数以至双曲函数，如何建立在严格的基础之上。

详细解答

STEP III 2001，第 3 题

纯数学 · 二次方程的根

STEP III 2001，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/01-s3-q3

关键思路

把关于根(root)的叙述转化为关于系数(coefficient)的叙述。设两根为 α 和 β，则 $\alpha+\beta=b$ 且 $\alpha\beta=c$，而判别式（discriminant）可告诉我们两根何时为实数。至于第 (ii) 部分，从 $f(x)=x^2-bx+c$ 的图像入手会较为容易：两根都严格介乎 -1 与 1 之间，当且仅当抛物线（parabola）在 ±1 处取正值，且其顶点（vertex）位于两者之间。

解答

以下一律记 $f(x)=x^2-bx+c$；当两根为实数时，称之为 $\alpha\le\beta$，于是

$$\alpha+\beta=b,\qquad \alpha\beta=c,\qquad \alpha,\beta=\tfrac{1}{2}\left(b\mp\sqrt{b^2-4c}\right).$$

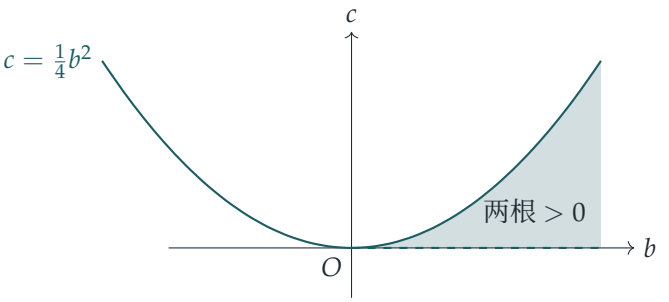
重根（repeated root）视作两个相等的实根。

第 (i) 部分

($\Rightarrow$) 假设两根皆为正实数。有实根便须 $b^2-4c\ge0$。又因两根皆为正数，故 $b=\alpha+\beta>0$ 且 $c=\alpha\beta>0$。合起来便得 $b>0$ 及 $b^2\ge4c>0$。

($\Leftarrow$) 假设 $b>0$ 且 $b^2\ge4c>0$。由于 $b^2-4c\ge0$，两根为实数。由于 $\alpha\beta=c>0$，两根均不为零且同号。由于 $\alpha+\beta=b>0$，这个共同的符号不可能是负号。所以两根皆为正数。

这就证明了两者等价。在 b-c 平面上，这些条件即 $b>0$ 及 $0<c\le\frac{1}{4}b^2$：也就是第一象限（first quadrant）内介乎正 b 轴与抛物线 $c=\frac{1}{4}b^2$ 之间的区域。抛物线包括在内（重根 $\frac{1}{2}b>0$），而 b 轴则不包括在内（该处 $c=0$，其中一根为 0）。



实线曲线（$b > 0$ 的部分）属于该区域；b 轴上的虚线线段和原点则不属于该区域。

第 (ii) 部分

我们需要两根皆为实数，且 $-1 < \alpha \le \beta < 1$。我们断言：在两根为实数的前提下，这情况成立当且仅当

$$f(1) = 1 - b + c > 0, \qquad f(-1) = 1 + b + c > 0, \qquad -1 < \tfrac{1}{2}b < 1.$$

($\Rightarrow$) 若 $-1 < \alpha \le \beta < 1$，则 $f(1) = (1-\alpha)(1-\beta) > 0$ 且 $f(-1) = (1+\alpha)(1+\beta) > 0$，因为每个括号内的数都是正数。此外，$\frac{1}{2}b = \frac{1}{2}(\alpha + \beta)$ 是 $(-1, 1)$ 内两个数的平均值，所以它也在 $(-1, 1)$ 内。

($\Leftarrow$) 因为 $f(x) = (x-\alpha)(x-\beta)$ 恰好在 $[\alpha, \beta]$ 上 ≤ 0，所以条件 $f(1) > 0$ 表示 1 不在 $[\alpha, \beta]$ 内。因此要么 $1 < \alpha$，要么 $1 > \beta$。顶点 $\frac{1}{2}b$ 位于 $[\alpha, \beta]$ 内，而 $\frac{1}{2}b < 1$，故 $\alpha \le \frac{1}{2}b < 1$。这就排除了 $1 < \alpha$，所以 $\beta < 1$。同理，由 $f(-1) > 0$ 和 $\frac{1}{2}b > -1$ 可得 $\alpha > -1$。

? **想一想** 为甚么不能只用 $f(1) > 0$ 和 $f(-1) > 0$ 这两个条件?

因为两根有可能都落在区间的同一侧。例如 $x^2 - 5x + 6 = 0$ 的根是 2 和 3，但 $f(\pm 1) > 0$。正是顶点条件 $|b| < 2$ 令两根落在 -1 与 1 之间，而不是同时落在某一端之外。

把这些条件与实根条件 $b^2 \ge 4c$ 合并，所求区域为

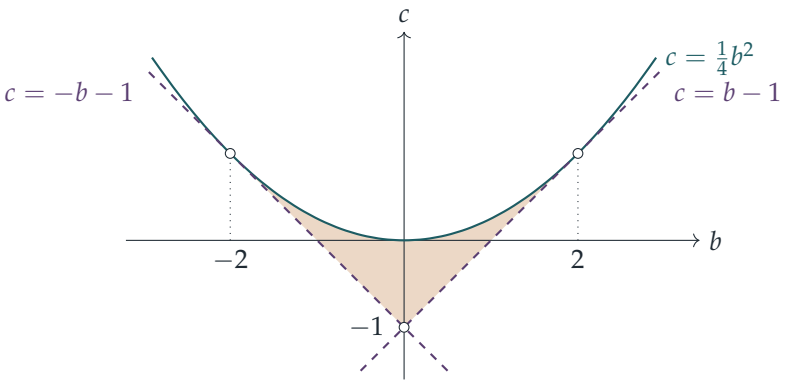
$$|b| < 2, \qquad c > b - 1, \qquad c > -b - 1, \qquad c \le \tfrac{1}{4}b^2,$$

即 $|b| < 2$ 且 $|b| - 1 < c \le \frac{1}{4}b^2$。

绘画草图时，要留意两条直线与抛物线如何相交。由于

$$\tfrac{1}{4}b^2 - (b-1) = \tfrac{1}{4}(b-2)^2 \ge 0,$$

直线 $c = b - 1$ 位于抛物线下方，且只在 $(2, 1)$ 与它相接。所以它就是抛物线在该点的切线 (tangent)。根据对称性 ($b \mapsto -b$)，$c = -b - 1$ 是在 $(-2, 1)$ 的切线。两条直线相交于 $(0, -1)$。所求区域是一个曲边三角形，由这两条切线和抛物线在两个切点之间的弧围成。



抛物线上 $|b| < 2$ 的弧包括在内（重根 $\frac{1}{2}b$，且 $|\frac{1}{2}b| < 1$）。两段虚线切线不包括在内，因为在它们上面有一根为 ± 1。三个角点 $(\pm 2, 1)$ 和 $(0, -1)$ 同样不包括在内。

验算

- $(b, c) = (0, -\frac{1}{4})$ 在第 (ii) 部分的区域之内：两根为 $\pm\frac{1}{2}$。$(b, c) = (0, -1)$ 是被排除的角点：两根恰好是 ± 1。
- $(b, c) = (2.7, 1.8)$ 满足 $c > |b| - 1$ 和 $c \leq \frac{1}{4}b^2$，但 $|b| > 2$。它的两根是 1.2 和 1.5，所以顶点条件把它排除在外是正确的。
- 我们以数值方法检验了两个区域：在超过 115 000 个有理点 (b, c) 上，把准确的根与所述条件互相比较，当中包括每条边界曲线上的点。结果完全吻合。

常见错误

- 在第 (ii) 部分遗漏顶点条件 $|b| < 2$，以致错误地加入了 $|b| = 2$ 以外、夹在每条切线与抛物线之间的狭长部分。
- 弄错边界：两部分的区域都包括抛物线，但第 (i) 部分的 b 轴和第 (ii) 部分的切线都不包括在内。

STEP 之外

关于 $f(\pm 1)$ 和顶点的条件，其实是一个一般问题的小特例：多项式（polynomial）的所有根何时都落在某个指定区域之内？对于单位圆盘（unit disc），答案由 Schur–Cohn 判别法（Schur–Cohn test）给出。对于左半平面（left half-plane），答案则是 Routh–Hurwitz 判据（Routh–Hurwitz criterion），控制工程师用它来判断一个系统是否稳定。第 (ii) 部分

的区域，正是递推关系（recurrence）$u_{n+2} = bu_{n+1} - cu_n$ 在特征根（characteristic root）为实数时的稳定区域。Siklos 的 *Advanced Problems in Mathematics* 一书中有更多同类的题目。

详细解答

STEP III 2001，第 4 题

纯数学 · 反三角函数

STEP III 2001，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/01-s3-q4

关键思路

看到 $\dfrac{2x}{1+x^2}$ 这个式子，应该马上有所联想：当 $x=\tan\theta$ 时，它正是 $\sin 2\theta$ 的“t 公式”（t-formula）。代入 $x=\tan\theta$ 之后，这个式子的 $\sin^{-1}$ 就变成 $\sin^{-1}(\sin 2\theta)$，于是整道题就化为：$\sin^{-1}(\sin\phi)$ 在甚么时候真的会还原为 ϕ？答案取决于 ϕ 是否落在主值范围（principal range）$[-\frac{1}{2}\pi, \frac{1}{2}\pi]$ 之内，而在 $x=\pm 1$ 处要分段处理，原因正在于此。

解答

第 (i) 部分

先看 $g(x)=\dfrac{2x}{1+x^2}$ 的一些基本性质。

- 分母永不为零，所以 g 在整条实数线上都有定义，而且处处连续（continuous）。
- $g(-x)=-g(x)$，所以 g 是奇函数（odd function），其图像关于原点旋转对称。
- $g(0)=0$；$g(x)$ 与 x 同号；而且当 $x\to\pm\infty$ 时 $g(x)\to 0$，因为分母的增长与 x^2 同阶，而分子只与 x 同阶。所以 x 轴是渐近线（asymptote）：当 $x\to\infty$ 时图像从上方趋近，当 $x\to-\infty$ 时从下方趋近。

由商法则（quotient rule），

$$g'(x)=\frac{2(1+x^2)-2x\cdot 2x}{(1+x^2)^2}=\frac{2(1-x^2)}{(1+x^2)^2}.$$

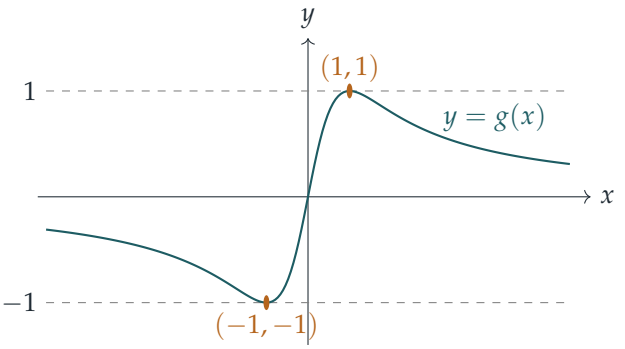
它只在 $x=\pm 1$ 处为零。当 $|x|<1$ 时它是正数，当 $|x|>1$ 时是负数，所以 $x=1$ 给出极大值 $g(1)=1$，$x=-1$ 给出极小值 $g(-1)=-1$。在原点的斜率（gradient）是 $g'(0)=2$。

不用微积分，也可以确认这两个转向点（turning points）给出的是全局的最大值和最小值：

$$1-g(x)=\frac{(1-x)^2}{1+x^2}\geq 0,\qquad 1+g(x)=\frac{(1+x)^2}{1+x^2}\geq 0,$$

所以对所有 x 都有 $-1\leq g(x)\leq 1$，等号只在 $x=\mp 1$ 时成立。由于 g 连续，并且取到 -1 和 1 这两个值，所以它取遍两者之间的每一个值（根据介值定理（intermediate value theorem））。因此

函数g 的值域（range）为 $-1\leq g(x)\leq 1$.



第 (ii) 部分

由第 (i) 部分，$\frac{2x}{1+x^2}$ 总是落在 $[-1,1]$ 之内，而这正是 $\sin^{-1}$ 的定义域（domain），所以 f 对每个实数 x 都有定义。

设 $\theta=\tan^{-1}x$，即 $x=\tan\theta$，其中 $-\frac{1}{2}\pi<\theta<\frac{1}{2}\pi$。那么

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

所以

$$f(x)=\sin^{-1}(\sin 2\theta).$$

想一想 $\sin^{-1}(\sin 2\theta)$ 不就是 2θ 吗?

只有当 2θ 落在 $\sin^{-1}$ 的值域，即 $[-\frac{1}{2}\pi,\frac{1}{2}\pi]$ 之内时才是如此。按定义，$\sin^{-1}y$ 是该区间内正弦值等于 y 的唯一一个角。所以要求 $f(x)$，就必须在 $[-\frac{1}{2}\pi,\frac{1}{2}\pi]$ 内找出与 2θ 正弦值相同的那个角。

情况 $-1\leq x\leq 1$。由于 tan 在 $(-\frac{1}{2}\pi,\frac{1}{2}\pi)$ 上递增，这表示 $-\frac{1}{4}\pi\leq\theta\leq\frac{1}{4}\pi$，所以 $-\frac{1}{2}\pi\leq 2\theta\leq\frac{1}{2}\pi$。这时 2θ 本身就是主值范围内正弦值为 $\sin 2\theta$ 的角，所以

$$f(x)=2\theta=2\tan^{-1}x\qquad(-1\leq x\leq 1).$$

情况 $x \geq 1$。这时 $\frac{1}{4}\pi \leq \theta < \frac{1}{2}\pi$，所以 $\frac{1}{2}\pi \leq 2\theta < \pi$，已经超出主值范围。不过

$$\sin(\pi - 2\theta) = \sin 2\theta \qquad 及 \qquad 0 < \pi - 2\theta \leq \tfrac{1}{2}\pi,$$

所以 $\pi - 2\theta$ 就是主值范围内正弦值相符的那个角。因此

$$f(x) = \pi - 2\theta = \pi - 2\tan^{-1} x \qquad (x \geq 1).$$

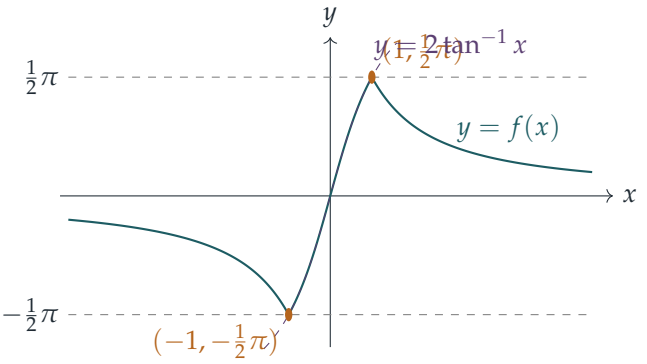
在 $x = 1$ 处，两条公式都给出 $\frac{1}{2}\pi$，这是理所当然的。至此，题目中“证明”（show that）的部分已经完成。

要画出图像，我们还需要知道 $x \leq -1$ 的情况。由于 g 是奇函数，$\sin^{-1}$ 也是奇函数，所以 f 是奇函数，于是当 $x \leq -1$ 时 $f(x) = -f(-x) = -\pi - 2\tan^{-1} x$。（也可以这样看：这时 $-\pi < 2\theta \leq -\frac{1}{2}\pi$，而 $-\pi - 2\theta$ 就是 $[-\frac{1}{2}\pi, 0)$ 内正弦值相同的那个角。）

图像的形状。

- 在 $[-1, 1]$ 上，$f(x) = 2\tan^{-1} x$ 由 $-\frac{1}{2}\pi$ 递增至 $\frac{1}{2}\pi$，并以斜率 2 通过原点。
- 当 $x \geq 1$ 时，$f(x) = \pi - 2\tan^{-1} x$ 由 $\frac{1}{2}\pi$ 开始递减，并且当 $x \to \infty$ 时从上方趋向 $\pi - 2 \cdot \frac{1}{2}\pi = 0$。由奇函数的性质，当 $x \leq -1$ 时，随着 $x \to -\infty$，它由 $-\frac{1}{2}\pi$ 上升，从下方趋向 0。所以 x 轴同样是渐近线。
- 当 $|x| < 1$ 时斜率是 $\dfrac{2}{1+x^2}$，当 $|x| > 1$ 时斜率是 $-\dfrac{2}{1+x^2}$。在 $x = 1$ 处，斜率由 $+1$ 跳到 -1（在 $x = -1$ 处情况类似），所以图像在 $(\pm 1, \pm\frac{1}{2}\pi)$ 处有尖角：f 在该处连续，但不可微（not differentiable）。

这些尖角可以直接从连锁律（chain rule）看出来：$f'(x) = g'(x)/\sqrt{1 - g(x)^2}$，而第 (i) 部分已给出 $1 - g^2 = (1-x^2)^2/(1+x^2)^2$，它的平方根是 $|1-x^2|/(1+x^2)$。所以 $f'(x) = \dfrac{2(1-x^2)}{(1+x^2)\,|1-x^2|} = \pm\dfrac{2}{1+x^2}$，正负号与 $1 - x^2$ 相同。



虚线显示 $2\tan^{-1} x$ 越过 $x = 1$ 之后继续延伸的样子：它会升到 $\frac{1}{2}\pi$ 以上，而 $\sin^{-1}$ 的值永远不可能这样，所以 f 改为向下“反射”回来。

验算

- 在 $x=\sqrt{3}$ 处：$g(\sqrt{3})=\frac{2\sqrt{3}}{4}=\frac{\sqrt{3}}{2}$，所以 $f(\sqrt{3})=\frac{1}{3}\pi$；而公式给出 $\pi-2\cdot\frac{1}{3}\pi=\frac{1}{3}\pi$。两者相符。
- 在 $x=10$ 处：$f(10)=\sin^{-1}(20/101)\approx 0.19934$，而 $\pi-2\tan^{-1}10\approx 0.19934$。用电脑在 $[-1,1]$、$[1,10^4]$ 和 $[-10^4,-1]$ 这三个区间内各随机测试 10^5 个 x 值，结果两者相符至 10^{-7} 以内，而在 $x=1$ 处的两个单侧斜率则分别算得 $+1$ 和 -1。
- 假如 f 处处都等于 $2\tan^{-1}x$，那么当 x 很大时它会接近 π，而 $\sin^{-1}$ 的值不可能这样；反射正好解决了这个问题。

常见错误

- 没有检查 2θ 是否在 $[-\frac{1}{2}\pi,\frac{1}{2}\pi]$ 之内，就写下 $\sin^{-1}(\sin 2\theta)=2\theta$。
- 把 f 在 $x=1$ 处画成平滑的圆顶。图像在该处其实有一个尖角，因为 $\sqrt{1-g^2}$ 涉及的是 $|1-x^2|$，而不是 $1-x^2$。
- 忘记当 $x\to\pm\infty$ 时 f 趋向 0（而不是 $\pm\pi$）。

STEP 之外

反三角函数(inverse trigonometric functions)是选取分支(branch)的第一个例子：sin 不是一对一(one-to-one)的，所以在求反函数之前必须先限制它的定义域，而像 $\sin^{-1}(\sin\phi)=\phi$ 这样的恒等式，也就只在该分支上成立。同样的问题在复分析（complex analysis）中成为核心课题，在那里 $\log z$ 和 $z^{1/2}$ 都需要分支切割（branch cuts）。$x=1$ 处的尖角也是一个精巧的例子：一个由平滑的片段拼合而成的函数，可以在某一点连续却不可微。想了解分支和多值函数（multi-valued functions），并看到大量图解，可参阅 Needham 的 *Visual Complex Analysis*（以图像讲解复分析）；至于连续性和可微性，可参阅 Abbott 的 *Understanding Analysis*（数学分析入门）。

详细解答

STEP III 2001，第 5 题

纯数学 · 抛物线的法线

STEP III 2001，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/01-s3-q5

关键思路

问有哪些法线（normal）通过定点 (h,k)，其实就是问一个以参数（parameter）t 为未知数的三次方程（cubic）有多少个实根。这个三次方程没有 t^2 项，所以它正好具有题目开首部分的形式 t^3+pt+q。实根的数目只会在两个根合并成重根（repeated root）时才改变，因此“一条法线”与“三条法线”之间的分界，就是该三次方程有重根之处。这条分界正是我们要找的轨迹（locus）。

解答

三次方程

设 $f(x)=x^3+px+q$。当 $p\geq 0$ 时，对所有 x 都有 $f'(x)=3x^2+p\geq 0$，而 $f'(x)=0$ 最多只会在 $x=0$ 这一点出现（且仅当 $p=0$ 时）。所以 f 在 $\mathbb{R}$ 上严格递增（strictly increasing）：若 $x_1<x_2$，则 $f(x_2)-f(x_1)=\int_{x_1}^{x_2} f'(x)\,\mathrm{d}x>0$，因为被积函数（integrand）非负，而且最多只在一点为零。

此外，当 $x\to-\infty$ 时 $f(x)\to-\infty$，当 $x\to+\infty$ 时 $f(x)\to+\infty$。由于 f 连续（continuous），它必在某处取值 0；又因为它严格递增，所以只会取这个值一次。因此方程恰有一个实数解。

通过 (h,k) 的法线

在 C 上，$\dfrac{\mathrm{d}y}{\mathrm{d}x}=\dfrac{\mathrm{d}y/\mathrm{d}t}{\mathrm{d}x/\mathrm{d}t}=\dfrac{2a}{2at}=\dfrac{1}{t}$，所以在参数为 t 的点上，法线的斜率（gradient）是 $-t$。（这在 $t=0$ 时同样成立：该处的切线（tangent）是铅垂的，而法线就是 x 轴。）法

线的方程是

$$y - 2at = -t\,(x - at^2), \qquad \text{即} \qquad y + tx = 2at + at^3.$$

它通过 (h,k)，当且仅当 $k + th = 2at + at^3$，即

$$\boxed{at^3 + (2a - h)\,t - k = 0} \qquad \text{或者，两边同除以} a > 0, \qquad t^3 + \frac{2a-h}{a}\,t - \frac{k}{a} = 0.$$

想一想 两个不同的 t 值会否给出同一条法线，以致数根的数目时把法线多算了？

不会。在 t 处的法线斜率是 $-t$，所以不同的 t 值给出斜率不同的法线，它们是不同的直线。因此，通过 (h,k) 的法线数目，正好等于该三次方程相异实根的数目。

若 $h \le 2a$，则系数 $p = \dfrac{2a-h}{a}$ 为 ≥ 0，所以由第一部分可知，该三次方程恰有一个实根。因此恰有一条 C 的法线通过 (h,k)，这正是所要证明的。

恰有两条法线通过的点的轨迹

把三次方程写成 $t^3 + Pt + Q = 0$，其中 $P = \dfrac{2a-h}{a}$，$Q = -\dfrac{k}{a}$。实系数三次方程的实根，按重数（multiplicity）计算，不是一个就是三个。所以它恰有两个相异实根，当且仅当它有一个二重根（double root）s 和另一个与之不同的单根（simple root）。

由于没有 t^2 项，各根之和为 0，所以单根是 $-2s$。我们需要 $s \neq 0$，才能保证 $-2s \neq s$。比较系数，

$$t^3 + Pt + Q = (t-s)^2(t+2s) = t^3 - 3s^2 t + 2s^3,$$

所以 $P = -3s^2$ 及 $Q = 2s^3$。换回原来的变量：

$$h = 2a + 3as^2, \qquad k = -2as^3, \qquad s \neq 0.$$

要消去 s，留意 $(h-2a)^3 = 27a^3s^6$ 及 $k^2 = 4a^2s^6$，所以 $4(h-2a)^3 = 108a^3s^6 = 27a\,k^2$。

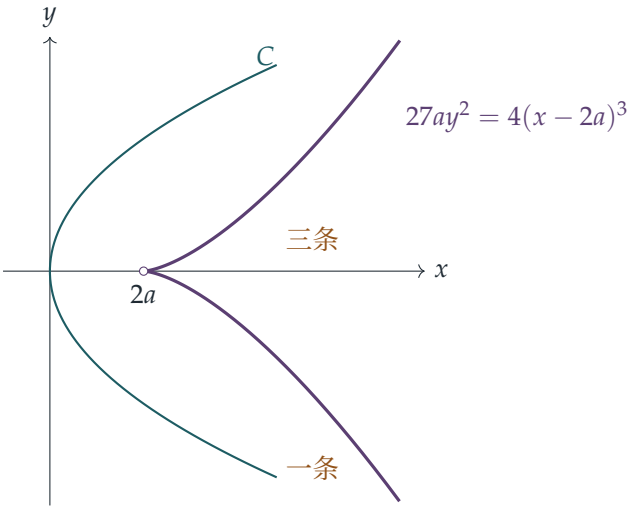
反过来，假设 $27ak^2 = 4(h-2a)^3$，且 $(h,k) \neq (2a,0)$。那么 $h - 2a \ge 0$，而事实上 $h > 2a$（若 $h = 2a$ 则 $k = 0$）。设 s 为满足 $2as^3 = -k$ 的实数。那么 $4a^2s^6 = k^2 = \frac{4}{27a}(h-2a)^3$，所以 $(h-2a)^3 = (3as^2)^3$，因而 $h - 2a = 3as^2$。所以 (h,k) 具有上述形式，其中 $s \neq 0$，并且恰有两条法线通过它。（这等价于重根条件 $4P^3 + 27Q^2 = 0$。）

点 $(2a,0)$ 本身要排除。在该处 $P = Q = 0$，三次方程是 $t^3 = 0$，有一个三重根（triple root），只有一条法线（即 x 轴）通过该点。

所以所求的轨迹是

$$\boxed{27a\,y^2 = 4\,(x-2a)^3}, \qquad \text{不包括点} (2a,0).$$

这是一条半立方抛物线（semicubical parabola），在 $(2a,0)$ 处有一个尖点（cusp），关于 x 轴对称，并位于 $x \geq 2a$ 的范围内，其中 $y = \pm\dfrac{2}{3\sqrt{3a}}(x-2a)^{3/2}$。在它的内部（即包含正 x 轴上 $2a$ 以外部分的区域），三次方程有三个相异实根，所以每一点都有三条法线通过。在它的外部，则只有一条法线通过。



此草图取 $a = 1$。空心圆圈标示被排除的尖点，而图中的标签则表示通过各区域内的点的法线数目。

验算

- 取 $a = 1$ 及 $s = 1$，得出点 $(5,-2)$。三次方程是 $t^3 - 3t + 2 = (t-1)^2(t+2)$，所以在 $(1,2)$ 和 $(4,-4)$ 处的法线都通过 $(5,-2)$。事实上，在 $t = 1$ 处的法线是 $y + x = 3$，而在 $t = -2$ 处的法线是 $y - 2x = -12$。两者都包含 $(5,-2)$，而且 $27 \cdot 4 = 108 = 4 \cdot 27$。
- 在 $(5,0.1)$（取 $a = 1$），$4P^3 + 27Q^2 = -108 + 0.27 < 0$，得出三条法线。在 $(5,4)$，其值为 $-108 + 432 > 0$，得出一条法线。这与草图吻合。
- 我们以数百个随机例子，用数值方法检验了法线方程以及各个“只有一个根”的结论。

常见错误

- 忘记排除尖点 $(2a,0)$。该点满足方程，但只有一条法线通过。
- 声称当 $p \geq 0$ 时 $f'(x) > 0$。当 $p = 0$ 时，$f'(0) = 0$，所以需要用稍强一点的论证，说明 f 仍然严格递增。

▶ 把法线的斜率写成 $1/t$（那是切线的斜率），而不是 $-t$。

STEP 之外

这条轨迹是抛物线的*渐屈线*（evolute）：即其曲率中心（centre of curvature）所描出的曲线，同时也是其法线的包络线（envelope）。越过渐屈线时，通过一点的法线数目会改变两条，这是突变理论（catastrophe theory）和三次方程判别式（discriminant）背后思想的一个入门例子。Siklos 的 *Advanced Problems in Mathematics* 收录了更多同类的 STEP 题目。Bruce 与 Giblin 的 *Curves and Singularities* 则严谨而有系统地阐述渐屈线和包络线。

详细解答

STEP III 2001，第 6 题

纯数学 · 向量与三角形的心

> **STEP III 2001，第 6 题**
>
> 继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：
>
> step.maths.org/questions/01-s3-q6

关键思路

一切都由平面的法向量（normal vector）$\mathbf{n} = \left(\frac{1}{a}, \frac{1}{b}, \frac{1}{c}\right)$ 决定：点 $\mathbf{r}$ 位于平面上，当且仅当 $\mathbf{r}\cdot\mathbf{n} = 1$，而由 O 和 M 所作的垂线都沿 $\mathbf{n}$ 的方向。垂心（orthocentre）和外心（circumcentre）来自两项立体几何事实：OA、OB、OC 两两互相垂直；而 M 是以 OA、OB、OC 为棱的长方体的中心，所以它与 O、A、B、C 等距。至于最后一部分，两条垂线互相平行，因此只需一对相似三角形（或一行向量代数）便可完成余下的工作。

解答

全题中 a、b、c 均不为零（否则平面不会与三条坐标轴都相交），所以

$$A = (a,0,0), \qquad B = (0,b,0), \qquad C = (0,0,c),$$

并记 $S = \dfrac{1}{a^2} + \dfrac{1}{b^2} + \dfrac{1}{c^2} = |\mathbf{n}|^2 > 0$。

形心

直线 OM 上的点是 $t\left(\frac{1}{2}a, \frac{1}{2}b, \frac{1}{2}c\right)$。代入平面方程得 $\frac{1}{2}t + \frac{1}{2}t + \frac{1}{2}t = 1$，所以 $t = \frac{2}{3}$，交点为

$$G = \left(\tfrac{1}{3}a, \tfrac{1}{3}b, \tfrac{1}{3}c\right) = \tfrac{1}{3}(A + B + C),$$

这正是三角形 ABC 的形心（centroid）。留意 $\overrightarrow{OG} = \frac{2}{3}\overrightarrow{OM}$，稍后会用到。

由 O 所作的垂线：垂心

由 O 所作的垂线是直线 $\lambda\mathbf{n}$。它与平面相交于 $\lambda\,\mathbf{n}\cdot\mathbf{n}=1$ 之处，即 $\lambda=1/S$，交点为

$$H=\frac{1}{S}\left(\frac{1}{a},\frac{1}{b},\frac{1}{c}\right).$$

我们证明 $AH\perp BC$。利用 $\overrightarrow{BC}=(0,-b,c)$，

$$\overrightarrow{AH}\cdot\overrightarrow{BC}=\left(\frac{1}{aS}-a\right)\cdot 0+\frac{1}{bS}\cdot(-b)+\frac{1}{cS}\cdot c=-\frac{1}{S}+\frac{1}{S}=0.$$

想一想 这个点积（dot product）是否有理由必定为零，而不只是碰巧如此?

有的。BC 位于该平面上，所以它垂直于法线 OH；而 BC 又位于 yz 平面上，所以它垂直于沿 x 轴的 OA。一条直线若垂直于两条不平行的直线 OA 和 OH，便垂直于整个平面 OAH，当然也垂直于 AH。

把字母轮换，用同样的论证可得 $BH\perp CA$ 和 $CH\perp AB$。由于 H 位于三角形所在的平面上，$AH\perp BC$ 即表示 H 在由 A 所作的高（altitude）上，另外两条高亦同理。所以 H 是三角形 ABC 的垂心。

由 M 所作的垂线：外心

由 M 所作的垂线是 $M+\mu\mathbf{n}$。它与平面相交于 $\frac{3}{2}+\mu S=1$ 之处，所以 $\mu=-\dfrac{1}{2S}$，交点为

$$P=M-\frac{1}{2S}\,\mathbf{n}.$$

现在，M 与 O、A、B、C 四点全部等距：例如

$$MA^2=\left(\tfrac{1}{2}a-a\right)^2+\left(\tfrac{1}{2}b\right)^2+\left(\tfrac{1}{2}c\right)^2=\tfrac{1}{4}\left(a^2+b^2+c^2\right),$$

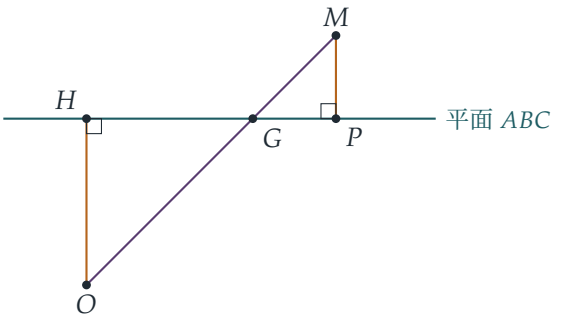
而根据对称性，MB^2、MC^2（以及 MO^2）都得出相同的值。由于 MP 垂直于平面而 PA 位于平面上，三角形 MPA 在 P 处有一直角，所以根据毕氏定理（Pythagoras' theorem）

$$PA^2=MA^2-MP^2,$$

同样地 $PB^2=MB^2-MP^2$ 及 $PC^2=MC^2-MP^2$。各式右方相等，所以 $PA=PB=PC$。平面上与三个顶点等距的点只有外心，所以 P 是三角形 ABC 的外心。

最后一部分

垂线 OH 和 MP 都平行于 $\mathbf{n}$，所以 O、H、M、P 位于同一平面上，而直线 OM 在 G 处穿过三角形所在的平面。此外，O 和 M 分别位于三角形所在平面的两侧，因为 $O\cdot\mathbf{n}=0<1$ 而 $M\cdot\mathbf{n}=\frac{3}{2}>1$。以下是通过这四点的截面：



三角形 OHG 和 MPG 分别在 H 和 P 处有直角，在 G 处的对顶角亦相等，所以两者相似（similar），比为 $OG : GM = 2 : 1$。因此 $HG : GP = 2 : 1$，而且 G 位于线段 HP 上。

同一结论也可直接由向量得出。以 $H = \frac{1}{S}\mathbf{n}$、$P = M - \frac{1}{2S}\mathbf{n}$ 及 $G = \frac{2}{3}M$ 代入，

$$\tfrac{1}{3}(H + 2P) = \tfrac{1}{3}\left(\tfrac{1}{S}\mathbf{n} + 2M - \tfrac{1}{S}\mathbf{n}\right) = \tfrac{2}{3}M = G.$$

所以 G 位于由垂心 H 至外心 P 的线段上，并把它按 $HG : GP = 2 : 1$ 的比分割，正如题目所求。（若三角形是等边三角形，则 $H = P = G$，无需证明。）

想一想 这是否已为每一个三角形证明了这个结果?

并不尽然。要令这个论证适用，三角形必须与某个上述类型的三角形 ABC 全等（congruent）。这时它的边满足 $BC^2 = b^2 + c^2$、$CA^2 = c^2 + a^2$ 及 $AB^2 = a^2 + b^2$，所以

$$AB^2 + AC^2 - BC^2 = 2a^2 > 0,$$

在 B 和 C 处亦同理。根据余弦定理（cosine rule），每个角都是锐角。反过来说，任何锐角三角形都可以这样实现：由这三条方程解出 a^2、b^2、c^2，它们全是正数。所以这个方法为锐角三角形证明了结果。至于直角或钝角三角形，以下是一个简短的平面论证，可涵盖所有情况。把外心置于原点，则顶点 $\mathbf{a}$、$\mathbf{b}$、$\mathbf{c}$ 的长度全部相同。设 $\mathbf{h} = \mathbf{a} + \mathbf{b} + \mathbf{c}$。那么

$$(\mathbf{h} - \mathbf{a}) \cdot (\mathbf{b} - \mathbf{c}) = (\mathbf{b} + \mathbf{c}) \cdot (\mathbf{b} - \mathbf{c}) = |\mathbf{b}|^2 - |\mathbf{c}|^2 = 0,$$

所以 $\mathbf{h}$ 在由 $\mathbf{a}$ 所作的高上，根据对称性亦在另外两条高上：它就是垂心。形心是 $\frac{1}{3}\mathbf{h} = \frac{1}{3}(\mathbf{h} + 2 \cdot \mathbf{0})$，这正是同一个 $2 : 1$ 的结论。

验算

取 $a = 1$、$b = c = 2$，则 $S = \frac{3}{2}$。于是 $H = \left(\frac{2}{3}, \frac{1}{3}, \frac{1}{3}\right)$、$P = \left(\frac{1}{6}, \frac{5}{6}, \frac{5}{6}\right)$ 及 $G = \left(\frac{1}{3}, \frac{2}{3}, \frac{2}{3}\right)$。各点均满足 $x + \frac{1}{2}y + \frac{1}{2}z = 1$。你可以验证 $\overrightarrow{AH} \cdot \overrightarrow{BC} = 0$ 和 $PA^2 = PB^2 = PC^2 = \frac{25}{12}$，以及 $\frac{1}{3}(H + 2P) = \left(\frac{1}{3}, \frac{2}{3}, \frac{2}{3}\right) = G$。我们亦曾随机选取 3000 组正负皆有的 a、b、c，以精确算术验证了以上所有事实。

常见错误

- 只证明 $AH \perp BC$ 便停下来。你需要证明全部三条高，或清楚写明“根据对称性”。你还需要指出 H 位于三角形所在的平面上：平面以外的点也可以满足 $AH \perp BC$，却不在高上。
- 把比的次序弄反。形心与垂心的距离，是它与外心的距离的两倍。
- 声称“由此”（hence）的部分涵盖每一个三角形。如上文所示，它只适用于锐角三角形。

STEP 之外

通过垂心、形心和外心的直线称为欧拉线（Euler line）。$\mathbf{h} = \mathbf{a} + \mathbf{b} + \mathbf{c}$（以外心为原点）这个事实，是研究九点圆（nine-point circle）的起点；九点圆的圆心同样在欧拉线上，位于 H 与外心的正中间。把平面图形看成一个简单立体图形的截面，这种技巧在几何中屡见不鲜。考克斯特（Coxeter）的 *Introduction to Geometry*（《几何学导论》），以及考克斯特与 Greitzer 合著的 *Geometry Revisited*（《几何学重温》），对欧拉线和九点圆都有精彩的论述。

详细解答

STEP III 2001，第 7 题

纯数学 · 正交轨线

> **STEP III 2001，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/01-s3-q7

关键思路

这条微分方程（differential equation）是可分离变量（separable）的，所以只要把它解出来，就能直接看出这族曲线是甚么。至于正交（orthogonal）曲线，切线（tangent）互相垂直即表示两个斜率（gradient）的乘积为 -1。由此得到第二条可分离变量的方程，而对它积分时，开首所描绘的那个函数便会出现。题目先要我们画第一幅略图，正是为了让我们几乎不费工夫就能画出最后的曲线。

解答

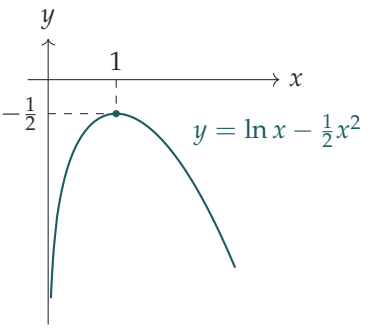
$\ln x - \frac{1}{2}x^2$ 的图像

设 $f(x) = \ln x - \frac{1}{2}x^2$，定义域为 $x > 0$。则

$$f'(x) = \frac{1}{x} - x = \frac{1 - x^2}{x}, \qquad f''(x) = -\frac{1}{x^2} - 1 < 0.$$

当 $x > 0$ 时，唯一的驻点（stationary point）位于 $x = 1$，该处 $f(1) = -\frac{1}{2}$。由于 $f'' < 0$ 处处成立，这是一个极大值（maximum），而且整个图像都是向下凹（concave）的。所以对任何 $x > 0$ 都有 $f(x) \le -\frac{1}{2} < 0$，图像永不与 x 轴相交。

在定义域的两端，当 $x \to 0^+$ 时 $f(x) \to -\infty$（因为 $\ln x \to -\infty$），所以 y 轴是一条垂直渐近线（vertical asymptote）。另外，当 $x \to \infty$ 时也有 $f(x) \to -\infty$，因为 $\frac{1}{2}x^2$ 的增长远比 $\ln x$ 快。



抛物线族

常数函数 $y = 0$ 是一个解。在其他情况下，我们分离变量，并避开 $x = \pm 1$，因为方程在该处没有定义：

$$\int \frac{\mathrm{d}y}{y} = \int \frac{2x}{x^2-1}\,\mathrm{d}x \quad \Longrightarrow \quad \ln|y| = \ln|x^2-1| + c.$$

所以 $|y| = \mathrm{e}^c\,|x^2-1|$。把正负号并入常数，再连同 $y = 0$ 一并考虑，方程的解为

$$y = A(x^2-1), \qquad A \text{ 为常数.}$$

反过来，每一条这样的曲线都满足该方程，因为

$$y' = 2Ax = \frac{2x \cdot A(x^2-1)}{x^2-1} = \frac{2xy}{x^2-1}.$$

当 $A \neq 0$ 时，每条曲线都是抛物线（parabola）。它通过 $(1,0)$ 和 $(-1,0)$，因为在该两点 $x^2-1=0$。它的斜率 $2Ax$ 只在 $x=0$ 时为零，所以它的顶点（vertex）是 $(0,-A)$，位于 y 轴上。这就是题目所求的曲线族。

正交曲线

取一点 (x,y)，其中 $x \neq \pm 1$ 且 $y \neq 0$。曲线族中恰好有一条抛物线通过这一点，即 $A = y/(x^2-1)$ 的那一条，而它在该点的斜率为 $\dfrac{2xy}{x^2-1}$。所求的曲线若要与这条抛物线成直角相交，两个斜率的乘积必须为 -1，所以该曲线必须满足

$$\frac{\mathrm{d}y}{\mathrm{d}x} = -\frac{x^2-1}{2xy} = \frac{1-x^2}{2xy}.$$

> ? **想一想** 为甚么可以用抛物线的微分方程，而不用公式 $y = A(x^2-1)$ 来求斜率?

微分方程只用 x 和 y 就表示出通过 (x,y) 的那条族中曲线的斜率，完全不涉及 A。这正是我们所需要的，因为正交曲线在它的每一点上所遇到的都是不同的抛物线。

再次分离变量，

$$2y\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{x} - x \quad \Longrightarrow \quad y^2 = \ln x - \tfrac{1}{2}x^2 + C,$$

这里取 $x > 0$，因为曲线通过 $(1,1)$。右边是 $f(x) + C$，即我们在开首所描绘的函数。在 $(1,1)$ 处须有 $1 = 0 - \frac{1}{2} + C$，所以 $C = \frac{3}{2}$，曲线为

$$\boxed{y^2 = \ln x - \tfrac{1}{2}x^2 + \tfrac{3}{2}} \qquad \text{或等价地} \qquad y^2 = f(x) + \tfrac{3}{2}.$$

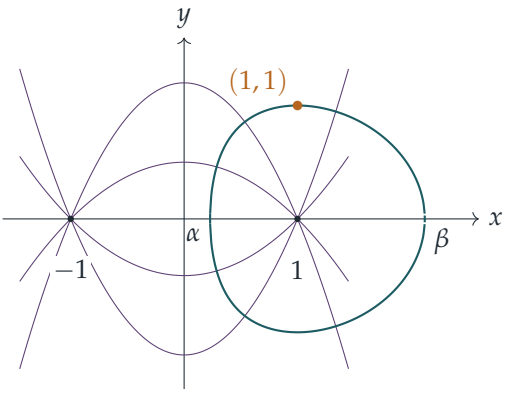
严格来说，微分方程通过 $(1,1)$ 的解是上半部分 $y = \sqrt{f(x) + \frac{3}{2}}$。它对 x 轴的反射同样与每条抛物线正交，所以自然的答案是整条闭曲线。

描绘曲线

实数点须满足 $f(x) \geq -\frac{3}{2}$。我们的第一幅略图显示，f 由 $-\infty$ 上升至 $x = 1$ 处的极大值 $-\frac{1}{2}$，然后再下降回 $-\infty$。所以 $f(x) = -\frac{3}{2}$ 恰好有一个根 $\alpha \in (0,1)$ 和一个根 $\beta > 1$。数值上，$\alpha \approx 0.229$，$\beta \approx 2.123$。因此：

- 曲线位于带状区域 $\alpha \leq x \leq \beta$ 之内，并且对 x 轴对称；
- y^2 在 $x = 1$ 处最大，其值为 1，由此得到两点 $(1, \pm 1)$，该处的切线是水平的；
- 在 $(\alpha, 0)$ 和 $(\beta, 0)$ 处，我们有 $y^2 = 0$，而 $\mathrm{d}(y^2)/\mathrm{d}x = f'(x) \neq 0$，所以 $\mathrm{d}y/\mathrm{d}x$ 为无穷大，切线是垂直的。

所以这条曲线是一个不对称的闭合卵形：左边陡峭上升，右边则较平缓地下降。图中也画出了几条抛物线以作比较。



假如我们容许 $x < 0$，并把 $\ln x$ 换成 $\ln|x|$，同样的运算还会在左半平面得出一个镜像的卵形。那个卵形并不通过 $(1,1)$，所以不属于答案的一部分。

验算

- 在 $(1,1)$ 处，曲线的斜率为 $\frac{1-1}{2}=0$。没有任何抛物线通过 $(1,1)$ 这一点本身，因为每条抛物线与直线 $x=1$ 只相交于 $(1,0)$。在 $x=1$ 附近，抛物线几乎是垂直的，这与曲线上的水平切线吻合。
- 数值上，在 $x=0.5$ 处，曲线上 $y=\sqrt{\ln 0.5-0.125+1.5}\approx 0.826$，斜率为 $\frac{0.75}{2(0.5)(0.826)}\approx 0.908$。通过该点的抛物线的斜率为 $\frac{2(0.5)(0.826)}{-0.75}\approx -1.101$，而 $0.908\times(-1.101)\approx -1$。
- 我们在卵形上数千个随机点验证了正交性，并以二分法（bisection）计算出两个根 $\alpha\approx 0.2291$ 和 $\beta\approx 2.1226$。

常见错误

- 使用某一条固定抛物线的斜率，例如取特定 A 值的 $2Ax$，而不是通过一般点 (x,y) 的那条族中曲线的斜率。
- 忘记 $y^2=f(x)+\frac{3}{2}$ 需要 $f(x)\geq -\frac{3}{2}$，因而画出一条无界的曲线，而不是闭合的卵形。

STEP 之外

求一族曲线，使它与给定的曲线族成直角相交，这就是**正交轨线**（orthogonal trajectories）问题。在物理学中，它们以场线和等势线的形式出现，例如在静电学或二维流体流动之中。解析（analytic）复变函数的实部和虚部总会给出这样的两族曲线，这是通往保角映射（conformal mapping）的其中一条途径。Boyce 与 DiPrima 的 *Elementary Differential Equations* 涵盖可分离变量方程和正交轨线。Needham 的 *Visual Complex Analysis* 则展示了复变量的图景。

详细解答

STEP III 2001，第 8 题

纯数学 · 复数轨迹

STEP III 2001，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/01-s3-q8

关键思路

设 $z = x + \mathrm{i}y$ 并展开：(i) 中的两条方程都会化简成同一个圆。至于 (ii) 中的辐角（argument）条件，可直接计算 $\frac{z-2}{z}$；辐角为 $\frac{\pi}{4}$ 的意思是“实部（real part）等于虚部（imaginary part），而且两者皆为正数”，这就是同一个圆再加上一个不等式。在 (iii) 中，把 $z = 2/w$ 代入圆的两种形式之中运算最简洁的那一种。(i) 中的第二种形式正是为此而设。

解答

第 (i) 部分

设 $z = x + \mathrm{i}y$。第一条方程是

$$(x-1)^2 + (y-1)^2 = 2 \iff x^2 + y^2 - 2x - 2y = 0.$$

第二条方程是

$$\begin{aligned} & (x-1)^2 + (y+1)^2 = 2\left[(x-1)^2 + y^2\right] \\ \iff\ & x^2 + y^2 - 2x + 2y + 2 = 2x^2 + 2y^2 - 4x + 2 \\ \iff\ & 0 = x^2 + y^2 - 2x - 2y. \end{aligned}$$

每一步都可逆，所以两条方程都等价于 $x^2 + y^2 - 2x - 2y = 0$，描述的是同一个轨迹（locus）。这个轨迹是以 $1 + \mathrm{i}$ 为圆心、半径为 $\sqrt{2}$ 的圆。它通过 0、2 和 2i（以及 $2 + 2\mathrm{i}$），最低点是 $1 + (1 - \sqrt{2})\mathrm{i}$，刚好在实轴（real axis）之下。

第 (ii) 部分

当 $z \neq 0$ 时，

$$\frac{z-2}{z} = \frac{(z-2)\bar{z}}{|z|^2} = \frac{x^2+y^2-2x+2\mathrm{i}y}{x^2+y^2},$$

因为 $(z-2)\bar{z} = |z|^2 - 2\bar{z} = x^2+y^2-2x+2\mathrm{i}y$。

? **想一想** $\arg(\text{某数}) = \frac{\pi}{4}$ 究竟说明了它的实部和虚部有甚么关系?

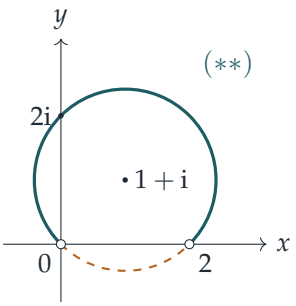
一个非零复数（complex number）$a+\mathrm{i}b$ 的辐角为 $\frac{\pi}{4}$，当且仅当它位于射线 $a = b > 0$ 上。单有相等并不足够：$a = b < 0$ 时辐角是 $-\frac{3\pi}{4}$。分母 x^2+y^2 为正数，所以 $(**)$ 成立当且仅当

$$x^2+y^2-2x = 2y \quad \text{及} \quad 2y > 0,$$

即当且仅当 z 位于 (i) 的圆上而且 $y > 0$。条件 $y > 0$ 自动排除了 $z = 0$（该处商没有定义）和 $z = 2$（该处商为零，没有辐角）。

因此 $(**)$ 描述的是圆在实轴以上的部分：由 0 出发，绕经 2i 和 $2+2\mathrm{i}$ 到达 2 的优弧（major arc），两个端点都不包括在内。实轴以下的短弧所对应的辐角则是 $-\frac{3\pi}{4}$。

这与几何上的理解吻合。$\arg\frac{z-2}{z}$ 是在 z 处、连往 0 和连往 2 的两条弦之间的夹角。由 0 到 2 的弦在圆心 $1+\mathrm{i}$ 所对的角是直角，所以根据圆心角定理（angle-at-the-centre theorem），它在优弧上每一点所对的角都是 $\frac{\pi}{4}$。



整个圆是 $(*)$；实线弧（不包括端点 0 和 2）是 $(**)$，虚线弧则不属于 $(**)$。

第 (iii) 部分

由于 $w = 2/z$，我们有 $z \neq 0$ 及 $z = 2/w$。采用圆的第二种形式 $|z-(1-\mathrm{i})|^2 = 2|z-1|^2$。代入后得

$$z-1 = \frac{2-w}{w}, \qquad z-(1-\mathrm{i}) = \frac{2-(1-\mathrm{i})w}{w} = \frac{(1-\mathrm{i})\big[(1+\mathrm{i})-w\big]}{w},$$

其中用到 $\frac{2}{1-\mathrm{i}} = 1+\mathrm{i}$。由于 $|1-\mathrm{i}|^2 = 2$，方程变成

$$\frac{2\,|w-(1+\mathrm{i})|^2}{|w|^2} = \frac{2\,|w-2|^2}{|w|^2} \iff |w-(1+\mathrm{i})| = |w-2|.$$

所以 w 与 $1+\mathrm{i}$ 和 2 等距：轨迹是连接这两点的线段的垂直平分线（perpendicular bisector）。中点是 $\frac{3}{2}+\frac{1}{2}\mathrm{i}$，线段的方向是 $1-\mathrm{i}$，所以平分线的方向是 $1+\mathrm{i}$。设 $w = u+\mathrm{i}v$，轨迹就是直线

$$v = u-1.$$

这条直线上的每一点都会出现。线上每个 w 都非零，因为 0 不在这条直线上，而 $z = 2/w$ 则由上面各个可逆的步骤可知满足 $(*)$。圆上唯一没有像（image）的点是 $z=0$，它被“送到无穷远”。一如所料，通过原点的圆在 $w = 2/z$ 之下映射成一条直线。

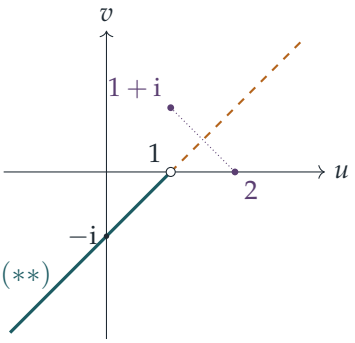
至于对应 $(**)$ 的部分，留意

$$\frac{z-2}{z} = 1-\frac{2}{z} = 1-w,$$

所以 $(**)$ 变成 $\arg(1-w) = \frac{\pi}{4}$，即 $1-w = r\,\mathrm{e}^{\mathrm{i}\pi/4}$，其中 $r>0$。因此

$$w = 1-\frac{r}{\sqrt{2}}(1+\mathrm{i}), \qquad r>0,$$

这是直线 $v = u-1$ 上由 $w=1$（不包括在内）出发、向左下方延伸的半直线（half-line）：即 $u<1$ 的点，等价地说是 $v<0$ 的点。这与 (ii) 吻合：$\operatorname{Im} w = -2y/|z|^2$，所以 $y>0$ 对应 $v<0$。例如，$z = 2\mathrm{i}$ 映射到 $w=-\mathrm{i}$，而 $z=2$ 则映射到被排除的端点 $w=1$。



整条直线 $v = u-1$ 是 $(*)$ 的像；实线半直线（$u<1$，不包括 $w=1$）是 $(**)$ 的像。点线线段连接的是被这条直线平分的两点。

验算

- $z = 2\mathrm{i}$ 在圆上，而且在实轴之上：$\frac{2\mathrm{i}-2}{2\mathrm{i}} = 1+\mathrm{i}$，其辐角为 $\frac{\pi}{4}$。它的像 $w=-\mathrm{i}$ 满足 $v = u-1$，且 $u<1$。

- 圆的最高点 $z = 1 + (1 + \sqrt{2})\mathrm{i}$ 映射到 $w \approx 0.2929 - 0.7071\mathrm{i}$，而确实有 $0.2929 - 1 = -0.7071$。
- 实轴以下的一点 $z = 1 + (1 - \sqrt{2})\mathrm{i}$ 给出 $\arg \frac{z-2}{z} = -\frac{3\pi}{4}$。它的像有 $u > 1$，位于直线的另一半。

常见错误

- 把 $\arg Z = \frac{\pi}{4}$ 仅仅当作 $\operatorname{Re} Z = \operatorname{Im} Z$，于是声称答案是整个圆。实部和虚部还须同为正数。
- 在 (ii) 中忘记排除 $z = 0$ 和 $z = 2$，或在 (iii) 中忘记排除 $w = 1$。
- 以为圆的像必定是另一个圆。在 $w = 2/z$ 之下，通过原点的圆会变成直线。

STEP 之外

映射 $w = 2/z$ 是一个莫比乌斯变换（Möbius transformation）。这类映射把圆和直线映射成圆和直线，而一个圆变成直线，当且仅当它通过那个被送到无穷远的点。这里弧的两个端点 0 和 2 分别去到 ∞ 和 1。商 $\frac{z-2}{z-0}$ 量度的是这两个端点所对的角，它变成了 $1 - w$：由像点指向 1 的方向。这就是 (ii) 中的弧在 (iii) 中变成半直线的原因。Needham 的 *Visual Complex Analysis*（《复分析：可视化方法》）把这种几何观点阐述得十分精彩。

详细解答

STEP III 2001，第 9 题

力学 · 摩擦力与简谐运动

> **STEP III 2001，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/01-s3-q9

关键思路

两条横梁沿一条与水平成 α 角的直线相距 d，因此木板的倾角就是 α。木板滑动时，它的重量由 B_1 逐渐转移到 B_2，而两个法向反作用力（normal reaction）随重心（centre of gravity）的位置线性地变化。由于两个摩擦系数（coefficient of friction）不同，总摩擦力也随之线性变化；回复力（restoring force）与位移成正比，这就是简谐运动（simple harmonic motion）。条件 $\mu_1 + \mu_2 = 2\tan\alpha$ 的作用，正是令这个运动的中心刚好落在两条横梁的正中间。

解答

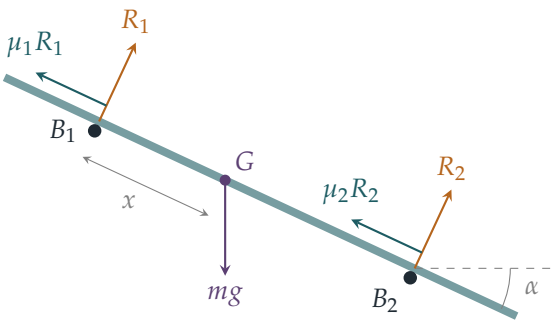
模型

把木板视为质量为 m 的刚性杆（rigid rod），它始终与两条横梁保持接触，而横梁则视为点。由 B_1 到 B_2 的直线长度为

$$\sqrt{d^2\sin^2\alpha + d^2\cos^2\alpha} = d$$

并与水平成 α 角，所以木板沿这条直线摆放，由 B_1 向下倾斜至 B_2。设 x 为重心 G 与 B_1 的距离，沿木板向斜下方量度，于是释放时 $x = 0$，而当 $x = d$ 时 G 位于 B_2 之上。

当 $0 \le x \le d$ 时，G 位于两条横梁之间。作用于木板的力有：作用于 G 的重量 mg；在 B_1、B_2 处垂直于木板的法向反作用力 R_1、R_2；以及当木板沿斜面向下运动时，沿斜面向上的摩擦力 $\mu_1 R_1$ 和 $\mu_2 R_2$。



法向反作用力

木板在垂直于自身的方向上没有运动，所以沿垂直于木板的方向分解力，得

$$R_1 + R_2 = mg\cos\alpha.$$

木板也没有转动，所以对 G 取力矩（moment）（重量和两个摩擦力的作用线或通过 G，或沿着木板，因此它们没有力矩），得

$$R_1 x = R_2(d - x).$$

解这两条方程，得

$$R_1 = \frac{mg\cos\alpha\,(d-x)}{d}, \qquad R_2 = \frac{mg\cos\alpha\,x}{d}.$$

当 $0 \le x \le d$ 时两者皆非负，所以木板确实一直搁在两条横梁上。

运动方程

当木板沿斜面向下运动时，沿斜面向下分解力，得

$$\begin{aligned} m\ddot{x} &= mg\sin\alpha - \mu_1 R_1 - \mu_2 R_2 \\ &= mg\cos\alpha\left[\tan\alpha - \mu_1 - \frac{(\mu_2-\mu_1)x}{d}\right]. \end{aligned}$$

现在使用题目给出的条件：由 $\mu_1 + \mu_2 = 2\tan\alpha$ 得 $\tan\alpha - \mu_1 = \frac{1}{2}(\mu_2 - \mu_1)$。因此

$$\ddot{x} = -\omega^2\left(x - \tfrac{1}{2}d\right), \qquad \text{其中} \qquad \omega^2 = \frac{g(\mu_2-\mu_1)\cos\alpha}{d} > 0,$$

不等式成立是因为 $\mu_2 > \mu_1$ 且 $\cos\alpha > 0$。

? **想一想** 木板真的会开始运动吗？摩擦力是否一直指向斜面上方？

在 $x=0$ 处，$\ddot{x}=\frac{1}{2}\omega^2 d>0$，所以木板确实会开始向下滑。上面的方程描述的是以 $x=\frac{1}{2}d$ 为中心的简谐运动，而由于 $t=0$ 时 $x=0$、$\dot{x}=0$，它的解为

$$x=\tfrac{1}{2}d\,(1-\cos\omega t),\qquad \dot{x}=\tfrac{1}{2}d\,\omega\sin\omega t.$$

当 $0<t<\pi/\omega$ 时有 $\dot{x}>0$，所以木板在整段时间内的确一直沿斜面向下运动，假设摩擦力沿斜面向上是正确的。此外，整段时间内 $0\le x\le d$，所以 G 一直位于两条横梁之间，R_1、R_2 的表达式仍然有效。

静止下来

速度第一次再度为零是在 $\omega t=\pi$ 之时，这时 $x=\frac{1}{2}d(1-\cos\pi)=d$：重心正好位于 B_2 之上。在这一刻 $R_1=0$ 而 $R_2=mg\cos\alpha$，所以木板只搁在 B_2 上。它会停留在那里，因为 B_2 处的摩擦力最大可达 $\mu_2 mg\cos\alpha$，而重量沿斜面向下的分量是 $mg\sin\alpha$，并且

$$\mu_2=\tfrac{1}{2}(\mu_1+\mu_2)+\tfrac{1}{2}(\mu_2-\mu_1)>\tfrac{1}{2}(\mu_1+\mu_2)=\tan\alpha.$$

所以木板静止下来时，重心位于 B_2 之上，所需时间为

$$\frac{\pi}{\omega}=\pi\left(\frac{d}{g(\mu_2-\mu_1)\cos\alpha}\right)^{\frac{1}{2}},$$

这正是所要证明的结果。

验算

- 若 $\mu_2-\mu_1$ 很小，木板几乎不减速，所需时间很长，与公式相符。这时每个 μ 都接近 $\tan\alpha$，所以合力在各处都接近零。
- 我们随机选取 α、d、μ_1（并取 $\mu_2=2\tan\alpha-\mu_1$），以数值方法逐步求解运动方程，直至速度第一次回到零为止。例如，取 $\alpha=0.7$、$d=2$、$\mu_1=0.3$ 及 $g=9.8$，预测的时间约为 1.56 秒。在每一个情况中，停止的时间和位置都与 π/ω 和 $x=d$ 吻合，（相对）误差在 10^{-5} 以内。

常见错误

- 取 $R_1=R_2=\frac{1}{2}mg\cos\alpha$。反作用力取决于 G 的位置，而正是这种依赖关系产生了振荡。
- 只做到“速度为零”便停笔，没有检查木板其后是否保持静止：你需要 $\mu_2\ge\tan\alpha$。

STEP 之外

这个运动是一个振子（oscillator）的半个周期，而它的“弹簧”来自随负载而变的摩擦力。摩擦力取决于位置或速度的系统会产生丰富多变的现象，包括黏滑振荡（stick–slip oscillation），门铰吱吱作响和琴弓拉动小提琴弦线，背后都是这种振荡。Kleppner 与 Kolenkow 所著的 *An Introduction to Mechanics*（力学入门）对摩擦、刚体平衡和简谐运动都有细致的论述。

详细解答

STEP III 2001，第 10 题

力学 · 相对速度

> **STEP III 2001，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/01-s3-q10

关键思路

相对速度（relative velocity）就是两个速度之差：A 相对于 C 的速度是 $\mathbf{v}_1 - \mathbf{u}$。只要把两个条件都写成点积（dot product），一切便可从中点 $\frac{1}{2}(\mathbf{v}_1 + \mathbf{v}_2)$ 量起，而两个条件就化为平方差的形式。从几何角度看，这两条方程描述的是一个圆和一条直线，这正是“两个速度”这个数目的来由。

解答

两个条件

我们把船只视为在海面这个水平面上运动的质点（particle），因此所有速度都是二维向量。A 和 B 相对于 C 的速度分别是

$$\mathbf{a} = \mathbf{v}_1 - \mathbf{u}, \qquad \mathbf{b} = \mathbf{v}_2 - \mathbf{u}.$$

大小（magnitude）相等和互相垂直（perpendicular）即是

$$|\mathbf{v}_1 - \mathbf{u}|^2 = |\mathbf{v}_2 - \mathbf{u}|^2 \qquad 及 \qquad (\mathbf{v}_1 - \mathbf{u}) \cdot (\mathbf{v}_2 - \mathbf{u}) = 0.$$

两条方程

设

$$\mathbf{w} = \mathbf{u} - \tfrac{1}{2}(\mathbf{v}_1 + \mathbf{v}_2), \qquad \mathbf{d} = \tfrac{1}{2}(\mathbf{v}_1 - \mathbf{v}_2).$$

于是 $\mathbf{v}_1 = \frac{1}{2}(\mathbf{v}_1 + \mathbf{v}_2) + \mathbf{d}$ 及 $\mathbf{v}_2 = \frac{1}{2}(\mathbf{v}_1 + \mathbf{v}_2) - \mathbf{d}$，所以

$$\mathbf{a} = \mathbf{d} - \mathbf{w}, \qquad \mathbf{b} = -\mathbf{d} - \mathbf{w}.$$

现在利用 $\mathbf{w}\cdot\mathbf{d}=\mathbf{d}\cdot\mathbf{w}$，把两个条件逐一展开：

$$\mathbf{a}\cdot\mathbf{b}=-(\mathbf{d}-\mathbf{w})\cdot(\mathbf{d}+\mathbf{w})=|\mathbf{w}|^2-|\mathbf{d}|^2,$$
$$|\mathbf{a}|^2-|\mathbf{b}|^2=(|\mathbf{d}|^2-2\mathbf{d}\cdot\mathbf{w}+|\mathbf{w}|^2)-(|\mathbf{d}|^2+2\mathbf{d}\cdot\mathbf{w}+|\mathbf{w}|^2)=-4\,\mathbf{w}\cdot\mathbf{d}.$$

因此，垂直条件 $\mathbf{a}\cdot\mathbf{b}=0$ 恰好等价于 $|\mathbf{w}|^2=|\mathbf{d}|^2$，即

$$\left|\mathbf{u}-\tfrac{1}{2}(\mathbf{v}_1+\mathbf{v}_2)\right|^2=\left|\tfrac{1}{2}(\mathbf{v}_1-\mathbf{v}_2)\right|^2,$$

而大小相等的条件 $|\mathbf{a}|^2=|\mathbf{b}|^2$ 恰好等价于 $\mathbf{w}\cdot\mathbf{d}=0$，（乘以 2 后）即

$$\Big(\mathbf{u}-\tfrac{1}{2}(\mathbf{v}_1+\mathbf{v}_2)\Big)\cdot(\mathbf{v}_1-\mathbf{v}_2)=0.$$

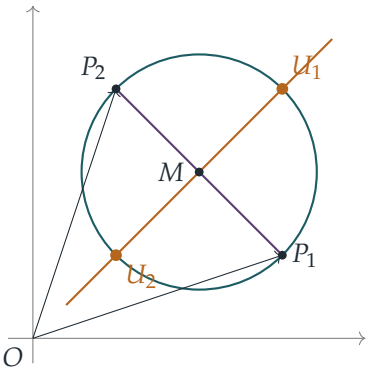
这就是所求的两条方程。以上每一步都是恒等式（identity），所以论证也可以倒过来走：两条方程成立，当且仅当两个条件成立。

为何 C 有两个可能的速度

把各个速度画成从同一个原点出发的位置向量（position vector），设 P_1、P_2、U 分别为 $\mathbf{v}_1$、$\mathbf{v}_2$、$\mathbf{u}$ 的终点，并设 M 为 P_1P_2 的中点。

想一想 每条方程分别说明点 U 可以在甚么位置?

第一条方程说明 U 与 M 的距离是 $\frac{1}{2}|\mathbf{v}_1-\mathbf{v}_2|$：它位于以 P_1P_2 为直径的圆上。第二条方程说明 MU 垂直于 P_1P_2（或 $U=M$）：U 位于 P_1P_2 的垂直平分线（perpendicular bisector）上。



在平面上，一条通过圆心的直线与一个半径为正数的圆恰好相交于两点，这两点位于一条直径的两端。这里的半径是 $\frac{1}{2}|\mathbf{v}_1-\mathbf{v}_2|$，它是正数，正正因为 $\mathbf{v}_1\neq\mathbf{v}_2$。所以 C 恰好有两个可能的速度：

$$\mathbf{u}=\tfrac{1}{2}(\mathbf{v}_1+\mathbf{v}_2)\pm\tfrac{1}{2}\,\mathbf{d}^{\perp},$$

其中 $\mathbf{d}^{\perp}$ 是把 $\mathbf{v}_1-\mathbf{v}_2$ 旋转一个直角所得的向量。（若 $\mathbf{v}_1=\mathbf{v}_2$，圆便收缩成单独一点 M，这时两个相对速度都是零，两个条件就没有甚么实质内容。）在图中，直角 P_1UP_2 就是半圆上的圆周角，而 $UP_1=UP_2$ 是因为 U 在垂直平分线上。

最后一部分

现在 $|\mathbf{v}_1|=|\mathbf{v}_2|=k$ 且 $\mathbf{v}_1\cdot\mathbf{v}_2=0$，其中 $k>0$（只有非零向量才谈得上垂直）。于是 $\mathbf{v}_1$ 和 $\mathbf{v}_2$ 互相垂直而且非零，所以它们张成（span）整个平面，我们可以写 $\mathbf{u}=\alpha\mathbf{v}_1+\beta\mathbf{v}_2$。

利用 $\mathbf{v}_1\cdot\mathbf{v}_2=0$ 和 $|\mathbf{v}_1|^2=|\mathbf{v}_2|^2$，第二条方程给出

$$\mathbf{u}\cdot(\mathbf{v}_1-\mathbf{v}_2)=\tfrac{1}{2}\left(|\mathbf{v}_1|^2-|\mathbf{v}_2|^2\right)=0,\qquad\text{所以}\qquad \alpha k^2-\beta k^2=0,$$

因此 $\beta=\alpha$。至于第一条方程，把它展开：

$$|\mathbf{u}|^2-\mathbf{u}\cdot(\mathbf{v}_1+\mathbf{v}_2)+\tfrac{1}{4}|\mathbf{v}_1+\mathbf{v}_2|^2=\tfrac{1}{4}|\mathbf{v}_1-\mathbf{v}_2|^2.$$

由于 $\mathbf{v}_1\cdot\mathbf{v}_2=0$，$|\mathbf{v}_1+\mathbf{v}_2|^2$ 和 $|\mathbf{v}_1-\mathbf{v}_2|^2$ 都等于 $2k^2$，所以这两项互相抵消，得 $|\mathbf{u}|^2=\mathbf{u}\cdot(\mathbf{v}_1+\mathbf{v}_2)$。代入 $\mathbf{u}=\alpha(\mathbf{v}_1+\mathbf{v}_2)$，这就是

$$2\alpha^2k^2=2\alpha k^2\quad\Longrightarrow\quad\alpha(\alpha-1)=0.$$

所以 $\alpha=0$，得 $\mathbf{u}=\mathbf{0}$；或 $\alpha=1$，得 $\mathbf{u}=\mathbf{v}_1+\mathbf{v}_2$。因此若 $\mathbf{u}\neq\mathbf{0}$，则 $\mathbf{u}=\mathbf{v}_1+\mathbf{v}_2$，正如所求。

这在物理上也合情合理。若 C 静止，相对速度就是 $\mathbf{v}_1$ 和 $\mathbf{v}_2$ 本身，它们本来就满足两个条件。若 $\mathbf{u}=\mathbf{v}_1+\mathbf{v}_2$，相对速度是 $-\mathbf{v}_2$ 和 $-\mathbf{v}_1$，它们同样长度相等而且互相垂直。

验算

取 $\mathbf{v}_1=(3,1)$，$\mathbf{v}_2=(1,3)$，与图中相同。公式给出 $\mathbf{u}=(2,2)\pm(1,1)$，即 $(3,3)$ 或 $(1,1)$。当 $\mathbf{u}=(3,3)$ 时，相对速度是 $(0,-2)$ 和 $(-2,0)$；当取 $(1,1)$ 时，相对速度是 $(2,0)$ 和 $(0,2)$。两对向量都长度相等而且互相垂直。用电脑测试 10^4 对随机的 $\mathbf{v}_1,\mathbf{v}_2$，结果确认了两个恒等式和两个解；另外对 10^3 对随机的等长垂直向量，电脑求得的解恰好是 $\mathbf{0}$ 和 $\mathbf{v}_1+\mathbf{v}_2$。

常见错误

- 把 A 相对于 C 的速度写成 $\mathbf{u}-\mathbf{v}_1$。正负号不影响本题的两个条件，但在大部分相对速度的题目中却至关重要。
- 忘记说明为何恰好有两个速度。这个数目取决于我们是在平面上讨论（在三维空间中，所有解构成一整个圆），也取决于半径不为零。

STEP 之外

由 $\mathbf{u}$ 转到 $\mathbf{w} = \mathbf{u} - \frac{1}{2}(\mathbf{v}_1 + \mathbf{v}_2)$，其实是转换参考系（reference frame）；而所有速度都平移同一个向量时相对速度不变，这个事实就是伽利略不变性（Galilean invariance），它是经典力学（classical mechanics）的起点。这里所用的“速度图”，即把速度画成点的做法，也能把碰撞和追逐问题化为几何问题。Kleppner 与 Kolenkow 合著的 *An Introduction to Mechanics*（力学入门）对相对运动和参考系有仔细的阐述。

详细解答

STEP III 2001，第 11 题

力学 · 转动能量守恒

STEP III 2001，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/01-s3-q11

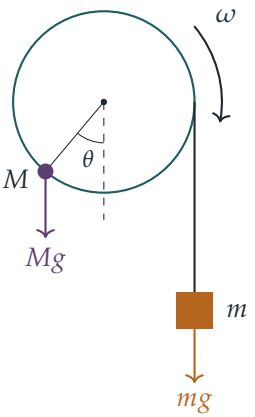
关键思路

整个系统都是光滑的（轮轴无摩擦，绳子是轻绳），所以机械能守恒（conservation of mechanical energy），只需一条能量方程便可把 ω 表示成 θ 的函数，完全不必考虑张力或其他力。圆柱体恰好在 ω^2 降至零时停下。于是第二部分变成关于单一个 θ 的函数的问题：悬挂质量所失去的势能（potential energy），必须始终大于质点所增加的势能。整理之后，这就是比较 m/M 与 $(1-\cos\theta)/\theta$ 的最大值，而求这个最大值并不困难。

解答

角速度

我们把 m 和 M 视为质点（particle），绳子视为不可伸长的轻绳（light and inextensible），且不在圆柱体上打滑，轮轴则视为光滑。设圆柱体沿着令 m 下降的方向转过角度 θ，并设 $\omega = \dot{\theta}$。



当圆柱体转过 θ 时，有长度为 $a\theta$ 的绳子松开，所以 m 下降了 $a\theta$，并以速率 $a\omega$ 运动。质点 M 位于圆柱体的边缘，所以它的速率也是 $a\omega$；它起初位于轴线下方深度 a 处，现在位于深度 $a\cos\theta$ 处，所以它上升了 $a(1-\cos\theta)$。

动能（kinetic energy）为

$$\tfrac{1}{2}I\omega^2+\tfrac{1}{2}M(a\omega)^2+\tfrac{1}{2}m(a\omega)^2=\tfrac{1}{2}\big(I+(M+m)a^2\big)\omega^2.$$

光滑的轮轴不作功，而绳子既轻且不可伸长，传递能量时并无损耗。所以动能的增加等于势能的净减少：

$$\tfrac{1}{2}\big(I+(M+m)a^2\big)\omega^2=mga\theta-Mga(1-\cos\theta),$$

由此得

$$\boxed{\omega^2=\frac{2ga\big(m\theta-M(1-\cos\theta)\big)}{I+(M+m)a^2}}\qquad\text{及}\qquad\omega=\sqrt{\frac{2ga\big(m\theta-M(1-\cos\theta)\big)}{I+(M+m)a^2}}.$$

建模上有一点值得验证：绳子必须保持拉紧。把能量方程对时间求导，得 $\big(I+(M+m)a^2\big)\ddot\theta=ga(m-M\sin\theta)$，而张力（tension）为 $T=m(g-a\ddot\theta)$。现在 $a\ddot\theta<g$ 等价于 $a^2(m-M\sin\theta)<I+(M+m)a^2$，即 $-Ma^2\sin\theta<I+Ma^2$，而这恒成立。所以在整个运动过程中 $T>0$，能量方程对所有 θ 都有效。

永不停下

开始时，$\ddot\theta=gam/(I+(M+m)a^2)>0$，所以圆柱体确实沿我们所选的方向开始转动。由于 ω 连续，它只有经过零才能改变正负号，所以只要对每个 $\theta>0$ 都有 $\omega^2>0$，圆柱体便会永远转动下去。设 $k=m/M$，由 ω^2 的表达式可见，这种情况恰好在以下条件成立时出现：

$$k\theta>1-\cos\theta\quad\text{对所有}\theta>0,\qquad\text{即}\qquad k>h(\theta)=\frac{1-\cos\theta}{\theta}\quad\text{对所有}\theta>0.$$

所以我们需要求出 h 在 $\theta > 0$ 上的最大值。

驻点（stationary points）。我们有

$$h'(\theta) = \frac{\theta\sin\theta - (1-\cos\theta)}{\theta^2}.$$

利用 $\sin\theta = 2\sin\frac{1}{2}\theta\cos\frac{1}{2}\theta$ 和 $1-\cos\theta = 2\sin^2\frac{1}{2}\theta$，

$$h'(\theta) = \frac{2\sin\frac{1}{2}\theta\left(\theta\cos\frac{1}{2}\theta - \sin\frac{1}{2}\theta\right)}{\theta^2}.$$

在 $0 < \theta < \pi$ 上，$\sin\frac{1}{2}\theta$ 和 $\cos\frac{1}{2}\theta$ 皆为正，所以 $h'(\theta)$ 的正负号与 $\theta - \tan\frac{1}{2}\theta$ 相同。设 $u(\theta) = \tan\frac{1}{2}\theta - \theta$。那么 $u(0) = 0$，$u'(\theta) = \frac{1}{2}\sec^2\frac{1}{2}\theta - 1$ 在 0 附近为负，而 $u''(\theta) = \frac{1}{2}\sec^2\frac{1}{2}\theta\tan\frac{1}{2}\theta > 0$，所以 u 在 $[0,\pi)$ 上是凸函数（convex function）。它先降至零以下，然后当 $\theta \to \pi$ 时趋向 $+\infty$。凸函数最多只能穿过零两次，所以 u 在 $(0,\pi)$ 内恰好有一个零点，那就是 α。因此 h 在 $(0,\alpha)$ 上递增，在 (α,π) 上递减。

> ? **想一想** h 在最大值处的值是多少？对于更大的 θ 又如何？

在 $\theta = \alpha$ 处，h' 的分子为零，所以 $1-\cos\alpha = \alpha\sin\alpha$，因此

$$h(\alpha) = \frac{1-\cos\alpha}{\alpha} = \sin\alpha.$$

对于 $\pi \le \theta \le 2\pi$，我们有 $\sin\frac{1}{2}\theta \ge 0$ 和 $\cos\frac{1}{2}\theta \le 0$，所以 $\theta\cos\frac{1}{2}\theta - \sin\frac{1}{2}\theta \le 0$，从而 $h' \le 0$：h 一直递减至 2π，所以在 $(\alpha, 2\pi]$ 上 $h < \sin\alpha$。对于 $\theta > 2\pi$，$h(\theta) \le 2/\theta < 1/\pi$。最后，$u(\frac{3}{4}\pi) = \tan\frac{3}{8}\pi - \frac{3}{4}\pi \approx 2.414 - 2.356 > 0$ 而 $u(\frac{2}{3}\pi) = \sqrt{3} - \frac{2}{3}\pi < 0$，所以 $\frac{2}{3}\pi < \alpha < \frac{3}{4}\pi$，由此得 $\sin\alpha > \sin\frac{3}{4}\pi = \frac{1}{\sqrt{2}} > \frac{1}{\pi}$。所以 $\sin\alpha$ 是 h 在 $\theta > 0$ 上的最大值，而且只在 $\theta = \alpha$ 处取得。

结论。若 $m/M > \sin\alpha$，则对每个 $\theta > 0$，

$$m\theta - M(1-\cos\theta) = M\theta\big(k - h(\theta)\big) \ge M\theta(k - \sin\alpha) > 0,$$

所以对所有 $\theta > 0$ 都有 $\omega^2 > 0$，圆柱体会一直转动而不会停下，这正是所要证明的。

（反过来说，若 $m/M < \sin\alpha$，则在 $\theta = \alpha$ 处 $\omega^2 < 0$，所以 ω 在此之前已降至零，圆柱体会停下并往回摆动。在临界情况 $m/M = \sin\alpha$ 下，系统会缓缓趋近平衡位置 $\theta = \alpha$，却永远不会真正到达。）

验算

- 当 $M = 0$ 时，公式给出 $\omega^2 = 2gma\theta/(I + ma^2)$，这是质量从滑轮上松开绳子下降的标准结果：$\omega^2$ 线性增长，与恒定角加速度的情况一致。

- 用二分法（bisection）解 $\alpha = \tan\frac{1}{2}\alpha$，得 $\alpha \approx 2.3311$（约 133.6°）和 $\sin\alpha \approx 0.7246$；在 $0 < \theta < 100$ 上对 $(1-\cos\theta)/\theta$ 作数值搜索，得出相同的最大值 0.7246。
- 对随机选取的 I, a, M, m 以数值方法逐步求解运动方程，结果与 ω^2 的公式吻合；而且当 $m/M = 0.735$ 时圆柱体永不停下，当 $m/M = 0.715$ 时则会停下。

常见错误

- 遗漏 M 或 m 的动能：两者都以速率 $a\omega$ 运动，所以有效转动惯量（moment of inertia）是 $I + (M+m)a^2$。
- 弄错 M 的高度：它上升了 $a(1-\cos\theta)$，而不是 $a\sin\theta$。
- 只证明 h 在 α 处有局部极大值，却没有检查对所有更大的 θ，h 的值都较小。

STEP 之外

这是在势场中的单自由度（one degree of freedom）运动：$\frac{1}{2}J\omega^2 + V(\theta) = 0$，其中 $V(\theta) = Mga(1-\cos\theta) - mga\theta$，是一个倾斜的“洗衣板”势（washboard potential）。从静止开始，运动会一去不返，当且仅当对每个 $\theta > 0$，V 都保持低于其初始值；同样的图像也可描述物理学中的受驱摆（driven pendulum）和约瑟夫森结（Josephson junction）。Kleppner 与 Kolenkow 所著的 *An Introduction to Mechanics*（力学入门）有讨论这类能量图。

详细解答

STEP III 2001，第 12 题

概率与统计 · 以指示变量求期望值

STEP III 2001，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/01-s3-q12

关键思路

在第 (i) 部分，不要尝试求出黑球数目的分布：把这个数目拆开，每个黑球对应一个指示变量（indicator），然后运用期望值的线性性质（linearity of expectation）。不放回地抽球直至袋空，其实就是把所有球随机排成一列，所以每个指示变量都归结为一个对称性论证（symmetry argument）。在第 (ii) 部分，两次抽到白球之间袋中的球并无改变，所以等候每个白球出现的次数服从几何分布（geometric distribution）；而题目中的“hence”，则来自这个观察：取走一个白球之后，我们回到同样的处境，只是白球少了一个。

解答

第 (i) 部分

设想一直抽下去，直至袋空为止。由于每次都是随机抽取，而且抽出的球不放回，这 $b+w$ 个球被抽出的次序，就是它们的一个均匀随机排列（uniformly random ordering）。到最后一个白球被取走时已抽出的黑球，正好就是在这个排列中位于最后一个白球之前的黑球。

把黑球编号为 $1, 2, \ldots, b$；若黑球 j 在最后一个白球之前被抽出，令 $I_j = 1$，否则令 $I_j = 0$。设 N 为最后一个白球被取走时已抽出的黑球数目，则

$$N = I_1 + I_2 + \cdots + I_b, \qquad \text{所以} \qquad \mathrm{E}(N) = \sum_{j=1}^{b} \mathrm{P}(I_j = 1)$$

这是根据期望值的线性性质（它并不需要各个 I_j 之间互相独立）。

想一想 黑球 j 排在所有白球之后的概率是多少?

只看由黑球 j 和 w 个白球组成的这 $w+1$ 个球。在所有球的均匀随机排列中，这 $w+1$ 个球的相对次序同样是均匀随机的，所以黑球 j 处于它们当中 $w+1$ 个位置的任何一个，机会均等。它排在这组球最后的概率是 $\frac{1}{w+1}$，而这正是事件 $I_j = 0$。因此

$$\mathrm{P}(I_j = 1) = 1 - \frac{1}{w+1} = \frac{w}{w+1},$$

从而

$$\boxed{\mathrm{E}(N) = \frac{bw}{w+1}.}$$

第 (ii) 部分

现在黑球会放回袋中，所以在抽到白球之前，袋中始终有 b 个黑球和 w 个白球。因此每次抽到黑球的概率是 $p = \frac{b}{b+w}$，抽到白球的概率是 $q = \frac{w}{b+w}$，而且与之前各次抽取独立，因为每次袋中的球都相同。

设 X 为抽到第一个白球之前抽到黑球的次数。那么 $X = n$ 表示先抽到 n 次黑球，然后抽到一个白球，所以

$$\mathrm{P}(X = n) = p^n q, \qquad n = 0, 1, 2, \ldots$$

若 $b = 0$，则 $X = 0$，结果显然成立，故取 $b \geq 1$；这时 $0 < p < 1$，利用当 $|x| < 1$ 时 $\sum_{n\geq 1} n x^{n-1} = \frac{1}{(1-x)^2}$（把等比级数（geometric series）求导即得），

$$\mathrm{E}(X) = \sum_{n=0}^{\infty} n p^n q = pq \sum_{n=1}^{\infty} n p^{n-1} = \frac{pq}{(1-p)^2} = \frac{p}{q} = \frac{b/(b+w)}{w/(b+w)} = \frac{b}{w},$$

因为 $1 - p = q$。这就是所要证明的结果。

至于最后一个白球，把抽到黑球的总次数写成 $X_w + X_{w-1} + \cdots + X_1$，其中 X_k 是袋中恰好有 k 个白球期间抽到黑球的次数。当袋中尚余 k 个白球时，袋中有 b 个黑球和 k 个白球，所以这个阶段正是把 w 换成 k 的“第一个白球”问题，而刚才证明的结果给出 $\mathrm{E}(X_k) = \frac{b}{k}$。根据期望值的线性性质，最后一个白球被取走时抽到黑球的期望次数是

$$\boxed{\sum_{k=1}^{w} \frac{b}{k} = b\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{w}\right).}$$

验算

当 $w = 1$ 时，两部分的结果都很容易直接看出。在 (i) 中，每一个黑球排在白球之前的概率都是 $\frac{1}{2}$，得出 $\frac{b}{2}$，与 $\frac{bw}{w+1}$ 相符。在 (ii) 中，只有一个阶段，而总和就是 $\frac{b}{1} = b$。在 (i) 中当 $w \to \infty$ 时，几乎每个黑球都在最后一个白球之前被抽出，而确实有 $\frac{bw}{w+1} \to b$。

我们对所有 $b \le 4$、$w \le 4$ 的情况逐一列出每个排列来验算 (i)；当 $b = 3$、$w = 4$ 时，结果恰好是 $\frac{12}{5}$。至于 (ii)，以 $b = 3$、$w = 4$ 进行 2×10^5 次模拟，所得平均值约为 6.245，而 $3\left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4}\right) = \frac{25}{4} = 6.25$。

常见错误

- 在 (i) 中，假设各个黑球是否被抽出是互相独立的。期望值的线性性质并不需要独立性，这正是指示变量如此好用的原因。
- 在 (ii) 中，采用了把抽到白球那一次也计算在内的几何分布，结果得出 $\frac{1}{q} = \frac{b+w}{w}$，比答案多了一。
- 在最后一部分的每个阶段都以 $b + w$ 作分母：袋中白球的数目每过一个阶段就减少一个。

STEP 之外

第二部分其实是乔装了的优惠券收集问题（coupon collector's problem）：把各个阶段的几何等候时间加起来，便得到一个调和和（harmonic sum），其增长速度与 $\ln w$ 相若。像第 (i) 部分那样把一个计数写成指示变量之和，是概率论中最强而有力的技巧之一；日后计算随机置换（random permutation）中不动点（fixed point）、纪录（record）和循环（cycle）的期望数目时，你会再次遇上它。Blitzstein 与 Hwang 合著的 *Introduction to Probability*（概率论入门）中，有许多关于这两个想法的例子。

详细解答

STEP III 2001，第 13 题

概率与统计 · 硬币序列游戏

STEP III 2001，第 13 题

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

step.maths.org/questions/01-s3-q13

本题在 *Beyond STEP* 第 12.8 节 *Penney 游戏*中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch12/penney.html

关键思路

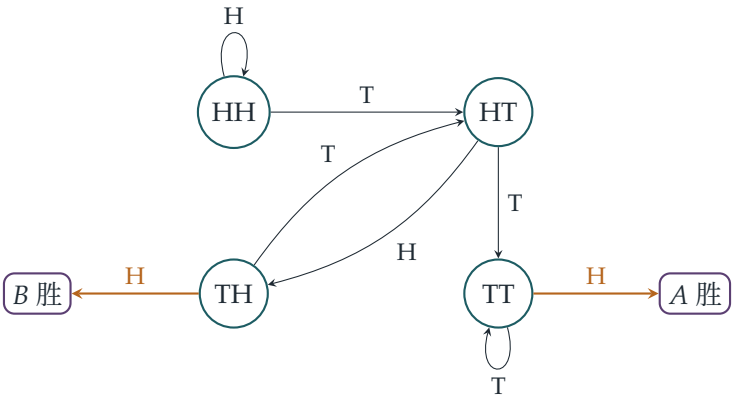
两个目标序列的长度都是三，所以没有人能在第三次掷币之前获胜；而从那时起，对往后有影响的就只有*最后两次掷币的结果*。这样便只有四个“状态”（state）（HH、HT、TH、TT），而以下一次掷币的结果作条件（conditioning），就可把这个游戏化为一个关于各状态下获胜概率的小型线性方程组。第 (ii) 和第 (iii) 部分也可以用同样的方法处理，不过在四场游戏之中，有两场用简短的直接论证会更快。

解答

在整个解答中，各次掷币互相独立（independent），每次出现 H 或 T 的概率均为 $\frac{1}{2}$。只要已经掷了最少两次而仍未有人获胜，下一次掷币连同最后两次的结果，就决定了此刻是否有人获胜；若没有，也决定了新的最后两次结果。更早的掷币结果并不相干，因为每个目标序列只有三个字母长。所以，从某一刻起获胜的概率只取决于最后两次掷币的结果，而且与把这两次当作一场新游戏的头两次掷币时的概率相同。

第 (i) 部分

设 p_{XY} 为已知头两次掷币结果是 XY 时，A（TTH）击败 B（THH）的概率。下图显示下一次掷币如何使我们在各状态之间转移。



为何 $p_{\mathrm{TT}} = 1$。假设最后两次掷币的结果是 TT。正面终会出现的概率为 1（连续出现 n 次反面的机会是 $2^{-n} \to 0$）。在正面出现之前，掷出的全是反面，所以 THH 不可能出现；而第一个正面之前最少有两个反面，所以它完成了 TTH。因此 A 必定获胜：$p_{\mathrm{TT}} = 1$。

为何 $p_{\mathrm{HT}} = \frac{1}{2}p_{\mathrm{TH}} + \frac{1}{2}p_{\mathrm{TT}}$。在 HT 之后，HTH 和 HTT 都不是目标序列，所以没有人会在第三次掷币时获胜。若第三次是 H（概率为 $\frac{1}{2}$），最后两次的结果便是 TH；若是 T，则是 TT。以第三次掷币的结果作条件，并运用上文“只有最后两次掷币才有影响”这一点，

$$p_{\mathrm{HT}} = \tfrac{1}{2}p_{\mathrm{TH}} + \tfrac{1}{2}p_{\mathrm{TT}}.$$

其余两个状态。在 TH 之后，掷出正面便得 THH，B 获胜；掷出反面则变成 HT。在 HH 之后，掷出正面仍是 HH，掷出反面则变成 HT（而 HHH、HHT 在这里都不是目标序列）。所以

$$p_{\mathrm{TH}} = \tfrac{1}{2}\cdot 0 + \tfrac{1}{2}p_{\mathrm{HT}}, \qquad p_{\mathrm{HH}} = \tfrac{1}{2}p_{\mathrm{HH}} + \tfrac{1}{2}p_{\mathrm{HT}}.$$

由第二条方程得 $p_{\mathrm{HH}} = p_{\mathrm{HT}}$。把 $p_{\mathrm{TH}} = \frac{1}{2}p_{\mathrm{HT}}$ 和 $p_{\mathrm{TT}} = 1$ 代入 p_{HT} 的方程，

$$p_{\mathrm{HT}} = \tfrac{1}{4}p_{\mathrm{HT}} + \tfrac{1}{2} \quad\Longrightarrow\quad \tfrac{3}{4}p_{\mathrm{HT}} = \tfrac{1}{2} \quad\Longrightarrow\quad p_{\mathrm{HT}} = \tfrac{2}{3}.$$

因此 $p_{\mathrm{HH}} = p_{\mathrm{HT}} = \frac{2}{3}$，而 $p_{\mathrm{TH}} = \frac{1}{2}\cdot\frac{2}{3} = \frac{1}{3}$，正如题目所求。

想一想 假如 TH 其实是在第 5 和第 6 次掷币时出现的，为何可以把游戏从 TH “重新开始”，当作 TH 就是头两次掷币的结果?

因为游戏往后的发展只取决于往后的掷币结果和最后两次掷币的结果：任何尚未完成的 TTH 或 THH，最多只能用到已掷出的最后两次结果。更早的历史对两个序列都没有帮助，而往后的掷币结果亦与它独立。

整体概率。没有人能在头两次掷币内获胜，而四种可能的组合各有 $\frac{1}{4}$ 的概率。根据全概率定理（law of total probability），

$$\mathrm{P}(A \text{ 胜}) = \tfrac{1}{4}\left(\tfrac{2}{3} + \tfrac{2}{3} + \tfrac{1}{3} + 1\right) = \tfrac{1}{4}\cdot\tfrac{8}{3} = \tfrac{2}{3}.$$

第 (ii) 部分

现在 B 的序列是 THH，C 的是 HHT。这里用直接论证最为简洁。

若头两次掷币的结果是 HH，则 C 获胜：其后出现的第一个反面（它出现的概率为 1）之前是 HH，于是完成了 HHT；而在它之前每次掷出的都是正面，所以 THH 不可能出现过。

*否则*头两次掷币之中有一个反面。HH 这个模式终会出现的概率为 1。考虑它第一次出现的位置。它并非由第 1 次掷币开始，所以紧接在它之前的一次掷币是存在的，而且结果不是 H（否则便有一个更早的 HH），所以是 T。于是第一个 HH 完成了一个 THH。同时，HHT 不可能在此之前出现，因为它需要一个更早的 HH。所以 B 获胜。

因此，除非头两次掷币的结果是 HH，否则 B 获胜：

$$\mathrm{P}(B\text{ 胜}) = 1 - \tfrac{1}{4} = \tfrac{3}{4}.$$

第 (iii) 部分

C（HHT）*对* D（HTT）。设 q_{XY} 为最后两次掷币结果是 XY 时 C 获胜的概率。由 HH 出发，C 必定获胜，理由与 (ii) 相同：第一个反面完成了 HHT，而 HTT 需要在 HT 之后再有一个反面，这不可能先出现。所以 $q_{\mathrm{HH}} = 1$。在其余各状态下以下一次掷币的结果作条件：

HT:	T 得 HTT（D 胜），H 得 TH:	$q_{\mathrm{HT}} = \frac{1}{2}q_{\mathrm{TH}}$,
TH:	H 得 HH，T 得 HT:	$q_{\mathrm{TH}} = \frac{1}{2} + \frac{1}{2}q_{\mathrm{HT}}$,
TT:	H 得 TH，T 得 TT:	$q_{\mathrm{TT}} = \frac{1}{2}q_{\mathrm{TH}} + \frac{1}{2}q_{\mathrm{TT}}$.

由最后一条得 $q_{\mathrm{TT}} = q_{\mathrm{TH}}$。把第一条代入第二条，得 $q_{\mathrm{TH}} = \frac{1}{2} + \frac{1}{4}q_{\mathrm{TH}}$，所以 $q_{\mathrm{TH}} = \frac{2}{3}$、$q_{\mathrm{TT}} = \frac{2}{3}$ 及 $q_{\mathrm{HT}} = \frac{1}{3}$。因此

$$\mathrm{P}(C\text{ 胜}) = \tfrac{1}{4}\left(1 + \tfrac{1}{3} + \tfrac{2}{3} + \tfrac{2}{3}\right) = \tfrac{1}{4}\cdot\tfrac{8}{3} = \tfrac{2}{3} > \tfrac{1}{2},$$

所以 C 比 D 更有可能获胜。

D（HTT）*对* A（TTH）。这与第 (ii) 部分互为镜像，只是 H 和 T 对调了。若头两次掷币的结果是 TT，则 A 获胜：第一个正面完成了 TTH，而 HTT 需要一个正面之后接两个反面，这不可能在那第一个正面之前发生。否则头两次掷币之中有一个正面。TT 这个模式终会出现的概率为 1；它第一次出现时并非由第 1 次掷币开始，而它之前的一次掷币不可能是 T（否则便有一个更早的 TT），所以是 H，于是第一个 TT 完成了 HTT。TTH 不可能在更早时出现，因为它需要一个更早的 TT。所以 D 获胜，而

$$\mathrm{P}(D\text{ 胜}) = 1 - \tfrac{1}{4} = \tfrac{3}{4} > \tfrac{1}{2}.$$

所以 D 比 A 更有可能获胜。

想一想 A 胜 B，B 胜 C，C 胜 D，而 D 又胜 A。这怎么可能?

“较有可能击败”并不是一个传递关系（transitive relation），所以并无矛盾：四位玩家形成一个循环，A、B、C、D 在各自的对局中获胜的概率分别是 $\frac{2}{3}$、$\frac{3}{4}$、$\frac{2}{3}$、$\frac{3}{4}$。后选的一方总可以挑一个能击败先选者的序列。

验算

- 在第 (i) 部分，代回验证：$\frac{1}{2}\cdot\frac{1}{3}+\frac{1}{2}\cdot 1=\frac{2}{3}=p_{\mathrm{HT}}$，以及 $\frac{1}{2}\cdot\frac{2}{3}=\frac{1}{3}=p_{\mathrm{TH}}$。
- 状态法同样适用于第 (ii) 部分和 D–A 这场游戏：在 B 对 C 的游戏中，由 HH、HT、TH、TT 出发时 B 获胜的概率分别是 $0, 1, 1, 1$，平均值为 $\frac{3}{4}$。
- 为每组对局模拟 200 000 场游戏，所得的获胜频率为 0.664、0.752、0.666 和 0.749，而精确值是 $\frac{2}{3}$、$\frac{3}{4}$、$\frac{2}{3}$、$\frac{3}{4}$。

常见错误

- 忘记游戏可以在第三次掷币时结束，例如写成 $p_{\mathrm{TH}}=\frac{1}{2}p_{\mathrm{HH}}+\frac{1}{2}p_{\mathrm{HT}}$，而没有留意到 TH 之后掷出 H 就是 B 即时获胜。
- 把四个状态的概率取平均，却没有说明头两次掷币的每种组合概率均为 $\frac{1}{4}$，以及没有人能在第三次掷币之前获胜。

STEP 之外

这就是 Penney 游戏（Penney’s game），是非传递概率（non-transitive probabilities）的著名例子。状态法其实就是吸收马可夫链（absorbing Markov chains）的理论，当中的吸收概率所满足的，正是这里所解的那类线性方程。任意两个序列之间的胜负比（odds）甚至有一条闭合公式，即康威（Conway）的“领先数”（leading numbers）算法，可以用鞅（martingales）来证明。关于马可夫链和首步分析（first-step analysis），可参阅 Grimmett 与 Welsh 的 *Probability: An Introduction*，或 Blitzstein 与 Hwang 的 *Introduction to Probability*。

详细解答

STEP III 2001，第 14 题

概率与统计 · 方差的无偏估计量

STEP III 2001，第 14 题

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

step.maths.org/questions/01-s3-q14

关键思路

整道题的关键只有一点：“样本的方差”所用的除数是 $n = 2$，因此会低估总体方差（population variance），而把它乘以二便可修正过来。只有两个值时，样本平均值（sample mean）就是两者的中点，所以每个值与它相距都是两值差距的一半。于是 V 只是单一个差 $X_1 - X_2$ 的函数，而最后要求的概率就变成正方形 $[0, a]^2$ 内的一块面积。

解答

X 的方差

X 的密度函数（density）在 $[0, a]$ 上是 $f(x) = \frac{1}{a}$。由对称性可知 $\mathrm{E}(X) = \frac{1}{2}a$，而

$$\mathrm{E}(X^2) = \int_0^a \frac{x^2}{a}\,\mathrm{d}x = \frac{a^2}{3}.$$

所以

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2 = \frac{a^2}{3} - \frac{a^2}{4} = \frac{a^2}{12},$$

正如题目所求。

样本方差

对于样本 X_1, X_2，样本平均值是 $\bar{X} = \frac{1}{2}(X_1 + X_2)$，而样本的方差就是各值与它的偏差平方的平均值：

$$V = \tfrac{1}{2}\big[(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2\big].$$

由于 $X_1 - \bar{X} = \frac{1}{2}(X_1 - X_2)$ 及 $X_2 - \bar{X} = -\frac{1}{2}(X_1 - X_2)$，两个偏差平方都等于 $\frac{1}{4}(X_1 - X_2)^2$。因此

$$V = \tfrac{1}{2} \cdot 2 \cdot \tfrac{1}{4}(X_1 - X_2)^2 = \tfrac{1}{4}(X_1 - X_2)^2,$$

这正是题目给出的形式。

? **想一想** 怎样不用二重积分（double integral）而求出 $\mathrm{E}\big((X_1 - X_2)^2\big)$?

由于 X_1 和 X_2 的分布相同，$\mathrm{E}(X_1 - X_2) = 0$，所以 $\mathrm{E}\big((X_1 - X_2)^2\big) = \mathrm{Var}(X_1 - X_2)$。因为 X_1 和 X_2 独立（independent），它们的方差可以相加（负号并无影响，因为 $\mathrm{Var}(-X_2) = \mathrm{Var}(X_2)$）：

$$\mathrm{Var}(X_1 - X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) = \frac{a^2}{12} + \frac{a^2}{12} = \frac{a^2}{6}.$$

因此

$$\mathrm{E}(2V) = \tfrac{1}{2}\,\mathrm{E}\big((X_1 - X_2)^2\big) = \tfrac{1}{2} \cdot \frac{a^2}{6} = \frac{a^2}{12} = \mathrm{Var}(X).$$

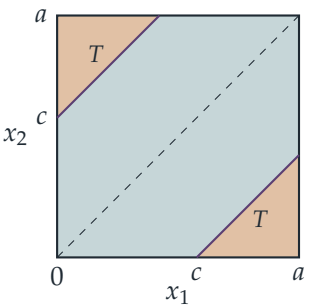
无论 a 取甚么值，$2V$ 的期望值（expected value）都等于它所估计的量，所以 $2V$ 是 X 的方差的无偏估计量（unbiased estimator）。（相比之下，$\mathrm{E}(V) = \frac{a^2}{24}$，只有真正方差的一半。）

概率

由于 $V \geq 0$ 而 $\frac{1}{12}a^2 > 0$，我们可以取平方根：

$$V < \frac{a^2}{12} \iff (X_1 - X_2)^2 < \frac{a^2}{3} \iff |X_1 - X_2| < \frac{a}{\sqrt{3}}.$$

因为 X_1 和 X_2 独立，而且各自服从均匀分布（uniform distribution），所以数对 (X_1, X_2) 均匀分布于正方形 $[0, a]^2$ 上。因此所求概率就是正方形面积中落在带状区域 $|x_1 - x_2| < c$ 内的比例，其中 $c = a/\sqrt{3} < a$。



余下的部分由角落两个着色的三角形 T 组成，每个都是等腰直角三角形，两条直角边的

长度均为 $a-c$。它们的总面积是 $2\cdot\frac{1}{2}(a-c)^2=(a-c)^2$，所以

$$\mathrm{P}\Big(V<\frac{a^2}{12}\Big)=1-\frac{(a-c)^2}{a^2}=1-\Big(1-\frac{1}{\sqrt{3}}\Big)^2=1-\Big(\frac{4}{3}-\frac{2}{\sqrt{3}}\Big)=\frac{2}{\sqrt{3}}-\frac{1}{3}.$$

也就是说，概率的准确值是

$$\mathrm{P}\Big(V<\frac{a^2}{12}\Big)=\frac{2\sqrt{3}-1}{3}.$$

取 $\sqrt{3}\approx 1.732$，这个值是 $\frac{2.464}{3}\approx 0.821$，所以准确至一位有效数字，概率是 0.8。

验算

- 答案与 a 无关，这是理所当然的：把 X 按某个倍数缩放，V 和 $\frac{1}{12}a^2$ 都会按该倍数的平方缩放。
- 取 $a=3$ 模拟 4×10^5 个数对，得出 $2V$ 的平均值为 0.751（目标值 0.75），而 $V<\frac{1}{12}a^2$ 所占的比例为 0.821，与 $\frac{1}{3}(2\sqrt{3}-1)=0.8214$ 吻合。
- 合理性检查：概率落在 $(0,1)$ 之内；若把 $\frac{1}{12}a^2$ 换成 $\frac{1}{4}a^2$（V 可取的最大值），则 $c=a$，而公式 $1-(1-c/a)^2$ 给出 1。

常见错误

- 计算“样本的方差”时用了除数 $n-1$：这样会得到 $\frac{1}{2}(X_1-X_2)^2$，而不是题目所述的 $\frac{1}{4}(X_1-X_2)^2$。题目要的是除数 n。
- 忘记了绝对值，结果只计算了两个三角形的其中一个（这样得出的概率会小于 $\frac{1}{2}$）。
- 写成 $\mathrm{Var}(X_1-X_2)=\mathrm{Var}(X_1)-\mathrm{Var}(X_2)=0$。

STEP 之外

这里的结果是一条一般定理在 $n=2$ 时的情况：从方差为 σ^2 的分布中抽取任何大小为 n 的样本，$\sum(X_i-\bar{X})^2$ 的平均值都是 $(n-1)\sigma^2$。这就是为甚么用除数 $n-1$（贝塞尔校正，Bessel's correction）会得到无偏估计量。大学一年级的统计学会进一步探讨：在各个无偏估计量之中，哪一个的方差最小。Grimmett 与 Welsh 的 *Probability: An Introduction*，以及 Blitzstein 与 Hwang 的 *Introduction to Probability*，都以本文所用的方式讲解和的期望值与方差。

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

2001 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。

