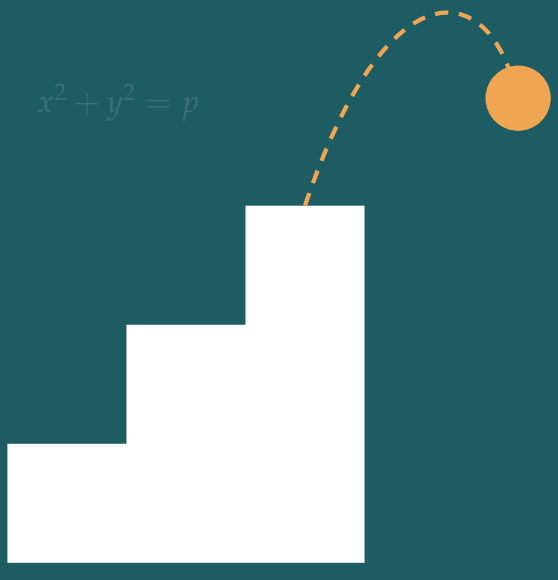
STEP 备试 · 大学数学

Beyond STEP

STEP I, II 及 III

2011 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

2011 年详细解答

Keith McNulty

Beyond STEP：以历届 STEP 试题为门，走进大学数学

2011 年详细解答

2011 年 STEP 试卷（STEP I, II 及 III）全部 39 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 隐函数微分法 纯数学 . . . 3

第 2 题 代换积分法 纯数学 . . . 7

第 3 题 三角恒等式 纯数学 . . . 10

第 4 题 抛物线的切线 纯数学 . . . 14

第 5 题 含绝对值的积分 纯数学 . . . 18

第 6 题 二项式级数求和 纯数学 . . . 21

第 7 题 可分离变量微分方程 纯数学 . . . 24

第 8 题 丢番图方程 纯数学 . . . 28

第 9 题 抛体运动 力学 . . . 31

第 10 题 弹性碰撞 力学 . . . 35

第 11 题 刚体的平衡 力学 . . . 39

第 12 题 排队与选票问题 概率与统计 . . . 43

第 13 题 连续随机变量 概率与统计 . . . 47

STEP II

第 1 题 含平方根的方程 纯数学 . . . 51

第 2 题 丢番图方程 纯数学 . . . 55

第 3 题 以单调性证明不等式 纯数学 . . . 58

第 4 题 三角方程 纯数学 . . . 62

第 5 题 向量与反射 纯数学 . . . 66

第 6 题 分部积分法 纯数学 . . . 69

第 7 题 数列与级数 纯数学 . . . 73

第 8 题 圆的渐伸线 纯数学 . . . 77

第 9 题 碰撞与恢复系数 力学 . . . 81

第 10 题 抛体与冲量 力学 85
第 11 题 二维空间中的平衡 力学 88
第 12 题 博弈中的等比级数 概率与统计 92
第 13 题 偏度的量度 概率与统计 96

STEP III

第 1 题 降阶法 纯数学 100
第 2 题 多项式的有理根 纯数学 104
第 3 题 解三次方程 纯数学 107
第 4 题 杨氏不等式 纯数学 111
第 5 题 以积分求面积 纯数学 115
第 6 题 积分代换 纯数学 119
第 7 题 数学归纳法 纯数学 122
第 8 题 复数轨迹 纯数学 126
第 9 题 圆柱上的运动 力学 130
第 10 题 弹性绳与简谐运动 力学 134
第 11 题 刚体转动 力学 138
第 12 题 以概率生成函数处理复合分布 概率与统计 142
第 13 题 离散分布的众数 概率与统计 145

详细解答

STEP I 2011，第 1 题

纯数学 • 隐函数微分法

> **STEP I 2011，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/11-s1-q1

关键思路

用隐函数求导（implicit differentiation），毋须把 y 解出来，就能得到每条曲线的斜率（gradient）。之后每一部分都只需写出交点 (p,q) 须满足的条件：它在直线上、它在曲线上，而两者的斜率之间有某种关系（相切时相等，法线（normal）时乘积为 -1）。斜率条件给出 p 与 q 之间一个简单的关系；把这关系代入另外两条方程，便可用两种不同方法确定 p，而两者必须一致。

解答

第 (i) 部分

把 y 视为 x 的函数（function），将 $\dfrac{a}{x}+\dfrac{b}{y}=1$ 对 x 求导，得

$$-\frac{a}{x^2}-\frac{b}{y^2}\frac{\mathrm{d}y}{\mathrm{d}x}=0.$$

由于 $b\neq 0$，可移项得

$$\frac{\mathrm{d}y}{\mathrm{d}x}=-\frac{ay^2}{bx^2},$$

此即所求。（曲线上每一点的 x 和 y 均不为零，因为方程中含有 a/x 和 b/y。）

斜率相等。直线 $ax+by=1$ 的斜率为 $-\frac{a}{b}$。曲线在 (p,q) 的斜率为 $-\dfrac{aq^2}{bp^2}$，因此

$$-\frac{aq^2}{bp^2}=-\frac{a}{b}\quad\Longrightarrow\quad\frac{q^2}{p^2}=1\quad\Longrightarrow\quad q^2=p^2,$$

其中我们约去了 a/b，这是容许的，因为 $a\neq 0$。故 $p=\pm q$。

两种情况。点 (p,q) 亦满足

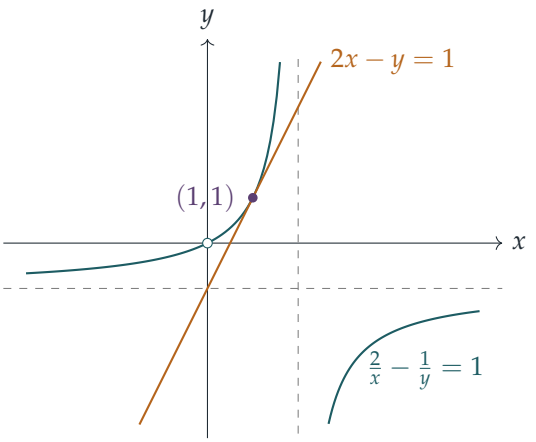
$$ap+bq=1 \qquad 及 \qquad \frac{a}{p}+\frac{b}{q}=1.$$

- 若 $q=p$：由直线得 $p(a+b)=1$，所以 $p=\dfrac{1}{a+b}$（特别地，$a+b\neq 0$）。由曲线得 $\dfrac{a+b}{p}=1$，所以 $p=a+b$。比较两者，得 $a+b=\dfrac{1}{a+b}$，即 $(a+b)^2=1$。
- 若 $q=-p$：由直线得 $p(a-b)=1$，所以 $p=\dfrac{1}{a-b}$。由曲线得 $\dfrac{a-b}{p}=1$，所以 $p=a-b$。比较两者，得 $(a-b)^2=1$。

所以 $(a-b)^2=1$ 或 $(a+b)^2=1$，此即所求。

> **想一想** 这样的点真的存在吗？还是我们只找到了一个必要条件（necessary condition）？

它们确实存在。取 $a=2$、$b=-1$，则 $a+b=1$。点 $(1,1)$ 既在直线 $2x-y=1$ 上，亦在曲线 $\frac{2}{x}-\frac{1}{y}=1$ 上，而两者在该点的斜率都是 2。这里曲线为 $y=\dfrac{x}{2-x}$，即去掉原点的一条直角双曲线（rectangular hyperbola）。在公共点上斜率相等，表示该直线是切线（tangent）：



空心圆圈标示原点：它在双曲线 $y=\frac{x}{2-x}$ 上，但不在原曲线上。

第 (ii) 部分

将 $\dfrac{a}{x}-\dfrac{b}{y}=1$ 隐函数求导，

$$-\frac{a}{x^2}+\frac{b}{y^2}\frac{\mathrm{d}y}{\mathrm{d}x}=0 \quad\Longrightarrow\quad \frac{\mathrm{d}y}{\mathrm{d}x}=\frac{ay^2}{bx^2}.$$

假设直线在 (p,q) 是曲线的法线。那么 $p\neq 0$、$q\neq 0$，而且

$$ap+bq=1 \qquad 及 \qquad \frac{a}{p}-\frac{b}{q}=1.$$

切线斜率 $\dfrac{aq^2}{bp^2}$ 与直线斜率 $-\dfrac{a}{b}$ 都是有限且非零的，因为 a、b、p、q 全不为零。所以“法线”即表示两者乘积为 -1：

$$\frac{aq^2}{bp^2}\cdot\left(-\frac{a}{b}\right)=-1 \quad\Longrightarrow\quad a^2q^2=b^2p^2 \quad\Longrightarrow\quad aq=\pm bp.$$

想一想 哪一个正负号是可能的?

把每个符号代入曲线方程检验。

- 若 $aq=bp$，则 $\dfrac{b}{q}=\dfrac{a}{p}$，所以 $\dfrac{a}{p}-\dfrac{b}{q}=0\neq 1$。这个符号不可能。
- 若 $aq=-bp$,则 $\dfrac{b}{q}=-\dfrac{a}{p}$,于是曲线方程变为 $\dfrac{2a}{p}=1$,得 $p=2a$,从而 $q=-\dfrac{bp}{a}=-2b$。

把 $p=2a$、$q=-2b$ 代入直线方程，得

$$2a^2-2b^2=1, \qquad 即 \qquad a^2-b^2=\tfrac{1}{2},$$

此即所求。法线的垂足为点 $(2a,-2b)$。

验算

- 对于第 (i) 部分, 上面 $a=2$、$b=-1$ 的例子满足 $(a+b)^2=1$, 而直线在 $(1,1)=(p,p)$ 相切。同样, $a=2$、$b=1$ 满足 $(a-b)^2=1$, 而直线 $2x+y=1$ 在 $(1,-1)$ 与 $\frac{2}{x}+\frac{1}{y}=1$ 相切。
- 对于第 (ii) 部分，取 $a=\frac{\sqrt{3}}{2}$、$b=\frac{1}{2}$，使 $a^2-b^2=\frac{1}{2}$，并取点 $(2a,-2b)=(\sqrt{3},-1)$。直线给出 $\frac{3}{2}-\frac{1}{2}=1$，曲线给出 $\frac{1}{2}+\frac{1}{2}=1$。该点的切线斜率为 $\dfrac{aq^2}{bp^2}=\frac{1}{\sqrt{3}}\approx 0.577$，直线斜率为 $-\sqrt{3}\approx -1.732$，乘积为 -1。
- 数值上，在数以百计的随机点，斜率公式都与有限差分（finite difference）所得的斜率吻合；而对 (a,b) 作随机搜索，只有当 a^2-b^2 在舍入误差范围内等于 0.5 时才会找到法线。

常见错误

- 在第 (i) 部分遗漏 $q=-p$ 的情况：正是这个情况给出 $(a-b)^2=1$。

► 在第 (ii) 部分停在 $aq = \pm bp$，没有检查曲线容许哪一个符号。

STEP 之外

隐函数求导让我们初窥隐函数定理（implicit function theorem）：在任何满足 $\partial F/\partial y \neq 0$ 的点附近，方程 $F(x,y) = 0$ 把 y 定义为 x 的可微函数，且 $\mathrm{d}y/\mathrm{d}x = -F_x/F_y$。这里的相切论证亦有一个代数上的“孪生”版本：一条直线与一条圆锥曲线（conic）相切，当且仅当求两者交点的二次方程有重根（repeated root）；你可以用这方法重新推导第 (i) 部分。若想了解这一切背后的分析（analysis），可读 Abbott 的 *Understanding Analysis*（《理解分析》）。

详细解答

STEP I 2011，第 2 题

纯数学 · 代换积分法

STEP I 2011，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/11-s1-q2

关键思路

积分 E 没有初等闭式（elementary closed form），所以我们根本不打算求它的值，而是把其余每个积分变形，直到 E 本身出现为止。对于头两个积分，做法是用代数除法（algebraic division）把分式拆开，从而把麻烦的 $\frac{1}{1+x}$ 分离出来。对于 (i) 和 (ii)，做法是选取一个代换（substitution），使被积函数（integrand）在 $[0,1]$ 上变回 $\frac{\mathrm{e}^u}{1+u}$：观察指数部分，想想怎样才能把它变成新的变量。

解答

头两个积分

把 x 写成 $x=(1+x)-1$，

$$\frac{x\,\mathrm{e}^x}{1+x}=\mathrm{e}^x-\frac{\mathrm{e}^x}{1+x},$$

因此

$$\int_0^1\frac{x\,\mathrm{e}^x}{1+x}\,\mathrm{d}x=\int_0^1\mathrm{e}^x\,\mathrm{d}x-E=\mathrm{e}-1-E,$$

正如所求。

至于第二个积分，把 x^2 除以 $1+x$：由于 $x^2=(x-1)(1+x)+1$，

$$\frac{x^2\,\mathrm{e}^x}{1+x}=x\,\mathrm{e}^x-\mathrm{e}^x+\frac{\mathrm{e}^x}{1+x}.$$

取 $u=x$、$\mathrm{d}v=\mathrm{e}^x\,\mathrm{d}x$ 作分部积分（integration by parts），

$$\int_0^1 x\,\mathrm{e}^x\,\mathrm{d}x=\left[x\,\mathrm{e}^x\right]_0^1-\int_0^1\mathrm{e}^x\,\mathrm{d}x=\mathrm{e}-(\mathrm{e}-1)=1.$$

因此

$$\int_0^1 \frac{x^2 \mathrm{e}^x}{1+x}\,\mathrm{d}x = 1-(\mathrm{e}-1)+E = 2-\mathrm{e}+E.$$

（另一做法是写成 $\frac{x^2}{1+x} = x - \frac{x}{1+x}$，然后直接利用第一个结果：$1-(\mathrm{e}-1-E)$ 得出相同答案。）

第 (i) 部分

指数是最棘手的部分，所以我们把它设为新的变量：

$$u = \frac{1-x}{1+x}.$$

解出 x：$u(1+x) = 1-x$，所以 $x = \dfrac{1-u}{1+u}$。于是

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

当 $x=0$ 时，$u=1$；当 $x=1$ 时，$u=0$。因此

$$\begin{aligned}\int_0^1 \frac{\mathrm{e}^{\frac{1-x}{1+x}}}{1+x}\,\mathrm{d}x &= \int_1^0 \mathrm{e}^u \cdot \frac{1+u}{2} \cdot \left(-\frac{2}{(1+u)^2}\right)\mathrm{d}u \\ &= \int_0^1 \frac{\mathrm{e}^u}{1+u}\,\mathrm{d}u = E.\end{aligned}$$

$\mathrm{d}x/\mathrm{d}u$ 带来的负号，用来把积分上下限调回递增次序。

> **想一想** 为甚么我们应该预期答案恰好是 E，而完全没有 e?

映射 $x \mapsto \frac{1-x}{1+x}$ 把 $[0,1]$ 映到自身（它把两个端点互换），而因子 $\frac{1}{1+x}$ 与 $\mathrm{d}x$ 结合后恰好得到 $\frac{\mathrm{d}u}{1+u}$。事实上 $\frac{\mathrm{d}x}{1+x} = -\frac{\mathrm{d}u}{1+u}$，这是记住整个计算的一个巧妙方法。

第 (ii) 部分

现在目标区间 $[0,1]$ 必须来自 $x \in [1,\sqrt{2}]$，而在这区间上 x^2 由 1 变到 2。所以我们试

$$u = x^2 - 1, \qquad \mathrm{d}u = 2x\,\mathrm{d}x,$$

当 $x=1$ 时 $u=0$，当 $x=\sqrt{2}$ 时 $u=1$。于是

$$\frac{\mathrm{d}x}{x} = \frac{x\,\mathrm{d}x}{x^2} = \frac{\mathrm{d}u}{2(1+u)}, \qquad \mathrm{e}^{x^2} = \mathrm{e}^{1+u} = \mathrm{e}\cdot\mathrm{e}^u,$$

因此

$$\int_1^{\sqrt{2}} \frac{\mathrm{e}^{x^2}}{x}\,\mathrm{d}x = \frac{\mathrm{e}}{2}\int_0^1 \frac{\mathrm{e}^u}{1+u}\,\mathrm{d}u = \frac{\mathrm{e}E}{2}.$$

想一想 为甚么不用代换 $u = x^2$?

这样也行得通，会得到 $\frac{1}{2}\int_1^2 \frac{\mathrm{e}^u}{u}\,\mathrm{d}u$，但区间是 $[1,2]$ 而不是 $[0,1]$。再平移一个单位，即 $u \mapsto u-1$，便得到相同的答案；$u = x^2 - 1$ 只不过是一步做完这两步。

验算

辛普森法则 (Simpson's rule) 给出 $E \approx 1.125386$。以数值计算，第一个积分是 $0.592896 = \mathrm{e} - 1 - E$，第二个是 $0.407104 = 2 - \mathrm{e} + E$，(i) 中的积分是 $1.125386 = E$，而 (ii) 中的积分是 $1.529558 = \frac{1}{2}\mathrm{e}E$。不用电脑也可以快速检查是否合理：头两个被积函数都是正的，而的确 $\mathrm{e} - 1 - E \approx 0.59 > 0$ 及 $2 - \mathrm{e} + E \approx 0.41 > 0$；此外第二个比第一个小，这也理所当然，因为在 $[0,1]$ 上 $x^2 \le x$。

常见错误

- 在 (i) 中忘记把上下限倒转（或漏掉负号），结果得到 $-E$；对于正的被积函数，这是不可能的答案。
- 在 (ii) 中漏掉 $\mathrm{e}^{x^2} = \mathrm{e}^{1+u}$ 带来的因子 e，或漏掉 $\mathrm{d}u = 2x\,\mathrm{d}x$ 带来的因子 $\frac{1}{2}$。

STEP 之外

像 E 这样的积分与指数积分（exponential integral）$\mathrm{Ei}(x) = \int_{-\infty}^{x} \frac{\mathrm{e}^t}{t}\,\mathrm{d}t$ 密切相关：代入 $t = 1 + x$ 可得 $E = \mathrm{e}^{-1}\big(\mathrm{Ei}(2) - \mathrm{Ei}(1)\big)$。刘维尔（Liouville）的理论证明了 $\int \frac{\mathrm{e}^x}{x}\,\mathrm{d}x$ 没有初等原函数（elementary antiderivative），这正是题目把 E 当作一个具名常数、并要求我们把一切化归为它的原因。Siklos 的 *Advanced Problems in Mathematics*（《数学进阶问题》）收录了更多同类型的题目。

详细解答

STEP I 2011，第 3 题

纯数学 · 三角恒等式

STEP I 2011，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/11-s1-q3

关键思路

在 $\frac{1}{3}\pi \mp \theta$ 处的两个正弦相乘，可以用平方差（difference of squares）的技巧干净地化简，于是一行就能证明 $(*)$。之后，余切（cotangent）的和式来自对乘积求导：三个因子乘积的导数除以该乘积本身，就是三个“对数导数”（logarithmic derivative）之和，而对正弦而言，这些对数导数正是余切。对余弦的版本做同样的事就得到正切，而 $\tan x + \cot x$ 是一个余割（cosecant）。

解答

恒等式（identity）$(*)$

记 $a = \frac{1}{3}\pi$。由复角公式（compound-angle formulae），

$$\begin{aligned}\sin(a-\theta)\sin(a+\theta) &= (\sin a\cos\theta - \cos a\sin\theta)(\sin a\cos\theta + \cos a\sin\theta)\\ &= \sin^2 a\cos^2\theta - \cos^2 a\sin^2\theta\\ &= \sin^2 a\,(1-\sin^2\theta) - (1-\sin^2 a)\sin^2\theta = \sin^2 a - \sin^2\theta.\end{aligned}$$

由于 $\sin^2 a = \frac{3}{4}$，

$$4\sin\theta\sin(\tfrac{1}{3}\pi-\theta)\sin(\tfrac{1}{3}\pi+\theta) = 4\sin\theta\left(\tfrac{3}{4}-\sin^2\theta\right) = 3\sin\theta - 4\sin^3\theta = \sin 3\theta,$$

最后一步用了三倍角公式（triple-angle formula）。这就对所有 θ 证明了 $(*)$。

第 (i) 部分

由于 $(*)$ 对所有 θ 成立，我们可以对两边求导。由乘积法则（product rule），

$$4\Big[\cos\theta\sin(a-\theta)\sin(a+\theta)-\sin\theta\cos(a-\theta)\sin(a+\theta)\\ +\sin\theta\sin(a-\theta)\cos(a+\theta)\Big]=3\cos 3\theta,$$

其中中间一项的负号来自对 $a-\theta$ 应用链式法则（chain rule）。

? **想一想** 怎样把它变成余切?

两边除以 $\sin 3\theta$，而由 $(*)$，它等于 $4\sin\theta\sin(a-\theta)\sin(a+\theta)$。左边每一项中，除了被求导的那个因子外，其余的正弦都会约去。只要 $\sin 3\theta\neq 0$，这样做便是容许的，而此时三个正弦亦全不为零。我们得到

$$\cot\theta-\cot(\tfrac{1}{3}\pi-\theta)+\cot(\tfrac{1}{3}\pi+\theta)=3\cot 3\theta\qquad(\sin 3\theta\neq 0).\tag{A}$$

现在代入 $\theta=\frac{1}{9}\pi$，此时 $\sin 3\theta=\sin\frac{1}{3}\pi\neq 0$。由于 $\frac{1}{3}\pi-\frac{1}{9}\pi=\frac{2}{9}\pi$ 及 $\frac{1}{3}\pi+\frac{1}{9}\pi=\frac{4}{9}\pi$，

$$\cot\tfrac{1}{9}\pi-\cot\tfrac{2}{9}\pi+\cot\tfrac{4}{9}\pi=3\cot\tfrac{1}{3}\pi=\frac{3}{\sqrt{3}}=\sqrt{3},$$

即为所求。

第 (ii) 部分

在 $(*)$ 中代入 $\theta=\frac{1}{6}\pi-\phi$。四个正弦变成

$$\begin{aligned}\sin\theta=\sin(\tfrac{1}{6}\pi-\phi)&=\cos(\tfrac{1}{3}\pi+\phi),\\ \sin(\tfrac{1}{3}\pi-\theta)=\sin(\tfrac{1}{6}\pi+\phi)&=\cos(\tfrac{1}{3}\pi-\phi),\\ \sin(\tfrac{1}{3}\pi+\theta)=\sin(\tfrac{1}{2}\pi-\phi)&=\cos\phi,\\ \sin 3\theta=\sin(\tfrac{1}{2}\pi-3\phi)&=\cos 3\phi,\end{aligned}$$

每次都用了 $\sin(\frac{1}{2}\pi-x)=\cos x$。于是，把 ϕ 改记为 θ，

$$4\cos\theta\cos(\tfrac{1}{3}\pi-\theta)\cos(\tfrac{1}{3}\pi+\theta)=\cos 3\theta\qquad\text{对所有}\theta.\tag{**}$$

把 $(*)$ 除以 $(**)$，因子 4 互相抵消，而每个正弦除以余弦都是正切；取倒数，

$$\cot\theta\cot(\tfrac{1}{3}\pi-\theta)\cot(\tfrac{1}{3}\pi+\theta)=\cot 3\theta,$$

此式在两边皆有定义之处成立，即 $\sin 3\theta\neq 0$ 及 $\cos 3\theta\neq 0$ 之处（每个条件都令相应的三个因子不为零）。

至于余割的和，我们如第 (i) 部分一样对 $(**)$ 求导：

$$4\Big[-\sin\theta\cos(a-\theta)\cos(a+\theta)+\cos\theta\sin(a-\theta)\cos(a+\theta)\\ -\cos\theta\cos(a-\theta)\sin(a+\theta)\Big]=-3\sin 3\theta.$$

除以 $\cos 3\theta = 4\cos\theta\cos(a-\theta)\cos(a+\theta)$（当它不为零时），再乘以 -1，

$$\tan\theta-\tan(\tfrac{1}{3}\pi-\theta)+\tan(\tfrac{1}{3}\pi+\theta)=3\tan 3\theta. \tag{B}$$

(A) 与 (B) 的正负号一致，所以可以把两式相加。

> **想一想** 为甚么把正切与余切相加会有帮助?

因为 $\tan x+\cot x$ 是单一个余割：

$$\tan x+\cot x=\frac{\sin^2 x+\cos^2 x}{\sin x\cos x}=\frac{2}{\sin 2x}=2\operatorname{cosec}2x.$$

所以只要 $\sin 6\theta\neq 0$，把 (A) 与 (B) 相加便得

$$2\operatorname{cosec}2\theta-2\operatorname{cosec}(\tfrac{2}{3}\pi-2\theta)+2\operatorname{cosec}(\tfrac{2}{3}\pi+2\theta)=6\operatorname{cosec}6\theta.$$

现在取 $\theta=\frac{1}{18}\pi$，使 $2\theta=\frac{1}{9}\pi$ 及 $6\theta=\frac{1}{3}\pi$（且 $\sin 6\theta\neq 0$）。由于 $\frac{2}{3}\pi-\frac{1}{9}\pi=\frac{5}{9}\pi$ 及 $\frac{2}{3}\pi+\frac{1}{9}\pi=\frac{7}{9}\pi$，除以 2 得

$$\operatorname{cosec}\tfrac{1}{9}\pi-\operatorname{cosec}\tfrac{5}{9}\pi+\operatorname{cosec}\tfrac{7}{9}\pi=3\operatorname{cosec}\tfrac{1}{3}\pi=3\cdot\frac{2}{\sqrt{3}}=2\sqrt{3},$$

即为所求。

验算

- 在 $\theta=\frac{1}{6}\pi$ 时，$(*)$ 变成 $4\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot 1=1=\sin\frac{1}{2}\pi$；而 $(**)$ 在 $\theta=0$ 时变成 $4\cdot 1\cdot\frac{1}{2}\cdot\frac{1}{2}=1=\cos 0$。
- 数值上，$\cot 20^\circ-\cot 40^\circ+\cot 80^\circ\approx 2.7475-1.1918+0.1763=1.7321\approx\sqrt{3}$，而 $\operatorname{cosec}20^\circ-\operatorname{cosec}100^\circ+\operatorname{cosec}140^\circ\approx 2.9238-1.0154+1.5557=3.4641\approx 2\sqrt{3}$。
- 恒等式 (A)、(B) 以及余切乘积式都可以用计算机在任意随机的 θ 测试；我们在数千个随机值上验证过。

常见错误

- 对 $\sin(\frac{1}{3}\pi-\theta)$ 求导时忘记链式法则带来的负号，以致破坏了 (A) 中正负交替的符号。
- 除以 $\sin 3\theta$ 时没有指出它在所代入的值上必须不为零。
- 试图从余切的乘积得出余割的和：真正能够合并的是和式 (A) 与 (B)。

STEP 之外

把导数除以函数本身的技巧就是对数微分法（logarithmic differentiation）：$\frac{\mathrm{d}}{\mathrm{d}\theta}\ln|f| = f'/f$ 把乘积变成和。再推进一步，sin 本身可以写成遍及其零点的无穷乘积（infinite product），而对数微分法随即给出著名的部分分式展开（partial-fraction expansion）$\pi\cot\pi x = \sum_n \frac{1}{x-n}$；恒等式 $(*)$ 正是它的一个有限的影子。这些想法在复分析（complex analysis）中有所发展；Needham 的 *Visual Complex Analysis*（《直观复分析》）是一个美妙的起点。

详细解答

STEP I 2011，第 4 题

纯数学 · 抛物线的切线

> **STEP I 2011，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/11-s1-q4

关键思路

把抛物线（parabola）上一般点的切线（tangent）写成不含分数的形式，两条切线的交点只需相减便可得出。至于角度，则用 $\overrightarrow{FT}$ 与 $\overrightarrow{FP}$ 的纯量积（scalar product）。各段长度其实比表面看来容易处理：FP 是焦半径（focal distance），而 FT^2 可以因式分解。求得 $\cos\phi$ 之后，留意答案对 p 和 q 是对称的。这种对称性就是证明角平分的全部关键。

解答

点 T

对 $y^2 = 4ax$ 作隐函数求导（implicit differentiation），得 $2y\,\dfrac{\mathrm{d}y}{\mathrm{d}x} = 4a$。因此在 P 点，即 $y = 2ap \neq 0$ 之处（因为 $p \neq 0$），斜率为 $\dfrac{4a}{4ap} = \dfrac{1}{p}$。所以 P 点的切线为

$$y - 2ap = \frac{x - ap^2}{p} \quad \Longleftrightarrow \quad py = x + ap^2,$$

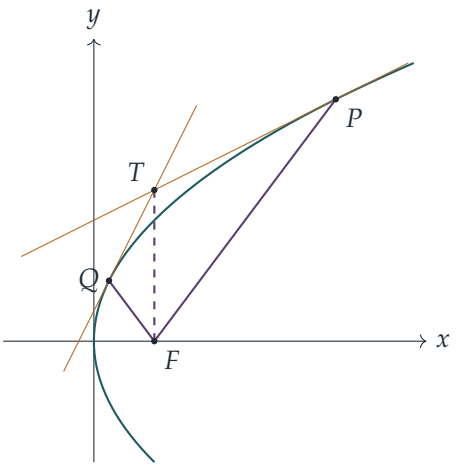
同理，Q 点的切线为 $qy = x + aq^2$。两式相减，得

$$(p - q)\,y = a(p^2 - q^2) = a(p - q)(p + q).$$

由于 P 和 Q 是相异的点，$p \neq q$，故可以除以 $p - q$，得 $y = a(p + q)$。于是

$$x = py - ap^2 = ap^2 + apq - ap^2 = apq,$$

所以 T 是点 $\big(apq,\, a(p + q)\big)$，与所求相符。



图中取 $a=1$、$p=2$、$q=\frac{1}{2}$，此时 FT 恰好是铅垂的。

角 ϕ

首先注意 $T\neq F$：否则需要 $p+q=0$ 及 $pq=1$，从而 $p^2=-1$，这是不可能的。所以角 TFP 是有意义的。设 $F=(a,0)$，则

$$\overrightarrow{FP}=a\begin{pmatrix}p^2-1\\2p\end{pmatrix},\qquad \overrightarrow{FT}=a\begin{pmatrix}pq-1\\p+q\end{pmatrix}.$$

两者的长度为

$$FP^2=a^2\left[(p^2-1)^2+4p^2\right]=a^2(p^2+1)^2,$$
$$FT^2=a^2\left[(pq-1)^2+(p+q)^2\right]=a^2\left[p^2q^2+p^2+q^2+1\right]=a^2(p^2+1)(q^2+1),$$

因为交叉项 $-2pq$ 与 $+2pq$ 互相抵消。所以 $FP=|a|(p^2+1)$，而 $FT=|a|\sqrt{(p^2+1)(q^2+1)}$。

想一想 我们怎样看出纯量积可以因式分解?

我们预期 (p^2+1) 会与 FP 约去，所以刻意寻找这个因式。展开得

$$\begin{aligned}\overrightarrow{FP}\cdot\overrightarrow{FT}&=a^2\left[(p^2-1)(pq-1)+2p(p+q)\right]\\&=a^2\left[p^3q-p^2-pq+1+2p^2+2pq\right]\\&=a^2\left[p^3q+pq+p^2+1\right]\\&=a^2(p^2+1)(pq+1).\end{aligned}$$

因此，利用 $|a|^2=a^2$，

$$\cos\phi=\frac{\overrightarrow{FP}\cdot\overrightarrow{FT}}{FP\cdot FT}=\frac{a^2(p^2+1)(pq+1)}{a^2(p^2+1)\sqrt{(p^2+1)(q^2+1)}}=\frac{pq+1}{\sqrt{(p^2+1)(q^2+1)}},$$

这正是所求的结果。

角平分

T 的坐标对 p 和 q 是对称的，所以在上面的计算中把 P 和 Q 的角色对调，便知角 TFQ 满足

$$\cos\angle TFQ = \frac{qp+1}{\sqrt{(q^2+1)(p^2+1)}} = \cos\phi.$$

两个角都在 $[0,\pi]$ 之内，而余弦在此区间上是一一对应（one-to-one）的，所以 $\angle TFP = \angle TFQ = \phi$。

因此，射线（ray）FP 和 FQ 各自与射线 FT 成 ϕ 角。在平面上，从 F 出发而与 FT 成 ϕ 角的射线只有两条，而且它们以直线 FT 为镜互为镜像。射线 FP 和 FQ 并不是同一条射线：一条穿过焦点（focus）的直线与抛物线相交于 F 两侧的两点（或只相交一次），所以从 F 出发的一条射线不可能同时经过 P 和 Q。因此 FP 和 FQ 互为对直线 FT 的反射，而这正好表示直线 FT 平分角 PFQ。

有一个细节可以解释为何题目说的是直线 FT。若 $pq+1>0$，则 $\phi<\frac{\pi}{2}$，射线 FT 本身位于角 $PFQ=2\phi$ 之内。若 $pq+1<0$，则 $\phi>\frac{\pi}{2}$，角 PFQ 为 $2\pi-2\phi$，平分它的是 TF 越过 F 的延长线。若 $pq=-1$，则 PQ 是焦弦（focal chord），$\phi=\frac{\pi}{2}$，而 FT 与它垂直，平分这个平角。在所有情况下，直线 FT 都平分角 PFQ。

验算

取图中的 $a=1$、$p=2$、$q=\frac{1}{2}$：$T=(1,\frac{5}{2})$，$\overrightarrow{FP}=(3,4)$，$\overrightarrow{FQ}=(-\frac{3}{4},1)$，$\overrightarrow{FT}=(0,\frac{5}{2})$。于是 $\cos\angle TFP=\frac{4}{5}$，而 $\cos\angle TFQ=\frac{1}{5/4}=\frac{4}{5}$，与 $\dfrac{2}{\sqrt{5\cdot\frac{5}{4}}}=\dfrac{4}{5}$ 吻合。用电脑测试 10^5 组随机选取的 a（正负皆有）、p 和 q，在 10^{-10} 的误差范围内证实了 T 的坐标、$\cos\phi$ 的公式以及角平分的结论。测试包括许多 $pq<-1$ 的情况，此时是 FT 向后的延长线起平分作用。

常见错误

- 除以 $p-q$ 时没有说明它为何不等于零（因为两点相异）。
- 写成 $FP=a(p^2+1)$，这在 $a<0$ 时是错的。$\cos\phi$ 中只出现 a^2，所以最终 a 的正负并无影响。
- 只说“余弦相等，所以 FT 平分该角”，却没有解释为何 FP 和 FQ 位于直线 FT 的两侧。

STEP 之外

这其实是抛物线的反射性质（reflective property）换了个样子出现：从任何外部点引出的两条切线段，在焦点处所张的角相等。椭圆（ellipse）和双曲线（hyperbola）也有同样的性质，三者都源于圆锥曲线（conic）的焦点－准线（focus–directrix）定义，大学一年级的几何及射影几何（projective geometry）课程会对此作更深入的发展。如想做更多同类风格的题目，Siklos 的 *Advanced Problems in Mathematics*（《高等数学难题》）可免费取得，非常值得逐题钻研。

详细解答

STEP I 2011，第 5 题

纯数学 · 含绝对值的积分

STEP I 2011，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/11-s1-q5

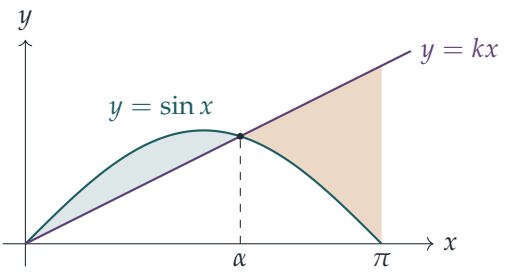
关键思路

只要知道 $\sin x - kx$ 在哪里变号，绝对值符号（modulus）就不再麻烦，而草图正告诉我们：只变号一次，就在 $x = \alpha$ 处。在该处把积分拆开，绝对值符号便会消失。最后一步是不再把 k 看作变量：由于 $k = \sin\alpha/\alpha$，所有东西都可以用 α 表示，而 I 的导数（derivative）便能整齐地因式分解。

解答

唯一的根

设 $g(x) = \sin x - kx$。直线 $y = kx$ 经过原点，斜率为 $k < 1$；而 $y = \sin x$ 离开原点时的斜率为 $\cos 0 = 1$。因此在 0 的右边附近，正弦曲线位于直线之上。在 $x = \pi$ 处，正弦曲线回到 0，但直线的高度为 $k\pi > 0$，所以在那里直线位于上方。由于 $y = \sin x$ 在 $(0, \pi)$ 上向下弯曲，而直线是直的，两者在其间恰好相交一次。



要令“恰好一次”无懈可击：$g(0) = 0$、$g'(x) = \cos x - k$，而且在 $(0, \pi)$ 上 $g''(x) = -\sin x < 0$。所以 g' 严格递减（strictly decreasing），由 0 处的 $1 - k > 0$ 减至 π 处的

$-1-k<0$，并恰好在一点 β 等于零。因此 g 在 $(0,\beta]$ 上严格递增，故在该处为正（因为它由 0 开始），并在 $[\beta,\pi)$ 上严格递减，由正值减至 $g(\pi)=-k\pi<0$。所以它在 $(0,\pi)$ 中恰有一个零点 α，而且

$$\sin x > kx \text{ 当} 0<x<\alpha, \qquad \sin x < kx \text{ 当} \alpha<x<\pi.$$

计算 I

利用这些正负号（即草图中的蓝绿色和琥珀色区域），

$$I=\int_0^{\alpha}(\sin x-kx)\,\mathrm{d}x-\int_{\alpha}^{\pi}(\sin x-kx)\,\mathrm{d}x.$$

$\sin x-kx$ 的一个原函数（antiderivative）是 $F(x)=-\cos x-\frac{1}{2}kx^2$，所以

$$\begin{aligned} I&=\big(F(\alpha)-F(0)\big)-\big(F(\pi)-F(\alpha)\big)=2F(\alpha)-F(0)-F(\pi)\\ &=\big(-2\cos\alpha-k\alpha^2\big)+1-\big(1-\tfrac{1}{2}k\pi^2\big)\\ &=\tfrac{1}{2}k\pi^2-2\cos\alpha-k\alpha^2. \end{aligned}$$

由于 α 满足 $(*)$，$k=\dfrac{\sin\alpha}{\alpha}$，所以 $k\alpha^2=\alpha\sin\alpha$，而且

$$I=\frac{\pi^2\sin\alpha}{2\alpha}-2\cos\alpha-\alpha\sin\alpha,$$

即为所求。

驻值（stationary value）

想一想 α 可以取甚么值?

每个 $k\in(0,1)$ 都给出一个 $\alpha\in(0,\pi)$。反过来，每个 $\alpha\in(0,\pi)$ 都来自 $k=\sin\alpha/\alpha$，而它位于 $(0,1)$ 之中，因为在该范围内 $0<\sin\alpha<\alpha$。因此我们必须研究 $I(\alpha)$ 在开区间 $(0,\pi)$ 内所有 α 的情况。

对第一项使用商法则（quotient rule）求导，得

$$\begin{aligned} \frac{\mathrm{d}I}{\mathrm{d}\alpha}&=\frac{\pi^2(\alpha\cos\alpha-\sin\alpha)}{2\alpha^2}+2\sin\alpha-(\sin\alpha+\alpha\cos\alpha)\\ &=\frac{\pi^2(\alpha\cos\alpha-\sin\alpha)}{2\alpha^2}-(\alpha\cos\alpha-\sin\alpha)\\ &=(\alpha\cos\alpha-\sin\alpha)\left(\frac{\pi^2}{2\alpha^2}-1\right). \end{aligned}$$

第一个因子在 $(0,\pi)$ 上永不为零。设 $h(\alpha)=\alpha\cos\alpha-\sin\alpha$。那么 $h(0)=0$，而且在 $(0,\pi)$ 上 $h'(\alpha)=\cos\alpha-\alpha\sin\alpha-\cos\alpha=-\alpha\sin\alpha<0$，所以在整个区间上 $h(\alpha)<0$。（等价地，在 $(0,\frac{1}{2}\pi)$ 上 $\tan\alpha>\alpha$，而在此之后 $\cos\alpha\le 0$。）

所以 $\mathrm{d}I/\mathrm{d}\alpha = 0$ 当且仅当 $\pi^2 = 2\alpha^2$，即在 $\alpha = \pi/\sqrt{2}$ 处，而它确实位于 $(0, \pi)$ 之中。这是唯一的驻值。由于 $h < 0$，$\mathrm{d}I/\mathrm{d}\alpha$ 的正负号与 $\frac{\pi^2}{2\alpha^2} - 1$ 的正负号相反：

- 当 $0 < \alpha < \pi/\sqrt{2}$ 时，$\frac{\pi^2}{2\alpha^2} > 1$，所以 $\mathrm{d}I/\mathrm{d}\alpha < 0$；
- 当 $\pi/\sqrt{2} < \alpha < \pi$ 时，$\frac{\pi^2}{2\alpha^2} < 1$，所以 $\mathrm{d}I/\mathrm{d}\alpha > 0$。

所以 I 先减后增：该驻值是极小值，而由于它是区间上唯一的驻点，它就是 I 的最小值。

当 $\alpha = \pi/\sqrt{2}$ 时，我们有 $\dfrac{\pi^2}{2\alpha} = \dfrac{\pi^2}{\sqrt{2}\,\pi} = \dfrac{\pi}{\sqrt{2}} = \alpha$，所以 I 的第一项与第三项互相抵消：

$$I_{\min} = \alpha \sin\alpha - 2\cos\alpha - \alpha\sin\alpha = -2\cos\frac{\pi}{\sqrt{2}}.$$

这个值是正数，理应如此，因为 $\pi/\sqrt{2} \approx 2.22$ 位于 $\frac{1}{2}\pi$ 与 π 之间，而余弦在该范围内为负。

验算

- 当 $k \to 0$ 时，$\alpha \to \pi$，公式给出 $I \to 2$，这正是 $\int_0^\pi \sin x\,\mathrm{d}x$。当 $k \to 1$ 时，$\alpha \to 0$，而 $I \to \frac{1}{2}\pi^2 - 2$，这正是 $\int_0^\pi (x - \sin x)\,\mathrm{d}x$。
- 数值上，$-2\cos(\pi/\sqrt{2}) \approx 1.2114$，比两个端值 2 和 $\frac{1}{2}\pi^2 - 2 \approx 2.935$ 都小。直接以中点法数值积分（midpoint quadrature）计算 $\int_0^\pi |\sin x - kx|\,\mathrm{d}x$，取 $k = \sin\alpha/\alpha \approx 0.3582$，同样得到 1.2114。

常见错误

- 对 α 求导时把 k 当作常数：k 依赖于 α，这正是我们要先代入 $k = \sin\alpha/\alpha$ 的原因。
- 只证明了 $\mathrm{d}I/\mathrm{d}\alpha = 0$ 仅在 $\alpha = \pi/\sqrt{2}$ 处成立，却忘记证明 $\alpha\cos\alpha - \sin\alpha$ 在 $(0, \pi)$ 中没有零点。

STEP 之外

唯一根的论证用到介值定理（intermediate value theorem）以及严格单调性，而求最小值时则默默用到一个事实：$\alpha \mapsto k$ 是 $(0, \pi)$ 与 $(0, 1)$ 之间的双射（bijection），因此对 α 求最小值等同于对 k 求最小值。要把这类重新参数化（reparametrisation）和存在性论证变得严谨，正是实分析（real analysis）入门课程的工作；Abbott 的 *Understanding Analysis*（《理解分析》）是极佳的起点。

详细解答

STEP I 2011，第 6 题

纯数学 · 二项式级数求和

> **STEP I 2011，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/11-s1-q6

关键思路

只要知道 $(1-x)^{-3}$ 的系数（coefficients），再乘上一个多项式（polynomial），其实只是把几个相邻的系数混合起来，所以较复杂的有理函数（rational function）的系数几乎可以不费吹灰之力得出。把这个过程倒过来看，一个各项为 (多项式，自变量为r) $\times\, x^r$ 的级数（series），只要看出是哪一个有理函数生成它，再代入适当的 x 值（这里是 $x=\frac{1}{2}$），便可求和。

解答

开首的结果

当 $|x|<1$ 时，指数（index）为 -3 的二项式展开（binomial expansion）给出

$$(1-x)^{-3}=\sum_{r=0}^{\infty}\frac{(-3)(-4)\cdots(-3-r+1)}{r!}(-x)^r.$$

分子共有 r 个因子，每个都是负数，所以它们的积是 $(-1)^r\,\dfrac{(r+2)!}{2!}$，而这个 $(-1)^r$ 与 $(-x)^r$ 中的 $(-1)^r$ 互相抵消。因此 x^r 的系数是

$$a_r=\frac{(r+2)!}{2!\,r!}=\tfrac{1}{2}(r+1)(r+2),$$

正如所求。（所以 $a_0=1$、$a_1=3$、$a_2=6$、$a_3=10$、...）

第 (i) 部分

写成

$$\frac{1-x+2x^2}{(1-x)^3} = (1-x+2x^2)\sum_{k\geq 0} a_k x^k.$$

x^r 项来自 $1\cdot a_r x^r$、$-x\cdot a_{r-1}x^{r-1}$ 及 $2x^2\cdot a_{r-2}x^{r-2}$，所以当 $r\geq 2$ 时，x^r 的系数是

$$\begin{aligned} a_r - a_{r-1} + 2a_{r-2} &= \tfrac{1}{2}(r+1)(r+2) - \tfrac{1}{2}r(r+1) + (r-1)r \\ &= (r+1) + r^2 - r \\ &= r^2 + 1. \end{aligned}$$

想一想 当 $r=0$ 及 $r=1$ 时，a_{r-1} 或 a_{r-2} 并不存在，这条公式是否仍然成立?

成立。若设 $a_{-1} = a_{-2} = 0$，公式 $\frac{1}{2}(r+1)(r+2)$ 在 $r=-1$ 及 $r=-2$ 时亦给出 0，所以上面的计算对所有 $r\geq 0$ 都成立。直接验证：常数项是 $a_0 = 1 = 0^2+1$，而 x 的系数是 $a_1 - a_0 = 2 = 1^2+1$。所以对所有 $r\geq 0$，x^r 的系数都是 r^2+1，正如所求。

至于该级数，分子 $1,2,5,10,17,26,37,50,\ldots$ 是 $r=0,1,2,\ldots$ 时的 r^2+1，而分母是 2^r。所以该级数是

$$\sum_{r=0}^{\infty}(r^2+1)\left(\tfrac{1}{2}\right)^r,$$

即上述展开式在 $x=\frac{1}{2}$ 时的值。这样做是合理的，因为 $(1-x)^{-3}$ 的二项式展开在 $|x|<1$ 时收敛（converges）。因此

$$1+\frac{2}{2}+\frac{5}{4}+\frac{10}{8}+\cdots = \frac{1-\frac{1}{2}+2\cdot\frac{1}{4}}{\left(\frac{1}{2}\right)^3} = \frac{1}{1/8} = 8.$$

第 (ii) 部分

首要任务是找出规律。把各项改写成分母为 $1,2,4,8,16,32,64$：

$$1,\ \frac{4}{2},\ \frac{9}{4},\ \frac{16}{8},\ \frac{25}{16},\ \frac{36}{32},\ \frac{49}{64},\ \cdots$$

所以该级数是 $\displaystyle\sum_{r=0}^{\infty}\frac{(r+1)^2}{2^r}$。

想一想 $(1-x)^3$ 上面放甚么分子，才会得出系数 $(r+1)^2$?

第 (i) 部分提示我们试用分子 $1+x$（同样是一个简短的多项式）。$(1+x)(1-x)^{-3}$ 中 x^r 的系数是

$$a_r + a_{r-1} = \tfrac{1}{2}(r+1)(r+2) + \tfrac{1}{2}r(r+1) = \tfrac{1}{2}(r+1)(2r+2) = (r+1)^2,$$

对所有 $r \geq 0$ 均成立（与之前一样取 $a_{-1} = 0$）。代入 $x = \frac{1}{2}$，

$$1 + 2 + \frac{9}{4} + 2 + \frac{25}{16} + \cdots = \frac{1 + \frac{1}{2}}{\left(\frac{1}{2}\right)^3} = \frac{3}{2} \cdot 8 = 12.$$

*另一方法：利用第 (i) 部分。*以 $s = r + 1$ 平移下标，使得 $2^{-r} = 2 \cdot 2^{-s}$：

$$\sum_{r \geq 0} \frac{(r+1)^2}{2^r} = 2 \sum_{s \geq 1} \frac{s^2}{2^s} = 2 \left(\sum_{s \geq 0} \frac{s^2 + 1}{2^s} - \sum_{s \geq 0} \frac{1}{2^s} \right) = 2(8 - 2) = 12,$$

这里用了第 (i) 部分的答案及几何级数（geometric series）$\sum 2^{-s} = 2$。

验算

- 以人手把 $(1 - x)^{-3}$ 展开为 $(1 + x + x^2 + \cdots)^3$，或把头几项乘开，得到 $1 + 3x + 6x^2 + 10x^3 + \cdots$，与 $\frac{1}{2}(r+1)(r+2)$ 吻合。
- 第 (i) 部分级数的部分和（partial sums）趋近 8：取 200 项后，和与 8 在机器精度内一致。第 (ii) 部分同样如此，极限为 12。
- 第 (ii) 部分的两个方法得出相同答案 12。

常见错误

- 二项式系数的正负号弄错：$(-3)(-4)\cdots$ 中有 r 个负号，$(-x)^r$ 中亦有 r 个，两者互相抵消。
- 在第 (ii) 部分，因为部分分数已约至最简而看错规律：$2 = \frac{4}{2} = \frac{16}{8}$，以及 $\frac{9}{8} = \frac{36}{32}$。
- 代入 $x = \frac{1}{2}$ 时没有指出展开式在该处收敛（$|x| < 1$）。

STEP 之外

我们一直在做的，其实是运用*生成函数*（*generating functions*）：把数列 $c_0, c_1, c_2, \ldots$ 编码成幂级数（power series）$\sum c_r x^r$，于是对数列的运算便变成对函数的代数运算。事实上，任何“r 的多项式乘以 x^r”之和，都是一个分母为 $(1 - x)$ 的幂的有理函数，这就是为何这类和总是得出有理数。要证明代入某个 x 值是合理的，需要依靠幂级数及其收敛半径（radius of convergence）的理论，可参阅 Abbott 的 *Understanding Analysis*（《理解分析》）；至于把生成函数作为解题工具，可参阅 Siklos 的 *Advanced Problems in Mathematics*（《数学进阶难题》）。

详细解答

STEP I 2011，第 7 题

纯数学 · 可分离变量微分方程

STEP I 2011，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/11-s1-q7

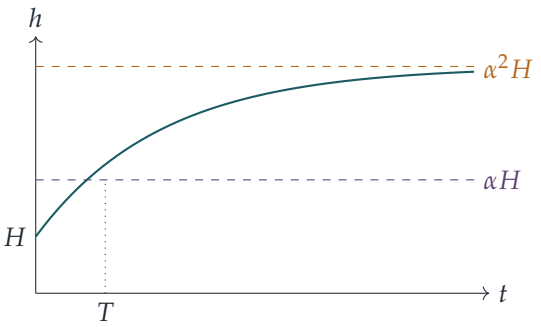
关键思路

体积等于（固定的截面面积）× 高度，所以高度的变化率就是（流入率 − 流出率）除以面积。题目中“高度保持不变”一句是确定流入率的线索：在那个高度，流入与流出恰好平衡。之后每一部分都是一条可分离变量（separable）的微分方程（differential equation）；而在第 (ii) 部分，代换（substitution）$u = \sqrt{h}$ 会把积分化为常规的计算。

解答

设固定的截面面积（cross-sectional area）为 A，固定的流入率为 q（单位时间内的体积）。水的体积为 Ah，所以

$$A\frac{\mathrm{d}h}{\mathrm{d}t} = q - (\text{流出率}).$$



高度从 H 开始上升，趋向流入与流出平衡的水平 $\alpha^2 H$，愈来愈接近却永远不会到达；问题是要多久才升到 αH，由于 $1 < \alpha < \alpha^2$，这个高度位于两者之间。（图中所示为第 (i)

部分的情况。）

第 (i) 部分

流出率为 λh，其中 $\lambda > 0$ 为某常数，所以

$$A\frac{\mathrm{d}h}{\mathrm{d}t} = q - \lambda h.$$

当 $h = \alpha^2 H$ 时高度保持不变，所以在该处 $\mathrm{d}h/\mathrm{d}t = 0$，得 $q = \lambda\alpha^2 H$。代入后得

$$\frac{\mathrm{d}h}{\mathrm{d}t} = \frac{\lambda}{A}\left(\alpha^2 H - h\right) = k\left(\alpha^2 H - h\right), \qquad k = \frac{\lambda}{A} > 0,$$

此即所求。

由于在整个上升过程中 $H \le h < \alpha^2 H$，所以 $\alpha^2 H - h > 0$，我们可以分离变量：

$$\begin{aligned} kT &= \int_H^{\alpha H} \frac{\mathrm{d}h}{\alpha^2 H - h} = \Big[-\ln\left(\alpha^2 H - h\right) \Big]_H^{\alpha H} = \ln\frac{\alpha^2 H - H}{\alpha^2 H - \alpha H} \\ &= \ln\frac{(\alpha-1)(\alpha+1)}{\alpha(\alpha-1)} = \ln\frac{\alpha+1}{\alpha} = \ln\left(1 + \frac{1}{\alpha}\right), \end{aligned}$$

其中因为 $\alpha > 1$，所以可约去 $\alpha - 1 \neq 0$。这就是所求的结果。

当 α 很大时，$x = 1/\alpha$ 很小，所以

$$kT \approx \frac{1}{\alpha} - \frac{1}{2\alpha^2} \approx \frac{1}{\alpha},$$

因为当 α 很大时，第二项相对于第一项可以忽略。

第 (ii) 部分

现在流出率为 $\mu\sqrt{h}$，其中 $\mu > 0$ 为某常数，所以 $A\,\mathrm{d}h/\mathrm{d}t = q - \mu\sqrt{h}$。高度在 $h = \alpha^2 H$ 时保持不变，所以 $q = \mu\sqrt{\alpha^2 H} = \mu\alpha\sqrt{H}$（利用了 $\alpha > 0$）。因此

$$\frac{\mathrm{d}h}{\mathrm{d}t} = c\left(\alpha\sqrt{H} - \sqrt{h}\right), \qquad c = \frac{\mu}{A} > 0.$$

同样，当 $h < \alpha^2 H$ 时括号内的值为正，所以

$$cT' = \int_H^{\alpha H} \frac{\mathrm{d}h}{\alpha\sqrt{H} - \sqrt{h}}.$$

想一想 分母中的 $\sqrt{h}$ 该如何处理?

代换 $u = \sqrt{h}$，于是 $h = u^2$ 及 $\mathrm{d}h = 2u\,\mathrm{d}u$；积分上下限变为 $u = \sqrt{H}$ 和 $u = \sqrt{\alpha H}$。为简便起见记 $a = \alpha\sqrt{H}$，被积函数便是一个有理函数（rational function），作除法后得

$$\frac{2u}{a-u} = -2 + \frac{2a}{a-u}.$$

因此

$$\begin{aligned} cT' &= \int_{\sqrt{H}}^{\sqrt{\alpha H}} \left(-2 + \frac{2a}{a-u}\right) \mathrm{d}u = \Big[-2u - 2a\ln(a-u)\Big]_{\sqrt{H}}^{\sqrt{\alpha H}} \\ &= -2\left(\sqrt{\alpha H} - \sqrt{H}\right) + 2a\sqrt{H}\,\ln\frac{\alpha\sqrt{H} - \sqrt{H}}{\alpha\sqrt{H} - \sqrt{\alpha}\sqrt{H}} \\ &= 2\sqrt{H}\left(1 - \sqrt{\alpha} + \alpha\ln\frac{\alpha-1}{\alpha - \sqrt{\alpha}}\right). \end{aligned}$$

由于 $u \le \sqrt{\alpha H} < \alpha\sqrt{H} = a$（因为当 $\alpha > 1$ 时 $\sqrt{\alpha} < \alpha$），对数内的量都是正数。最后，因式分解 $\alpha - 1 = (\sqrt{\alpha}-1)(\sqrt{\alpha}+1)$ 及 $\alpha - \sqrt{\alpha} = \sqrt{\alpha}(\sqrt{\alpha}-1)$，得

$$\frac{\alpha-1}{\alpha-\sqrt{\alpha}} = \frac{\sqrt{\alpha}+1}{\sqrt{\alpha}} = 1 + \frac{1}{\sqrt{\alpha}},$$

所以

$$cT' = 2\sqrt{H}\left(1 - \sqrt{\alpha} + \alpha\ln\left(1 + \frac{1}{\sqrt{\alpha}}\right)\right),$$

此即所求。

当 α 很大时，$1/\sqrt{\alpha}$ 很小，所以

$$\alpha\ln\left(1 + \frac{1}{\sqrt{\alpha}}\right) \approx \alpha\left(\frac{1}{\sqrt{\alpha}} - \frac{1}{2\alpha}\right) = \sqrt{\alpha} - \frac{1}{2}.$$

两个大项 $\pm\sqrt{\alpha}$ 互相抵消，剩下

$$cT' \approx 2\sqrt{H}\left(1 - \sqrt{\alpha} + \sqrt{\alpha} - \tfrac{1}{2}\right) = \sqrt{H}.$$

留意这里我们需要用到对数展开中的 x^2 项：若只用 $\ln(1+x) \approx x$，括号内的值会是 1，于是得到 $2\sqrt{H}$，这是错的。这正是题目给出两项近似式的原因。

验算

- 当 $\alpha \to 1^+$ 时，(i) 给出 $kT \to \ln 2$，而不是 0。这是对的：与平衡点（equilibrium）的差距起初为 $(\alpha^2-1)H \approx 2(\alpha-1)H$，而须缩小至 $(\alpha^2-\alpha)H \approx (\alpha-1)H$，即差距减半；而以速率 k 进行的指数衰减（exponential decay）减半所需时间为 $(\ln 2)/k$。在 (ii) 中，$cT' \to 2\sqrt{H}\ln 2$，这与在 $h = H$ 附近把 $\sqrt{h}$ 线性化（linearising）所得的等效速率 $c/(2\sqrt{H})$ 下同样的减半相吻合。
- 取 $\alpha = 4$ 及 $H = 1$：$kT = \ln 1.25 \approx 0.2231$，$cT' = 2(-1 + 4\ln 1.5) \approx 1.2437$。对多组随机的 H, α, k, c 用数值方法（龙格—库塔法（Runge–Kutta））逐步求解两条微分方程，结果与两条公式吻合至 10^{-9}。
- 取 $\alpha = 10^4$、$H = 2$：$\alpha\ln(1 + 1/\alpha) \approx 0.99995$，$cT'/\sqrt{H} \approx 1.0066$，与极限值一致（下一项给出 $cT' \approx \sqrt{H}\,(1 + \frac{2}{3\sqrt{\alpha}})$）。

常见错误

- 忘记利用平衡条件消去流入率，结果剩下太多常数。
- 在第 (ii) 部分只用 $\ln(1+x) \approx x$，因而丢失了关键的 $-\frac{1}{2}$。
- 把 $\sqrt{\alpha^2 H}$ 写成 $\alpha^2\sqrt{H}$。

STEP 之外

两个模型都是一阶自治微分方程（first-order autonomous differential equation）$\dot{h} = f(h)$，并在 $h = \alpha^2 H$ 处有一个稳定平衡点；线性的情况其实就是换了包装的牛顿冷却定律（Newton's law of cooling），而 $\sqrt{h}$ 的情况则是描述液体从小孔流出的托里拆利定律（Torricelli's law）。不必明确求解而去研究平衡点及其稳定性，正是常微分方程（ODE）定性理论（qualitative theory）的起点。Boyce 与 DiPrima 合著的 *Elementary Differential Equations*（《初等微分方程》）在前几章便讨论了自治方程和水箱混合模型。

详细解答

STEP I 2011，第 8 题

纯数学 · 丢番图方程

> **STEP I 2011，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/11-s1-q8

关键思路

立方函数是严格递增（strictly increasing）的，所以比较 m^3 与 n^3 及 $(n+1)^3$，等同于比较 m 与 n 及 $n+1$。若 m 严格位于两个连续整数（consecutive integers）之间，它就不可能是整数。这样便只剩下少数几个 n 值需要逐一验证。第 (ii) 部分则把同样的夹逼（squeeze）方法用于一条新的方程。

解答

全题均利用 $x \mapsto x^3$ 在整个 $\mathbb{R}$ 上严格递增这一事实，故对实数 a, b，

$$a < b \iff a^3 < b^3.$$

第 (i)(a) 部分

由 $(*)$ 得

$$m^3 - n^3 = n^2 + 1 > 0,$$

因为 $n^2 \geq 0$。所以 $m^3 > n^3$，因此 $m > n$。

其次，$m < n+1$ 当且仅当 $m^3 < (n+1)^3$。利用 $(*)$ 并展开，

$$\begin{aligned} m^3 < (n+1)^3 &\iff n^3 + n^2 + 1 < n^3 + 3n^2 + 3n + 1 \\ &\iff 0 < 2n^2 + 3n, \end{aligned}$$

这正是所要求的等价关系。每一步都是可逆的，所以“当且仅当”成立。

现在 $2n^2 + 3n = n(2n+3)$。这是一个二次式（quadratic），其根为 $n = -\frac{3}{2}$ 及 $n = 0$，且首项系数（leading coefficient）为正，所以它恰好在 $n < -\frac{3}{2}$ 或 $n > 0$ 时为正。对任何

这样的 n，我们同时有 $m > n$ 及 $m < n+1$。因此

$$n < m < n+1 \quad \text{除非} \quad -\tfrac{3}{2} \le n \le 0.$$

第 (i)(b) 部分

设 m 和 n 是满足 $(*)$ 的整数。若 $n < -\frac{3}{2}$ 或 $n > 0$，则由 (a) 部分得 $n < m < n+1$。连续整数 n 与 $n+1$ 之间严格来说没有任何整数，所以这是不可能的。

? **想一想** (a) 部分对所有实数 n 都成立。那么“是整数”这个条件在哪里用上?

它用上了两次。一是如上所述，排除区间以外的所有 n；二是把区间 $-\frac{3}{2} \le n \le 0$ 缩减为 $n \in \{-1, 0\}$。我们逐一验证这两个值：

- $n = 0$：$m^3 = 0 + 0 + 1 = 1$，所以 $m = 1$。
- $n = -1$：$m^3 = -1 + 1 + 1 = 1$，所以 $m = 1$。

两者都给出整数 m。所以唯一的整数解是 $(m, n) = (1, 0)$ 及 $(1, -1)$，得证。

第 (ii) 部分

我们仿照第 (i) 部分的做法。首先比较 p 与 q：

$$p^3 - q^3 = 2q^2 - 1,$$

只要 $q^2 > \frac{1}{2}$，即对每个整数 $q \neq 0$，此式皆为正。对这些 q，我们有 $p > q$。

其次比较 p 与 $q+1$：

$$\begin{aligned} p^3 < (q+1)^3 &\iff q^3 + 2q^2 - 1 < q^3 + 3q^2 + 3q + 1 \\ &\iff 0 < q^2 + 3q + 2 = (q+1)(q+2). \end{aligned}$$

此式恰好在 $q > -1$ 或 $q < -2$ 时成立。

把两者结合起来。若 q 是整数，且 $q \neq 0$、$q \neq -1$ 及 $q \neq -2$，则 $q < p < q+1$，所以 p 不可能是整数。（若 $q \ge 1$ 或 $q \le -3$，两个条件均成立。）剩下 $q \in \{-2, -1, 0\}$：

- $q = 0$：$p^3 = -1$，所以 $p = -1$。
- $q = -1$：$p^3 = -1 + 2 - 1 = 0$，所以 $p = 0$。
- $q = -2$：$p^3 = -8 + 8 - 1 = -1$，所以 $p = -1$。

三者都给出整数 p，所以整数解为

$$(p, q) = (-1, 0), \quad (0, -1), \quad (-1, -2).$$

验算

把每个答案代回原式。例如，$(-1)^3 = -1 = (-2)^3 + 2(-2)^2 - 1 = -8 + 8 - 1$。用电脑搜索所有 $|n|, |q| < 10^5$ 的整数，恰好找到 $(*)$ 的两个解及第 (ii) 部分的三个解，再无其他。你也可以用数值方法测试夹逼的结果。当 $n = 5$ 时，$m = \sqrt[3]{151} \approx 5.32$，严格位于 5 与 6 之间。

常见错误

- 在第 (ii) 部分遗漏 $q = 0$。此时 $p^3 - q^3 = -1 < 0$，所以下界 $p > q$ 不成立，这个情况需要另行验证。
- 写“$m^3 > n^3$，所以 $m > n$”而不给理由。这是成立的，因为立方运算是严格递增的（而平方运算对负数则不然），这一点值得写明。
- 只写到 $n < m < n + 1$ 便停下，没有说明为何这意味着 m 不是整数。

STEP 之外

像 $y^3 = f(x)$ 这类求整数点的方程称为*丢番图方程*（Diophantine equations）。把未知数夹在相邻的幂之间，是处理这类方程的标准 first 步。较难的版本，例如 $y^2 = x^3 + k$，会引出椭圆曲线（elliptic curves）以及 Baker 的对数线性型（linear forms in logarithms）理论。这些都是深奥的领域，但仍有不少情况可以用这种初等的估界方法解决。Davenport 的 *The Higher Arithmetic*（《高等算术》）是初等数论（number theory）的一本好入门书。

详细解答

STEP I 2011，第 9 题

力学 • 抛体运动

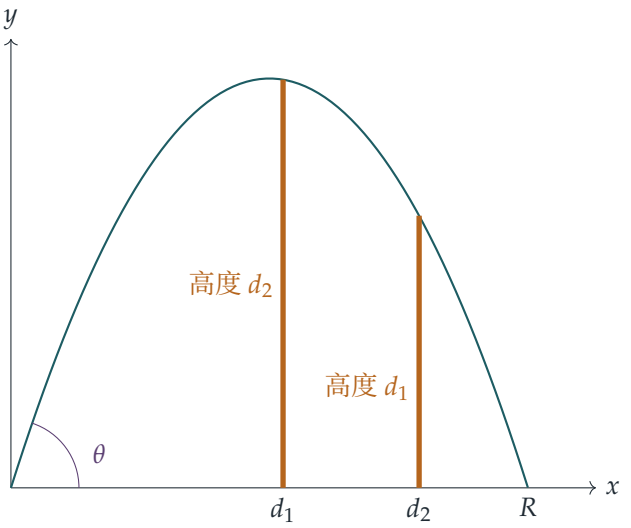
> **STEP I 2011，第 9 题**
>
> 继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：
>
> step.maths.org/questions/11-s1-q9

关键思路

把轨迹（trajectory）写成单一方程 $y = x\tan\theta - ax^2$，其中常数 a 把 g、速率和角度合并在一起。“刚好越过”一幅墙，意思是路径经过墙顶的角位，所以每幅墙都给出一条关于两个未知数 $\tan\theta$ 和 a 的线性方程（linear equation）。解这对方程便得出角度，而射程（range）就只是 $y = 0$ 的另一个根（root）。

解答

我们把质点模型化为只受重力作用而运动的一点（不计空气阻力（air resistance）），从原点 O 以速率 u、仰角 θ 抛出，x 为水平方向，y 为竖直向上方向。两幅墙的高度不同，因此我们假设 $d_1 \neq d_2$。



（图中采用 $d_1 < d_2$；但代数推导并不依赖哪一个较大。）

轨迹

在时间 t 时，$x = ut\cos\theta$ 及 $y = ut\sin\theta - \frac{1}{2}gt^2$。消去 $t = x/(u\cos\theta)$，得

$$y = x\tan\theta - ax^2, \qquad \text{其中} a = \frac{g}{2u^2\cos^2\theta} > 0.$$

设 $T = \tan\theta$。质点刚好越过每幅墙，所以它经过墙顶 (d_1, d_2) 和 (d_2, d_1)：

$$d_2 = d_1 T - ad_1^2, \tag{1}$$
$$d_1 = d_2 T - ad_2^2. \tag{2}$$

求 $\tan\theta$

为了消去 a，把 (1) 乘以 d_2^2，把 (2) 乘以 d_1^2，使两条式中的 a 项都是 $ad_1^2d_2^2$：

$$d_2^3 = d_1d_2^2\,T - ad_1^2d_2^2,$$
$$d_1^3 = d_1^2d_2\,T - ad_1^2d_2^2.$$

相减，得

$$d_2^3 - d_1^3 = d_1d_2(d_2 - d_1)\,T.$$

? **想一想** 我们可以除以 $d_2 - d_1$ 吗?

可以：两幅墙的高度 d_2 和 d_1 不同，所以 $d_1 \neq d_2$。（若两者相等，就只有一幅墙，而单一个点无法同时确定 θ 和 u。）由于 $d_2^3 - d_1^3 = (d_2 - d_1)(d_2^2 + d_1d_2 + d_1^2)$，除以 $d_1d_2(d_2 - d_1)$

便得

$$\tan\theta = \frac{d_1^2 + d_1 d_2 + d_2^2}{d_1 d_2},$$

此即所求。

射程

质点落地之处满足 $y = 0$ 且 $x \neq 0$，即 $x = T/a$。因此我们也需要求出 a。把 (1) 除以 d_1，(2) 除以 d_2，得

$$\frac{d_2}{d_1} = T - ad_1, \qquad \frac{d_1}{d_2} = T - ad_2.$$

以第二式减第一式，消去 T：

$$a(d_2 - d_1) = \frac{d_2}{d_1} - \frac{d_1}{d_2} = \frac{(d_2 - d_1)(d_2 + d_1)}{d_1 d_2}, \qquad \text{所以} \qquad a = \frac{d_1 + d_2}{d_1 d_2},$$

这里同样用到 $d_1 \neq d_2$。因此射程为

$$R = \frac{T}{a} = \frac{d_1^2 + d_1 d_2 + d_2^2}{d_1 d_2} \cdot \frac{d_1 d_2}{d_1 + d_2} = \frac{d_1^2 + d_1 d_2 + d_2^2}{d_1 + d_2}.$$

作为合理性检查，$R - d_2 = \dfrac{d_1^2}{d_1 + d_2} > 0$，同理 $R - d_1 = \dfrac{d_2^2}{d_1 + d_2} > 0$，所以质点落在两幅墙之外，这是必然的。

> ? **想一想** g 和 u 到哪里去了？

它们始终只通过单一组合 a 出现，而两幅墙完全确定了 a：$u^2 = g/(2a\cos^2\theta)$。所以两幅墙决定了整条轨迹，而射程只取决于 d_1 和 d_2。

验算

- 对称性（symmetry）：互换 d_1 和 d_2 只是把两幅墙对调，而两个答案关于 d_1, d_2 都是对称的。
- 代入数值：取 $d_1 = 2$、$d_2 = 3$，得 $\tan\theta = \frac{19}{6}$、$a = \frac{5}{6}$ 及 $R = \frac{19}{5}$。于是 $y(2) = \frac{19}{3} - \frac{10}{3} = 3$，$y(3) = \frac{19}{2} - \frac{15}{2} = 2$，符合要求。以 $g = 9.8$ 对运动作数值逐步模拟，同样落在 $x = 3.8$。
- 由于 $d_1^2 + d_2^2 \geq 2d_1 d_2$，得 $\tan\theta \geq 3$，所以抛射角总是很陡（至少约 71.6°）。这是合理的：要先越过较高的墙，然后这么快就降到较矮的墙，路径必须是一道窄拱。

常见错误

- 混淆哪幅墙有哪个高度：距离为 d_1 的墙，高度是 d_2。

- 把 a 写成含 $\cos\theta$ 而非 $\cos^2\theta$，或没有把 a 当作单一未知数处理，结果留下一个关于 u 和 θ 的混乱方程组。
- 除以 $d_2 - d_1$ 时没有说明它为何不等于零。

STEP 之外

这里的技巧在于：一个带有两个自由参数（parameter）（此处为 $\tan\theta$ 和 a）的曲线族，可由它必须经过的两个点确定下来，这是插值（interpolation）的一个小例子。选取适当的未知数组合这个想法贯穿整个力学，例如求所有以固定抛射速率发射的轨迹的“包络线”（envelope）（即安全抛物线（parabola of safety））。如想学习更多大学一年级程度的抛射体及质点力学，可参阅 Kleppner 和 Kolenkow 的 *An Introduction to Mechanics*（《力学导论》）。

详细解答

STEP I 2011，第 10 题

力学 · 弹性碰撞

STEP I 2011，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/11-s1-q10

关键思路

两个质点（particle）都由同一点从静止开始运动，而在释放与碰撞之间发生的一切（自由下落、一次完全弹性（perfectly elastic）的反弹）都不会损失任何能量。因此在任一高度，每个质点的速率都相同，而它们会在同一高度相遇。之后，碰撞就是一道标准的两条方程问题（动量（momentum）守恒，以及 $e = 1$ 的牛顿碰撞定律（Newton's law of restitution）），而 B 的最大高度再次由能量求得。

解答

我们把 A 和 B 视为质点，忽略空气阻力（air resistance），并全程取向上为正方向。

为何两者速率相等

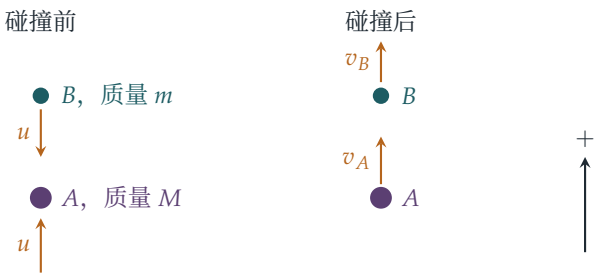
每个质点都在高度 h 处由静止开始运动。自由飞行期间只有重力作用，所以机械能守恒（conservation of mechanical energy）；A 在平面上的反弹是完全弹性的，因此它只会令 A 的速度反向而不改变其速率，那里也没有损失能量。故此，在第一次碰撞之前，每个质点的单位质量总能量都是 gh，而处于高度 y 的质点的速率 v 满足

$$\tfrac{1}{2}v^2 + gy = gh, \qquad \text{所以} \qquad v = \sqrt{2g(h-y)}.$$

速率只取决于高度，与质量或质点过去的运动经历无关。在第一次碰撞的一刻，两个质点处于同一高度 y，所以它们的速率相等。此外，A 尚未回到 P，所以 $y < h$，而这个共同速率 $u = \sqrt{2g(h-y)}$ 是正数。由于 A 已经反弹，它正向上运动；而 B（从未反弹）则正向下运动。

碰撞

碰撞前一刻，A 的速度为 $+u$，B 的速度为 $-u$。设碰撞后一刻两者的速度分别为 v_A 和 v_B。



碰撞中的冲力（impulsive force）是这对质点之间的内力，而在撞击的一瞬间重力不产生任何冲量（impulse），所以动量守恒：

$$Mu - mu = Mv_A + mv_B. \tag{1}$$

碰撞是完全弹性的（$e = 1$），所以分离速率等于接近速率：

$$v_B - v_A = u - (-u) = 2u. \tag{2}$$

把由 (2) 得到的 $v_B = v_A + 2u$ 代入 (1)：

$$(M-m)u = (M+m)v_A + 2mu \quad \Longrightarrow \quad v_A = \frac{(M-3m)u}{M+m},$$

然后

$$v_B = v_A + 2u = \frac{(M-3m)u + 2(M+m)u}{M+m} = \frac{(3M-m)u}{M+m}.$$

由于 $M > 3m$，我们有 $M - 3m > 0$，所以 $v_A > 0$；而 $v_B = v_A + 2u > 0$ 亦成立。因此第一次碰撞后两个质点都向上运动，正如题目所要求。

想一想 我们怎样知道 B 在下一次与 A 相遇之前，确实能到达其飞行的最高点？

碰撞后两个质点的加速度都是 $-g$，所以两者都在飞行时，相对速度一直保持 $v_B - v_A = 2u$：B 以稳定的速率离 A 而去，因此在 A 再次撞到平面之前，两者不可能相遇。若 A 在 B 仍在上升时反弹，就比较两者的单位质量能量：B 的是 $\frac{1}{2}v_B^2 + gy$，A 的是 $\frac{1}{2}v_A^2 + gy$；由于 $v_B > v_A > 0$，后者较小（而 A 的反弹不改变其能量）。所以每当两者处于同一高度时，A 的速率都小于 B 的速率。要在 B 仍在上升时从下方追上 B，A 在相遇一刻向上的速率必须至少与 B 一样快，而我们刚刚看到这是不可能的。因此第二次碰撞只会在 B 经过其最高点之后才发生，而这个最高点正是题目所求的最大高度。

B 的最大高度

碰撞发生在高度 y 处，其中 $u^2 = 2g(h-y)$，即

$$y = h - \frac{u^2}{2g}.$$

碰撞后 B 以初速率 v_B 只在重力作用下上升，所以它会再上升 $v_B^2/(2g)$（由 $v^2 = v_B^2 - 2gs$ 并取 $v = 0$ 得出）。因此它的最大高度是

$$\begin{aligned} y + \frac{v_B^2}{2g} &= h - \frac{u^2}{2g} + \frac{(3M-m)^2u^2}{2g(M+m)^2} \\ &= h + \frac{u^2}{2g(M+m)^2}\left[(3M-m)^2 - (M+m)^2\right] \\ &= h + \frac{u^2}{2g(M+m)^2}(2M-2m)(4M) \\ &= h + \frac{4M(M-m)u^2}{(M+m)^2 g}, \end{aligned}$$

其中第三行用了平方差（difference of two squares）公式。这正是所求的结果。

验算

- 能量：以上述的 v_A 和 v_B，有 $Mv_A^2 + mv_B^2 = (M+m)u^2$，正如弹性碰撞所要求。
- 极限情况（limiting case）：若 $m \to 0$，则 $v_A \to u$ 而 $v_B \to 3u$（一个极轻的球撞上一个重球，就像撞上一道移动的墙一样反弹，相对于它以速度 $2u$ 离开），而高度变为 $h + 4u^2/g$，与 $y + 9u^2/(2g)$ 吻合。
- 由于 $M > m$，额外高度是正数：B 最终到达 P 之上，这是理所当然的，因为它从 A 获得了能量。
- 数值方面，我们在随机时间释放 B，取随机的 h 及满足 $M > 3m$ 的质量，逐步模拟运动（包括 A 其后的反弹）直至第二次碰撞，发现 B 的最大高度与公式吻合，相对误差约为 10^{-10}。

常见错误

- 正负号错误：把碰撞前两个速度都取作 $+u$，就忘记了 A 正在上升时 B 正在下落。
- 把额外上升的高度 $v_B^2/(2g)$ 从平面或从 P 起计，而不是从碰撞高度 $h - u^2/(2g)$ 起计。

STEP 之外

“速率只取决于高度”这个论证，其实就是保守力（conservative force）下的能量守恒：某一点的速率只取决于该点的势能（potential energy），与所走的路径无关。碰撞本身通常在质心参考系（centre-of-mass frame）中处理最为简单：在该参考系中，弹性碰撞只是把每个质点的速度反向；再变换回原参考系，一行便得到 v_A 和 v_B。Kleppner 与 Kolenkow 的 *An Introduction to Mechanics*（《力学导论》）仔细地发展了这两个概念。

详细解答

STEP I 2011，第 11 题

力学 • 刚体的平衡

STEP I 2011，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：

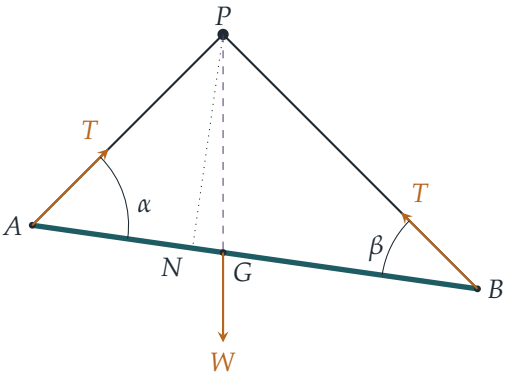
step.maths.org/questions/11-s1-q11

关键思路

杆上只受三个力：绳在 A 和 B 两点的张力（tension），以及作用于 G 的重量。由于钉子光滑（smooth），两个张力相等，因此对 G 取力矩（moments）（重量对 G 没有力矩），便立即得出两个角之间的关系。第二部分还需要一个事实：处于平衡（equilibrium）而又不全部平行的三个力，必定通过同一点。这一点就是钉子，所以 PG 是铅垂的。

解答

我们把杆视为刚体杆（rigid rod），绳为轻（light）而不可伸长（inextensible）的，钉子为光滑的。设杆的重量为 W。



证明 $3\sin\alpha = 4\sin\beta$

绳是轻的，并绕过光滑的钉子，所以整条绳的张力 T 都相同。因此杆在 A 受到沿 AP 方向的力 T，在 B 受到沿 BP 方向的力 T，另外还有在 G 铅垂向下作用的 W。

对 G 取力矩。重量没有力矩。由 G 到直线 AP 的垂直距离是 $AG\sin\alpha = 3d\sin\alpha$，由 G 到直线 BP 的垂直距离是 $GB\sin\beta = 4d\sin\beta$，因为 $GB = 7d - 3d = 4d$。A 点的张力使杆绕 G 向一个方向转动，B 点的张力则使杆向另一个方向转动，所以平衡时

$$T\cdot 3d\sin\alpha = T\cdot 4d\sin\beta \quad\Longrightarrow\quad 3\sin\alpha = 4\sin\beta,$$

（两边除以 $Td \neq 0$）即为所求。

绳的长度

由 $\cos\beta = \frac{4}{5}$ 得 $\sin\beta = \frac{3}{5}$（β 是三角形的一个角，所以 $\sin\beta > 0$）。于是

$$\sin\alpha = \tfrac{4}{3}\sin\beta = \tfrac{4}{5}, \qquad \cos\alpha = \tfrac{3}{5},$$

这里取正根，因为 α 是锐角（acute）。接着

$$\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta = \tfrac{3}{5}\cdot\tfrac{4}{5} - \tfrac{4}{5}\cdot\tfrac{3}{5} = 0,$$

而 $0 < \alpha+\beta < \pi$，所以 $\alpha+\beta = \frac{\pi}{2}$。因此三角形 PAB 在 P 处是直角，斜边（hypotenuse）为 $AB = 7d$。故此

$$PA = 7d\cos\alpha = \tfrac{21}{5}d, \qquad PB = 7d\cos\beta = \tfrac{28}{5}d,$$

而绳的长度是

$$PA + PB = \tfrac{49}{5}d.$$

杆的倾角

想一想 我们知道 P 相对于杆的位置，但“铅垂”方向在哪里？

两个张力都沿通过 P 的直线作用，所以它们的合力（resultant）也通过 P。平衡时这个合力必须与重量抵消，因此它必须与 W 大小相等、方向相反，并且在同一直线上。（处于平衡的三个不平行的力必定共点（concurrent）。）所以通过 G 的铅垂线经过 P：PG 是铅垂的。

设 N 为由 P 到 AB 的垂线的垂足。在直角三角形 PAN 中，

$$AN = PA\cos\alpha = \tfrac{21}{5}d\cdot\tfrac{3}{5} = \tfrac{63}{25}d, \qquad PN = PA\sin\alpha = \tfrac{21}{5}d\cdot\tfrac{4}{5} = \tfrac{84}{25}d.$$

由于 $AN = \frac{63}{25}d < 3d = AG$，点 N 位于 A 与 G 之间，且

$$NG = 3d - \tfrac{63}{25}d = \tfrac{12}{25}d.$$

在直角三角形 PNG 中，

$$\tan\angle PGN = \frac{PN}{NG} = \frac{84}{12} = 7.$$

角 PGN 是杆与铅垂线 GP 之间、在 A 一侧量度的角。它是锐角，所以端点 A 在通过 G 的水平线之上，而 B 在其之下，与题目给出的位置相符。杆与水平线之间的角 θ 等于 $\frac{\pi}{2} - \angle PGN$，所以

$$\tan\theta = \frac{1}{\tan\angle PGN} = \frac{1}{7}, \qquad \text{即} \qquad \theta = \arctan\tfrac{1}{7},$$

即为所求。

验算

- 绳长 $\frac{49}{5}d = 9.8d$ 比杆长 $7d$ 长，符合三角形 PAB 中三角不等式（triangle inequality）的要求。
- 两个相等张力的合力平分它们之间的夹角，所以 PG 平分角 APB。根据角平分线定理（angle bisector theorem），$PA : PB$ 应等于 $AG : GB = 3 : 4$，而 $\frac{21}{5} : \frac{28}{5} = 3 : 4$，确实如此。
- 作为数值验算，我们把 P 放在以 A 和 B 为焦点（foci）的椭圆（ellipse）上（即在这个绳长下钉子所有可能的位置），然后找出使 G 在 P 之下的深度取驻定值（stationary）的倾角。结果是 $8.13°$，与 $\arctan\frac{1}{7} \approx 8.130°$ 吻合。

常见错误

- 给绳的两段设置不同的张力。光滑的钉子不能改变轻绳中的张力。
- 把两段距离对调，得出 $4\sin\alpha = 3\sin\beta$。较长的一段 $GB = 4d$ 应配 B 处的角。
- 忘记 α 是锐角，而容许 $\cos\alpha = -\frac{3}{5}$。这会得出另一个三角形，而在其中 $\alpha + \beta$ 并不是直角。

STEP 之外

在绳长固定的情况下，钉子必定位于以 A 和 B 为焦点的椭圆上。于是，PG 平分角 APB 这个事实正是椭圆的反射性质（reflection property）：在 P 点的法线（normal）平分两条

焦半径（focal radii）之间的夹角。此外，平衡位置恰好就是势能（potential energy）取驻定值的位置，这正是大学一年级课程中虚功原理（principle of virtual work）和拉格朗日力学（Lagrangian mechanics）的起点。Kleppner 和 Kolenkow 所著的 *An Introduction to Mechanics*（《力学导论》）是学习平衡、力矩和能量方法的好读物。

详细解答

STEP I 2011，第 12 题

概率与统计 • 排队与选票问题

> **STEP I 2011，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/11-s1-q12

本题在 *Beyond STEP* 第 9.10 节总能找赎中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch09/ballot.html

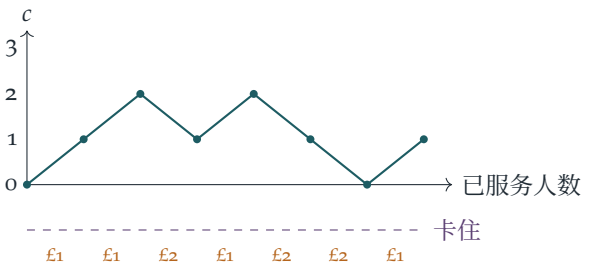
关键思路

只需追踪一个数字：我手上有多少枚 £1 硬币。每位付 £1 的顾客令这个数字加一；每位付 £2 的顾客令它减一（我收下他们的 £2 硬币，但这枚硬币不能用来找零）。我会卡住，当且仅当一位付 £2 的顾客到来时，我手上一枚 £1 硬币也没有。由于付 £2 的顾客只有寥寥几位，危险全都集中在队伍的前端，所以我们只需看队伍最前面的几个人，逐一分析每种开头会带来甚么结果。

解答

重要的只是队伍中硬币的排列模式，而不是那些人是谁。由于所有人的排列都是等可能的，任何一个位置上的人，等可能是尚未排入队伍的任何一人，因此我们可以从队头开始逐个建构队伍：若尚余 a 位持 £1 硬币的人和 b 位持 £2 硬币的人，则下一位持 £1 硬币的概率（probability）为 $\frac{a}{a+b}$，持 £2 硬币的概率为 $\frac{b}{a+b}$。

设 c 为我手上 £1 硬币的数目。我能服务所有人，当且仅当 c 永远不需要跌至 0 以下，即在队伍中的每一点，已服务的 £1 顾客人数至少与已服务的 £2 顾客人数一样多。



上图显示一条成功的队伍：c 的路径从未触及虚线。

第 (i) 部分

若只有一位付 £2 的顾客，唯一会卡住的情况是此人排在队首，那时我手上没有零钱。若队首是其他人，那么到这位 £2 顾客到来时，我手上至少有一枚 £1 硬币。这位 £2 顾客等可能排在 $m+1$ 个位置中的任何一个，因此

$$\mathrm{P}(\text{卖给所有人}) = 1 - \frac{1}{m+1} = \frac{m}{m+1}.$$

第 (ii) 部分

以 1 表示付 £1 的顾客，以 2 表示付 £2 的顾客，并列出所有开头。

- 若第一人是 2，我立即卡住。
- 若队伍以 $1,2,2$ 开头，则 c 依次为 $1,0$，我在第三人时卡住。
- 若队伍以 $1,1$ 开头，则两人之后 $c=2$。付 £2 的顾客总共只有两位，而 c 只会在服务其中一人时下跌，因此 c 永远不会跌至 $2-2=0$ 以下。成功。
- 若队伍以 $1,2,1$ 开头，则 c 依次为 $1,0,1$，而只剩一位付 £2 的顾客，因此 c 永远不会跌至 0 以下。成功。

这些情况涵盖了所有可能，因此我能服务所有人，当且仅当队伍以 $1,1$ 或 $1,2,1$ 开头。从队头开始建构队伍，

$$\mathrm{P}(1,1) = \frac{m}{m+2}\cdot\frac{m-1}{m+1}, \quad \mathrm{P}(1,2,1) = \frac{m}{m+2}\cdot\frac{2}{m+1}\cdot\frac{m-1}{m} = \frac{2(m-1)}{(m+2)(m+1)}.$$

这两个事件互斥（disjoint），因此

$$\mathrm{P}(\text{卖给所有人}) = \frac{m(m-1)+2(m-1)}{(m+2)(m+1)} = \frac{(m-1)(m+2)}{(m+2)(m+1)} = \frac{m-1}{m+1},$$

即为所求。（我们需要 $m \geq 2$，使最后一个因子中尚余 $m-1 \geq 1$ 位 £1 顾客。）

第 (iii) 部分

现在有三位付 £2 的顾客。同样地，看队伍最前面的三个人。

- 以 2 开头，或以 1,2,2 开头：卡住，与 (ii) 完全相同。
- 以 1,1,1 开头：此时 $c=3$，三位 £2 顾客最多只能把 c 减至 0。成功。
- 以 1,1,2 或 1,2,1 开头：两种情况下，三人之后都有 $c=1$，而余下的是 $m-2$ 位持 £1 硬币的人和 2 位持 £2 硬币的人，以随机次序排列。

> **想一想** 在最后一种情况中，队伍余下的部分何时会出问题?

以 $c=1$ 开始处理队伍余下的部分，就好比在一条有两位 £2 顾客的队伍前面，多了一位免费的 £1 顾客，这正是改头换面的第 (ii) 部分。在 $c=1$ 且只剩两位 £2 顾客的情况下，我卡住当且仅当接下来两人都是 £2 顾客：c 先变成 0，然后我便无法服务。若接下来两人中有任何一人持 £1 硬币，则 c 永远不会跌至 0 以下（仿照 (ii) 检查 $1,\ldots$ 和 $2,1,\ldots$ 即可）。在首三人之后余下的 m 人中，

$$\mathrm{P}(\text{接下来两人都是2})=\frac{2}{m}\cdot\frac{1}{m-1},$$
$$\text{故}\quad \mathrm{P}(\text{其余部分成功})=1-\frac{2}{m(m-1)}=\frac{(m-2)(m+1)}{m(m-1)},$$

这里用了 $m^2-m-2=(m-2)(m+1)$。

设 $D=(m+3)(m+2)(m+1)$。各个有用开头的概率为

$$\mathrm{P}(1,1,1)=\frac{m(m-1)(m-2)}{D},$$
$$\mathrm{P}(1,1,2)=\frac{m\cdot(m-1)\cdot 3}{D},\qquad \mathrm{P}(1,2,1)=\frac{m\cdot 3\cdot(m-1)}{D}.$$

由于各个开头互斥，而且无论开头如何，队伍余下的部分同样是随机次序，

$$\begin{aligned}\mathrm{P}(\text{成功})&=\frac{m(m-1)(m-2)}{D}+\frac{6m(m-1)}{D}\cdot\frac{(m-2)(m+1)}{m(m-1)}\\&=\frac{(m-2)\left[m(m-1)+6(m+1)\right]}{D}=\frac{(m-2)(m^2+5m+6)}{D}\\&=\frac{(m-2)(m+2)(m+3)}{(m+3)(m+2)(m+1)}=\frac{m-2}{m+1},\end{aligned}$$

即为所求。（这里 $m\geq 3$ 确保 $m-1\geq 2$，使条件概率（conditional probability）有意义。）

验算

- 最小的情况：当 $m=n$ 时，对 $n=1,2,3$，公式分别给出 $\frac{1}{2}$、$\frac{1}{3}$、$\frac{1}{4}$。当 $m=n=2$ 时，由 1 和 2 组成的六种排列模式等可能，而只有 1122 和 1212 成功：$\frac{2}{6}=\frac{1}{3}$。✓

- m 很大时：三个概率都趋向 1，这在 £1 硬币充裕时理应如此。
- 通过完整列举所有队伍次序，当 $m = 5$ 时，对 $n = 1, 2, 3$ 我得到 $\frac{5}{6}$、$\frac{2}{3}$、$\frac{1}{2}$，与 $\frac{m}{m+1}$、$\frac{m-1}{m+1}$、$\frac{m-2}{m+1}$ 吻合；这些公式亦已对直至 14 的每个 m 作精确验证。

常见错误

- 忘记我收到的 £2 硬币不能用来找零：只有 £1 硬币才算数。
- 把每个人当作以固定概率 $\frac{m}{m+n}$ 独立抽出。队伍是以不放回抽样（without replacement）的方式产生的，所以概率会随着人逐一排入而改变。
- 在 (iii) 中，以为三人之后 $c = 1$ 必定安全，因而漏掉了失败情况 1, 1, 2, 2, 2 或 1, 2, 1, 2, 2。

STEP 之外

对每个 $n \leq m$，规律都是 $\frac{m+1-n}{m+1}$。这就是*选票问题*（ballot problem），最简洁的证明运用了*反射原理*（reflection principle）：一条触及卡住水平的路径可以经反射得到一条终点不同的路径，从而把坏队伍的数目化为单一个二项式系数（binomial coefficient）。当 $m = n$ 时，好队伍的数目是一个卡特兰数（Catalan number），它在组合数学中随处可见。关于随机游走（random walk）和反射原理，可参阅 Grimmett 与 Welsh 的 *Probability: An Introduction*（《概率导论》）。

详细解答

STEP I 2011，第 13 题

概率与统计 · 连续随机变量

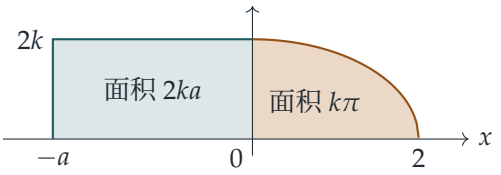
STEP I 2011，第 13 题

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/11-s1-q13

关键思路

先画出密度函数（density function）的图像。它由两部分拼成：在 $[-a,0)$ 上是一个高为 $2k$ 的长方形，在 $[0,2]$ 上则是一个半径为 2 的四分之一圆（垂直方向按比例 k 缩放）。图像下方的面积就是概率，因此单靠长方形和四分之一圆，大部分工作已经完成，毋须做任何积分。总面积必须等于 1，由此便可定出 k。之后所有关于 d 的问题，都归结为同一件事：某一点右方有多少面积?



解答

求 k

在 $[0,2]$ 上，$\sqrt{4-x^2}$ 的图像是圆 $x^2+y^2=4$ 的四分之一，所以

$$\int_0^2 \sqrt{4-x^2}\,\mathrm{d}x = \tfrac{1}{4}\pi\cdot 2^2 = \pi.$$

（用题目给出的原函数（antiderivative）计算，结果一致：$2\arcsin 1+0-0=\pi$。）因此总概率为 $2ka+k\pi=1$，即

$$k=\frac{1}{2a+\pi}.$$

第 (i) 部分

把 $\mathrm{E}(X)$ 的积分在 0 处分拆：

$$\int_{-a}^{0} 2kx\,\mathrm{d}x = k\left[x^2\right]_{-a}^{0} = -ka^2.$$

至于第二部分，$(4-x^2)^{3/2}$ 的导数是 $-3x(4-x^2)^{1/2}$，所以

$$\int_{0}^{2} kx\sqrt{4-x^2}\,\mathrm{d}x = k\left[-\tfrac{1}{3}(4-x^2)^{3/2}\right]_{0}^{2} = k\cdot\tfrac{1}{3}\cdot 8 = \tfrac{8}{3}k.$$

两部分相加，得

$$\mathrm{E}(X) = k\left(\frac{8}{3} - a^2\right) = \frac{8-3a^2}{3(2a+\pi)}.$$

第 (ii) 部分

0 右方的概率就是四分之一圆的面积：

$$\mathrm{P}(X>0) = k\pi = \frac{\pi}{2a+\pi}.$$

? **想一想** 知道了 $\mathrm{P}(X>0)$，为何就能判断 d 的正负？

由于密度函数在 $(-a,2)$ 上为正，函数 $t \mapsto \mathrm{P}(X>t)$ 在 $-a \le t \le 2$ 上连续（continuous）且严格递减（strictly decreasing）。它由 1 降至 0，所以恰好有一个 d 满足 $\mathrm{P}(X>d) = \frac{1}{10}$。由于函数递减，$d<0$ 当且仅当 $\mathrm{P}(X>0) < \frac{1}{10}$。换句话说，当 0 右方的概率不足十分之一时，$d$ 便是负数。

现在假设 $2a > 9\pi$，则 $2a+\pi > 10\pi$，所以

$$\mathrm{P}(X>0) = \frac{\pi}{2a+\pi} < \frac{1}{10},$$

因此 $d<0$，得证。（这个论证反过来亦成立：只有当 $2a>9\pi$ 时才有 $d<0$。）

在这情况下 $-a \le d < 0$，所以 d 右方的面积，就是长方形由 d 到 0、宽度为 $-d$ 的那一部分，再加上整个四分之一圆：

$$\mathrm{P}(X>d) = 2k(-d) + k\pi = \frac{\pi-2d}{2a+\pi}.$$

令它等于 $\frac{1}{10}$，得 $10\pi - 20d = 2a+\pi$，所以

$$d = \frac{9\pi-2a}{20}.$$

这个值确实位于 $[-a,0)$ 之内：因为 $2a>9\pi$，所以它是负数；而 $d \ge -a$ 等价于 $9\pi+18a \ge 0$，这显然成立。

第 (iii) 部分

现在 $d=\sqrt{2}>0$，所以 d 落在圆弧部分之下，而

$$\mathrm{P}(X>\sqrt{2})=k\int_{\sqrt{2}}^{2}\sqrt{4-x^2}\,\mathrm{d}x.$$

利用题目给出的结果，

$$\begin{aligned}\int_{\sqrt{2}}^{2}\sqrt{4-x^2}\,\mathrm{d}x&=\left[2\arcsin\left(\tfrac{1}{2}x\right)+\tfrac{1}{2}x\sqrt{4-x^2}\right]_{\sqrt{2}}^{2}\\&=\left(2\cdot\tfrac{\pi}{2}+0\right)-\left(2\cdot\tfrac{\pi}{4}+\tfrac{\sqrt{2}}{2}\cdot\sqrt{2}\right)\\&=\frac{\pi}{2}-1.\end{aligned}$$

（从几何角度看，这是被 $x=\sqrt{2}$ 处的弦（chord）割出的弓形（circular segment），它等于圆面积的八分之一，即 $\frac{\pi}{2}$，减去一个面积为 1 的直角三角形。）所以

$$\frac{\frac{\pi}{2}-1}{2a+\pi}=\frac{1}{10}\quad\Longrightarrow\quad 5\pi-10=2a+\pi\quad\Longrightarrow\quad a=2\pi-5.$$

这是正数（$a\approx 1.283$），并且满足 $2a\le 9\pi$；由于 $d>0$，这与第 (ii) 部分的结论吻合。

验算

- 在第 (i) 部分，取 $a=1$，均值为 $\frac{5}{3(2+\pi)}\approx 0.3242$。以 2×10^5 个样本进行蒙地卡罗模拟（Monte Carlo simulation），得出 0.3236。
- 在临界情况 $2a=9\pi$ 下，第 (ii) 部分的公式给出 $d=0$，与条件一致。取 $a=20$，公式给出 $d=\frac{9\pi-40}{20}\approx -0.586$，而以数值方法解 $\mathrm{P}(X>d)=\frac{1}{10}$ 亦得出相同的值。
- 在第 (iii) 部分，取 $a=2\pi-5$，以辛普森法则（Simpson's rule）计算得 $\mathrm{P}(X>\sqrt{2})=0.1000$。

常见错误

- 没有求出 k。归一化（normalisation）条件 $2ka+k\pi=1$ 是隐藏的第一步，而所有答案都只应以 a 表示。
- 在第 (ii) 部分漏掉四分之一圆，只写 $\mathrm{P}(X>d)=-2kd$。其实 $[0,2]$ 上的全部概率，同样位于负数 d 的右方。
- 把 $\int_{-a}^{0}2kx\,\mathrm{d}x$ 的正负号弄错。由于整个区间上 $x<0$，这个积分是负数。

STEP 之外

d 这个值就是 X 的上 10% 点，即 0.9 分位数（quantile）。要求出它，就要求累积分布函数（cumulative distribution function）F 的反函数；而这个反函数，正正是用均匀分布的随机变量去模拟另一个随机变量时所用的工具（逆变换抽样（inverse transform sampling））。这个密度函数形状混合，一边平坦、一边弯曲，说明了为何分位数往往要分段求出。Grimmett 和 Welsh 所著的 *Probability: An Introduction*（《概率导论》）对分布函数和分位数有仔细的论述。

详细解答

STEP II 2011，第 1 题

纯数学 · 含平方根的方程

STEP II 2011，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/11-s2-q1

关键思路

两部分都涉及平方根之和。只要找出定义域（domain）、端点的值和对称性（symmetry），这些函数的图像便很容易理解。在第 (i) 部分，图像给出一个不等式（inequality）：曲线永远不会低于某个高度，而直线只能在一点达到该高度，所以根本不用解方程。在第 (ii) 部分，图像告诉我们应该预期有多少个根，从而让我们剔除平方时引入的多余根。

解答

第 (i) 部分

设 $f(x)=\sqrt{1-x}+\sqrt{3+x}$。两个平方根只有在 $1-x\geq 0$ 及 $3+x\geq 0$ 时才是实数，所以定义域为 $-3\leq x\leq 1$。在两个端点，$f(-3)=2+0=2$ 及 $f(1)=0+2=2$。

令 $x=-1+t$，得 $f=\sqrt{2-t}+\sqrt{2+t}$。以 $-t$ 代替 t 时此式不变，所以曲线关于 $x=-1$ 对称。求导数（derivative），

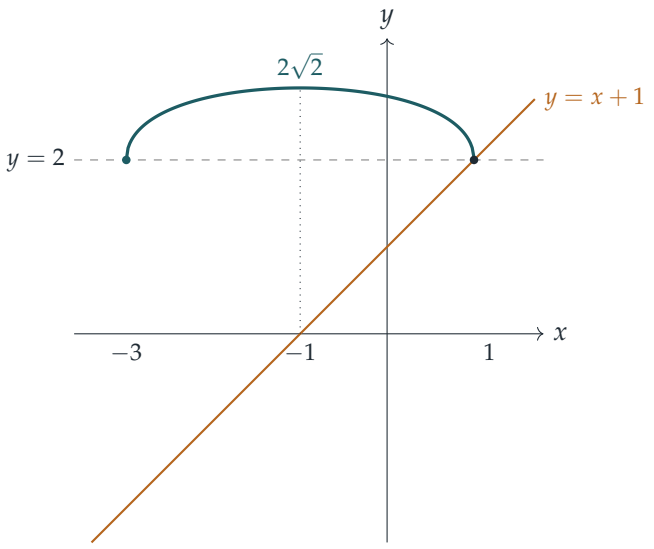
$$f'(x)=-\frac{1}{2\sqrt{1-x}}+\frac{1}{2\sqrt{3+x}},$$

它只在 $1-x=3+x$，即 $x=-1$ 时为零，此时 $f(-1)=2\sqrt{2}$。当 $x<-1$ 时，$3+x<1-x$，所以 $f'(x)>0$；当 $x>-1$ 时，$f'(x)<0$。因此 f 由 2 上升至极大值（maximum）$2\sqrt{2}$，再回落至 2。在 $x=-3$ 附近，f' 的第二项趋于无限大，所以曲线以垂直方向离开 $(-3,2)$；同理，它亦以垂直方向到达 $(1,2)$。f 的每一项都是凹函数（concave），所以整条曲线是一顶平滑的“帽子”。

要看出最小值，更简洁的方法是平方：

$$f(x)^2=(1-x)+(3+x)+2\sqrt{(1-x)(3+x)}=4+2\sqrt{(1-x)(3+x)}\ \geq\ 4,$$

所以在整个定义域上 $f(x) \geq 2$，等号只在 $x = -3$ 及 $x = 1$ 时成立。



现在引入直线 $y = x + 1$。在定义域上 $x \leq 1$，所以 $x + 1 \leq 2$，等号只在 $x = 1$ 时成立。另一方面 $f(x) \geq 2$。因此 $f(x) = x + 1$ 迫使两边都等于 2，而这只在 $x = 1$ 时发生。验证：$\sqrt{0} + \sqrt{4} = 2 = 1 + 1$。所以恰好有一个实数值满足方程：

$$x = 1.$$

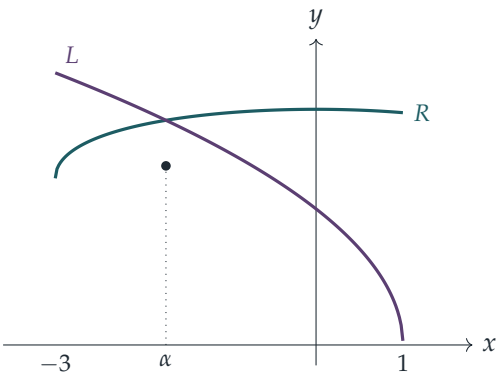
（在 $[-3, 1]$ 以外，左边没有定义，所以没有其他候选值。）

第 (ii) 部分

同样，我们需要 $x \leq 1$、$x \geq -3$ 及 $x \leq 3$，所以定义域为 $[-3, 1]$。比较

$$L(x) = 2\sqrt{1 - x} \qquad \text{和} \qquad R(x) = \sqrt{3 + x} + \sqrt{3 - x}.$$

L 严格递减（strictly decreasing），由 $L(-3) = 4$ 降至 $L(1) = 0$。R 的形状与第 (i) 部分的曲线相同：它关于 $x = 0$ 对称（互换 x 与 $-x$），$R(\pm 3) = \sqrt{6}$，极大值为 $R(0) = 2\sqrt{3}$，在 $[-3, 0]$ 上递增，在 $[0, 3]$ 上递减。在我们的定义域上，只看到由 $x = -3$ 至 $x = 1$ 的一段，其中 $R(1) = 2 + \sqrt{2}$。



在 $x=-3$，$L=4>\sqrt{6}=R$；在 $x=0$，$L=2<2\sqrt{3}=R$。在 $[-3,0]$ 上，L 严格递减而 R 递增，所以 $L-R$ 严格递减，并恰好变号一次。在 $[0,1]$ 上，$L\le 2$ 而 $R\ge 2+\sqrt{2}$，所以两者不相交。因此方程**恰好有一个**实数解，且它位于 $-3<x<0$。

要求出这个解，就要平方。两边都是非负的，所以平方不会损失任何东西：

$$\begin{aligned} & 4(1-x)=(3+x)+(3-x)+2\sqrt{9-x^2} \\ \iff\quad & -1-2x=\sqrt{9-x^2}. \end{aligned}$$

想一想 再次平方时，新方程的哪些解才是真正的解?

右边是平方根，所以它 ≥ 0；真正的解必须满足 $-1-2x\ge 0$，即 $x\le -\frac{1}{2}$。在这个条件下，平方是可逆的：

$$1+4x+4x^2=9-x^2 \iff 5x^2+4x-8=0 \iff x=\frac{-2\pm 2\sqrt{11}}{5}.$$

根 $\frac{-2+2\sqrt{11}}{5}\approx 0.93$ 违反 $x\le -\frac{1}{2}$，所以我们舍去它。另一个根 $\frac{-2-2\sqrt{11}}{5}\approx -1.73$ 满足条件，并位于 $[-3,0]$ 内，正如图像所预测。所以唯一解是

$$x=-\frac{2+2\sqrt{11}}{5}.$$

验算

- 在 $x\approx -1.72665$：$2\sqrt{1-x}\approx 3.3025$，而 $\sqrt{3+x}+\sqrt{3-x}\approx 1.1284+2.1741\approx 3.3025$。两边相符。
- 被舍去的根 $x\approx 0.92665$ 给出 $L\approx 0.542$，但 $R\approx 3.42$。它其实是 $\sqrt{3+x}-\sqrt{3-x}=2\sqrt{1-x}$ 的解，而这条方程平方后得出同一结果。
- 在第 (i) 部分，$f(-1)=2\sqrt{2}\approx 2.83>0=(-1)+1$，与直线除了在 $(1,2)$ 外处处位于曲线下方一致。

常见错误

- 在第 (i) 部分用平方和代数方法硬解。这既繁复，又可能引入增根（extraneous root），而界 $f \geq 2 \geq x+1$ 却能立即解决问题。
- 保留 $5x^2+4x-8=0$ 的两个根。每一步平方都需要附带其符号条件，这里是 $-1-2x \geq 0$。
- 忘记定义域：对于 $x>1$，两条方程都没有意义。

STEP 之外

第 (i) 部分是分析（analysis）中一个常见技巧的小例子：与其解 $f(x)=g(x)$，不如证明 $f \geq c \geq g$，再看等号何时成立。第 (ii) 部分在求出任何根之前，就结合介值定理（intermediate value theorem）与单调性（monotonicity）来数算根的个数，这正是大学第一门分析课程中证明存在性与唯一性（existence and uniqueness）的方法。Abbott 的 *Understanding Analysis*（《理解分析》）对连续性（continuity）与介值定理有细致的论述。

详细解答

STEP II 2011，第 2 题

纯数学 · 丢番图方程

STEP II 2011，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/11-s2-q2

关键思路

每当见到 x^3+y^3，就应把它因式分解（factorise）：$x^3+y^3=(x+y)(x^2-xy+y^2)$。一旦 $x+y$ 被确定（在 (i) 中为 k，在 (ii) 中为 z^2），因式 $x+y$ 便可与 kz^3 的一部分约去，剩下一个二次条件。对这个二次式配方（completing the square），可把它化成“某式 $=3\times$ 完全平方数（perfect square）”的形式，再利用上下界（bounds），便只剩下少数几个立方数需要逐一尝试。这正是题目先要求列出不超过 10^3 的立方数的原因。

以供参考：$1, 8, 27, 64, 125, 216, 343, 512, 729, 1000$。

解答

第 (i) 部分

由于 $x+y=k$，

$$kz^3=x^3+y^3=(x+y)(x^2-xy+y^2)=k(x^2-xy+y^2),$$

而 $k>0$，故 $z^3=x^2-xy+y^2$。代入 $y=k-x$：

$$\begin{aligned} z^3 &= x^2-x(k-x)+(k-x)^2 \\ &= x^2-kx+x^2+k^2-2kx+x^2 \\ &= k^2-3kx+3x^2, \end{aligned}$$

得证。

完全平方数。两边乘以 4 并配方：

$$4z^3-k^2=12x^2-12kx+3k^2=3\left(4x^2-4kx+k^2\right)=3(k-2x)^2.$$

所以 $\dfrac{4z^3-k^2}{3}=(k-2x)^2$，即整数 $k-2x$ 的平方。

上下界。由上式，$4z^3=k^2+3(k-2x)^2\geq k^2$，故 $\frac{1}{4}k^2\leq z^3$。至于上界，注意 $3kx-3x^2=3x(k-x)=3xy$，所以

$$z^3=k^2-3xy<k^2,$$

因为 x 和 y 都是正数。因此 $\frac{1}{4}k^2\leq z^3<k^2$。

$k=20$ 的情况。我们寻找满足 $x+y=20$ 的解。上下界给出 $100\leq z^3<400$，从立方数列表可知 $z^3\in\{125,216,343\}$。接着检验 $(4z^3-400)/3$ 是否为完全平方数：

- $z=5$：$(500-400)/3=100/3$，不是整数；
- $z=6$：$(864-400)/3=464/3$，不是整数；
- $z=7$：$(1372-400)/3=324=18^2$。✓

所以在 $k=20$ 时 $(k-2x)^2=18^2$。由于 $x<y$ 意味着 $2x<x+y=k$，故 $k-2x>0$，因此 $k-2x=18$，得 $x=1$、$y=19$。验算：$1^3+19^3=1+6859=6860=20\times343=20\times7^3$。所以

$$x=1,\quad y=19,\quad z=7$$

是 $(*)$ 在 $k=20$ 时的一个解。

第 (ii) 部分

现设 $x+y=z^2$。则

$$kz^3=(x+y)(x^2-xy+y^2)=z^2(x^2-xy+y^2),$$

两边除以 z^2（由于 $z>0$，这是容许的），得 $x^2-xy+y^2=kz$。

> ? **想一想** 怎样才能令同样的“完全平方”技巧在这里奏效？

利用恒等式 $4(x^2-xy+y^2)=(x+y)^2+3(x-y)^2$，这其实就是 (i) 中所得结果的另一种面貌（展开右边即可验证）。代入 $x+y=z^2$ 和 $k=19$，得

$$76z=z^4+3(y-x)^2,\qquad \text{即}\qquad 3(y-x)^2=z\,(76-z^3).$$

由于 $y>x$，左边为正，故 $z^3<76$，因此 $z\in\{1,2,3,4\}$。我们需要 $z(76-z^3)/3$ 为完全平方数：

- $z=1$：$75/3=25=5^2$，所以 $y-x=5$；但 $x+y=1$，对正整数而言不可能。
- $z=2$：$2\times68/3=136/3$，不是整数。

- $z=3$：$3\times 49/3=49=7^2$，所以 $y-x=7$ 且 $x+y=9$，得 $x=1$、$y=8$。
- $z=4$：$4\times 12/3=16=4^2$，所以 $y-x=4$ 且 $x+y=16$，得 $x=6$、$y=10$。

验算：$1^3+8^3=513=19\times 27=19\times 3^3$，以及 $6^3+10^3=1216=19\times 64=19\times 4^3$。所以 $(*)$ 在 $k=19$ 时的两个解为

$$(x,y,z)=(1,8,3)\quad 及\quad (x,y,z)=(6,10,4).$$

（事实上，上述论证显示在 $x+y=z^2$ 的条件下，只有这两个解。）

验算

- 用一个简单例子检验 (i) 的恒等式：$k=5$、$x=2$、$y=3$ 给出 $x^3+y^3=35=5\times 7$，而 $k^2-3kx+3x^2=25-30+12=7$。✓
- 如上文所做，把每个答案代回 $(*)$：$6860=20\times 343$、$513=19\times 27$、$1216=19\times 64$。
- 用电脑搜索所有满足 $x+y=20$ 的 $x<y$，以及在 $k=19$ 时所有满足 $x+y=z^2$ 的 $x<y$，结果正好找到这些解，再无其他。

常见错误

- 忘记正负号：$(k-2x)^2=18^2$ 也容许 $k-2x=-18$，但这会给出 $x=19>y$，被条件 $x<y$ 排除。
- 在 (ii) 中把 $z=1$ 保留为解：$y-x=5$ 本身没有问题，但它不能与 $x+y=1$ 同时成立。

STEP 之外

像 $x^3+y^3=kz^3$ 这类方程，问的是哪些整数 k 可以写成两个有理数（rational number）立方之和。这个问题直接通向椭圆曲线（elliptic curves），至今仍未完全解决；著名的的士数（taxicab number）$1729=1^3+12^3=9^3+10^3$ 也属于同一领域。这里用到的技巧（因式分解、配方，再利用上下界把搜索范围收窄至有限）是初等数论（elementary number theory）的基本功。Davenport 的 *The Higher Arithmetic*（《高等算术》）是很好的下一步读物。

详细解答

STEP II 2011，第 3 题

纯数学 · 以单调性证明不等式

> **STEP II 2011，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/11-s2-q3

关键思路

这里每一条不等式（inequality）都源自同一个原则：如果一个函数（function）从零出发，而其导数（derivative）非负，那么它会一直保持非负。最后两个结果再加了一点变化。一旦知道某函数是递增（increasing）的，我们便可以代入两个早已知道大小次序的输入值（例如 $x \leq \arcsin x$），从而读出输出值之间的不等式。看出要比较哪两个输入值，就是整个技巧所在。

解答

第 (i)(a) 部分

求导，并对 $x\cos x$ 使用乘积法则（product rule），

$$f'(x) = \cos x - (\cos x - x\sin x) = x\sin x.$$

当 $0 \leq x \leq \frac{1}{2}\pi$ 时，$x \geq 0$ 及 $\sin x \geq 0$ 均成立，所以 $f'(x) \geq 0$，因此 f 在此区间上递增。

故此，当 $0 \leq x \leq \frac{1}{2}\pi$ 时，我们有 $x \geq 0$，所以 $f(x) \geq f(0) = \sin 0 - 0 = 0$，得证。

第 (i)(b) 部分

设 $h(x) = \arcsin x - x$，其中 $0 \leq x < 1$。利用题目给出的事实，

$$h'(x) = \frac{\mathrm{d}}{\mathrm{d}x}(\arcsin x) - 1 \geq 1 - 1 = 0,$$

所以 h 递增，且 $h(x) \geq h(0) = 0$。即是说，当 $0 \leq x < 1$ 时，$\arcsin x \geq x$。

（题目给出的事实其实就是 $\dfrac{1}{\sqrt{1-x^2}} \geq 1$，这成立是因为 $0 < \sqrt{1-x^2} \leq 1$。）

第 (i)(c) 部分

记 $g(x) = \dfrac{x}{\sin x}$。当 $0 < x < \frac{1}{2}\pi$ 时，由商法则（quotient rule）得

$$g'(x) = \frac{\sin x - x\cos x}{\sin^2 x} = \frac{f(x)}{\sin^2 x}.$$

由 (a) 部分，此处 $f(x) \geq 0$，而 $\sin^2 x > 0$，所以 $g'(x) \geq 0$，因此 g 在 $(0, \frac{1}{2}\pi)$ 上递增。

> **想一想** 知道 g 是递增的，对处理一条涉及 $\arcsin x$ 的不等式有甚么帮助?

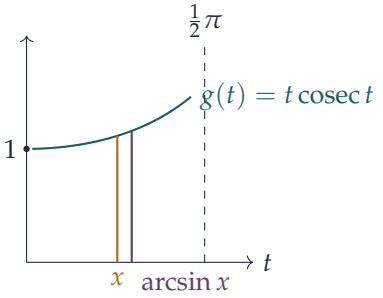
看看目标：$\dfrac{\arcsin x}{x} \geq \dfrac{x}{\sin x}$。右方是 $g(x)$。左方同样是 g 的一个值。若 $y = \arcsin x$，则 $\sin y = x$，所以

$$g(y) = \frac{y}{\sin y} = \frac{\arcsin x}{x}.$$

因此我们需要证明 $g(\arcsin x) \geq g(x)$；由于 g 递增，这可由 (b) 部分的 $\arcsin x \geq x$ 推出。

最后，取 $0 < x < 1$，并设 $y = \arcsin x$。于是 $0 < y < \frac{1}{2}\pi$。另外 $0 < x < 1 < \frac{1}{2}\pi$，所以 x 和 y 都位于 g 递增的区间内。由 (b) 部分，$y \geq x$，所以 $g(y) \geq g(x)$：

$$(\arcsin x)\, x^{-1} = \frac{y}{\sin y} \geq \frac{x}{\sin x} = x \operatorname{cosec} x \qquad (0 < x < 1).$$



上图展示 $x = 0.8$ 时的论证：输入值 $\arcsin x$ 位于 x 的右方，所以递增曲线在那里较高。

第 (ii) 部分

我们照搬 (i) 的模式。首先是 (b) 的类比。设 $m(x) = x - \arctan x$，其中 $x \geq 0$。由题目给出的事实，$m'(x) = 1 - \dfrac{\mathrm{d}}{\mathrm{d}x}(\arctan x) \geq 0$，所以 $m(x) \geq m(0) = 0$。因此

$$\arctan x \leq x \qquad (x \geq 0).$$

接着是 (c) 的类比。设 $k(x) = x^{-1}\tan x$，其中 $0 < x < \frac{1}{2}\pi$。则

$$k'(x) = \frac{x\sec^2 x - \tan x}{x^2} = \frac{x - \sin x\cos x}{x^2\cos^2 x} = \frac{2x - \sin 2x}{2x^2\cos^2 x},$$

这里先将分子分母同乘以 $\cos^2 x$（它是正数），再同乘以 2。

> **想一想** 为甚么 $2x - \sin 2x \geq 0$?

这就是熟悉的“当 $t \geq 0$ 时 $\sin t \leq t$”，取 $t = 2x$ 即可。它可以用与其他部分相同的方式证明：$p(t) = t - \sin t$ 满足 $p'(t) = 1 - \cos t \geq 0$ 及 $p(0) = 0$，所以当 $t \geq 0$ 时 $p(t) \geq 0$。

所以 $k'(x)$ 的分子非负，分母为正，从而 $k'(x) \geq 0$。因此 k 在 $(0, \frac{1}{2}\pi)$ 上递增。

现在取 $0 < x < \frac{1}{2}\pi$，并设 $y = \arctan x$。于是 $0 < y < \frac{1}{2}\pi$，所以 x 和 y 都位于 k 递增的区间内，且 $\tan y = x$。由于 $y \leq x$，我们得到 $k(y) \leq k(x)$：

$$\frac{x}{\arctan x} = \frac{\tan y}{y} \leq \frac{\tan x}{x}.$$

x 和 $\arctan x$ 都是正数，所以两边同乘以 $x \arctan x$ 不会改变不等号的方向：

$$(\tan x)(\arctan x) \geq x^2 \qquad (0 < x < \tfrac{1}{2}\pi),$$

得证。

验算

- 在 (c) 中取 $x = \frac{1}{2}$：$\arcsin(\frac{1}{2}) / \frac{1}{2} = \pi/3 \approx 1.0472$，而 $\frac{1}{2}\operatorname{cosec}\frac{1}{2} \approx 1.0429$。不等式成立，而且只是勉强成立，这正符合预期：当 $x \to 0$ 时两边都趋向 1。
- 在 (ii) 中取 $x = 1$：$\tan 1 \cdot \arctan 1 = \frac{1}{4}\pi \tan 1 \approx 1.2232 \geq 1$。
- 当 x 很小时，$\tan x \approx x + \frac{1}{3}x^3$ 及 $\arctan x \approx x - \frac{1}{3}x^3$，所以两者之积为 $x^2 + O(x^6)$：(ii) 中的不等式在 0 附近极之紧。我们在 20,000 个随机点作数值验算，确认了上述每一条不等式和每一条导数公式。

常见错误

- 在 (c) 中比较 g 在错误的一对点上的值，或忘记检查 $\arcsin x$ 是否位于已知 g 递增的区间 $(0, \frac{1}{2}\pi)$ 之内。
- 在 (ii) 中，输入值之间的不等式方向相反（$\arctan x \leq x$），所以输出值之间的不等式方向也相反。若把方向弄反，就会得出错误的不等式。
- 将不等式两边同乘以某数时，没有说明该乘数（$x \arctan x$ 或 $\cos^2 x$）是正数。

STEP 之外

贯穿全题的原则——非负导数迫使函数递增——是均值定理（mean value theorem）的推论，任何分析（analysis）入门课程都会证明这条定理。比较 $g(x)$ 与 $g(\arcsin x)$ 的技巧（其中 sin 与 arcsin 互为反函数），与你日后在反函数（inverse functions）和凸性（convexity）中会遇到的想法一脉相承。Abbott 的 *Understanding Analysis*（《理解分析》）对均值定理及其推论给出了严谨而易读的证明。

详细解答

STEP II 2011，第 4 题

纯数学 · 三角方程

STEP II 2011，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/11-s2-q4

关键思路

把 $\cos\theta$ 写成某个相关角的正弦（或余弦），这样“$\cos\theta = \sin 4\theta$”就变成两个余弦相等的方程，可以精确求解。将 $\sin 4\theta$ 展开成 $\sin\theta$ 的幂，会得到一个三次方程（cubic），其根正是这些角的正弦值，像 $\sin 18^\circ$ 这样的根式（surd）就是这样出现的。第 (ii) 部分是一个关于 $\sin^2 x$ 的二次方程（quadratic）；而第 (iii) 部分只要找到一个正弦平方等于 $\frac{1}{4}$ 的角，就会发现它其实是乔装了的第 (ii) 部分。

解答

第 (i) 部分

为了把同类的东西放在一起比较，写 $\sin 4\theta = \cos(90^\circ - 4\theta)$。两个角的余弦相等，当且仅当它们在相差 360° 的倍数之下相等或互为相反数，因此

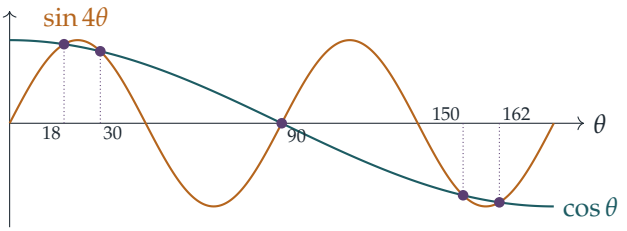
$$\cos\theta = \cos(90^\circ - 4\theta) \iff \theta = \pm(90^\circ - 4\theta) + 360^\circ k \quad \text{对某整数} k.$$

- 取 $+$ 号：$5\theta = 90^\circ + 360^\circ k$，所以 $\theta = 18^\circ + 72^\circ k$。在 $0^\circ < \theta < 180^\circ$ 内得 $\theta = 18^\circ, 90^\circ, 162^\circ$。
- 取 $-$ 号：$3\theta = 90^\circ - 360^\circ k$，所以 $\theta = 30^\circ - 120^\circ k$，得 $\theta = 30^\circ, 150^\circ$。

因此解为

$$\theta = 18^\circ, 30^\circ, 90^\circ, 150^\circ, 162^\circ.$$

下图证实两条曲线恰好相交五次。



接下来处理“由此”的部分。设 $s = \sin\theta$。两次运用倍角公式（double-angle formulae），

$$\sin 4\theta = 2\sin 2\theta\cos 2\theta = 4\sin\theta\cos\theta\,(1 - 2\sin^2\theta),$$

于是方程 $\cos\theta = \sin 4\theta$ 变成

$$\cos\theta\big(1 - 4s(1 - 2s^2)\big) = 0.$$

想一想 为甚么可以约去 $\cos\theta$?

在五个解之中，只有 $\theta = 90°$ 使 $\cos\theta = 0$。其余四个解 $18°, 30°, 150°, 162°$ 都有 $\cos\theta \neq 0$，所以对它们可以除以 $\cos\theta$，得

$$8s^3 - 4s + 1 = 0.$$

我们知道 $s = \sin 30° = \frac{1}{2}$ 是其中一个根，因式分解（factorising）得

$$8s^3 - 4s + 1 = (2s - 1)(4s^2 + 2s - 1).$$

$\sin 18°$ 是这个三次方程的根，而由于正弦在 $[0°, 90°]$ 上递增，$0 < \sin 18° < \sin 30° = \frac{1}{2}$。所以 $\sin 18° \neq \frac{1}{2}$，它必定是 $4s^2 + 2s - 1 = 0$ 的根，而这个方程的根为

$$s = \frac{-2 \pm \sqrt{4 + 16}}{8} = \frac{-1 \pm \sqrt{5}}{4}.$$

只有取 $+$ 号的根是正数，所以

$$\sin 18° = \tfrac{1}{4}\big(\sqrt{5} - 1\big),$$

得证。

第 (ii) 部分

设 $s = \sin x$。由于 $\sin^2 2x = 4\sin^2 x\cos^2 x = 4s^2(1 - s^2)$，方程只取决于 s：

$$4s^2 + 1 = 16s^2(1 - s^2) \iff 16s^4 - 12s^2 + 1 = 0.$$

这是关于 s^2 的二次方程：

$$s^2 = \frac{12 \pm \sqrt{144 - 64}}{32} = \frac{12 \pm 4\sqrt{5}}{32} = \frac{3 \pm \sqrt{5}}{8}.$$

要开平方根，留意

$$\left(\frac{1+\sqrt{5}}{4}\right)^2 = \frac{6+2\sqrt{5}}{16} = \frac{3+\sqrt{5}}{8}, \qquad \left(\frac{\sqrt{5}-1}{4}\right)^2 = \frac{6-2\sqrt{5}}{16} = \frac{3-\sqrt{5}}{8}.$$

s^2 的两个值都介乎 0 与 1 之间，所以四个平方根每一个都确实是 $\sin x$ 可取的值；又因为以上每一步都可逆，它们都满足原方程。可能的值为

$$\sin x = \tfrac{1}{4}+\tfrac{1}{4}\sqrt{5}, \qquad -\tfrac{1}{4}-\tfrac{1}{4}\sqrt{5}, \qquad -\tfrac{1}{4}+\tfrac{1}{4}\sqrt{5}, \qquad \tfrac{1}{4}-\tfrac{1}{4}\sqrt{5}.$$

由第 (i) 部分，$\frac{1}{4}(\sqrt{5}-1) = \sin 18°$；而 $\frac{1}{4}(1+\sqrt{5}) = \sin 54°$（它等于 $\cos 36° = 1 - 2\sin^2 18°$，你可以验证两者结果相同）。

第 (iii) 部分

把第 (ii) 部分的方程除以 4：

$$\sin^2 x + \tfrac{1}{4} = \sin^2 2x.$$

将它与 $\sin^2 3\alpha + \sin^2 5\alpha = \sin^2 6\alpha$ 比较。若设 $x = 3\alpha$，则自然有 $2x = 6\alpha$，而只要

$$\sin^2 5\alpha = \tfrac{1}{4} \qquad 及 \qquad \sin 3\alpha \text{ 是第 (ii) 部分四个值之一},$$

两条方程便吻合。

想一想 $\sin 3\alpha$ 是那些值之一就足够了吗?

足够：在第 (ii) 部分我们看到，方程 $\sin^2 x + \frac{1}{4} = \sin^2 2x$ 只通过 $s = \sin x$ 依赖于 x，而那四个 s 值都满足它。所以任何正弦取这些值之一的 x 都可行。

取 $\alpha = 6°$。则 $5\alpha = 30°$，所以 $\sin^2 5\alpha = \frac{1}{4}$；而 $3\alpha = 18°$，所以 $\sin 3\alpha = \frac{1}{4}(\sqrt{5}-1)$，正是我们的值之一。因此

$$\sin^2 18° + \sin^2 30° = \sin^2 36°,$$

$\alpha = 6°$ 是一个解。

取 $\alpha = 42°$。则 $5\alpha = 210°$，所以 $\sin 5\alpha = -\frac{1}{2}$，$\sin^2 5\alpha = \frac{1}{4}$；而 $3\alpha = 126°$，所以 $\sin 3\alpha = \sin 54° = \frac{1}{4}(1+\sqrt{5})$，同样是我们的值之一。所以 $\alpha = 42°$ 是一个解。

因此两个合适的值是 $\alpha = 6°$ 和 $\alpha = 42°$。

同样的方法亦给出 $\alpha = 66°$（$5\alpha = 330°$，$\sin 3\alpha = \sin 198° = -\sin 18°$）及 $\alpha = 78°$（$5\alpha = 390°$，$\sin 3\alpha = \sin 234° = -\sin 54°$）。范围内另一个满足 $\sin^2 5\alpha = \frac{1}{4}$ 的角是 $\alpha = 30°$,但它不成立,因为 $\sin 90° = 1$ 不是我们的值之一。(事实上 $1 + \frac{1}{4} \neq \sin^2 180° = 0$。)

验算

► $\sin 18° \approx 0.30902$，而 $\frac{1}{4}(\sqrt{5}-1) \approx 0.30902$。

- 对 $\alpha = 6°$：$\sin^2 18° + \sin^2 30° \approx 0.09549 + 0.25 = 0.34549$，而 $\sin^2 36° \approx 0.34549$。
- 在 $(0°, 90°)$ 上以数值方法扫描 $\sin^2 3\alpha + \sin^2 5\alpha - \sin^2 6\alpha$，找到的零点恰好是 $6°, 42°, 45°, 66°, 78°$。多出来的 $45°$ 并非来自第 (ii) 部分：在该处 $\frac{1}{2} + \frac{1}{2} = 1$。

常见错误

- 忘记 $\cos A = \cos B$ 的第二组解（即 $A = -B + 360°k$），因而在第 (i) 部分漏掉 $30°$ 和 $150°$。
- 不加说明就约去 $\cos\theta$。这里这样做没问题，但只是因为使 $\cos\theta = 0$ 的解是 $90°$，而不是 $18°$。
- 在第 (ii) 部分只给出 $\sin x$ 的正值。四个值全部都是可能的。

STEP 之外

$\sin 18°$ 之所以有精确值，是因为 $18°$ 是直角的五分之一，因此它与正五边形（regular pentagon）及黄金比例（golden ratio）息息相关：$\frac{1}{4}(1+\sqrt{5}) = \frac{1}{2}\varphi$。更一般地，$\cos(360°/n)$ 能只用平方根表示，当且仅当正 n 边形可以用直尺和圆规作出（尺规作图，ruler and compasses）；高斯（Gauss）证明了这取决于整除 n 的费马质数（Fermat primes）。这是域论（field theory）和伽罗瓦理论（Galois theory）的经典入门应用。Allenby 的 *Rings, Fields and Groups*（《环、域与群》）逐步建立所需的代数，而 Needham 的 *Visual Complex Analysis*（《直观复分析》）则说明为甚么单位根（roots of unity）与多边形密不可分。

详细解答

STEP II 2011，第 5 题

纯数学 · 向量与反射

> **STEP II 2011，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/11-s2-q5

关键思路

要把一点关于一条经过原点的直线作反射（reflection），可把它的位置向量（position vector）分解为沿该直线的部分和与该直线垂直的部分：反射保留前者，并把后者反向。$\mathbf{b}$ 沿 $\mathbf{a}$ 方向的分量（component）就是我们熟悉的投影（projection）$\frac{\mathbf{a}\cdot\mathbf{b}}{\mathbf{a}\cdot\mathbf{a}}\,\mathbf{a}$，其余的都只是点积（dot product）的运算整理。至于共线（collinear）问题，记住：一点位于直线 AB 上，当且仅当它的位置向量可写成 $s\mathbf{a}+t\mathbf{b}$，其中 $s+t=1$。

解答

整个解答中，$\mathbf{a}$ 和 $\mathbf{b}$ 均非零且互不平行，因为 O、A、B 不共线。特别地，$\mathbf{a}\cdot\mathbf{a}\neq 0$ 且 $\mathbf{b}\cdot\mathbf{b}\neq 0$，所以可以除以它们。记 $\theta=\angle AOB$。

反射点 C

设 P 为从 B 到直线 OA 的垂足（foot of the perpendicular）。它的位置向量是某个倍数 $k\mathbf{a}$，而 $\mathbf{b}-k\mathbf{a}$ 与 $\mathbf{a}$ 垂直：

$$(\mathbf{b}-k\mathbf{a})\cdot\mathbf{a}=0 \quad\Longrightarrow\quad k=\frac{\mathbf{a}\cdot\mathbf{b}}{\mathbf{a}\cdot\mathbf{a}}.$$

关于直线 OA 的反射保持 P 不动，并把 B 映到另一侧的点 C，使 P 为 BC 的中点。所以 $\mathbf{c}=2\overrightarrow{OP}-\mathbf{b}$，即

$$\mathbf{c}=\frac{2\,\mathbf{a}\cdot\mathbf{b}}{\mathbf{a}\cdot\mathbf{a}}\,\mathbf{a}-\mathbf{b}=\lambda\mathbf{a}-\mathbf{b},\qquad \lambda=\frac{2\,\mathbf{a}\cdot\mathbf{b}}{\mathbf{a}\cdot\mathbf{a}},$$

即为所求。

反射点 D

以直线 OB 代替 OA，同样的论证给出

$$\mathbf{d} = \frac{2\,\mathbf{b}\cdot\mathbf{c}}{\mathbf{b}\cdot\mathbf{b}}\,\mathbf{b} - \mathbf{c}.$$

而 $\mathbf{b}\cdot\mathbf{c} = \lambda\,\mathbf{a}\cdot\mathbf{b} - \mathbf{b}\cdot\mathbf{b}$，所以

$$\begin{aligned}\mathbf{d} &= \left(\frac{2\lambda\,\mathbf{a}\cdot\mathbf{b}}{\mathbf{b}\cdot\mathbf{b}} - 2\right)\mathbf{b} - (\lambda\mathbf{a} - \mathbf{b}) \\ &= \left(\frac{2\lambda\,\mathbf{a}\cdot\mathbf{b}}{\mathbf{b}\cdot\mathbf{b}} - 1\right)\mathbf{b} - \lambda\mathbf{a}.\end{aligned}$$

这就是 $\mathbf{d} = \mu\mathbf{b} - \lambda\mathbf{a}$，其中

$$\mu = \frac{2\lambda\,\mathbf{a}\cdot\mathbf{b}}{\mathbf{b}\cdot\mathbf{b}} - 1 = \frac{4(\mathbf{a}\cdot\mathbf{b})^2}{(\mathbf{a}\cdot\mathbf{a})(\mathbf{b}\cdot\mathbf{b})} - 1 = 4\cos^2\theta - 1,$$

最后一步用了 $\mathbf{a}\cdot\mathbf{b} = |\mathbf{a}||\mathbf{b}|\cos\theta$。下面需要用到的正是 $4\cos^2\theta - 1$ 这个形式。

A、B、D 共线

点 D 位于直线 AB 上，当且仅当对某个纯量（scalar）t 有 $\mathbf{d} = \mathbf{a} + t(\mathbf{b} - \mathbf{a}) = (1-t)\mathbf{a} + t\mathbf{b}$。另一方面，我们亦有 $\mathbf{d} = -\lambda\mathbf{a} + \mu\mathbf{b}$。

> ? **想一想** 为甚么可以比较 $\mathbf{a}$ 和 $\mathbf{b}$ 的系数（coefficient）？

因为 $\mathbf{a}$ 和 $\mathbf{b}$ 互不平行。若 $s\mathbf{a} + t\mathbf{b} = s'\mathbf{a} + t'\mathbf{b}$，则 $(s - s')\mathbf{a} = (t' - t)\mathbf{b}$；除非两个括号都等于零，否则这会使 $\mathbf{a}$ 和 $\mathbf{b}$ 平行。这正是用到 O、A、B 不共线的地方。

所以 $-\lambda = 1 - t$ 且 $\mu = t$。消去 t，可知 A、B、D 共线当且仅当

$$\mu = \lambda + 1.$$

$\lambda = -\frac{1}{2}$ 的情况

此时共线条件给出 $\mu = \frac{1}{2}$，所以 $4\cos^2\theta - 1 = \frac{1}{2}$，即 $\cos^2\theta = \frac{3}{8}$。由于 $\lambda < 0$，我们有 $\mathbf{a}\cdot\mathbf{b} < 0$，所以 $\cos\theta < 0$，故

$$\cos\angle AOB = -\sqrt{\tfrac{3}{8}} = -\frac{\sqrt{6}}{4}.$$

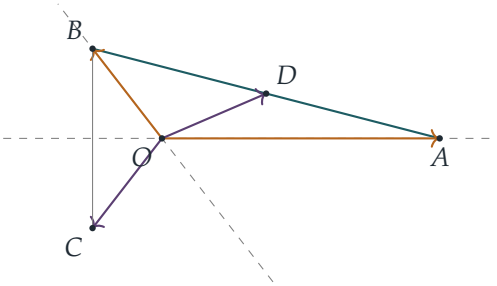
于是

$$\mathbf{d} = \tfrac{1}{2}\mathbf{b} + \tfrac{1}{2}\mathbf{a} = \tfrac{1}{2}(\mathbf{a} + \mathbf{b}),$$

所以 D **是** AB **的中点**。关于图形的形状，我们还可以说得更多。由 $\lambda = \dfrac{2|\mathbf{b}|\cos\theta}{|\mathbf{a}|} = -\frac{1}{2}$，

$$\frac{|\mathbf{a}|}{|\mathbf{b}|} = -4\cos\theta = \sqrt{6},$$

所以 $OA = \sqrt{6}\,OB$，而角 AOB 是钝角（obtuse angle）（约 127.8°）。此外，反射保持与 O 的距离不变，所以 $OD = OC = OB$：AB 的中点到 O 的距离恰好与 B 相同。



验算

- 先关于 OA、再关于 OB 反射两次，等同于绕 O 旋转（rotation）2θ。所以 D 应是 B 旋转 2θ 后的位置，且 $\cos\angle BOD = \cos 2\theta = 2\cos^2\theta - 1$。在最后的情况中，这等于 $2\cdot\frac{3}{8} - 1 = -\frac{1}{4}$。直接计算，$\mathbf{b}\cdot\frac{1}{2}(\mathbf{a}+\mathbf{b}) = \frac{1}{2}(-\frac{3}{2}+1)|\mathbf{b}|^2 = -\frac{1}{4}|\mathbf{b}|^2$，其中用了 $\mathbf{a}\cdot\mathbf{b} = -\frac{1}{4}\,\mathbf{a}\cdot\mathbf{a} = -\frac{3}{2}\,\mathbf{b}\cdot\mathbf{b}$，结果吻合。
- 若 $\theta = 90°$，则 $\lambda = 0$，C 是 $-\mathbf{b}$，而 D 亦是 $-\mathbf{b}$。公式给出 $\mu = 4\cdot 0 - 1 = -1$，与预期相符。
- 数值上，取 $\mathbf{a} = (\sqrt{6}, 0)$ 及 $\mathbf{b} = (-0.6124,\ 0.7906)$，反射两次得 $\mathbf{d} = (0.9186,\ 0.3953)$，这正是 AB 的中点，而且 $|\mathbf{d}| = 1 = |\mathbf{b}|$。

常见错误

- 把反射写成 $\mathbf{b} - 2\times$（投影）。这其实是关于与 OA 垂直的直线作反射。可用 $\mathbf{b} = \mathbf{a}$ 来测试：它应保持不动。
- 取 $\cos\theta = +\sqrt{6}/4$。λ 的正负决定了 $\mathbf{a}\cdot\mathbf{b}$ 的正负。
- 比较系数时没有说明 $\mathbf{a}$ 和 $\mathbf{b}$ 线性无关（linearly independent）。

STEP 之外

映射 $\mathbf{x} \mapsto \frac{2\,\mathbf{a}\cdot\mathbf{x}}{\mathbf{a}\cdot\mathbf{a}}\,\mathbf{a} - \mathbf{x}$ 是一个线性映射（linear map），它的矩阵是大学一年级课程中行列式（determinant）为 -1 的正交矩阵（orthogonal matrix）的标准例子。两次反射的复合是一个旋转，旋转角为两面镜子之间夹角的两倍。这是研究等距变换（isometry）以及由反射生成的群（group）——例如正多边形的对称群——的起点。Strang 的 *Introduction to Linear Algebra*（《线性代数导论》）对投影和反射矩阵有很好的讲解。

详细解答

STEP II 2011，第 6 题

纯数学 · 分部积分法

STEP II 2011，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/11-s2-q6

关键思路

被积函数（integrand）$[f'(x)]^2[f(x)]^n$ 只差一个 f' 的因子，就会变成一眼便能积出来的形式，因为 $f'(x)[f(x)]^n$ 正是 $[f(x)]^{n+1}/(n+1)$ 的导数（derivative）。所以我们用分部积分（integration by parts），把这一部分拆出来。剩下的积分含有 f''，而条件 $f''=kff'$ 又把它变回同样容易处理的形式 $f'f^m$。至于各个应用，关键就在于选取 f 和 n，使所给的被积函数等于 $[f']^2f^n$，再求出 k。

解答

一般结果

以 f 代表 $f(x)$，其余类推。用分部积分，取 $u=f'$ 及 $\dfrac{\mathrm{d}v}{\mathrm{d}x}=f'f^n$，于是 $\dfrac{\mathrm{d}u}{\mathrm{d}x}=f''$ 及 $v=\dfrac{f^{n+1}}{n+1}$（这没有问题，因为 $n\geq 1$ 表示 $n+1\neq 0$）：

$$\begin{aligned} I=\int f'\cdot f'f^n\,\mathrm{d}x &= \frac{f'f^{n+1}}{n+1}-\frac{1}{n+1}\int f''f^{n+1}\,\mathrm{d}x \\ &= \frac{f'f^{n+1}}{n+1}-\frac{k}{n+1}\int f'f^{n+2}\,\mathrm{d}x \qquad \text{利用}f''=kff' \\ &= \frac{f'f^{n+1}}{n+1}-\frac{kf^{n+3}}{(n+1)(n+3)}+C. \end{aligned}$$

因此

$$\boxed{I=\frac{f'(x)\,[f(x)]^{n+1}}{n+1}-\frac{k\,[f(x)]^{n+3}}{(n+1)(n+3)}+C}$$

这正是题目要求的、以 $f(x)$、$f'(x)$、k 和 n 表示的式子。

第 (i) 部分

对于 $f(x)=\tan x$，我们有 $f'(x)=\sec^2 x$ 及

$$f''(x)=2\sec x\cdot\sec x\tan x=2\tan x\sec^2 x=2f(x)f'(x),$$

所以条件成立，且 $k=2$。代入公式得

$$\int \sec^4 x\tan^n x\,\mathrm{d}x=\frac{\sec^2 x\tan^{n+1}x}{n+1}-\frac{2\tan^{n+3}x}{(n+1)(n+3)}+C. \tag{1}$$

验证。直接计算：利用 $\sec^2 x=1+\tan^2 x$ 及代换（substitution）$t=\tan x,\mathrm{d}t=\sec^2 x\,\mathrm{d}x$：

$$\int \sec^4 x\tan^n x\,\mathrm{d}x=\int(1+t^2)t^n\,\mathrm{d}t=\frac{t^{n+1}}{n+1}+\frac{t^{n+3}}{n+3}+C.$$

用 $\sec^2 x=1+t^2$ 改写 (1) 的右方，得

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

答案相同。所以对 $f=\tan$ 而言，结果得到验证。

所求积分。由于

$$\frac{\sin^4 x}{\cos^8 x}=\frac{\sin^4 x}{\cos^4 x}\cdot\frac{1}{\cos^4 x}=\tan^4 x\sec^4 x=[f'(x)]^2[f(x)]^4,$$

这是 (1) 中 $n=4$ 的情况。利用刚才得到的化简形式，

$$\int\frac{\sin^4 x}{\cos^8 x}\,\mathrm{d}x=\frac{\tan^5 x}{5}+\frac{\tan^7 x}{7}+C.$$

第 (ii) 部分

> ? **想一想** 应该取哪一个 f？那个 $\sec^2 x$ 又从何而来？

括号 $(\sec x+\tan x)$ 显然是 f 的候选，所以试取 $f(x)=\sec x+\tan x$。它的导数是

$$f'(x)=\sec x\tan x+\sec^2 x=\sec x(\tan x+\sec x)=\sec x\,f(x).$$

这就是关键事实：对 f 求导，只不过是把它乘以 $\sec x$。因此

$$\begin{aligned}f''(x)&=\sec x\tan x\,f(x)+\sec x\,f'(x)=\sec x\tan x\,f+\sec^2 x\,f\\&=\sec x\,f\,(\tan x+\sec x)=\sec x\,f^2,\end{aligned}$$

又因为 $f\,f'=f\cdot\sec x\,f=\sec x\,f^2$ 亦成立，所以 $f''=ff'$，即 $k=1$。

此外 $[f'(x)]^2 = \sec^2 x\,[f(x)]^2$，所以

$$\sec^2 x\,(\sec x + \tan x)^6 = \sec^2 x\, f^2 \cdot f^4 = [f'(x)]^2[f(x)]^4,$$

这是一般结果中 $n = 4$、$k = 1$ 的情况：

$$\begin{aligned}\int \sec^2 x\,(\sec x + \tan x)^6\,\mathrm{d}x &= \frac{f'f^5}{5} - \frac{f^7}{35} + C = \frac{\sec x\, f^6}{5} - \frac{f^7}{35} + C \\ &= \frac{f^6}{35}\big(7\sec x - (\sec x + \tan x)\big) + C \\ &= \frac{(\sec x + \tan x)^6\,(6\sec x - \tan x)}{35} + C.\end{aligned}$$

验算

- 对 (i) 的答案求导：$\sec^2 x\tan^4 x + \sec^2 x\tan^6 x = \sec^2 x\tan^4 x(1 + \tan^2 x) = \tan^4 x\sec^4 x$。
- 以数值计算，辛普森法则 (Simpson's rule) 给出 $\int_{-0.7}^{1.1} \sin^4 x/\cos^8 x\,\mathrm{d}x \approx 22.12954$，与 $\left[\frac{1}{5}\tan^5 x + \frac{1}{7}\tan^7 x\right]_{-0.7}^{1.1}$ 吻合；而 $\int_{-1}^{0.9} \sec^2 x(\sec x + \tan x)^6\,\mathrm{d}x \approx 133.68566$，与 (ii) 的答案吻合。
- 对于 $f = \tan$ 及 1 至 6 的每一个 n，一般公式都与数值积分（quadrature）的结果一致。

常见错误

- 在分部积分中选取 $u = f^n$：这样会产生 $f^{n-1}[f']^3$，情况只会更糟。应把一个 f' 与 f^n 配成一对，使它能恰好积出来。
- 对 $\tan x$ 取 $k = 1$：其实 $f'' = 2\sec^2 x\tan x$，所以 $k = 2$。
- 在 (ii) 中，忘记 $\sec^2 x$ 必须以 $[f']^2/f^2$ 的形式吸收进去，这正是指数由 6 降至 $n = 4$ 的原因。

STEP 之外

条件 $f'' = kff'$ 本身就是一条微分方程(differential equation)：积分一次得 $f' = \frac{1}{2}kf^2 + c$，这是一条一阶可分离变量方程（separable equation），其解包括 tan（$k = 2$，$c = 1$）及 $\sec x + \tan x (k = 1, c = \frac{1}{2})$。认识到某个被积函数之所以容易处理，是因为该函数满足某条常微分方程 (ODE)，这个主题在大学一年级有关特殊函数 (special functions) 和递推公式 (reduction formulae) 的课程中会反复出现。Boyce 和 DiPrima 的 *Elementary Differential*

Equations（《初等微分方程》）讲解了像这样的可分离方程和自治方程（autonomous equations）。

详细解答

STEP II 2011，第 7 题

纯数学 · 数列与级数

STEP II 2011，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/11-s2-q7

关键思路

切勿逐项处理这两个数列（sequence）：把每个和拆成 λ 的幂之和与 μ 的幂之和，再分别套用几何级数（geometric series）公式。这两个数满足 $\lambda+\mu=2$ 及 $\lambda\mu=-1$，所以它们是 $x^2=2x+1$ 的根。单凭这条二次方程（quadratic equation），所有棘手的分母都会化简掉。另一个工具是平方：b_k^2 和 a_k^2 都会化成 a_{2k} 加一个常数。

解答

首先列出一些全程都会用到的事实：

$$\lambda-1=\sqrt{2}, \qquad \mu-1=-\sqrt{2}, \qquad \lambda\mu=1-2=-1, \qquad \lambda-\mu=2\sqrt{2}.$$

此外 $a_0=2$ 及 $b_0=0$。

第 (i) 部分

由于 $\lambda\neq 1$ 且 $\mu\neq 1$，我们可以对每个和套用几何级数公式：

$$\sum_{r=0}^{n}\lambda^r=\frac{\lambda^{n+1}-1}{\lambda-1}=\frac{\lambda^{n+1}-1}{\sqrt{2}}, \qquad \sum_{r=0}^{n}\mu^r=\frac{\mu^{n+1}-1}{\mu-1}=-\frac{\mu^{n+1}-1}{\sqrt{2}}.$$

把第一个和减去第二个和，

$$\sum_{r=0}^{n}b_r=\frac{\lambda^{n+1}-1}{\sqrt{2}}+\frac{\mu^{n+1}-1}{\sqrt{2}}=\frac{a_{n+1}-2}{\sqrt{2}}=-\sqrt{2}+\frac{1}{\sqrt{2}}a_{n+1},$$

正是所求。若改为把两个和相加，两个 -1 便互相抵消：

$$\boxed{\sum_{r=0}^{n} a_r = \frac{\lambda^{n+1}-1}{\sqrt{2}} - \frac{\mu^{n+1}-1}{\sqrt{2}} = \frac{1}{\sqrt{2}}\, b_{n+1}.}$$

（快速检验：当 $n=0$ 时，这式表示 $a_0 = b_1/\sqrt{2} = 2\sqrt{2}/\sqrt{2} = 2$，是正确的。）

第 (ii) 部分

由第 (i) 部分，内层的和是 $b_{m+1}/\sqrt{2}$。因此

$$\sum_{m=0}^{2n}\left(\sum_{r=0}^{m} a_r\right) = \frac{1}{\sqrt{2}}\sum_{m=0}^{2n} b_{m+1} = \frac{1}{\sqrt{2}}\sum_{s=1}^{2n+1} b_s = \frac{1}{\sqrt{2}}\sum_{s=0}^{2n+1} b_s,$$

最后一步成立，是因为 $b_0 = 0$。现在以 $2n+1$ 代替 n 应用第 (i) 部分，得

$$\sum_{m=0}^{2n}\left(\sum_{r=0}^{m} a_r\right) = \frac{1}{\sqrt{2}}\left(-\sqrt{2} + \frac{a_{2n+2}}{\sqrt{2}}\right) = \tfrac{1}{2}a_{2n+2} - 1. \qquad (*)$$

? **想一想** 怎样把 a_{2n+2} 与 b_{n+1}^2 联系起来?

我们取平方，并利用 $\lambda\mu = -1$ 处理交叉项：

$$b_{n+1}^2 = \lambda^{2n+2} + \mu^{2n+2} - 2(\lambda\mu)^{n+1} = a_{2n+2} - 2(-1)^{n+1}, \qquad a_{n+1}^2 = a_{2n+2} + 2(-1)^{n+1}.$$

若 n 是奇数，则 $n+1$ 是偶数，所以 $b_{n+1}^2 = a_{2n+2} - 2$。于是 $(*)$ 等于 $\frac{1}{2}(a_{2n+2}-2) = \frac{1}{2}b_{n+1}^2$，正是所求。

若 n 是偶数，则 $(-1)^{n+1} = -1$，所以这时等于 $a_{2n+2} - 2$ 的是 a_{n+1}^2。因此

$$\boxed{n \text{ 为偶数}: \quad \sum_{m=0}^{2n}\left(\sum_{r=0}^{m} a_r\right) = \tfrac{1}{2}a_{n+1}^2}$$

或等价地写成 $\frac{1}{2}b_{n+1}^2 - 2$。

第 (iii) 部分

由第 (i) 部分，再用第 (ii) 部分的平方恒等式，

$$\left(\sum_{r=0}^{n} a_r\right)^2 = \tfrac{1}{2}b_{n+1}^2 = \tfrac{1}{2}a_{2n+2} - (-1)^{n+1}.$$

对于第二个和，我们同样把它拆成两个几何级数，这次公比分别是 λ^2 和 μ^2：

$$\sum_{r=0}^{n} \lambda^{2r+1} = \lambda\,\frac{\lambda^{2n+2}-1}{\lambda^2-1}.$$

想一想 λ^2-1 可以化简吗?

可以。由于 $\lambda^2=2\lambda+1$，得 $\lambda^2-1=2\lambda$，于是因子 λ 被约去，和为 $\frac{1}{2}(\lambda^{2n+2}-1)$。由于 μ 满足同一条二次方程，$\mu^2-1=2\mu$，而 μ 的和为 $\frac{1}{2}(\mu^{2n+2}-1)$（注意 $\mu^2\neq 1$，所以公式适用）。两者相加，

$$\sum_{r=0}^{n} a_{2r+1}=\tfrac{1}{2}\left(a_{2n+2}-2\right)=\tfrac{1}{2}a_{2n+2}-1.$$

因此

$$\left(\sum_{r=0}^{n} a_r\right)^2-\sum_{r=0}^{n} a_{2r+1}=\left(\tfrac{1}{2}a_{2n+2}-(-1)^{n+1}\right)-\left(\tfrac{1}{2}a_{2n+2}-1\right)=1+(-1)^n.$$

若 n 是偶数，此式等于 2，正是所求。若 n 是奇数，此式等于 0，所以

$$\boxed{n\text{ 为奇数}:\quad \left(\sum_{r=0}^{n} a_r\right)^2=\sum_{r=0}^{n} a_{2r+1}.}$$

验算

各 a_n 都是整数，并满足 $a_{n+1}=2a_n+a_{n-1}$：$2,2,6,14,34,82,198,478,\ldots$。在第 (ii) 部分取 $n=3$，部分和 $2,4,10,24,58,140,338$ 加起来是 576。由于 $b_4=24\sqrt{2}$，得 $\frac{1}{2}b_4^2=576$，两者吻合。在第 (iii) 部分取 $n=2$：$(2+2+6)^2-(2+14+82)=100-98=2$。取 $n=1$：$(2+2)^2-(2+14)=0$。我们用电脑验证了全部三个部分，包括奇偶两种情况（parity），直至 n 等于 24 的每一个值。

常见错误

- 忘记 $\mu-1=-\sqrt{2}$ 是负数，以致在第 (i) 部分把 a 和 b 的角色对调。
- 写成 $(\lambda\mu)^{n+1}=1$ 而不是 $(-1)^{n+1}$；这个正负号正是要按奇偶分情况的关键所在。
- 在第 (ii) 部分移动 b_{m+1} 之和的下标时，没有留意到正是 $b_0=0$ 才容许和由 0 开始。

STEP 之外

两个数列都满足线性递推关系（linear recurrence）$x_{n+1}=2x_n+x_{n-1}$，其特征方程（characteristic equation）为 $x^2=2x+1$。所以 λ 和 μ 是它的根，而 a_n 和 b_n 是特解。由于 $\lambda^n=\frac{1}{2}(a_n+b_n)$，恒等式 $\lambda^n\mu^n=(-1)^n$ 便化为 $a_n^2-b_n^2=4(-1)^n$，这是环（ring）$\mathbb{Z}[\sqrt{2}]$

中的一条佩尔型方程（Pell-type equation）。Davenport 的 *The Higher Arithmetic*（《高等算术》）阐述了这类单位（units）与 $\sqrt{2}$ 的连分数（continued fractions）之间如何互相配合。

详细解答

STEP II 2011，第 8 题

纯数学 · 圆的渐伸线

> **STEP II 2011，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/11-s2-q8

关键思路

在任何时刻，绳子都由两段组成：一段是圆上的弧，其长度可以由角度直接读出；另一段是沿切线（tangent）离开圆的直线段。由于总长度固定，直线段的长度就是剩下的部分；而切线与半径垂直，所以用向量（vector）就能轻易定出 B 的位置。至于最后的面积，留意整个被扫过的区域就是“外支下方的面积，减去内支下方的面积，再减去半圆”。

解答

参数方程

设 $P = (\cos t, \sin t)$ 为绳子离开圆的点，因此 t 是由正 x 轴到 OP 的角度。开始时 $P = A = (-1, 0)$，所以 $t = \pi$；当绳子从圆的顶部绕过去时，P 沿顺时针方向移动，t 递减至 0。

与圆接触的那段绳子是由角度 π 到角度 t 的弧。半径为 1，所以这段弧长为 $\pi - t$。绳子长度为 π，所以直线段 PB 的长度为

$$\pi - (\pi - t) = t.$$

由于绳子拉紧，PB 是圆在 P 点的切线，因此它垂直于 $OP = (\cos t, \sin t)$。与 OP 垂直的两个单位向量是 $\pm(\sin t, -\cos t)$。开始时（$t = \pi$）绳子由 A 垂直向上，方向为 $(0, 1) = (\sin \pi, -\cos \pi)$，而根据连续性（continuity），符号在整个过程中保持不变。因此

$$\overrightarrow{OB} = \overrightarrow{OP} + t\,(\sin t, -\cos t),$$

即

$$x = \cos t + t \sin t, \qquad y = \sin t - t \cos t,$$

正如所求。（验算：$t=\pi$ 给出 $(-1,\pi)$，即 B 的起始位置；$t=0$ 给出 $(1,0)$，即 B 碰到圆的位置。）

x 的最大值

求导后，$\pm\sin t$ 与 $\pm\cos t$ 各项互相抵消：

$$\frac{\mathrm{d}x}{\mathrm{d}t}=-\sin t+\sin t+t\cos t=t\cos t,\qquad \frac{\mathrm{d}y}{\mathrm{d}t}=\cos t-\cos t+t\sin t=t\sin t.$$

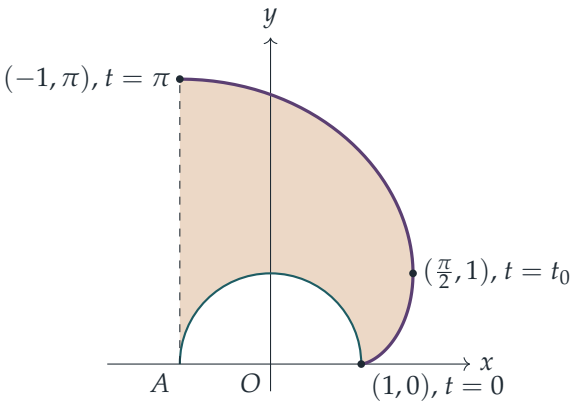
当 $0<t<\pi$ 时，$\frac{\mathrm{d}x}{\mathrm{d}t}=t\cos t$ 在 $t<\frac{\pi}{2}$ 时为正，在 $t>\frac{\pi}{2}$ 时为负。所以当 t 取遍 $[0,\pi]$ 时，x 先递增后递减，最大值出现在

$$t_0=\frac{\pi}{2},\qquad \text{此时}(x,y)=\left(\frac{\pi}{2},1\right).$$

草图

有两个事实有帮助。第一，当 $0<t\le\pi$ 时 $y>0$：若 $t\le\frac{\pi}{2}$，则由于 $\tan t>t$，$y=\cos t(\tan t-t)>0$（当 $t=\frac{\pi}{2}$ 时直接有 $y=1$）；若 $t>\frac{\pi}{2}$，则 $\cos t<0$，所以 y 的两项都非负，且 $\sin t-t\cos t>0$。第二，$x^2+y^2=1+t^2$，所以 B 永远在圆外，而且离圆的距离随 t 增大。此外 $\frac{\mathrm{d}y}{\mathrm{d}t}=t\sin t\ge 0$，所以 y 随 t 稳定递增。

因此曲线由 $(1,0)$ 出发（该处 $\frac{\mathrm{d}y}{\mathrm{d}x}=\tan t=0$，所以它水平地离开 x 轴），上升至最右点 $(\frac{\pi}{2},1)$，该处切线垂直，然后扫过圆的顶部回到 $(-1,\pi)$，该处再次水平。这条曲线是圆的渐伸线（involute）。



阴影区域介乎曲线、绳子的初始位置（虚线）与半圆之间，就是绳子扫过的面积。

$\pi \geq t \geq t_0$ 部分的面积

在曲线的这一部分，x 由 $\frac{\pi}{2}$（在 $t=\frac{\pi}{2}$）递减至 -1（在 $t=\pi$），且 $y>0$。所以这段弧与 x 轴之间的面积为

$$\int_{-1}^{\pi/2} y\,\mathrm{d}x = \int_{\pi}^{\pi/2} y\,\frac{\mathrm{d}x}{\mathrm{d}t}\,\mathrm{d}t = \int_{\pi/2}^{\pi} \left(t^2\cos^2 t - t\sin t\cos t\right)\mathrm{d}t,$$

其中我们对调了上下限以吸收负号。我们需要两个积分，都是先用倍角公式（double-angle formula），再用分部积分（integration by parts）。写 $\sin t\cos t = \frac{1}{2}\sin 2t$ 及 $\cos^2 t = \frac{1}{2}(1+\cos 2t)$。用分部积分，

$$\int t\sin 2t\,\mathrm{d}t = -\tfrac{1}{2}t\cos 2t + \tfrac{1}{4}\sin 2t,$$
$$\int t^2\cos 2t\,\mathrm{d}t = \tfrac{1}{2}t^2\sin 2t - \int t\sin 2t\,\mathrm{d}t = \tfrac{1}{2}t^2\sin 2t + \tfrac{1}{2}t\cos 2t - \tfrac{1}{4}\sin 2t.$$

在 $t=\pi$ 时，$\sin 2t=0$，$\cos 2t=1$；在 $t=\frac{\pi}{2}$ 时，$\sin 2t=0$，$\cos 2t=-1$。因此

$$\int_{\pi/2}^{\pi} t\sin t\cos t\,\mathrm{d}t = \tfrac{1}{2}\left[-\tfrac{1}{2}t\cos 2t\right]_{\pi/2}^{\pi} = \tfrac{1}{2}\left(-\frac{\pi}{2}-\frac{\pi}{4}\right) = -\frac{3\pi}{8},$$
$$\int_{\pi/2}^{\pi} t^2\cos^2 t\,\mathrm{d}t = \tfrac{1}{2}\left[\tfrac{1}{3}t^3\right]_{\pi/2}^{\pi} + \tfrac{1}{2}\left[\tfrac{1}{2}t\cos 2t\right]_{\pi/2}^{\pi} = \frac{7\pi^3}{48} + \tfrac{1}{2}\left(\frac{\pi}{2}+\frac{\pi}{4}\right) = \frac{7\pi^3}{48}+\frac{3\pi}{8}.$$

所以所求面积为

$$\frac{7\pi^3}{48}+\frac{3\pi}{8}-\left(-\frac{3\pi}{8}\right) = \frac{7\pi^3}{48}+\frac{3\pi}{4}.$$

绳子扫过的面积

> ? **想一想** $t \geq t_0$ 那段弧下方的区域比扫过的区域多。必须除去甚么?

刚求得的面积，是外弧（$t_0 \leq t \leq \pi$）与 x 轴之间、$-1 \leq x \leq \frac{\pi}{2}$ 范围内的全部区域。扫过的区域等于这个面积减去位于其下方的两块：上半圆（semicircle），面积为 $\frac{\pi}{2}$；以及内弧（$0 \leq t \leq t_0$）与 x 轴之间、$1 \leq x \leq \frac{\pi}{2}$ 范围内的狭长小块。内弧位于圆外（因为当 $t>0$ 时 $x^2+y^2 = 1+t^2 > 1$），并位于外弧下方，所以这两块互不重叠，相减是合理的。

在内弧上 x 随 t 递增，所以狭长小块的面积为

$$\int_0^{\pi/2} y\,\frac{\mathrm{d}x}{\mathrm{d}t}\,\mathrm{d}t = \int_0^{\pi/2}\left(t\sin t\cos t - t^2\cos^2 t\right)\mathrm{d}t.$$

使用同样的原函数（antiderivative），现在以 $t=0$ 为下限（该处 $\sin 2t=0$，$\cos 2t=1$），

$$\int_0^{\pi/2} t\sin t\cos t\,\mathrm{d}t = \tfrac{1}{2}\left[-\tfrac{1}{2}t\cos 2t\right]_0^{\pi/2} = \frac{\pi}{8},$$
$$\int_0^{\pi/2} t^2\cos^2 t\,\mathrm{d}t = \tfrac{1}{2}\cdot\frac{\pi^3}{24} + \tfrac{1}{2}\left[\tfrac{1}{2}t\cos 2t\right]_0^{\pi/2} = \frac{\pi^3}{48}-\frac{\pi}{8},$$

所以狭长小块的面积为 $\frac{\pi}{4} - \frac{\pi^3}{48}$（约 0.14，又小又正，令人放心）。因此绳子扫过的面积为

$$\left(\frac{7\pi^3}{48} + \frac{3\pi}{4}\right) - \left(\frac{\pi}{4} - \frac{\pi^3}{48}\right) - \frac{\pi}{2} = \frac{8\pi^3}{48} = \frac{\pi^3}{6}.$$

验算

- 最后的答案有更简洁的求法。当 t 减少 δt 时，绳子的直线段（长度为 t）绕着一个基本上固定的点转过角度 δt，扫出面积约为 $\frac{1}{2}t^2\,\delta t$ 的细小扇形（sector）。把这些加起来得 $\int_0^\pi \frac{1}{2}t^2\,\mathrm{d}t = \frac{\pi^3}{6}$，结果一致。
- 数值上：用辛普森法则（Simpson's rule）求外弧下方的面积得 6.87794，与 $\frac{7\pi^3}{48} + \frac{3\pi}{4}$ 吻合；对阴影区域的边界使用鞋带公式（shoelace formula）得 5.16771 $\approx \frac{\pi^3}{6}$。

常见错误

- 以为直线段的长度是 $\pi - t$（那是弧长），或者在两个切线方向中选错了一个；检查 $t = \pi$ 的情况可以同时发现这两种错误。
- 把积分上下限保留为 t_0 到 π，却没有留意 x 沿该段弧递减，结果得出负的“面积”。
- 忘记内弧下方的小区域，结果得出扫过面积为 $\frac{7\pi^3}{48} + \frac{\pi}{4}$。

STEP 之外

绳子从一条曲线上解开时，其末端描出的曲线称为渐伸线，而绳子所解开的那条曲线则称为它的渐屈线（evolute）：渐伸线在 B 点的曲率中心（centre of curvature）正是绳子离开圆的点 P。大多数齿轮的齿形都采用圆的渐伸线，因为两个这样的齿互相滚动时，速度比保持不变。“验算”一节中的“细小扇形”论证，其实就是极坐标面积公式（polar-area formula）$\frac{1}{2}\int r^2\,\mathrm{d}\theta$ 的化身。想进一步了解曲率及这类曲线，可参阅 Siklos 的 *Advanced Problems in Mathematics*（《数学进阶难题》）；至于背后严谨的微积分，可参阅 Abbott 的 *Understanding Analysis*（《理解分析》）。

详细解答

STEP II 2011，第 9 题

力学 · 碰撞与恢复系数

STEP II 2011，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：

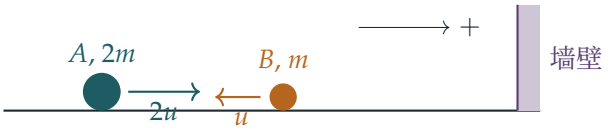
step.maths.org/questions/11-s2-q9

关键思路

这里每一次碰撞都是正碰（direct collision），所以每次都用同样两条方程解决：动量守恒（conservation of momentum）和牛顿（Newton）恢复定律（law of restitution）。真正的危险只在于正负号。一开始就固定一个正方向（向着墙壁），把每个速度都写成该方向上带正负号的数，代数运算便会自然理顺。最后的论断归结为证明某个表达式恒为正；由于它是 f 的线性（linear）函数，我们只需考察 f 取值范围的两个端点。

解答

模型。 我们把 A 和 B 视为光滑（smooth）平面上的质点（particle），所以除碰撞时刻外没有水平力作用；墙壁亦是光滑的，因此它只会令 B 的速度反向（并按 f 缩放）。第一次碰撞后 B 继续前进并撞向墙壁，所以墙壁位于 B 那一侧，即沿 A 原本运动方向、在 B 之外。我们取该方向为正方向。于是沿直线的次序始终是 A、B、墙壁。



第一次碰撞

碰撞前 A 的速度为 $2u$，B 的速度为 $-u$。设碰撞后两者的速度分别为 v_A 和 v_B。由动量守恒：

$$2m(2u) + m(-u) = 2m\,v_A + m\,v_B \quad \Longrightarrow \quad 2v_A + v_B = 3u.$$

接近速率为 $2u-(-u)=3u$。由于 B 在 A 的前方，分离速率为 v_B-v_A，牛顿恢复定律给出

$$v_B-v_A=3eu.$$

第一条方程减去第二条方程，得 $3v_A=3u-3eu$，所以

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

两者皆 ≥ 0，所以碰撞后 A 以速率 $(1-e)u$ 向墙壁移动（若 $e=1$ 则静止），而 B 以速率 $(1+2e)u$ 向墙壁移动。这与 B 随后撞到墙壁相符。

墙壁

墙壁光滑而碰撞为正碰，所以 B 沿直线以速率 $f(1+2e)u$ 反弹，其速度变为 $-f(1+2e)u$。与此同时，A 仍以速度 $(1-e)u\geq 0$ 运动，因为没有任何力作用于它。B 现正向 A 移动，所以两者必定再次相遇。

第二次碰撞

第二次碰撞前一刻，速度为

$$A:\ (1-e)u,\qquad B:\ -f(1+2e)u.$$

设碰撞后的速度为 w_A 和 w_B。由动量守恒得

$$2w_A+w_B=2(1-e)u-f(1+2e)u.$$

> ? **想一想** 这一次恢复方程应该怎样写，哪个减哪个？

B 仍然是较接近墙壁的质点，所以分离速度仍是 w_B-w_A。接近速率是 A 的速度减去 B 的速度，即 $(1-e)u+f(1+2e)u$。因此

$$w_B-w_A=e\big[(1-e)u+f(1+2e)u\big].$$

为求出 w_B，把这条方程的两倍加到动量方程上：

$$\begin{aligned}3w_B&=2(1-e)u-f(1+2e)u+2e(1-e)u+2ef(1+2e)u\\&=2(1-e)(1+e)u+f(1+2e)(2e-1)u\\&=2(1-e^2)u-(1-4e^2)fu.\end{aligned}$$

所以

$$w_B=\tfrac{2}{3}(1-e^2)u-\tfrac{1}{3}(1-4e^2)fu,$$

此速度以正方向（即向着墙壁）量度。这就是所求的结果。

B 总是向墙壁移动

我们须证明对所有 $0 < e \le 1$ 及 $0 < f \le 1$，都有 $w_B > 0$。由于 $u > 0$，这等同于证明

$$g(f) = 2(1-e^2) + (4e^2-1)f > 0.$$

$1-4e^2$ 一项可正可负，所以不能直接把它略去。改为固定 e，注意 g 是 f 的线性函数，而且

$$g(0) = 2(1-e^2) \ge 0, \qquad g(1) = 2-2e^2+4e^2-1 = 1+2e^2 > 0.$$

$[0,1]$ 上的线性函数是其两端值的加权平均：$g(f) = (1-f)\,g(0) + f\,g(1)$。当 $0 < f \le 1$ 时，第一项 ≥ 0，而第二项严格为正（因为 $f > 0$ 且 $g(1) > 0$），所以 $g(f) > 0$。因此 $w_B > 0$，无论 e 和 f 取何值，B 在第二次碰撞后都向墙壁移动。

验算

- 取 $e = f = 1$（全程完全弹性），A 在第一次碰撞时完全停下，B 以 $3u$ 离开，再以 $3u$ 返回，撞向一个质量为它两倍的静止质点。公式给出 $w_B = 0 + u = u$，与弹性碰撞的结果 $\frac{m-2m}{3m}(-3u) = u$ 吻合。
- 取 $e = f = \frac{1}{2}$，公式给出 $w_B = \frac{2}{3}\cdot\frac{3}{4}u = \frac{1}{2}u$。逐步模拟位置和碰撞，亦得到相同的值 $0.5u$。
- 对正负号而言，最坏情况是 $e = 1$ 而 f 很小。此时 $w_B = fu$，虽然很小，但仍然为正。

常见错误

- 在第二次碰撞中用速率代替带正负号的速度，因而把 B 的反弹速度符号弄错。
- 因为 B“从墙壁那边过来”而把恢复方程写成 $w_A - w_B$。质点沿直线的次序并没有改变，所以分离速度仍是 $w_B - w_A$。
- 假设 $1-4e^2 > 0$ 来断言 $w_B > 0$。当 $e > \frac{1}{2}$ 时此假设不成立，这正是论证必须用到 $f \le 1$ 的原因。

STEP 之外

恢复定律是一个经验模型（empirical model）。在大学一年级的课程中，你会同时接触一个更简洁的概念——质心参考系（centre-of-mass frame）：在该参考系中总动量为零，而恢复系数为 e 的碰撞只是把两个速度都乘以 $-e$。这令第一次碰撞之类的结果几乎一目了然。与墙壁的反复碰撞亦会在撞球问题（billiard problems）和计数论证中再次出现，例如

那个著名的问题：两个木块与一面墙之间的碰撞次数，竟然给出 π 的各位数字。Kleppner 和 Kolenkow 的 *An Introduction to Mechanics*（《力学导论》）正是以这种精神讲解动量与碰撞。

详细解答

STEP II 2011，第 10 题

力学 · 抛体与冲量

STEP II 2011，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/11-s2-q10

关键思路

由于 R 始终位于 P 的正下方，两颗质点在整个过程中水平速度相同，因此弦线保持铅垂，只需考虑铅垂方向的运动。当 P 恰好在平面上方 $\frac{1}{2}H$ 时，弦线绷紧（taut）；此时的急拉是一个铅垂方向的冲量（impulse），把 P 向上的动量（momentum）分给 R。此后两颗质点速度完全相同，因此作为同一个抛射体（projectile）一起飞行，直至 R 落地。

解答

最大高度

把质点视为只受重力作用的点。速度的铅垂分量为 $u\sin\theta - gt$，在最高点时为零。在铅垂方向使用 $v^2 = u^2 + 2as$，末铅垂速率为 0，

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

值得记下以备后用：由此得 $u\sin\theta = \sqrt{2gH}$。

弦线何时绷紧

以 O 为原点，x 为水平方向，y 为铅垂向上方向。质点 P 的水平速度始终为 $u\cos\theta$。要令 R 保持在 P 的正下方，R 必须以速率 $u\cos\theta$ 射出；由于平面光滑（smooth），R 便以这个恒定速率滑行。因此弦线始终铅垂，而距离 PR 就是 P 的高度：

$$y_P = u\sin\theta\, t - \tfrac{1}{2}gt^2 = \sqrt{2gH}\, t - \tfrac{1}{2}gt^2.$$

当 y_P 首次达到 $\frac{1}{2}H$ 时，弦线绷紧（这必定会发生，因为 P 上升至 $H > \frac{1}{2}H$）：

$$\tfrac{1}{2}gt^2 - \sqrt{2gH}\,t + \tfrac{1}{2}H = 0 \quad\Longrightarrow\quad t = \frac{\sqrt{2gH} \pm \sqrt{2gH - gH}}{g} = (\sqrt{2} \pm 1)\sqrt{\frac{H}{g}}.$$

较小的根是首次到达的时刻，因此弦线在以下时间后绷紧：

$$t_1 = (\sqrt{2} - 1)\sqrt{H/g},$$

正如所求。（较大的根是 P 下降时再次经过高度 $\frac{1}{2}H$ 的时刻。）

急拉

在 t_1 时，P 的铅垂速度为

$$u\sin\theta - gt_1 = \sqrt{2gH} - (\sqrt{2} - 1)\sqrt{gH} = \sqrt{gH},$$

方向向上，而 R 没有铅垂速度。两者的水平速度均为 $u\cos\theta$。

> **想一想** 在弦线急拉的一瞬间，有哪些冲量作用，方向如何？

弦线铅垂，因此张力（tension）冲量 J 是铅垂的：对 P 向下，对 R 向上。平面光滑，只能向上推 R；由于 R 被向上拉离平面，平面不施加任何冲量。所以水平速度不变，而在铅垂方向上两颗质点的总动量守恒。弦线不可伸长（inextensible）且保持绷紧，因此紧接急拉之后，P 和 R 有相同的铅垂速度 w。两者质量同为 m，

$$m\sqrt{gH} + m\cdot 0 = 2mw \qquad\Longrightarrow\qquad w = \tfrac{1}{2}\sqrt{gH}.$$

距离 D

急拉之后，若张力为零，P 和 R 速度相同，加速度亦同为向下的 g。因此它们如同两个相同的抛射体，相隔固定距离 $\frac{1}{2}H$ 运动：弦线恰好保持绷紧而张力为零，这是自洽的。R 从高度 0 以向上速度 w 出发，因此在以下时间后回到平面：

$$t_2 = \frac{2w}{g} = \sqrt{\frac{H}{g}}.$$

所以 R 运动的总时间为

$$t_1 + t_2 = \sqrt{2}\sqrt{\frac{H}{g}} = \sqrt{\frac{2H}{g}} = \sqrt{\frac{u^2\sin^2\theta}{g^2}} = \frac{u\sin\theta}{g},$$

而在整段时间内 R 以 $u\cos\theta$ 水平移动。因此

$$D = u\cos\theta\cdot\frac{u\sin\theta}{g} = \frac{u^2\sin\theta\cos\theta}{g} = \frac{u^2\sin 2\theta}{2g}.$$

这恰好是 P 单独运动时射程（range）的一半：R 落地的时刻，正是没有弦线时 P 到达其轨迹最高点的时刻。

$D = H$ 的情况

设

$$\frac{u^2 \sin\theta\cos\theta}{g} = \frac{u^2 \sin^2\theta}{2g}$$

并除以 $u^2 \sin\theta / g$（它不为零，因为质点要上升，须有 $0 < \theta < \frac{1}{2}\pi$），得 $\cos\theta = \frac{1}{2}\sin\theta$，所以

$$\tan\theta = 2.$$

验算

- t 的两个根之积为 H/g，之和为 $2\sqrt{2H/g}$，即 P 到达最高点再回到平面所需的时间；根据对称性，理应如此。
- 以随机的 u 和 θ 进行数值模拟（逐步推进 P 的飞行直至其高度为 $\frac{1}{2}H$，施加急拉，再逐步推进直至 R 落地），结果与 t_1 和 D 吻合；例如 $u = 18.11$、$\theta = 0.711$、$g = 9.8$，两种方法均得 $D \approx 16.543$。
- 取 $\tan\theta = 2$、$u = 10$、$g = 9.8$：$D = H \approx 4.082$。

常见错误

- 取了较大的根 $(\sqrt{2}+1)\sqrt{H/g}$，那是 P 下降时再次经过高度 $\frac{1}{2}H$ 的时刻。
- 忘记 R 在急拉前滑行的时间 t_1，只用急拉后的飞行来计算 D。
- 在急拉时使用能量守恒：不可伸长的弦线急拉绷紧时会损失动能；只有动量守恒。

STEP 之外

急拉之后，这对质点的行为与位于其质心（centre of mass）的单一质点完全相同：内力（这里是张力）永远不会改变系统质心的运动。这是大学一年级课程中质点系及刚体（rigid body）力学的起点，而像这题一样的冲击约束亦以同样方法处理。Kleppner 和 Kolenkow 的 *An Introduction to Mechanics*（《力学导论》）对这些概念有仔细的阐述。

详细解答

STEP II 2011，第 11 题

力学 · 三维空间中的平衡

STEP II 2011，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/11-s2-q11

关键思路

一切都建基于为每条绳写出一个沿绳方向的单位向量（unit vector）。每个单位向量都可分拆为铅垂部分和水平部分：铅垂部分由绳与铅垂线所成的角决定；水平部分则由 O 指向绳在天花板上的端点，并由该方向与 OA 所成的角决定。一旦知道三个单位向量，平衡（equilibrium）条件便只是三条纯量方程，$\mathbf{i}$、$\mathbf{j}$、$\mathbf{k}$ 各一条。

解答

我们把 P 视为质点（particle），并把绳视为轻而不可伸长（light and inextensible），因此每条绳都以沿绳方向的力把 P 拉向其固定端。取 O 为原点，$\mathbf{j}$ 沿 OA，$\mathbf{k}$ 铅垂向上；$\mathbf{i}$ 为水平方向，并垂直于 OA。

先处理三角学。由于 $\tan\theta = \sqrt{2}$，以 $1, \sqrt{2}, \sqrt{3}$ 为边的直角三角形给出

$$\sin\theta = \frac{\sqrt{2}}{\sqrt{3}}, \qquad \cos\theta = \frac{1}{\sqrt{3}}.$$

由于 $\tan\phi = \frac{1}{4}\sqrt{2}$，两直角边为 $4, \sqrt{2}$，斜边为 $\sqrt{18} = 3\sqrt{2}$，所以

$$\sin\phi = \frac{1}{3}, \qquad \cos\phi = \frac{2\sqrt{2}}{3}.$$

第 (i) 部分

与铅垂线成角 α 的绳，其单位向量（由 P 指向天花板）的铅垂分量为 $\cos\alpha$，水平部分的长度为 $\sin\alpha$。

对于 PB，$\alpha = 90^\circ - \theta$，所以铅垂分量为 $\cos(90^\circ - \theta) = \sin\theta = \frac{\sqrt{2}}{\sqrt{3}}$，而水平部分的长度为 $\sin(90^\circ - \theta) = \cos\theta = \frac{1}{\sqrt{3}}$。

水平部分沿 OB 方向（因为 O 在 P 的正上方），与 $\mathbf{j}$ 成角 $90^\circ+\theta$。与 $\mathbf{j}$ 成角 $90^\circ+\theta$ 的水平单位向量为 $\pm\sin(90^\circ+\theta)\,\mathbf{i}+\cos(90^\circ+\theta)\,\mathbf{j}=\pm\cos\theta\,\mathbf{i}-\sin\theta\,\mathbf{j}$。我们选取 $\mathbf{i}$ 的指向，使 B 位于 OA 的负 $\mathbf{i}$ 一侧；题目所要求的形式决定了这个选择。于是沿 PB 的单位向量的水平部分为

$$\tfrac{1}{\sqrt{3}}(-\cos\theta\,\mathbf{i}-\sin\theta\,\mathbf{j})=\tfrac{1}{\sqrt{3}}\left(-\tfrac{1}{\sqrt{3}}\,\mathbf{i}-\tfrac{\sqrt{2}}{\sqrt{3}}\,\mathbf{j}\right)=-\tfrac{1}{3}\,\mathbf{i}-\tfrac{\sqrt{2}}{3}\,\mathbf{j}.$$

加上铅垂分量，沿 PB 方向的单位向量为

$$-\tfrac{1}{3}\,\mathbf{i}-\tfrac{\sqrt{2}}{3}\,\mathbf{j}+\tfrac{\sqrt{2}}{\sqrt{3}}\,\mathbf{k},$$

正如所求。（验算：$\frac{1}{9}+\frac{2}{9}+\frac{2}{3}=1$。）

第 (ii) 部分

用同样的方法可得另外两条绳的单位向量。

绳 PA：与铅垂线成 30°，水平部分沿 $\mathbf{j}$：

$$\mathbf{e}_A=\tfrac{1}{2}\,\mathbf{j}+\tfrac{\sqrt{3}}{2}\,\mathbf{k}.$$

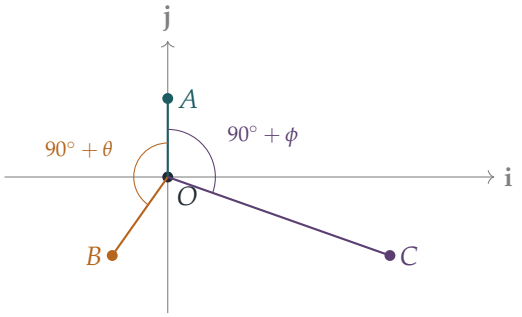
绳 PC：与铅垂线成 60°，所以铅垂分量为 $\frac{1}{2}$，水平长度为 $\frac{\sqrt{3}}{2}$。方向 OC 与 $\mathbf{j}$ 成角 $90^\circ+\phi$，所以其水平单位向量为 $\pm\cos\phi\,\mathbf{i}-\sin\phi\,\mathbf{j}$。

> ? **想一想** C 是在 OA 与 B 相同的一侧，还是在相反的一侧?

图中显示的是相反一侧，而且平衡条件本身亦要求如此。PA 中的张力（tension）没有 $\mathbf{i}$ 分量，而 PB 中的张力则有负分量。若 PC 也有负的 $\mathbf{i}$ 分量，便没有任何力能与它们平衡。所以符号为 $+$，且

$$\mathbf{e}_C=\tfrac{\sqrt{3}}{2}\left(\tfrac{2\sqrt{2}}{3}\,\mathbf{i}-\tfrac{1}{3}\,\mathbf{j}\right)+\tfrac{1}{2}\,\mathbf{k}=\tfrac{\sqrt{6}}{3}\,\mathbf{i}-\tfrac{\sqrt{3}}{6}\,\mathbf{j}+\tfrac{1}{2}\,\mathbf{k}.$$

下面的天花板平面图（取 $OP=1$，故 $OA=\tan30^\circ$、$OB=\cot\theta$、$OC=\tan60^\circ$）显示了各点的布局。



因此作用于 P 的力为

$$\begin{aligned}\text{张力，绳}PA:&\quad T\left(\tfrac{1}{2}\,\mathbf{j}+\tfrac{\sqrt{3}}{2}\,\mathbf{k}\right),\\ \text{张力，绳}PB:&\quad U\left(-\tfrac{1}{3}\,\mathbf{i}-\tfrac{\sqrt{2}}{3}\,\mathbf{j}+\tfrac{\sqrt{2}}{\sqrt{3}}\,\mathbf{k}\right),\\ \text{张力，绳}PC:&\quad V\left(\tfrac{\sqrt{6}}{3}\,\mathbf{i}-\tfrac{\sqrt{3}}{6}\,\mathbf{j}+\tfrac{1}{2}\,\mathbf{k}\right),\\ \text{重量:}&\quad -W\,\mathbf{k}.\end{aligned}$$

第 (iii) 部分

P 处于平衡状态，所以四个力之和为零。我们逐一取分量。

$\mathbf{i}$：

$$-\tfrac{1}{3}U+\tfrac{\sqrt{6}}{3}V=0\quad\Longrightarrow\quad U=\sqrt{6}\,V,$$

正如所求。

$\mathbf{j}$：

$$\tfrac{1}{2}T-\tfrac{\sqrt{2}}{3}U-\tfrac{\sqrt{3}}{6}V=0.$$

代入 $U=\sqrt{6}V$，并利用 $\sqrt{2}\sqrt{6}=2\sqrt{3}$：

$$\tfrac{1}{2}T=\tfrac{2\sqrt{3}}{3}V+\tfrac{\sqrt{3}}{6}V=\tfrac{5\sqrt{3}}{6}V\quad\Longrightarrow\quad T=\tfrac{5\sqrt{3}}{3}V.$$

$\mathbf{k}$：

$$\tfrac{\sqrt{3}}{2}T+\tfrac{\sqrt{2}}{\sqrt{3}}U+\tfrac{1}{2}V=W.$$

代入以上两个结果，

$$\tfrac{\sqrt{3}}{2}\cdot\tfrac{5\sqrt{3}}{3}V+\tfrac{\sqrt{2}\sqrt{6}}{\sqrt{3}}V+\tfrac{1}{2}V=\tfrac{5}{2}V+2V+\tfrac{1}{2}V=5V=W.$$

因此

$$V=\frac{W}{5},\qquad U=\frac{\sqrt{6}\,W}{5},\qquad T=\frac{5\sqrt{3}}{3}\cdot\frac{W}{5}=\frac{W}{\sqrt{3}}=\frac{\sqrt{3}\,W}{3}.$$

三者皆为正，所以每条绳都是拉紧的（taut），与模型一致。

验算

- 每个单位向量的长度都是 1：对于 $\mathbf{e}_C$，$\frac{6}{9}+\frac{3}{36}+\frac{1}{4}=1$。
- $\mathbf{e}_A$ 与 $\mathbf{e}_C$ 的水平部分之间的夹角，其余弦为 $-\frac{1}{3}=-\sin\phi$，所以夹角为 $90^\circ+\phi\approx 109.47^\circ$，理应如此。
- 取 $W=1$：$T\approx 0.5774$、$U\approx 0.4899$、$V=0.2$。于是铅垂分量 $0.5774\times 0.8660+0.4899\times 0.8165+0.2\times 0.5=0.5+0.4+0.1=1$ 与重量平衡。留意这个整齐的分配：PA、PB、PC 分别承受重量的 $\frac{1}{2}$、$\frac{2}{5}$、$\frac{1}{10}$。

常见错误

- 处理绳 PB 时混淆 sin 与 cos：它与铅垂线所成的角是 $90^\circ - \theta$，所以其铅垂分量为 $\sin\theta$，而不是 $\cos\theta$。
- 把 C 放在 OA 与 B 相同的一侧。这样 **i** 方程便会迫使张力为负（即 $V = -\frac{1}{3}W$），而绳是不可能有负张力的。
- 忘记 $\cos(90^\circ + \phi) = -\sin\phi$，以致 $\mathbf{e}_B$ 与 $\mathbf{e}_C$ 的 **j** 分量符号出错。

STEP 之外

三条沿三个独立方向的绳使系统成为静定（statically determinate）：平衡方程组成一个 3×3 线性方程组，并有唯一解。若再加第四条绳，单靠静力学便不能确定各张力，工程师必须引入绳的弹性。这正是结构分析（structural analysis）以及有关秩（rank）与零空间（null space）的线性代数的起点。Kleppner 与 Kolenkow 的 *An Introduction to Mechanics*（《力学导论》）就是以这种方式把力当作向量处理，而 Strang 的 *Introduction to Linear Algebra*（《线性代数导论》）则解释这类方程组何时有唯一解。

详细解答

STEP II 2011，第 12 题

概率与统计 · 博弈中的等比级数

STEP II 2011，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/11-s2-q12

关键思路

每局最多只有三分，所以我们可以列出一局所有可能的进程（只有八个序列），并把它们分为“Xavier 胜”、“Younis 胜”和“无人胜出，再打一局”三类。由于每一局都重新开始，整场比赛是一个几何过程（geometric process）：Younis 赢得整场比赛的概率（probability），等于 Younis 赢得一局的概率除以一局分出胜负的概率。

解答

把每一局写成一串得分者，例如 XYY 表示 Xavier 赢第一分，而 Younis 赢第二和第三分。第一分以概率 p 归 X。此后，每一分以概率 p 归于与上一分相同的球员，并以概率 $1-p$ 转换到另一位球员。

只要同一位球员连赢两分，该局便分出胜负，因此可能的局为：

序列	概率	结果
XX	$p \cdot p$	Xavier 胜
YXX	$(1-p)(1-p)p$	Xavier 胜
YY	$(1-p)p$	Younis 胜
XYY	$p(1-p)p$	Younis 胜
XYX	$p(1-p)(1-p)$	再打一局
YXY	$(1-p)(1-p)(1-p)$	再打一局

（一局一旦出现 XX 或 YY，便不需要第三分；我们设想第三分有没有打，都没有分别。）未分胜负的局，正好是得分交替出现的局，即后两分都“转换”。

第 (i) 部分

设 A 为 Younis 赢得某一局的事件（event），C 为该局未分胜负的事件。由上表，

$$\mathrm{P}(A) = (1-p)p + p^2(1-p) = p(1-p)(1+p) = p(1-p^2),$$
$$\mathrm{P}(C) = p(1-p)^2 + (1-p)^3 = (1-p)^2\big(p + (1-p)\big) = (1-p)^2.$$

各局在相同规则下独立（independent）进行，所以 Younis 赢得整场比赛，当且仅当对某个 $n \geq 1$，首 $n-1$ 局都未分胜负，而 Younis 赢第 n 局。因此

$$w = \sum_{n=1}^{\infty} \mathrm{P}(C)^{n-1}\,\mathrm{P}(A) = \frac{\mathrm{P}(A)}{1-\mathrm{P}(C)},$$

其中由于当 $p \neq 0$ 时 $\mathrm{P}(C) = (1-p)^2 < 1$，这个几何级数（geometric series）收敛（converges）。（等价地：若第一局未分胜负，我们便回到起点，所以 $w = \mathrm{P}(A) + \mathrm{P}(C)\,w$。）而 $1-(1-p)^2 = 2p - p^2 = p(2-p)$，所以

$$w = \frac{p(1-p^2)}{p(2-p)} = \frac{1-p^2}{2-p},$$

其中约去 p 用到了 $p \neq 0$。这就是所求的结果。

与 $\frac{1}{2}$ 比较。我们有

$$w - \tfrac{1}{2} = \frac{2(1-p^2) - (2-p)}{2(2-p)} = \frac{p - 2p^2}{2(2-p)} = \frac{p(1-2p)}{2(2-p)}.$$

当 $0 < p \leq 1$ 时，因子 p 和 $2-p$ 都是正数，所以 $w - \frac{1}{2}$ 与 $1-2p$ 同号。因此当 $0 < p < \frac{1}{2}$ 时 $w > \frac{1}{2}$，当 $p > \frac{1}{2}$ 时 $w < \frac{1}{2}$（而在 $p = \frac{1}{2}$ 时 $w = \frac{1}{2}$，正如对称性所要求）。

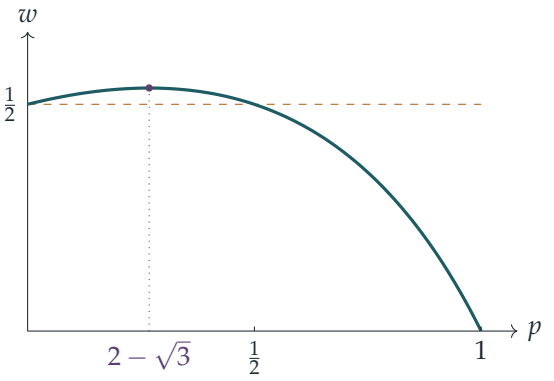
? **想一想** 降低 p 会令 Younis 更有可能赢得第一分。这是否必定对 Younis 有利?

w 是否随 p 递减? 不是。用商法则（quotient rule）求导，

$$\frac{\mathrm{d}w}{\mathrm{d}p} = \frac{-2p(2-p) + (1-p^2)}{(2-p)^2} = \frac{p^2 - 4p + 1}{(2-p)^2}.$$

分子的根为 $p = 2 \pm \sqrt{3}$,其中只有 $2-\sqrt{3} \approx 0.268$ 位于 $[0,1]$ 之内。所以当 $0 < p < 2-\sqrt{3}$ 时 $\mathrm{d}w/\mathrm{d}p > 0$，当 $2-\sqrt{3} < p \leq 1$ 时 $\mathrm{d}w/\mathrm{d}p < 0$。在区间 $0 < p < 2-\sqrt{3}$ 上，减小 p 实际上会减小 w。例如 $w(0.2) = \frac{8}{15} \approx 0.533$，但 $w(0.1) = \frac{99}{190} \approx 0.521$。

原因是 p 同时扮演两个角色。较小的 p 让 Younis 取得第一分，但同时也令连胜不太可能出现，因为赢得一分的球员只会以概率 p 继续得分。当 p 极小时，得分倾向交替出现，第一分的优势便被抵销。最大值为 $w(2-\sqrt{3}) = 4 - 2\sqrt{3} \approx 0.536$，而当 $p \to 0$ 时 $w \to \frac{1}{2}$。



第 (ii) 部分

当 $p = \frac{2}{3}$ 时，

$$w = \frac{1 - \frac{4}{9}}{2 - \frac{2}{3}} = \frac{5/9}{4/3} = \frac{5}{12}.$$

由于 $\mathrm{P}(C) = \frac{1}{9} < 1$，比赛以概率 1 结束，所以 Xavier 赢得比赛的概率为 $1 - \frac{5}{12} = \frac{7}{12}$。Xavier 的期望收益（expected gain）为

$$\tfrac{7}{12} \times 1 - \tfrac{5}{12} \times k,$$

而当此值为零时，比赛是公平的（fair），即

$$k = \frac{7}{5}.$$

所以当 Younis 胜出时，Xavier 应付给 Younis £1.40。

第 (iii) 部分

当 $p = 0$ 时，Younis 必定赢得每一局的第一分，而之后每一分都必定由输掉上一分的球员赢得。所以每一局都以概率 1 按 YXY 进行。没有人能连赢两分，比赛永不结束：两位球员都不会胜出。

这就是第 (i) 部分排除 $p = 0$ 的原因。此时 $\mathrm{P}(A) = 0$ 且 $1 - \mathrm{P}(C) = 0$，所以公式 $w = \mathrm{P}(A)/(1 - \mathrm{P}(C))$ 变成 $0/0$。当 $p \to 0$ 时 $\frac{1-p^2}{2-p} \to \frac{1}{2}$ 这个事实，描述的是那些邻近而最终会结束的比赛；它并不是 $p = 0$ 这个情况本身的概率。

验算

- 当 $p = \frac{1}{2}$ 时，每一分都等于掷一枚公平的硬币，所以 w 应为 $\frac{1}{2}$。公式给出 $\frac{3/4}{3/2} = \frac{1}{2}$。
- 当 $p = 1$ 时，Xavier 必定赢得首两分，所以 $w = 0$。公式给出 0。

- 表中六局的概率之和为 1。
- 对全部八个得分序列作精确枚举（p 由 0.01 至 0.99），重现了 w；而在 $p = \frac{2}{3}$ 下模拟 200 000 场比赛，Younis 胜出的比例为 41.6%，对比 $\frac{5}{12} \approx 41.7\%$。

常见错误

- 把转换规则弄反。概率 p 属于赢得上一分的球员，而不是属于 Xavier。
- 把错误的乘积设为相等，得出 $k = \frac{5}{7}$。公平的意思是 $\mathrm{P}(\text{X 胜}) \times 1 = \mathrm{P}(\text{Y 胜}) \times k$，所以较不可能胜出的一方得到较大的赔付。
- 在第 (iii) 部分把 $p = 0$ 代入公式，并声称每位球员的概率都是 $\frac{1}{2}$，但事实上比赛永不结束。

STEP 之外

这里的得分构成一条马可夫链（Markov chain）：每一分只取决于前一分。把比赛分拆为相同而独立的局，是更新（renewal）论证的一个例子；而方程 $w = \mathrm{P}(A) + \mathrm{P}(C)w$ 是“首步分析”（first-step analysis）最简单的例子，这是计算马可夫链中吸收概率（absorption probabilities）的标准方法。第 (iii) 部分展示了吸收失败时的情况：链永远循环下去。Grimmett 与 Welsh 的 *Probability: An Introduction*（《概率导论》）以大学一年级的程度介绍马可夫链及这些方法。

详细解答

STEP II 2011，第 13 题

概率与统计 · 偏度的量度

STEP II 2011，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/11-s2-q13

关键思路

偏度（skewness）描述的是分布的“不对称”程度。这里的三种量度从三个不同角度把它找出来：通过三阶中心矩（third central moment）、通过中位数（median）在两个十分位数（decile）之间的位置，以及通过平均值与中位数之间的差距。对于 $[0,1]$ 上的密度 $2x$，一切都可以明确算出：各阶矩都只是一行就能完成的积分，而累积分布函数（cumulative distribution function）是 x^2，所以它的反函数就是平方根。真正要花功夫的只是小心处理根式（surd）的运算，以及在最后找一个干净利落的方法比较它们的大小。

解答

偏度量度甚么。偏度量度的是分布相对其中心的*不对称性*。对称分布的偏度为零。若较长的尾部在右边，偏度为正；若较长的尾部在左边，偏度为负。

第 (i) 部分

把立方展开，并利用期望值的线性性质（linearity of expectation），记住 μ 是常数：

$$\begin{aligned}\mathrm{E}\big((X-\mu)^3\big) &= \mathrm{E}(X^3) - 3\mu\,\mathrm{E}(X^2) + 3\mu^2\,\mathrm{E}(X) - \mu^3 \\ &= \mathrm{E}(X^3) - 3\mu(\sigma^2+\mu^2) + 3\mu^3 - \mu^3 \\ &= \mathrm{E}(X^3) - 3\mu\sigma^2 - \mu^3,\end{aligned}$$

这里用了 $\mathrm{E}(X)=\mu$ 和 $\mathrm{E}(X^2)=\sigma^2+\mu^2$（即方差（variance）的常用公式移项后的形式）。除以 σ^3，便得到所求的 γ 表达式。

密度为 $2x$ 的分布。对于 $k \geq 0$,

$$\mathrm{E}(X^k) = \int_0^1 x^k \cdot 2x \,\mathrm{d}x = \frac{2}{k+2}.$$

所以 $\mu = \frac{2}{3}$、$\mathrm{E}(X^2) = \frac{1}{2}$ 及 $\mathrm{E}(X^3) = \frac{2}{5}$。因此

$$\sigma^2 = \tfrac{1}{2} - \tfrac{4}{9} = \tfrac{1}{18}, \qquad \sigma = \frac{1}{3\sqrt{2}}, \qquad \sigma^3 = \frac{1}{27 \cdot 2\sqrt{2}} = \frac{1}{54\sqrt{2}}.$$

分子为

$$\mathrm{E}(X^3) - 3\mu\sigma^2 - \mu^3 = \frac{2}{5} - 3 \cdot \frac{2}{3} \cdot \frac{1}{18} - \frac{8}{27} = \frac{2}{5} - \frac{1}{9} - \frac{8}{27} = \frac{54 - 15 - 40}{135} = -\frac{1}{135}.$$

因此

$$\gamma = -\frac{1}{135} \cdot 54\sqrt{2} = -\frac{54\sqrt{2}}{135} = -\frac{2\sqrt{2}}{5},$$

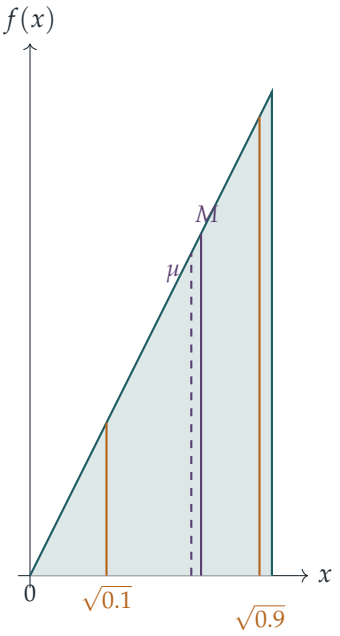
正如所求。负号是合理的：密度集中在 $x = 1$ 附近，在左边留下一条伸向 0 的长尾。

第 (ii) 部分

对于 $0 \leq x \leq 1$，累积分布函数为

$$\mathrm{F}(x) = \int_0^x 2t \,\mathrm{d}t = x^2,$$

它在 $[0,1]$ 上递增，所以对于 $0 \leq p \leq 1$，其反函数为 $\mathrm{F}^{-1}(p) = \sqrt{p}$。



十分位数偏度。把 D 的分子和分母同乘以 $\sqrt{10}$，便可消去所有根式，只剩下一个：

$$D = \frac{\sqrt{\frac{9}{10}} - 2\sqrt{\frac{1}{2}} + \sqrt{\frac{1}{10}}}{\sqrt{\frac{9}{10}} - \sqrt{\frac{1}{10}}} = \frac{3 - 2\sqrt{5} + 1}{3 - 1} = \frac{4 - 2\sqrt{5}}{2} = 2 - \sqrt{5},$$

因为 $\sqrt{10} \cdot \sqrt{\frac{1}{2}} = \sqrt{5}$。这正是要证明的结果。

皮尔逊偏度。中位数满足 $\mathrm{F}(M) = \frac{1}{2}$，所以 $M = \frac{1}{\sqrt{2}}$。利用 $\frac{1}{\sigma} = 3\sqrt{2}$，

$$P = 3\Big(\frac{2}{3} - \frac{1}{\sqrt{2}}\Big) \cdot 3\sqrt{2} = 9\sqrt{2} \cdot \frac{2}{3} - 9 = 6\sqrt{2} - 9.$$

数值上，$D \approx -0.236$、$P \approx -0.515$ 及 $\gamma \approx -0.566$，所以大小次序看来是对的；现在我们加以证明。

> ? **想一想** 不用计算机，怎样比较 $2 - \sqrt{5}$ 和 $6\sqrt{2} - 9$ 这类数的大小？

先移项使两边都是正数，然后平方：对于正数，$a < b$ 当且仅当 $a^2 < b^2$。以下每一步都是可逆的，所以最后的陈述成立，便证明了最初的陈述。

$D > P$。我们需要证明 $2 - \sqrt{5} > 6\sqrt{2} - 9$，即 $\sqrt{5} + 6\sqrt{2} < 11$。两边都是正数，所以这等价于

$$\begin{aligned}(\sqrt{5} + 6\sqrt{2})^2 = 77 + 12\sqrt{10} < 121 &\iff 12\sqrt{10} < 44 \\ &\iff 3\sqrt{10} < 11 \\ &\iff 90 < 121,\end{aligned}$$

其中最后一步再次把两个正数平方。由于 $90 < 121$ 成立，所以 $D > P$。

$P > \gamma$。我们需要证明 $6\sqrt{2} - 9 > -\frac{2\sqrt{2}}{5}$。乘以 5 并移项，即 $32\sqrt{2} > 45$。两边都是正数，所以这等价于 $32^2 \cdot 2 > 45^2$，即 $2048 > 2025$，这是成立的。所以 $P > \gamma$。

因此 $D > P > \gamma$。三种量度都是负数，一致认为这个分布向左偏，但对偏的程度却各有不同的看法。

验算

- (i) 中的恒等式对任何对称分布都给出 0：例如，若 X 以相等概率取 ± 1，则 $\mu = 0$、$\sigma = 1$ 及 $\mathrm{E}(X^3) = 0$。
- 数值上，用中点法则计算 $\mathrm{E}\big((X-\mu)^3\big)/\sigma^3$，得到 -0.56569，与 $-2\sqrt{2}/5$ 吻合。
- 以 U 为 $[0,1]$ 上的均匀分布，模拟 $X = \sqrt{U}$（它满足 $\mathrm{F}(x) = x^2$），得到的样本中位数接近 0.707，十分位数接近 0.316 和 0.949，与草图一致。

- 平均值 $\frac{2}{3}$ 位于中位数 $\frac{1}{\sqrt{2}} \approx 0.707$ 的左边，这是左偏分布的典型特征，也是 $P < 0$ 的原因。

常见错误

- 把 $\sigma^3 = (\frac{1}{18})^3$ 写成这样，而不是 $(\frac{1}{18})^{3/2}$：要立方的是标准差（standard deviation），不是方差。
- 在其中一边可能是负数的情况下把不等式平方。应先移项使两边都是正数。
- 在 γ 的公式或 P 的分母中，该用 $\sigma^2 = \frac{1}{18}$ 的地方却用了 $\mathrm{E}(X^2) = \frac{1}{2}$。

STEP 之外

$\mathrm{E}\big((X-\mu)^3\big)$ 这个量是三阶中心矩，而 (i) 中的展开正是矩与累积量（cumulant）之间关系的起点；对于独立随机变量之和，累积量可以简单地相加。像 D 这类以分位数（quantile）为基础的量度在统计学中备受重视，因为它们是稳健的（robust）：单一个极端的观测值几乎不会移动十分位数，却可以大幅改变三阶矩。Grimmett 和 Welsh 的 *Probability: An Introduction*（《概率导论》）以合适的程度讲解矩与矩母函数（moment generating function）。

详细解答

STEP III 2011，第 1 题

纯数学 · 降阶法

STEP III 2011，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/11-s3-q1

关键思路

第 (i) 部分是一条可分离方程（separable equation），而它的出现是有原因的：把 $y = z\mathrm{e}^{-x}$ 代入 $(*)$ 后，得出的正正就是这条方程。含 z 本身的项会互相抵消，只剩下一条只涉及 z' 和 z'' 的方程，而对 $u = z'$ 来说它是一阶的。这就是*降阶法*（reduction of order）：只要知道一个解（e^{-x}），便可以求出另一个解。第 (iii) 部分则在右方不为零的情况下作同样的代换，结果变成一条线性（linear）一阶方程，可用积分因子（integrating factor）求解。

解答

全题我们都在 $x \neq -1$ 的区间上讨论，因此可以除以 $x+1$。

第 (i) 部分

由于

$$\frac{x+2}{x+1} = 1 + \frac{1}{x+1},$$

分离变量（对 $u \neq 0$）得

$$\int \frac{\mathrm{d}u}{u} = \int \left(1 + \frac{1}{x+1}\right) \mathrm{d}x \quad \Longrightarrow \quad \ln|u| = x + \ln|x+1| + c.$$

取指数，得 $|u| = \mathrm{e}^{c}\,|x+1|\,\mathrm{e}^{x}$。正负号与因子 e^{c} 可合并成一个任意常数，而 $u = 0$ 亦是一个解，故通解（general solution）为

$$u = C(x+1)\mathrm{e}^{x}, \qquad C \text{ 为任意常数。}$$

第 (ii) 部分

设 $y = z\mathrm{e}^{-x}$，由乘积法则（product rule）得

$$y' = (z' - z)\mathrm{e}^{-x},$$
$$y'' = (z'' - z')\mathrm{e}^{-x} - (z' - z)\mathrm{e}^{-x} = (z'' - 2z' + z)\mathrm{e}^{-x}.$$

代入 $(*)$，并消去非零因子 e^{-x}：

$$\begin{aligned} 0 &= (x+1)(z'' - 2z' + z) + x(z' - z) - z \\ &= (x+1)z'' + \big(-2(x+1) + x\big)z' + \big((x+1) - x - 1\big)z \\ &= (x+1)z'' - (x+2)z'. \end{aligned}$$

所有 z 项都完全抵消了。记 $u = \dfrac{\mathrm{d}z}{\mathrm{d}x}$，再除以 $x+1$，得

$$\frac{\mathrm{d}u}{\mathrm{d}x} - \left(\frac{x+2}{x+1}\right)u = 0,$$

这是一条关于 $\dfrac{\mathrm{d}z}{\mathrm{d}x}$ 的一阶方程，正如题目所要求。它恰好就是第 (i) 部分的方程。

想一想 为甚么 z 项会消失？我们能否事先预料到？

z 的系数就是代入 $z = 1$ 所得的结果，即在 $(*)$ 中检验 $y = \mathrm{e}^{-x}$。由于 $(x+1)\mathrm{e}^{-x} - x\mathrm{e}^{-x} - \mathrm{e}^{-x} = 0$，$y = \mathrm{e}^{-x}$ 本身已经是一个解，所以常数 z 必定满足方程，因此不可能剩下单独的 z 项。这正是选用这个代换的原因。

由第 (i) 部分，$z' = C(x+1)\mathrm{e}^{x}$。用分部积分（integration by parts），

$$\int (x+1)\mathrm{e}^{x}\,\mathrm{d}x = (x+1)\mathrm{e}^{x} - \int \mathrm{e}^{x}\,\mathrm{d}x = x\mathrm{e}^{x} + \text{常数},$$

所以

$$z = Cx\mathrm{e}^{x} + B.$$

因此 $y = z\mathrm{e}^{-x} = Cx + B\mathrm{e}^{-x}$，把 C 改名为 A，得

$$y = Ax + B\mathrm{e}^{-x},$$

其中 A 和 B 为任意常数。$(*)$ 的每一个解都可以这样得到，因为任何 y 都可以写成 $z\mathrm{e}^{-x}$，其中 $z = y\mathrm{e}^{x}$；所以这就是通解，正如题目所要求。

第 (iii) 部分

同样的代换只会改变右方，右方变成 $(x+1)^2$ 除以 e^{-x}：

$$(x+1)z'' - (x+2)z' = (x+1)^2\mathrm{e}^{x}.$$

设 $u = z'$，再除以 $x+1$，得

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

这是线性方程。由第 (i) 部分，$(x+1)\mathrm{e}^x$ 是齐次（homogeneous）方程的解，所以它的倒数 $\dfrac{\mathrm{e}^{-x}}{x+1}$ 是一个积分因子。（它在相差正负号的意义下等于 $\exp\left(-\int \frac{x+2}{x+1}\,\mathrm{d}x\right)$。）两边同乘，得

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{u\,\mathrm{e}^{-x}}{x+1}\right) = 1 \quad\Longrightarrow\quad u = (x+1)\mathrm{e}^x(x+k)$$

其中 k 为常数。所以

$$z' = (x^2+x)\mathrm{e}^x + k(x+1)\mathrm{e}^x.$$

第二项积分后得 $kx\mathrm{e}^x$，与第 (ii) 部分相同。至于第一项，用两次分部积分，

$$\begin{aligned}\int (x^2+x)\mathrm{e}^x\,\mathrm{d}x &= (x^2+x)\mathrm{e}^x - \int (2x+1)\mathrm{e}^x\,\mathrm{d}x \\ &= (x^2+x)\mathrm{e}^x - (2x+1)\mathrm{e}^x + 2\mathrm{e}^x + \text{常数} = (x^2-x+1)\mathrm{e}^x + \text{常数}.\end{aligned}$$

因此 $z = (x^2-x+1)\mathrm{e}^x + kx\mathrm{e}^x + B$，而

$$y = z\mathrm{e}^{-x} = x^2 - x + 1 + kx + B\mathrm{e}^{-x}.$$

$-x$ 可并入任意的 kx 中，所以记 $A = k-1$，通解为

$$y = x^2 + 1 + Ax + B\mathrm{e}^{-x},$$

其中 A 和 B 为任意常数。

另一方法：由于右方是二次式，我们可以尝试特解（particular integral）$y = px^2+qx+r$。代入得 $(x+1)2p + x(2px+q) - (px^2+qx+r) = px^2+2px+2p-r$，与 x^2+2x+1 比较系数，得 $p=1$、$r=1$，而 q 可任意取值（因为 x 是一个余函数（complementary function））。再加上第 (ii) 部分的余函数，便得出相同的答案。

验算

直接代入 $y = x^2+1$：$y' = 2x$，$y'' = 2$，所以左方为 $2(x+1) + 2x^2 - x^2 - 1 = x^2+2x+1 = (x+1)^2$，符合要求。数值上，随机选取 A、B 及 $x \in (-0.9, 3)$，以有限差分（finite difference）计算 $y = x^2+1+Ax+B\mathrm{e}^{-x}$ 的导数，结果满足方程的误差约在 3×10^{-6} 以内；同样的测试亦验证了第 (i) 和第 (ii) 部分。

常见错误

- 把 y'' 算错：它是 $(z''-2z'+z)\mathrm{e}^{-x}$，$z$ 前面是加号；这里若出现正负号错误，便会留下一个多余的 z 项。

- 在第 (iii) 部分消去因子 e^{-x} 时，忘记把右方乘以 e^{x}。
- 漏掉常数：最终答案需要两个任意常数，每次积分各产生一个。

STEP 之外

降阶法是一种通用技巧：若 y_1 是某条线性二阶方程的解，代入 $y = y_1 v$ 必定会消去 v 项，剩下一条关于 v' 的一阶方程。这亦是*参数变易法*（variation of parameters）背后的思路，该方法对任何右方都能求出特解。两者在 Boyce 和 DiPrima 的 *Elementary Differential Equations*（《初等微分方程》）中都有清晰的讲解。

详细解答

STEP III 2011，第 2 题

纯数学 · 多项式的有理根

> **STEP III 2011，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/11-s3-q2

关键思路

将 $\mathrm{f}(p/q)=0$ 乘以 q^{n-1}，除了一个分母外，其余分母全部消去：首项变成 p^n/q，而其余各项都是整数。因此 p^n/q 必须是整数；由于 p 与 q 没有公因数，这就迫使 $q=1$。此后，每一部分都变成关于整数根的问题，而这些根可以从大小、整除性，或者更巧妙地从奇偶性（parity）加以排除。

解答

开首的结果

设 $\mathrm{f}(p/q)=0$，其中 p、q 为互质（coprime）整数，且 $q>0$。那么

$$0=q^{n-1}\,\mathrm{f}\left(\frac{p}{q}\right)=\frac{p^n}{q}+a_{n-1}p^{n-1}+a_{n-2}p^{n-2}q+\cdots+a_1p\,q^{n-2}+a_0q^{n-1},$$

这是因为项 $a_k(p/q)^k$ 变成 $a_kp^kq^{n-1-k}$，而对每个 $k\le n-1$ 都有 $n-1-k\ge 0$。首项以外的每一项都是整数，所以

$$\frac{p^n}{q}=-\left(a_{n-1}p^{n-1}+\cdots+a_0q^{n-1}\right)$$

是整数；即 q 整除 p^n。

> ? **想一想** 为何 $q\mid p^n$ 会迫使 $q=1$?

若 $q>1$，则它有某个质因数 r。于是 r 整除 p^n；由于 r 是质数，它必定整除 p 本身。但这样 r 便是 p 与 q 一个大于 1 的公因数，这是不可能的。所以 q 没有任何质因数，又因

$q > 0$，得

$$q = 1.$$

因此 $\mathrm{f}(x) = 0$ 的每个有理根 p/q 都等于 p，即一个整数。（留意这里正是用上了 f 是首一（monic）多项式这一点：若首项系数为 a_n，就会剩下 $a_n p^n/q$，而我们只能推断 $q \mid a_n$。）

以下两个简单事实会有帮助。若 m 是一个整数根，则

$$m\left(m^{n-1} + a_{n-1}m^{n-2} + \cdots + a_1\right) = -a_0,$$

所以 m 整除 a_0；又因 $a_0 \neq 0$，故 $m \neq 0$。

第 (i) 部分

数 $2^{1/n}$ 是 $x^n - 2 = 0$ 的根，而这个多项式是首一的、系数为整数，且 $a_0 = -2 \neq 0$。因此若 $2^{1/n}$ 是有理数，它便会是一个满足 $m^n = 2$ 的整数 m。但

$$1^n = 1 < 2 < 4 \leq 2^n,$$

所以正实数 $2^{1/n}$ 严格地位于相邻整数 1 与 2 之间，不可能是整数。因此对每个 $n \geq 2$，$2^{1/n}$ 都是无理数。

第 (ii) 部分

这里 $n = 3$，$a_0 = 1$。任何有理根都是整除 1 的整数 m，所以 $m = \pm 1$。但

$$1^3 - 1 + 1 = 1 \neq 0, \qquad (-1)^3 - (-1) + 1 = 1 \neq 0.$$

所以 $x^3 - x + 1 = 0$ 没有有理根。

（另一方法：$m^3 - m = (m-1)m(m+1)$ 是一个包含两个相邻整数的乘积，所以是偶数，于是 $m^3 - m + 1$ 是奇数，永不为零。这个奇偶性的想法正是第 (iii) 部分所需要的。）

第 (iii) 部分

多项式 $x^n - 5x + 7$ 是首一的、系数为整数，常数项 $7 \neq 0$，所以任何有理根都是整数 m。我们证明 $m^n - 5m + 7$ 恒为奇数。

- 若 m 为偶数，则 m^n 与 $5m$ 都是偶数，所以 $m^n - 5m + 7$ 是偶数 − 偶数 + 奇数，即奇数。
- 若 m 为奇数，则 m^n 与 $5m$ 都是奇数，所以 $m^n - 5m$ 是偶数，而 $m^n - 5m + 7$ 是奇数。

无论哪种情况，$m^n - 5m + 7$ 都是奇数，特别地它不等于 0。因此对任何 $n \geq 2$，$x^n - 5x + 7 = 0$ 都没有整数根，从而没有有理根。

（若改为逐一检查 7 的因数，便要证明对 $m = \pm 1, \pm 7$ 及每一个 n 都有 $m^n - 5m + 7 \neq 0$；这样做也行得通，但较为繁琐；奇偶性则一次过处理所有 n。）

验算

- 试一个确实有有理根的多项式，例如 $x^2 - x - 6 = (x-3)(x+2)$：根 3 与 -2 正如所料都是整数，而且都整除 $a_0 = -6$。
- 非首一的多项式可以有非整数的有理根，例如 $2x - 1$；这就是首项系数为 1 之所以重要的原因。
- 数值上，$2^{1/2} \approx 1.414$ 与 $2^{1/3} \approx 1.260$ 都位于 1 与 2 之间；而 $x^3 - x + 1$ 在 $x = \pm 1$ 时取值 1。以电脑搜索 3000 个随机的首一整系数多项式，所找到的有理根全是整数；并确认了对 $|m| \leq 200$、$2 \leq n \leq 14$，$m^n - 5m + 7$ 都是奇数。

常见错误

- 乘以 q^n 而非 q^{n-1}：你会得到 $p^n = -q(\cdots)$，这样也行得通，但之后必须说明 $q \mid p^n$；若乘以 q^{n-1}，分数 p^n/q 便令这一点一目了然。
- 断言“$q \mid p^n$，所以 $q \mid p$”而不说明原因；论证须经由 q 的一个质因数。
- 在 (iii) 中只检查了几个 n 值。论证必须同时对所有 $n \geq 2$ 成立。

STEP 之外

这是首一形式的*有理根定理*（rational root theorem）：若一个有理数是某个首一整系数多项式的根，则它是整数。用代数数论（algebraic number theory）的语言来说，整数是*整闭*（integrally closed）的，而首一整系数多项式的根称为代数整数（algebraic integers）。(iii) 中的奇偶性技巧，是将多项式模一个质数（modulo a prime）约化的最简单情况；这是证明多项式没有根或不可约（irreducible）的有力工具。Davenport 的 *The Higher Arithmetic*（《高等算术》）是很好的下一步读物，而 Allenby 的 *Rings, Fields and Groups*（《环、域与群》）则发展了代数方面的内容。

详细解答

STEP III 2011，第 3 题

纯数学 · 解三次方程

STEP III 2011，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/11-s3-q3

关键思路

两个立方 $(x-\alpha)^3$ 和 $(x-\beta)^3$ 为我们提供两个可自由选取的常数 a 和 b，而一个没有 x^2 项的首一三次多项式（monic cubic）要符合四个条件。因此我们用其中两个条件求出 a 和 b，其余两个条件就必须恰好是二次方程所给出、关于 $\alpha+\beta$ 和 $\alpha\beta$ 的事实。一旦把三次式写成两个立方之和（sum of two cubes），令它等于零便化为“某个比的立方等于一个常数”，只要取立方根（cube root）便可解出，而且三个立方根都要取。

解答

把三次式写成两个立方之和

由于 $p\neq 0$，二次方程 $pt^2-qt+p^2=0$ 满足

$$\alpha+\beta=\frac{q}{p},\qquad \alpha\beta=\frac{p^2}{p}=p,$$

而其判别式（discriminant）为 q^2-4p^3。因此条件 $q^2\neq 4p^3$ 正好表示 $\alpha\neq\beta$。此外 $\alpha\beta=p\neq 0$，故两个根都不为零。

展开得

$$a(x-\alpha)^3+b(x-\beta)^3=(a+b)x^3-3(a\alpha+b\beta)x^2+3(a\alpha^2+b\beta^2)x-(a\alpha^3+b\beta^3).$$

要与 $x^3-3px+q$ 的系数相符，我们需要

$$a+b=1,\qquad a\alpha+b\beta=0,\qquad a\alpha^2+b\beta^2=-p,\qquad a\alpha^3+b\beta^3=-q.$$

头两条是 a 和 b 的线性方程。把 $b=1-a$ 代入第二条，得 $a(\alpha-\beta)=-\beta$；由于 $\alpha\neq\beta$，

可以两边相除：

$$a = \frac{\beta}{\beta - \alpha}, \qquad b = \frac{\alpha}{\alpha - \beta}.$$

现在验证其余两个条件会自动成立：

$$a\alpha^2 + b\beta^2 = \frac{\beta\alpha^2 - \alpha\beta^2}{\beta - \alpha} = \frac{\alpha\beta(\alpha - \beta)}{\beta - \alpha} = -\alpha\beta = -p,$$

$$a\alpha^3 + b\beta^3 = \frac{\beta\alpha^3 - \alpha\beta^3}{\beta - \alpha} = \frac{\alpha\beta(\alpha - \beta)(\alpha + \beta)}{\beta - \alpha} = -\alpha\beta(\alpha + \beta) = -p \cdot \frac{q}{p} = -q.$$

所以，取这两个 a 和 b 的值，

$$x^3 - 3px + q = \frac{\beta}{\beta - \alpha}(x - \alpha)^3 + \frac{\alpha}{\alpha - \beta}(x - \beta)^3,$$

此即所求。

解 $x^3 - 24x + 48 = 0$

这里 $p = 8$、$q = 48$，而 $q^2 = 2304 \neq 2048 = 4p^3$，所以上述结果适用。对应的二次方程是 $8t^2 - 48t + 64 = 0$，即 $t^2 - 6t + 8 = 0$，其根为 2 和 4。取 $\alpha = 2$、$\beta = 4$；则 $a = 2$、$b = -1$，于是

$$x^3 - 24x + 48 = 2(x - 2)^3 - (x - 4)^3.$$

（快速展开即可确认：$2x^3 - 12x^2 + 24x - 16 - x^3 + 12x^2 - 48x + 64$。）

? **想一想** 我们怎样把 $2(x - 2)^3 = (x - 4)^3$ 变成可以求解的形式?

两边除以 $(x - 4)^3$。这样做是容许的，因为 $x = 4$ 不是根（$64 - 96 + 48 = 16$）。方程变成

$$\left(\frac{x - 2}{x - 4}\right)^3 = \frac{1}{2}.$$

记 $c = 2^{1/3}$。立方等于 $\frac{1}{2}$ 的复数为 $k = c^{-1}$、$c^{-1}\omega$ 和 $c^{-1}\omega^2$。对每一个值，解 $\dfrac{x - 2}{x - 4} = k$：

$$x - 2 = kx - 4k \quad \Longrightarrow \quad x = \frac{2 - 4k}{1 - k},$$

由于 $k \neq 1$，这一步是成立的。映射 $k \mapsto x$ 是一一对应（one-to-one）的，所以 k 的三个值给出三个不同的根，它们必然就是该三次方程的全部根。

当 $k = c^{-1}$ 时，分子分母同乘以 c：

$$x = \frac{2c - 4}{c - 1} = \frac{2(2 - 2^{1/3})}{1 - 2^{1/3}},$$

这正是题目所给的解。当 $k = c^{-1}\omega$ 时，分子分母同乘以 $c\,\omega^{-1} = c\,\omega^2$（利用 $\omega^3 = 1$）：

$$x = \frac{2c\omega^2 - 4}{c\omega^2 - 1} = \frac{2(2 - 2^{1/3}\omega^2)}{1 - 2^{1/3}\omega^2},$$

同理，$k = c^{-1}\omega^2$ 给出

$$x = \frac{2(2 - 2^{1/3}\omega)}{1 - 2^{1/3}\omega}.$$

所以三个解为 $x = \dfrac{2(2 - 2^{1/3}\omega^j)}{1 - 2^{1/3}\omega^j}$，其中 $j = 0, 1, 2$。

$p = r^2$、$q = 2r^3$ 的情况

此时 $q^2 = 4r^6 = 4p^3$，正是上述方法所排除的情况：二次式 $r^2t^2 - 2r^3t + r^4 = r^2(t - r)^2$ 有重根（repeated root）$t = r$，我们无法求出 a 和 b。不过，二次方程有重根，暗示三次方程也有重根。事实上，代入 $x = r$ 得 $r^3 - 3r^3 + 2r^3 = 0$，而导数（derivative）$3x^2 - 3r^2$ 在 $x = r$ 处亦为零，所以 $(x - r)^2$ 是一个因式。比较 x^3 项和常数项，得

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

根为 $x = r$（二重根）和 $x = -2r$。

想一想 我们能否从先前的公式得出这些根?

不能。当 $\alpha = \beta$ 时，两个立方合而为一，所以 $a(x - \alpha)^3 + b(x - \alpha)^3$ 只可能是 $(x - \alpha)^3$，它没有 x 项可与 $-3px$ 相配。这就是我们需要另一种方法的原因。把三次式写成 $(x - r)$ 的幂，便可看出实际情况：$x^3 - 3r^2x + 2r^3 = (x - r)^3 + 3r(x - r)^2$，所以二重根正好位于二次方程的重根 $t = r$ 处。

验算

- 实根为 $2(2 - 1.259921)/(1 - 1.259921) \approx -5.694644$，代入后得 $x^3 - 24x + 48 \approx 0$，误差约为 10^{-13}。另外两个根为 $2.847322 \pm 0.567212\,\mathrm{i}$，残差同样为零。
- 三个根之和为 0（因为没有 x^2 项），积为 -48：$-5.694644 \times (2.847322^2 + 0.567212^2) \approx -48.0$。
- 一个实根加两个复根，与理论相符：$x^3 - 3px + q$ 只有在 $q^2 < 4p^3$ 时才有三个实根，而这里 $q^2 > 4p^3$。
- 该恒等式（identity）可以用数值检验：我们对 2000 组随机的 (p, q)，在随机复数 x 处作了检验，并对多个 r 值检验了因式分解 $(x - r)^2(x + 2r)$。

常见错误

- 只取 $\frac{1}{2}$ 的实立方根，因而只找到一个解。三次方程有三个根，它们分别来自三个立方根。

- 搞错 a、b 各自对应哪个根：a（即 $(x-\alpha)^3$ 的系数）的分子是 β。
- 在消去分母时把 ω 和 ω^2 对调。由于这两个答案是一对共轭（conjugate pair），无论哪种做法，最终得到的根集合都相同。

STEP 之外

这是十六世纪解三次方程方法的一个版本，通常归功于卡尔达诺（Cardano）和塔尔塔利亚（Tartaglia）：先化为没有 x^2 项的“缺项”三次方程（depressed cubic），再解一个辅助二次方程（即预解式（resolvent））。伽罗瓦理论（Galois theory）解释了为何三次和四次方程有这样的预解式，而一般的五次方程却没有。Allenby 的 *Rings, Fields and Groups*（《环、域与群》）为通往这些概念提供了一条易读的路径。

详细解答

STEP III 2011，第 4 题

纯数学 · 杨氏不等式

STEP III 2011，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/11-s3-q4

本题在 *Beyond STEP* 第 3.4 节以面积证明不等式中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch03/areas-inequalities.html

关键思路

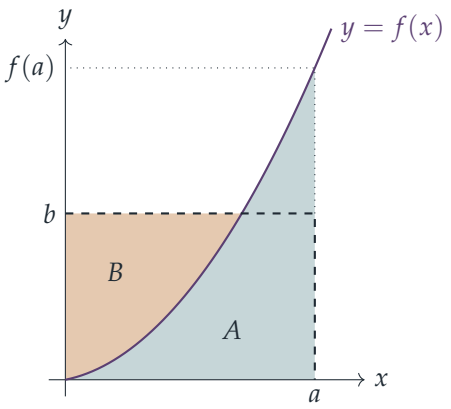
$(*)$ 中的两个积分（integral）都是同一幅图上的面积：$\int_0^a f(x)\,\mathrm{d}x$ 是曲线与 x 轴之间的面积，而 $\int_0^b f^{-1}(y)\,\mathrm{d}y$ 是曲线与 y 轴之间的面积。两者合起来总能覆盖以 $(0,0)$ 和 (a,b) 为顶角的矩形，而且除非顶角 (a,b) 刚好落在曲线上，否则还会有剩余。这样一来，第 (ii) 和第 (iii) 部分只不过是选取适当的 f，再计算两个积分而已。

解答

第 (i) 部分

由于 f 连续（continuous）且斜率（gradient）为正，它是严格递增（strictly increasing）的，因此反函数（inverse function）f^{-1} 存在，而且同样递增；又 $f(0)=0$ 表示曲线 $y=f(x)$ 从原点出发。

设 A 为 $0\le x\le a$ 范围内曲线与 x 轴之间的区域，其面积为 $\int_0^a f(x)\,\mathrm{d}x$。设 B 为 $0\le y\le b$ 范围内曲线与 y 轴之间的区域；对 y 积分，其面积为 $\int_0^b f^{-1}(y)\,\mathrm{d}y$。这两个区域位于曲线的两侧，所以不会重叠。



先假设 $b \le f(a)$，如图所示。矩形 $0 \le x \le a$、$0 \le y \le b$ 中的每一点，要么位于曲线上或其下方（因而属于 A），要么位于曲线上或其左方（因而属于 B）。所以 A 和 B 合起来覆盖了面积为 ab 的矩形，$(*)$ 成立。多出的部分是 A 中位于直线 $y = b$ 以上的部分，即 $f^{-1}(b) \le x \le a$ 范围内夹在该直线与曲线之间的狭长区域。

若改为 $b \ge f(a)$，只需把两者的角色对调，同样的论证仍然成立：矩形被覆盖，而这次多出的部分是 B 中位于直线 $x = a$ 右方的部分。

因此 $(*)$ 在所有情况下都成立，而等号成立需要多出部分的面积为零。当 $a, b > 0$ 且 $b \ne f(a)$ 时，多出的部分由曲线和一条直线围成，两者只在一端相交，所以其面积为正。因此，对于 $a > 0$ 及 $b > 0$，

$$\text{等号成立当且仅当} b = f(a),$$

即矩形的顶角 (a, b) 落在曲线上的时候。

第 (ii) 部分

当 $p > 1$ 时，$f(x) = x^{p-1}$ 在 $x \ge 0$ 上连续，$f(0) = 0$，并在该范围内严格递增，所以 $(*)$ 适用。（其斜率 $(p-1)x^{p-2}$ 在 $x > 0$ 时为正；若 $p < 2$，斜率在 0 处趋于无穷大，但这不影响面积的论证。）其反函数为 $f^{-1}(y) = y^{1/(p-1)}$。

条件 $\frac{1}{p} + \frac{1}{q} = 1$ 给出 $\frac{1}{q} = \frac{p-1}{p}$，所以

$$q = \frac{p}{p-1}, \qquad \frac{1}{p-1} + 1 = \frac{p}{p-1} = q.$$

于是 $f^{-1}(y) = y^{q-1}$，而两个积分分别为

$$\int_0^a x^{p-1}\,\mathrm{d}x = \frac{a^p}{p}, \qquad \int_0^b y^{q-1}\,\mathrm{d}y = \frac{b^q}{q}.$$

代入 $(*)$ 得

$$ab \le \frac{a^p}{p} + \frac{b^q}{q},$$

即为所求。

想一想 等号条件 $b = f(a)$ 在这里代表甚么？又应如何化简 b^q?

条件是 $b = a^{p-1}$。由于 $(p-1)q = p$，故 $b^q = a^{(p-1)q} = a^p$。所以

$$\frac{a^p}{p} + \frac{b^q}{q} = a^p\left(\frac{1}{p} + \frac{1}{q}\right) = a^p = a \cdot a^{p-1} = ab,$$

等号确实成立。

第 (iii) 部分

取 $0 \le x \le \frac{1}{2}\pi$ 上的 $f(x) = \sin x$。它是连续的，$f(0) = 0$，而其斜率 $\cos x$ 在 $0 \le x < \frac{1}{2}\pi$ 时为正，所以它严格递增（斜率只在端点处为零，这不影响论证）。它把 $[0, \frac{1}{2}\pi]$ 映射到 $[0,1]$ 上，反函数为 arcsin，而这正好是 a 和 b 容许的取值范围。

第一个积分是

$$\int_0^a \sin x \,\mathrm{d}x = 1 - \cos a.$$

至于第二个积分，取 $u = \arcsin y$ 及 $\mathrm{d}v = \mathrm{d}y$，用分部积分法（integration by parts）：

$$\begin{aligned}\int_0^b \arcsin y \,\mathrm{d}y &= \Big[y \arcsin y\Big]_0^b - \int_0^b \frac{y}{\sqrt{1-y^2}}\,\mathrm{d}y \\ &= b \arcsin b + \Big[\sqrt{1-y^2}\Big]_0^b \\ &= b \arcsin b + \sqrt{1-b^2} - 1.\end{aligned}$$

两者相加，两个 1 互相抵消，$(*)$ 变成

$$ab \le b \arcsin b + \sqrt{1-b^2} - \cos a,$$

这就是所求的不等式（inequality），等号在 $b = \sin a$ 时成立。

想一想 a 可以自由选取。选哪一个值才能得到我们想要的不等式?

对于 $t \ge 1$，令 $b = t^{-1}$，它按要求落在 $(0,1]$ 内。把不等式乘以 $t > 0$，并利用 $t\sqrt{1-t^{-2}} = \sqrt{t^2-1}$（因为 $t > 0$，此式成立），得

$$a \le \arcsin(t^{-1}) + \sqrt{t^2-1} - t\cos a \quad \text{对所有} a \in [0, \tfrac{1}{2}\pi].$$

目标不等式中没有 a，而右方有一个 t，这正是 $t\cos a$ 在 $a = 0$ 时的值。取 $a = 0$:

$$0 \le \arcsin(t^{-1}) + \sqrt{t^2-1} - t, \qquad \text{即} \qquad \arcsin(t^{-1}) \ge t - \sqrt{t^2-1} \quad (t \ge 1).$$

验算

- 当 $t = 1$ 时，最后的不等式变成 $\frac{1}{2}\pi \geq 1$；当 $t = 2$ 时，则变成 $\frac{1}{6}\pi \approx 0.5236 \geq 2 - \sqrt{3} \approx 0.2679$。
- 当 $t \to \infty$ 时，两边都是 $\frac{1}{t}$ 的量级：$\arcsin(t^{-1}) \approx t^{-1} + \frac{1}{6}t^{-3}$，而 $t - \sqrt{t^2 - 1} = \frac{1}{t + \sqrt{t^2 - 1}} \approx \frac{1}{2t}$。所以这个下界约为真实值的一半。
- 我们在 2×10^5 个随机点 (a, b) 上测试了第 (iii) 部分的不等式；它从未失效，而两边的差距只在曲线 $b = \sin a$ 附近才缩小至零。我们亦以随机的 p、a 和 b 验证了杨氏不等式（Young's inequality），并用辛普森法则（Simpson's rule）核对了两个积分。

常见错误

- 在第 (i) 部分只画出 $b \leq f(a)$ 的情况。$b \geq f(a)$ 的情况是对称的，但你应该明确指出这一点。
- 写出 $f^{-1}(y) = y^{1/(p-1)}$ 之后，没有察觉 $1/(p-1) = q - 1$，以致得不出积分 b^q/q。
- 在最后一步中，试图寻找使界限最佳的 a。那个 a 是 $\arcsin(t^{-1})$，但它只会给出一个平凡的恒等式。最简单的选择 $a = 0$ 才是行得通的那一个。

STEP 之外

第 (ii) 部分就是杨氏不等式,它是推导赫尔德(Hölder)不等式 $\sum |x_i y_i| \leq \left(\sum |x_i|^p\right)^{1/p} \left(\sum |y_i|^q\right)^{1/q}$ 的标准途径，而赫尔德不等式又是分析学中 L^p 空间（L^p spaces）的基础。把 $\int f$ 与 $\int f^{-1}$ 配对的做法，正是勒让德变换（Legendre transform）和凸对偶（convex duality）的雏形，这些概念在力学（由拉格朗日量（Lagrangian）过渡到哈密顿量（Hamiltonian））和最佳化（optimisation）中都会再次出现。有关分析学的背景知识，可参阅 Bartle 与 Sherbert 的 *Introduction to Real Analysis*（《实分析导论》）。

详细解答

STEP III 2011，第 5 题

纯数学 · 以积分求面积

STEP III 2011，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/11-s3-q5

本题在 *Beyond STEP* 第 4.9 节 *Holditch* 定理中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch04/holditch.html

关键思路

直角坐标（cartesian）公式 $(*)$ 之所以有用，是因为它是双线性（bilinear）的：若 A 是 P 平移一个只依赖 t 的向量所得，那么 OA 扫过的面积可拆成三部分：OP 扫过的面积、单由平移产生的面积，以及一个交叉项（cross term）。这根杆正好给出这类平移，即 $\mp$ 单位向量（unit vector）$(\cos t, \sin t)$ 的某个倍数。交叉项 f 看似无从计算，但我们根本不需要算它：A 和 B 都沿同一条曲线 $\mathcal{C}$ 走一圈，所以 $[A] = [B]$，而这条等式就确定了 f。

解答

面积公式

设 $x = r\cos\theta$、$y = r\sin\theta$，其中 r 和 θ 都是 t 的函数（function），由乘积法则（product rule）得

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \dot{r}\cos\theta - r\sin\theta\,\dot{\theta}, \qquad \frac{\mathrm{d}y}{\mathrm{d}t} = \dot{r}\sin\theta + r\cos\theta\,\dot{\theta},$$

其中上方的点表示 $\mathrm{d}/\mathrm{d}t$。于是

$$\begin{aligned} x\frac{\mathrm{d}y}{\mathrm{d}t} - y\frac{\mathrm{d}x}{\mathrm{d}t} &= r\cos\theta\,(\dot{r}\sin\theta + r\cos\theta\,\dot{\theta}) - r\sin\theta\,(\dot{r}\cos\theta - r\sin\theta\,\dot{\theta}) \\ &= r^2(\cos^2\theta + \sin^2\theta)\,\dot{\theta} = r^2\frac{\mathrm{d}\theta}{\mathrm{d}t}, \end{aligned}$$

因为含 $r\dot{r}\sin\theta\cos\theta$ 的项互相抵消。把积分（integral）变量由 θ 换成 t，

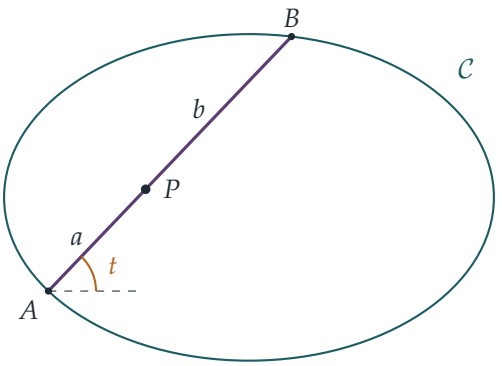
$$\tfrac{1}{2}\int r^2\,\mathrm{d}\theta = \tfrac{1}{2}\int r^2\frac{\mathrm{d}\theta}{\mathrm{d}t}\,\mathrm{d}t = \tfrac{1}{2}\int\left(x\frac{\mathrm{d}y}{\mathrm{d}t} - y\frac{\mathrm{d}x}{\mathrm{d}t}\right)\mathrm{d}t,$$

这正是 $(*)$。

A 和 B 的坐标

杆的方向是单位向量 $\mathbf{u} = (\cos t, \sin t)$，由 A 指向 B，而 P 位于两者之间，且 $AP = a$、$PB = b$。所以 A 是 $P - a\mathbf{u}$，B 是 $P + b\mathbf{u}$：

$$A = (x - a\cos t,\ y - a\sin t), \qquad B = (x + b\cos t,\ y + b\sin t).$$



OA 扫过的面积

同样以点表示 $\mathrm{d}/\mathrm{d}t$。对 A 有

$$x_A = x - a\cos t, \quad \dot{x}_A = \dot{x} + a\sin t, \qquad y_A = y - a\sin t, \quad \dot{y}_A = \dot{y} - a\cos t.$$

展开得

$$x_A\dot{y}_A = x\dot{y} - a\cos t\,(x + \dot{y}) + a^2\cos^2 t,$$
$$y_A\dot{x}_A = y\dot{x} + a\sin t\,(y - \dot{x}) - a^2\sin^2 t,$$

所以

$$x_A\dot{y}_A - y_A\dot{x}_A = (x\dot{y} - y\dot{x}) + a^2 - a\Big[(x + \dot{y})\cos t + (y - \dot{x})\sin t\Big].$$

现在以 t 由 0 至 2π 应用 $(*)$。第一个括号给出 $[P]$，常数 a^2 给出 $\frac{1}{2}\cdot 2\pi a^2 = \pi a^2$，而根据 f 的定义，最后一项给出 $-af$。因此

$$[A] = [P] + \pi a^2 - af,$$

即为所求。

对 B 而言，唯一的改变是处处把 $-a$ 换成 $+b$（即在上面的推导中以 $-b$ 代替 a）。所以

$$[B] = [P] + \pi b^2 + bf.$$

$\mathcal{C}$ 与 $\mathcal{D}$ 之间的面积

想一想 $[A]$、$[B]$ 和 $[P]$ 实际上等于甚么?

当 t 由 0 变到 2π，杆转了整整一圈。两个端点都沿 $\mathcal{C}$ 逆时针走一圈，并回到起点：转了半圈时，A 和 B 互换位置；转了一整圈后，两者都回到原位。原点位于 $\mathcal{D}$ 之内，因此也位于 $\mathcal{C}$ 之内。所以 OA 和 OB 各自恰好扫过 $\mathcal{C}$ 所围的区域一次，而 OP 扫过 $\mathcal{D}$ 所围的区域一次。以 $|\mathcal{C}|$ 和 $|\mathcal{D}|$ 表示这两个面积，

$$[A] = [B] = |\mathcal{C}|, \qquad [P] = |\mathcal{D}|.$$

令 $[A] = [B]$，便可消去 $[P]$：

$$\pi a^2 - af = \pi b^2 + bf \quad \Longrightarrow \quad (a+b)f = \pi(a^2 - b^2) = \pi(a-b)(a+b).$$

由于 $a + b > 0$（它是杆的长度），可以两边相除，得 $f = \pi(a - b)$。因此

$$|\mathcal{C}| - |\mathcal{D}| = [A] - [P] = \pi a^2 - a\pi(a-b) = \pi ab.$$

所以两曲线之间的面积是 πab。改用 $[B]$ 计算，同样得到 $\pi b^2 + b\pi(a-b) = \pi ab$，这是一个令人安心的一致性检查。

验算

- 圆形。若 $\mathcal{C}$ 是半径为 R 的圆,长度为 $a+b$ 的弦与圆心的距离恒为 $d = \sqrt{R^2 - \frac{1}{4}(a+b)^2}$。由毕氏定理（Pythagoras），$P$ 与圆心的距离是 $\sqrt{d^2 + \frac{1}{4}(a-b)^2}$，所以 $\mathcal{D}$ 是一个同心圆（concentric circle），其半径的平方为 $R^2 - ab$。两者之间的面积是 $\pi R^2 - \pi(R^2 - ab) = \pi ab$。
- 对称性。答案关于 a 和 b 对称，这是必然的，因为把杆倒转便会互换两者的角色。
- 数值验证。我们让一根 $a = 1$、$b = 1.5$ 的杆沿椭圆 $x^2/16 + y^2/9 = 1$ 移动，并用鞋带公式（shoelace formula）计算所描出曲线的面积。两曲线之间的面积算得 4.712390, 与 $\pi ab = 1.5\pi \approx 4.712389$ 吻合。我们亦对随机选取的曲线 $P(t)$,用数值积分（quadrature）验证了 $[A]$ 和 $[B]$ 的恒等式。

常见错误

- 把 $\dot{x}_A = \dot{x} + a\sin t$ 的正负号弄错。对 $-a\cos t$ 求导得 $+a\sin t$。
- 试图直接计算 f。它取决于 $\mathcal{D}$ 的形状，没有更多资料便无法算出。关键在于 $[A] = [B]$ 就能确定它。

STEP 之外

这就是 Holditch 定理（Holditch's theorem，1858 年）：$\mathcal{C}$ 与 $\mathcal{D}$ 之间的面积只取决于 a 和 b，而与 $\mathcal{C}$ 的形状无关。公式 $(*)$ 是格林定理（Green's theorem）的特例，该定理把区域上的面积分转化为沿其边界的线积分。它也是求积仪（planimeter）背后的原理；求积仪是一种沿边界描画来量度面积的机械装置。格林定理是大学一年级向量微积分（vector calculus）的核心内容，而 Needham 的 *Visual Complex Analysis*（《直观复分析》）对这类边界积分给出了优美的几何观点。

详细解答

STEP III 2011，第 6 题

纯数学 · 积分代换

STEP III 2011，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/11-s3-q6

关键思路

题目不容许我们计算任何积分的值，所以可用的工具只有那些把一个积分变成另一个积分的方法：代换法（substitution）和分部积分法（integration by parts）。积分上下限就是线索。数 $\frac{1}{3}, \frac{1}{2}$ 与 $\ln 2, \ln 3$ 由 $v = e^{-u}$ 联系起来，而 $\frac{1}{2}\ln 2, \frac{1}{2}\ln 3$ 则由 $t = e^{-2x}$ 与 $\frac{1}{2}, \frac{1}{3}$ 联系起来。积分 T 和 V 的上下限相同，而 $\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{artanh} t = \frac{1}{1-t^2}$ 提示我们可以在两者之间用分部积分。三条联系已足以使四个积分全部相等。



解答

全文使用

$$\operatorname{artanh} t = \tfrac{1}{2}\ln\frac{1+t}{1-t} \quad (|t|<1), \qquad \frac{\mathrm{d}}{\mathrm{d}t}\operatorname{artanh} t = \frac{1}{1-t^2},$$

特别地，$\operatorname{artanh}\frac{1}{2} = \frac{1}{2}\ln 3$ 及 $\operatorname{artanh}\frac{1}{3} = \frac{1}{2}\ln 2$。所有被积函数（integrand）在其（闭且有界的）区间上都是连续的，因此每个积分都存在，下面的运算都是合理的。

$T = V$：分部积分

在 T 中，我们对 $\frac{1}{t}$ 积分、对 $\operatorname{artanh} t$ 求导，取 $u = \operatorname{artanh} t$ 及 $\mathrm{d}w = \frac{\mathrm{d}t}{t}$，于是 $w = \ln t$（由于在区间上 $t > 0$，这没有问题）：

$$\begin{aligned} T &= \Big[\ln t \, \operatorname{artanh} t\Big]_{1/3}^{1/2} - \int_{1/3}^{1/2} \frac{\ln t}{1-t^2}\,\mathrm{d}t \\ &= \Big[\ln t \, \operatorname{artanh} t\Big]_{1/3}^{1/2} + V, \end{aligned}$$

因为余下的积分正正就是 V（哑变量（dummy variable）用甚么名称并不重要）。

> **想一想** 边界项（boundary term）不会破坏等式吗?

它等于零，而这正是本题巧妙之处。在上限，$\ln\frac{1}{2}\cdot\operatorname{artanh}\frac{1}{2} = (-\ln 2)(\frac{1}{2}\ln 3)$；在下限，$\ln\frac{1}{3}\cdot\operatorname{artanh}\frac{1}{3} = (-\ln 3)(\frac{1}{2}\ln 2)$。两者都等于 $-\frac{1}{2}\ln 2\ln 3$，所以

$$\Big[\ln t \, \operatorname{artanh} t\Big]_{1/3}^{1/2} = -\tfrac{1}{2}\ln 2\ln 3 + \tfrac{1}{2}\ln 2\ln 3 = 0,$$

因此 $T = V$。

$V = U$：代换 $v = e^{-u}$

设 $v = e^{-u}$，则 $\ln v = -u$ 及 $\mathrm{d}v = -e^{-u}\,\mathrm{d}u$。当 $v = \frac{1}{3}$ 时，$u = \ln 3$；当 $v = \frac{1}{2}$ 时，$u = \ln 2$。于是

$$\begin{aligned} V &= -\int_{\ln 3}^{\ln 2} \frac{-u}{1-e^{-2u}}\,(-e^{-u})\,\mathrm{d}u = -\int_{\ln 3}^{\ln 2} \frac{u\,e^{-u}}{1-e^{-2u}}\,\mathrm{d}u \\ &= \int_{\ln 2}^{\ln 3} \frac{u\,e^{-u}}{1-e^{-2u}}\,\mathrm{d}u \qquad \text{（交换上下限）} \\ &= \int_{\ln 2}^{\ln 3} \frac{u}{e^{u}-e^{-u}}\,\mathrm{d}u \qquad \text{（分子分母同乘以 } e^{u}\text{）} \\ &= \int_{\ln 2}^{\ln 3} \frac{u}{2\sinh u}\,\mathrm{d}u = U. \end{aligned}$$

$X = T$：代换 $t = e^{-2x}$

首先改写 X 的被积函数。对 $x > 0$，

$$\coth x = \frac{e^{x}+e^{-x}}{e^{x}-e^{-x}} = \frac{1+e^{-2x}}{1-e^{-2x}}, \qquad \text{所以} \qquad \ln(\coth x) = \ln\frac{1+e^{-2x}}{1-e^{-2x}} = 2\operatorname{artanh}\big(e^{-2x}\big),$$

由于 $0 < e^{-2x} < 1$，此式成立。现设 $t = e^{-2x}$，则 $\mathrm{d}t = -2t\,\mathrm{d}x$，即 $\mathrm{d}x = -\frac{\mathrm{d}t}{2t}$。当 $x = \frac{1}{2}\ln 2$ 时，$t = \frac{1}{2}$；当 $x = \frac{1}{2}\ln 3$ 时，$t = \frac{1}{3}$。因此

$$X = \int_{1/2}^{1/3} 2\operatorname{artanh} t\left(-\frac{1}{2t}\right)\mathrm{d}t = \int_{1/3}^{1/2} \frac{\operatorname{artanh} t}{t}\,\mathrm{d}t = T.$$

结论

把三条联系合起来，得 $X = T = V = U$，所以四个积分全部相等，得证。我们始终无须求出任何一个积分的值。

验算

对每个积分分别使用辛普森法则（Simpson's rule），得 $T \approx U \approx V \approx X \approx 0.177704$，大约在小数点后 12 位都吻合。用计算机快速验证边界项为零也令人安心：$\ln \frac{1}{2} \cdot \operatorname{artanh} \frac{1}{2}$ 和 $\ln \frac{1}{3} \cdot \operatorname{artanh} \frac{1}{3}$ 的结果都是 $-0.3807\ldots$。

常见错误

- 在 $V = U$ 中漏掉负号：共有三个负号（V 前面的一个、$\ln v = -u$ 中的一个及 $\mathrm{d}v = -e^{-u}\,\mathrm{d}u$ 中的一个），而交换上下限又带来第四个。
- 因为边界项看似不为零，便放弃分部积分，却没有在两端把它算出来。
- 写成 $\ln \coth x = \operatorname{artanh}(e^{-2x})$，漏掉了因子 2；这个因子稍后会被 $\mathrm{d}x = -\frac{\mathrm{d}t}{2t}$ 中的 $\frac{1}{2}$ 抵消。

STEP 之外

把 $\operatorname{artanh} t/t$ 展开成幂级数（power series），可见这个共同值等于 $\chi_2(\frac{1}{2}) - \chi_2(\frac{1}{3})$，其中 $\chi_2(z) = \sum_{k\geq 0} z^{2k+1}/(2k+1)^2$ 是勒让德（Legendre）chi 函数，与双对数函数（dilogarithm）$\mathrm{Li}_2(z) = \sum_{n\geq 1} z^n/n^2$ 关系密切。这些函数的许多恒等式，正是来自这类变量代换。不计算任何一个积分而证明两个积分相等，是数学分析（analysis）中的标准技巧；而支持代换法和分部积分法的定理（连续性（continuity）、微积分基本定理（fundamental theorem of calculus））会在第一门分析课程中得到严格的表述。Abbott 的 *Understanding Analysis*（《理解分析》）是看这些定理被仔细证明的好去处。

详细解答

STEP III 2011，第 7 题

纯数学 · 数学归纳法

STEP III 2011，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/11-s3-q7

关键思路

把 $\sqrt{a+1}+\sqrt{a}$ 平方，就能消去两个独立的平方根，剩下形如 $p+q\sqrt{a(a+1)}$ 的数；因此当 n 为偶数时，我们可以通过乘以 T_2，由 T_n 推进到 T_{n+2}。接着会发现，量 $A^2-a(a+1)B^2$ 在这个乘法下保持不变，而这正是归纳法（induction）所需要的。再多乘一个因子 $\sqrt{a+1}+\sqrt{a}$ 便能处理 n 为奇数的情况；而 (i) 和 (ii) 中的恒等式（identity）说明，两项分别是两个相邻整数的平方根。

解答

全文记 $N=a(a+1)$，于是 $\sqrt{a+1}\,\sqrt{a}=\sqrt{N}$。

第 (i) 部分

基本情况（*base case*），$n=2$。展开得

$$T_2=(a+1)+2\sqrt{a+1}\,\sqrt{a}+a=(2a+1)+2\sqrt{N},$$

所以 $A_2=2a+1$ 和 $B_2=2$ 都是整数，而且

$$A_2^2-NB_2^2=4a^2+4a+1-4a(a+1)=1.$$

归纳步骤（*inductive step*）。假设对某个偶数 $n\geq 2$，有 $T_n=A_n+B_n\sqrt{N}$，其中整数 A_n, B_n 满足 $A_n^2-NB_n^2=1$。为简便起见，记 $p=2a+1$，并注意 $p^2=4a^2+4a+1=4N+1$。

由于 $T_{n+2} = T_n T_2$，

$$\begin{aligned} T_{n+2} &= \left(A_n + B_n\sqrt{N}\right)\left(p + 2\sqrt{N}\right) \\ &= \left(pA_n + 2NB_n\right) + \left(2A_n + pB_n\right)\sqrt{N}. \end{aligned}$$

因此可取

$$A_{n+2} = pA_n + 2NB_n, \qquad B_{n+2} = 2A_n + pB_n,$$

由于 p、N、A_n 和 B_n 都是整数，它们也是整数。至于恒等式，交叉项 $\pm 4pNA_nB_n$ 互相抵消：

$$\begin{aligned} A_{n+2}^2 - NB_{n+2}^2 &= \left(p^2A_n^2 + 4pNA_nB_n + 4N^2B_n^2\right) - N\left(4A_n^2 + 4pA_nB_n + p^2B_n^2\right) \\ &= (p^2 - 4N)A_n^2 - N(p^2 - 4N)B_n^2 \\ &= A_n^2 - NB_n^2 = 1, \end{aligned}$$

这里先用了 $p^2 - 4N = 1$，再用归纳假设（inductive hypothesis）。所以命题对 $n+2$ 成立；由归纳法，命题对每个正偶数 n 都成立。

另外注意 $A_2, B_2 > 0$，而递推关系（recurrence）的系数都是正数，所以对每个偶数 n，都有 $A_n > 0$ 和 $B_n > 0$。第 (iii) 部分会用到这一点。

想一想 为甚么 $A^2 - NB^2$ 在乘以 T_2 之后仍然保持不变?

因为 $A^2 - NB^2 = (A + B\sqrt{N})(A - B\sqrt{N})$，而同样的计算显示，“共轭”（conjugate）$A - B\sqrt{N}$ 在每一步都被乘以 $p - 2\sqrt{N}$。因此两者之积每一步都被乘以 $(p + 2\sqrt{N})(p - 2\sqrt{N}) = p^2 - 4N = 1$。上面的代数运算只不过是把这个事实逐步写出来而已。

第 (ii) 部分

$n = 1$ 的情况。$T_1 = 1 \cdot \sqrt{a+1} + 1 \cdot \sqrt{a}$，所以 $C_1 = D_1 = 1$，且 $(a+1)C_1^2 = a + 1 = aD_1^2 + 1$。

$n \geq 3$ 为奇数的情况。此时 $m = n - 1$ 是正偶数，所以由第 (i) 部分，$T_m = A_m + B_m\sqrt{N}$，且 $A_m^2 - NB_m^2 = 1$。利用 $\sqrt{N}\sqrt{a+1} = (a+1)\sqrt{a}$ 和 $\sqrt{N}\sqrt{a} = a\sqrt{a+1}$，

$$\begin{aligned} T_n = \left(\sqrt{a+1} + \sqrt{a}\right)\left(A_m + B_m\sqrt{N}\right) &= A_m\sqrt{a+1} + A_m\sqrt{a} + (a+1)B_m\sqrt{a} + aB_m\sqrt{a+1} \\ &= \left(A_m + aB_m\right)\sqrt{a+1} + \left(A_m + (a+1)B_m\right)\sqrt{a}. \end{aligned}$$

所以 $C_n = A_m + aB_m$ 和 $D_n = A_m + (a+1)B_m$ 都是整数（而且是正数）。于是

$$\begin{aligned} (a+1)C_n^2 - aD_n^2 &= (a+1)\left(A_m^2 + 2aA_mB_m + a^2B_m^2\right) \\ &\quad - a\left(A_m^2 + 2(a+1)A_mB_m + (a+1)^2B_m^2\right) \\ &= A_m^2 + a(a+1)\left[a - (a+1)\right]B_m^2 \\ &= A_m^2 - NB_m^2 = 1, \end{aligned}$$

因为含 A_mB_m 的项互相抵消。因此 $(a+1)C_n^2 = aD_n^2 + 1$，得证。（$n=1$ 的情况其实是同一个计算，只是取 $T_0 = 1$，即 $A_0 = 1$、$B_0 = 0$。）

第 (iii) 部分

n 为偶数。由于 $A_n > 0$ 且 $B_n > 0$，

$$T_n = A_n + B_n\sqrt{N} = \sqrt{A_n^2} + \sqrt{NB_n^2},$$

而由第 (i) 部分，整数 NB_n^2 和 $A_n^2 = NB_n^2 + 1$ 是相邻的。

> **想一想** 为甚么我们需要知道 A_n 和 B_n 的正负号?

因为 $\sqrt{A^2} = |A|$，而不是 A。只有在知道 $A_n \geq 0$ 之后，写 $A_n = \sqrt{A_n^2}$ 才成立，这就是我们在第 (i) 部分记下正性的原因。

n 为奇数。由于 $C_n, D_n > 0$，

$$T_n = C_n\sqrt{a+1} + D_n\sqrt{a} = \sqrt{(a+1)C_n^2} + \sqrt{aD_n^2},$$

而由第 (ii) 部分，整数 aD_n^2 和 $(a+1)C_n^2 = aD_n^2 + 1$ 是相邻的。

所以对每个正整数 n，都存在某个正整数 k 使 $T_n = \sqrt{k+1} + \sqrt{k}$：具体而言，$n$ 为偶数时 $k = a(a+1)B_n^2$，n 为奇数时 $k = aD_n^2$。

验算

取 $a = 3$，则 $T_n = (2+\sqrt{3})^n$。于是 $T_2 = 7 + 4\sqrt{3} = 7 + 2\sqrt{12} = \sqrt{49} + \sqrt{48}$，且 $49 = 12 \cdot 4 + 1$。接着，由 $A_2 = 7$、$B_2 = 2$ 得 $C_3 = 13$、$D_3 = 15$，以及 $T_3 = 26 + 15\sqrt{3} = \sqrt{676} + \sqrt{675} \approx 51.9808$，与直接计算 $(2+\sqrt{3})^3$ 的结果一致。我们用电脑对 $1 \leq a \leq 39$ 和 $n \leq 23$ 运行了递推关系，以整数运算准确验证了两条恒等式，并以数值方式验证了 T_n 的展开式。

常见错误

- 忘记基本情况，或者在第 (i) 部分只涉及偶数 n 的情况下，却从 $n = 1$ 开始归纳。
- 在第 (iii) 部分写 $A_n = \sqrt{A_n^2}$，却没有指出 A_n（以及 B_n、C_n、D_n）是正数。

STEP 之外

方程 $x^2 - Ny^2 = 1$ 就是佩尔方程（Pell's equation），而第 (i) 部分说明，$(2a+1) + 2\sqrt{a(a+1)}$ 的各次幂为 $N = a(a+1)$ 生成无限多组解。关键事实是：$x + y\sqrt{N}$ 的“范数”（norm）$x^2 - Ny^2$ 具有积性（multiplicative），这正是二次数域（quadratic number field）及其单位（unit）理论的起点。达文波特（Davenport）的 *The Higher Arithmetic*（《高等算术》）对佩尔方程和连分数（continued fraction）有非常易读的论述。

详细解答

STEP III 2011，第 8 题

纯数学 · 复数轨迹

STEP III 2011，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/11-s3-q8

关键思路

将分子和分母同乘以分母的共轭（conjugate），便得到 u 和 v；之后每一部分只需代入题目给出的直线，再认出一条熟悉的曲线。半角代换（half-angle substitution）把 $\frac{2x}{1+x^2}$ 和 $\frac{x^2-1}{x^2+1}$ 变成正弦和余弦，这正是题目建议用它的原因。整个过程中，要时刻留意曲线上哪些点真正被取到：端点和缺失的点占了答案的一半。

解答

设 $z = x + \mathrm{i}y$。则 $1 + \mathrm{i}z = (1-y) + \mathrm{i}x$ 及 $\mathrm{i} + z = x + \mathrm{i}(1+y)$，因此把分子和分母同乘以 $x - \mathrm{i}(1+y)$，得

$$\begin{aligned} w &= \frac{\big((1-y)+\mathrm{i}x\big)\big(x-\mathrm{i}(1+y)\big)}{x^2+(1+y)^2} \\ &= \frac{(1-y)x + x(1+y) + \mathrm{i}\big(x^2-(1-y)(1+y)\big)}{x^2+(1+y)^2}. \end{aligned}$$

因此

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

适用于 $z \neq -\mathrm{i}$（在该点 w 无定义）。

知道逆映射（inverse map）亦很有帮助。由 $w(\mathrm{i}+z) = 1 + \mathrm{i}z$ 得 $z(w-\mathrm{i}) = 1 - \mathrm{i}w$，所以

$$z = \frac{1-\mathrm{i}w}{w-\mathrm{i}}, \qquad w \neq \mathrm{i}.$$

因此每个 $w \neq \mathrm{i}$ 都恰好来自一个 z，而 $w = \mathrm{i}$ 永远取不到。这就是我们会一再遇到的那个缺失点。

第 (i) 部分

当 $y=0$ 时，

$$u=\frac{2x}{1+x^2},\qquad v=\frac{x^2-1}{x^2+1}.$$

令 $x=\tan(\theta/2)$。当 x 取遍所有实数时，θ 取遍 $-\pi<\theta<\pi$，而每个 x 恰好对应一个这样的 θ。由半角公式，

$$u=\frac{2\tan(\theta/2)}{1+\tan^2(\theta/2)}=\sin\theta,\qquad v=-\frac{1-\tan^2(\theta/2)}{1+\tan^2(\theta/2)}=-\cos\theta.$$

所以 $u^2+v^2=1$。此外，当 $-\pi<\theta<\pi$ 时，$(\sin\theta,-\cos\theta)$ 描出整个单位圆（unit circle），只缺 $\theta=\pm\pi$ 所对应的点，即 $(u,v)=(0,1)$。因此 w 的轨迹（locus）是圆 $u^2+v^2=1$ 去掉单一点 $w=\mathrm{i}$，正如所求。（缺失点是 $x\to\pm\infty$ 时 w 的极限。）

第 (ii) 部分

现在 $-1<x<1$，对应 $-\frac{\pi}{2}<\theta<\frac{\pi}{2}$。在此范围内 $v=-\cos\theta<0$，而 $u=\sin\theta$ 由 -1 递增至 1（不包括两端）。所以 w 的轨迹是

$$u^2+v^2=1,\quad v<0,$$

即单位圆的下半部分，不包括其端点 $w=\pm1$（它们来自 $x=\pm1$）。

第 (iii) 部分

当 $x=0$ 时，得 $u=0$ 及

$$v=\frac{y^2-1}{(1+y)^2}=\frac{(y-1)(y+1)}{(y+1)^2}=\frac{y-1}{y+1}=1-\frac{2}{y+1},$$

由于 $y>-1$，约去 $y+1$ 是没有问题的。

> ? **想一想** 是否每个负的 v 值都能取到，还是只取到其中一部分?

当 y 由 -1 增加至 1，$y+1$ 在 $(0,2)$ 中递增，所以 $\frac{2}{y+1}$ 在 $(1,\infty)$ 中递减，而 v 则在 $(-\infty,0)$ 中连续且严格递增。当 $y\to-1^+$ 时，$v\to-\infty$；当 $y\to1^-$ 时，$v\to0^-$。所以每个负的 v 都恰好取到一次。w 的轨迹是负虚轴

$$u=0,\quad v<0,$$

不包括原点（它来自 $y=1$）。

第 (iv) 部分

当 $y = 1$ 时，

$$u = \frac{2x}{x^2+4}, \qquad v = \frac{x^2}{x^2+4}.$$

于是

$$u^2 + v^2 = \frac{4x^2 + x^4}{(x^2+4)^2} = \frac{x^2(x^2+4)}{(x^2+4)^2} = \frac{x^2}{x^2+4} = v,$$

所以 w 位于 $u^2 + v^2 = v$ 上，即

$$u^2 + \left(v - \tfrac{1}{2}\right)^2 = \tfrac{1}{4},$$

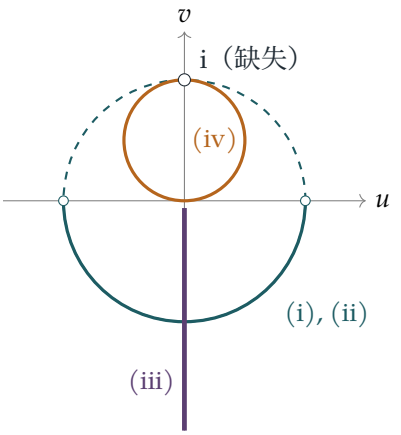
这是圆心为 $\frac{1}{2}\mathrm{i}$、半径为 $\frac{1}{2}$ 的圆。

> ? **想一想** 这个圆上哪些点真正被取到?

由于 $v = 1 - \frac{4}{x^2+4}$，我们有 $0 \le v < 1$，所以最高点 $w = \mathrm{i}$ 永远取不到。其余每一点都能取到：令 $x = 2\tan(\phi/2)$，其中 $-\pi < \phi < \pi$，得

$$u = \frac{\tan(\phi/2)}{1+\tan^2(\phi/2)} = \tfrac{1}{2}\sin\phi, \qquad v = \frac{\tan^2(\phi/2)}{1+\tan^2(\phi/2)} = \tfrac{1}{2}(1-\cos\phi),$$

它绕整个圆一周，只缺 $\phi = \pm\pi$，即点 $(0,1)$。（另一种方法是用逆映射，它说明每个 $w \ne \mathrm{i}$ 都恰好来自一个 z。）所以轨迹是圆 $u^2 + v^2 = v$ 去掉点 $w = \mathrm{i}$。它在 $w = 0$ 与实轴相切，该点是 $z = \mathrm{i}$ 的像。



由虚线和实线组成的青色圆是实轴的像，即第 (i) 部分；实线的下半部分是第 (ii) 部分。紫红色射线是第 (iii) 部分，琥珀色圆是第 (iv) 部分。

验算

- $z = 0$ 给出 $w = 1/\mathrm{i} = -\mathrm{i}$：它既在下半圆 (ii) 上，也在射线 (iii) 上，而 $z = 0$ 确实同时位于这两段线上。
- $z = 1 + \mathrm{i}$ 给出 $w = \frac{\mathrm{i}}{1+2\mathrm{i}} = 0.4 + 0.2\mathrm{i}$，而 $0.4^2 + 0.2^2 = 0.2 = v$，所以它位于 (iv) 的圆上。
- 很大的实数 z：$w(10^8) \approx 2 \times 10^{-8} + \mathrm{i}$，趋近缺失点 i。
- 我们亦以数值方法对数以千计的随机点作了验证，既验证了 u, v 的公式，也验证了每个所声称的轨迹上的每一点，都在正确的直线或线段上有原像（preimage）。

常见错误

- 只写出圆或半圆，却没有说明哪些点被去掉或不包括在内：这里是 (i) 和 (iv) 中的 $w = \mathrm{i}$，以及 (ii) 中的端点 ± 1。
- 声称 (iii) 的轨迹是整条虚轴：对每个 $-1 < y < 1$，$v = \frac{y-1}{y+1}$ 都是负数。

STEP 之外

这个映射是一个默比乌斯变换（Möbius transformation）；事实上 $w = \mathrm{i}\,\dfrac{z-\mathrm{i}}{z+\mathrm{i}}$，所以 $|w| = |z-\mathrm{i}|/|z+\mathrm{i}|$，而它等于 1 当且仅当 z 与 $\pm\mathrm{i}$ 等距，亦即当它是实数时。默比乌斯映射把直线和圆映成直线和圆（直线被视为“经过 ∞ 的圆”），而缺失点 i 正是 ∞ 的像。这个映射（即凯莱变换（Cayley transform））把上半平面映成单位圆盘，是复分析（complex analysis）和双曲几何（hyperbolic geometry）中的重要工具。Needham 的 *Visual Complex Analysis*（《直观复分析》）以图像为主导，对默比乌斯变换有十分精彩的论述。

详细解答

STEP III 2011，第 9 题

力学 • 圆柱上的运动

STEP III 2011，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：

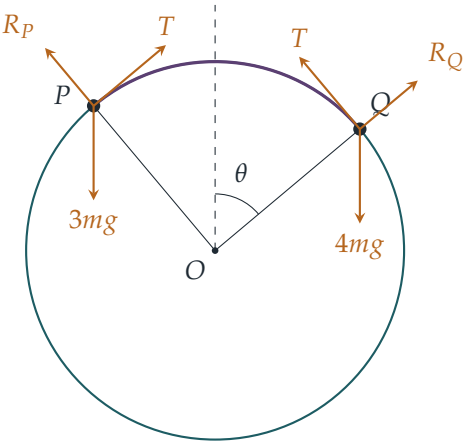
step.maths.org/questions/11-s3-q9

关键思路

长度为 $\frac{1}{2}\pi a$ 的绳子贴在圆柱面上，在 O 点所张的角是直角，因此 OP 与 OQ 始终互相垂直，整个系统只需一个坐标 θ 来描述。由能量守恒（conservation of energy）可直接得出 $\dot{\theta}^2$，再沿指向圆心的方向应用牛顿第二定律（Newton's second law），便可把每个法向反作用力（normal reaction）以 $\dot{\theta}^2$ 表示。质点离开表面，即表示法向反作用力降至零。至于张力（tension），则用切向（tangential）方程：把两条方程相加可消去 T 并得出 $\ddot{\theta}$，之后任取其中一条便可求出 T。

解答

模型。P 和 Q 是质点 (particle), 绳子轻而不可伸长 (inextensible), 圆柱面光滑 (smooth)。轻绳绕过光滑表面时，各处张力同为 T。两质点之间的弧长为 $\frac{1}{2}\pi a$，所以 $\angle POQ = \frac{1}{2}\pi$。设 θ 为 OQ 与向上铅垂线所成的角，Q 位于最高点的一侧。那么 OP 在另一侧与向上铅垂线所成的角为 $\frac{1}{2}\pi - \theta$。只要两质点仍留在表面上，它们都在半径为 a 的圆上以角速率（angular speed）$\dot{\theta}$ 运动，因此各自的速率均为 $a\dot{\theta}$。



平衡。在每个质点处沿表面方向分解力（法向反作用力没有切向分量，而绳子沿切向拉向最高点）：对 Q 得 $T = 4mg\sin\theta$，对 P 得 $T = 3mg\sin(\frac{1}{2}\pi - \theta) = 3mg\cos\theta$。因此在平衡（equilibrium）位置时 $\tan\theta = \frac{3}{4}$，即 $\sin\theta = \frac{3}{5}$ 及 $\cos\theta = \frac{4}{5}$。

能量。取 O 所在高度为势能（potential energy）零点。Q 的高度为 $a\cos\theta$，P 的高度为 $a\cos(\frac{1}{2}\pi - \theta) = a\sin\theta$。法向反作用力不做功（它们与运动方向垂直），而张力对整个系统不做净功（绳子不可伸长），所以能量守恒。由于扰动轻微，系统由静止开始运动，故

$$\tfrac{1}{2}(3m+4m)a^2\dot\theta^2 + 4mga\cos\theta + 3mga\sin\theta = 4mga\cdot\tfrac{4}{5} + 3mga\cdot\tfrac{3}{5} = 5mga.$$

两边乘以 $2/(ma)$，得

$$7a\dot\theta^2 + 8g\cos\theta + 6g\sin\theta = 10g,$$

此即所求。当 Q 向下移动时，θ 由 $\tan^{-1}\frac{3}{4}$ 开始增大。

第 (i) 部分

Q 在半径为 a 的圆上以速率 $a\dot\theta$ 运动，所以它有指向 O 的加速度 $a\dot\theta^2$。其重量指向 O 的分量为 $4mg\cos\theta$，而 R_Q 向外作用。绳子沿切向，因此在径向上没有贡献。于是

$$4mg\cos\theta - R_Q = 4ma\dot\theta^2, \qquad \text{所以} \qquad R_Q = 4mg\cos\theta - 4ma\dot\theta^2.$$

当 R_Q 降至零，即 $a\dot\theta^2 = g\cos\theta$ 时，Q 离开表面。把此式代入 $\theta = \beta$ 时的能量方程：

$$7g\cos\beta + 8g\cos\beta + 6g\sin\beta = 10g \quad\Longrightarrow\quad 15\cos\beta + 6\sin\beta = 10,$$

此即所求。

> ? **想一想** 为甚么能量方程一直到 $\theta = \beta$ 仍然成立?

只要两个质点都留在圆柱面上,能量方程便成立,而题目告诉我们 Q 先离开。你也可以自行验证这一点。对 P 作同样的推导，得 $R_P = 3mg\sin\theta - 3ma\dot\theta^2$，它在 $13\sin\theta + 8\cos\theta = 10$ 时为零。在 $\beta \approx 73.6^\circ$ 时，左边约为 14.7，所以 R_P 仍然是正数。

第 (ii) 部分

现在沿切向分解力，以运动方向为正方向。Q 以切向加速度 $a\ddot{\theta}$ 远离最高点，而 P 以同样的切向加速度 $a\ddot{\theta}$ 移向最高点（绳子使两者速率保持相等）。对 Q 而言，重力有向前的分量 $4mg\sin\theta$，张力则向后拉：

$$4ma\ddot{\theta} = 4mg\sin\theta - T. \tag{1}$$

对 P 而言，张力把它向前拉，重力有向后的分量 $3mg\cos\theta$：

$$3ma\ddot{\theta} = T - 3mg\cos\theta. \tag{2}$$

把 (1) 和 (2) 相加可消去 T：

$$7ma\ddot{\theta} = 4mg\sin\theta - 3mg\cos\theta.$$

（这与把能量方程对时间求导再除以 $2\dot{\theta}$ 的结果一致。）代入 (1)：

$$\begin{aligned} T &= 4mg\sin\theta - 4ma\ddot{\theta} \\ &= 4mg\sin\theta - \tfrac{4}{7}m(4g\sin\theta - 3g\cos\theta) \\ &= \tfrac{12}{7}mg\sin\theta + \tfrac{12}{7}mg\cos\theta = \tfrac{12}{7}mg(\sin\theta + \cos\theta), \end{aligned}$$

此即所求。当 $0 < \theta < \frac{1}{2}\pi$ 时此值为正，所以绳子在整个过程中都保持拉紧。

验算

- 在平衡位置时，$\sin\theta + \cos\theta = \frac{7}{5}$，所以 $T = \frac{12}{5}mg$。这与 $4mg\cdot\frac{3}{5}$ 及 $3mg\cdot\frac{4}{5}$ 相符。
- 写成 $15\cos\beta + 6\sin\beta = \sqrt{261}\cos(\beta - \alpha)$，其中 $\tan\alpha = \frac{2}{5}$，可得 $\beta \approx 1.2839$ 弧度 $\approx 73.56°$。我们从刚越过平衡位置之处开始以数值方法逐步模拟运动，发现 R_Q 正是在这个角度首次降至零，而 R_P 则保持为正。在整条路径上，能量关系式和张力公式在舍入误差范围内均成立。

常见错误

- 把 P 的高度当作 $a\cos\theta$ 而不是 $a\sin\theta$。绳长的设置，关键正在于直角 $\angle POQ$。
- 给两个质点不同的张力，或把 T 放进径向方程。绳子沿切向接触质点，所以 T 只出现在切向方程中。

STEP 之外

由于一个坐标 θ 便足以描述一切，这是一个单自由度（one degree of freedom）系统。拉格朗日力学（Lagrangian mechanics）可以直接由动能和势能得出运动方程，完全不需要考虑约束力（forces of constraint）。法向反作用力和张力之后会以约束条件的“拉格朗日乘数”（Lagrange multipliers）的形式出现。在接触拉格朗日（Lagrange）的理论之前，Kleppner 和 Kolenkow 所著的 *An Introduction to Mechanics*（《力学导论》）是加深你对约束运动和能量方法理解的好起点。

详细解答

STEP III 2011，第 10 题

力学 · 弹性绳与简谐运动

STEP III 2011，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/11-s3-q10

关键思路

两个质点组成的系统不受任何外来的水平力，所以它们的质心（centre of mass）以由冲量（impulse）决定的恒定速率运动。相对于质心，弦的伸长量（extension）遵从简谐运动（simple harmonic motion），而弦恰好在半个振动周期内保持拉紧。把运动拆分为“质心”加“伸长量”两部分，每个答案几乎不费功夫便会自然得出。

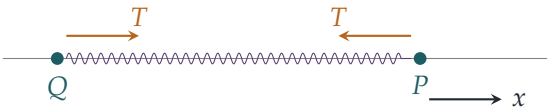
解答

模型与设置

两质点在光滑水平面上运动，因此它们所受的水平力只有轻弦的张力（tension），而这两个力大小相等、方向相反。取直线 QP 为 x 轴，方向由 Q 指向 P，并设 Q 起初位于 $x = 0$，P 起初位于 $x = a$。设 x_P 和 x_Q 为施加冲量后时间 t 时两者的位置，并设

$$e = (x_P - x_Q) - a$$

为弦的伸长量。施加冲量后一刻，$\dot{x}_P = u$，$\dot{x}_Q = 0$ 及 $e = 0$，且 $\dot{e} = u > 0$，因此弦立即变得拉紧。



弦拉紧期间，由胡克定律（Hooke's law）得张力

$$T = \frac{\lambda e}{a} = \frac{\frac{1}{2}ma\omega^2 e}{a} = \frac{1}{2}m\omega^2 e.$$

张力把 P 往回拉向 Q，并把 Q 向前拉向 P：

$$m\ddot{x}_P = -\tfrac{1}{2}m\omega^2 e, \qquad m\ddot{x}_Q = +\tfrac{1}{2}m\omega^2 e.$$

质心

把两条方程相加，得 $\ddot{x}_P + \ddot{x}_Q = 0$。因此位于 $\bar{x} = \frac{1}{2}(x_P + x_Q)$ 的质心 G 以恒定速度运动，该速度等于其初始值：

$$\dot{\bar{x}} = \tfrac{1}{2}(u + 0) = \tfrac{1}{2}u.$$

（这其实就是动量守恒（conservation of momentum）：总动量 mu 由质量 $2m$ 共同分担。）

伸长量

把两条方程相减，

$$\ddot{e} = \ddot{x}_P - \ddot{x}_Q = -\omega^2 e.$$

这是角频率（angular frequency）为 ω 的简谐运动。由 $e(0) = 0$ 及 $\dot{e}(0) = u$，得

$$e = \frac{u}{\omega}\sin\omega t.$$

想一想 为甚么角频率是 ω，而不是 $\omega/\sqrt{2}$ 或 $\sqrt{2}\,\omega$?

每个质点所受的力都是 $\frac{1}{2}m\omega^2 e$，但 e 的变化源于两个质点都在运动：把方程相减时，两者的加速度相加。换一个角度看，两个质量同为 m 的质点，其约化质量（reduced mass）为 $\frac{1}{2}m$，而弦的劲度（stiffness）为 $\lambda/a = \frac{1}{2}m\omega^2$，所以 $\omega^2 = \frac{1}{2}m\omega^2 / \frac{1}{2}m$。题目特意选取这个弹性模量（modulus），令结果整齐漂亮。

弦的长度下一次等于 a，是在 $e = 0$ 再次成立之时，即 $t > 0$ 时首次出现 $\sin\omega t = 0$ 之时：

$$t_1 = \frac{\pi}{\omega}.$$

在整段 $0 < t < \pi/\omega$ 期间都有 $e > 0$，所以弦确实一直拉紧，上述方程亦一直成立。

所走的距离

把每个位置写成质心位置加上或减去间距的一半：

$$x_P = \bar{x} + \tfrac{1}{2}(a + e), \qquad x_Q = \bar{x} - \tfrac{1}{2}(a + e).$$

在 t_1 时，$e = 0$ 再次成立，所以每个质点移动的量都与 G 完全相同，即

$$\tfrac{1}{2}u \cdot \frac{\pi}{\omega} = \frac{\pi u}{2\omega}.$$

为确定这个位移就是所走的距离，我们要验证两个质点都从不折返。求导得

$$\dot{x}_P = \tfrac{1}{2}u + \tfrac{1}{2}\dot{e} = \tfrac{1}{2}u(1+\cos\omega t), \qquad \dot{x}_Q = \tfrac{1}{2}u - \tfrac{1}{2}\dot{e} = \tfrac{1}{2}u(1-\cos\omega t),$$

而对所有 t，两者皆 ≥ 0。所以每个质点都只向前移动，各自走了距离 $\dfrac{\pi u}{2\omega}$，正如所求。

速率

在 $t_1 = \pi/\omega$ 时，$\cos\omega t_1 = -1$，所以

$$\dot{x}_P = 0, \qquad \dot{x}_Q = u.$$

因此 P 瞬间静止，而 Q 以速率 u 向 P 移动。两个质点恰好交换了速度，就如两个等质量物体发生完全弹性碰撞（perfectly elastic collision）一样。这并非巧合：动量守恒，而储存的弹性势能已全数转回动能。

到碰撞为止的时间

在 t_1 之后，Q 向 P 移动，所以间距小于 a。弦变得松弛，不施加任何力，两个质点都不受任何水平力。因此 P 保持静止，而 Q 由距离 P 为 a 之处开始，以恒定速率 u 移动。它再经过时间 a/u 便到达 P。由施加冲量到碰撞的总时间为

$$\frac{\pi}{\omega} + \frac{a}{u}.$$

验算

- 动量：在 t_1 时，总动量为 $m\cdot 0 + mu = mu$，与开始时相同。能量：两个时刻的动能都是 $\frac{1}{2}mu^2$，而两次弦都处于自然长度。
- 量纲：u/ω 是长度，a/u 是时间，理应如此。
- 我们取 $m = a = \omega = u = 1$，以数值方法逐步计算运动方程。弦在 $t = 3.14159$ 时变得松弛，此时每个质点都移动了 $1.57080 = \pi/2$；两者的速率分别为 0 和 1，而碰撞发生在 $t = 4.14159 = \pi + 1$。

常见错误

- 把 P 当作系于固定点，这会得出角频率 $\omega/\sqrt{2}$ 和错误的时间。弦的两端都在移动。
- 在 t_1 之后仍沿用简谐运动。弦一旦短于 a 便会松弛，运动变为匀速。
- 完全忘记最后一段：碰撞发生在弦回复自然长度之后再过 a/u。

STEP 之外

把二体问题（two-body problem）拆分为质心的运动和间距的运动，而后者由约化质量 $\mu = m_1 m_2/(m_1 + m_2)$ 支配，是大学一年级力学中最有用的技巧之一。它把行星轨道、振动中的双原子分子以及碰撞的粒子都化为单体问题。在本题中，弹性弦就像一次“柔软”的碰撞，历时 π/ω，但结果与瞬间发生的弹性碰撞完全相同。Kleppner 与 Kolenkow 所著的 *An Introduction to Mechanics*（《力学导论》）对二体约化和约化质量有细致的推导。

详细解答

STEP III 2011，第 11 题

力学 • 刚体转动

STEP III 2011，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/11-s3-q11

关键思路

扭转之后，每条绳仍然由天花板上的固定点连到圆盘边缘上已转过 θ 的那一点。因此每条绳的水平部分是圆的一条弦（chord），所以几何关系要从上方俯视来看。由对称性，每条绳的情况都一样，所以铅直方向的平衡（equilibrium）决定张力（tension），而各张力的水平分量对轴的力矩（moment）之和就给出力偶（couple）。圆盘一经释放，余下的就交给能量守恒（energy conservation）。

解答

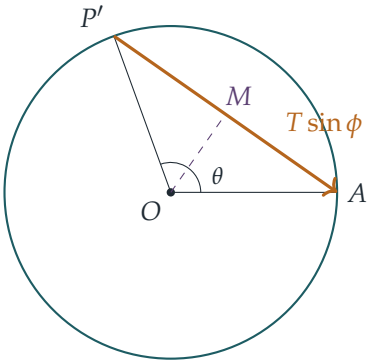
我们把圆盘模型化为刚性均匀薄片（lamina），把绳模型化为轻而不可伸长（inextensible）的绳。由对称性，所有绳都是拉紧的，张力同为 T，倾角同为 ϕ，而圆盘保持水平，圆心在 z 轴上。当圆盘转过 θ 时，它会上升至天花板下方深度 h 处。

绳的角度

经过 P 的绳系于天花板上的 $A = (a, 0, 0)$。转动后，绳的下端位于

$$P' = (a\cos\theta,\ a\sin\theta,\ -h).$$

从上方看，A 和 P' 都在以轴为圆心、半径为 a 的圆上，两者在圆心 O 所夹的角为 θ。所以绳的水平投影（projection）是一条长度为 $2a\sin\frac{1}{2}\theta$ 的弦。



从上方俯视的图。M 是弦 $P'A$ 的中点，所以 OM 平分角 θ，与弦垂直，且长度为 $OM = a\cos\frac{1}{2}\theta$。绳本身长 b，与铅直方向成角 ϕ，所以它的水平投影长度为 $b\sin\phi$。令两者相等，得

$$b\sin\phi = 2a\sin\tfrac{1}{2}\theta,$$

此即所求。绳的铅直投影则给出圆盘的新深度：

$$h = b\cos\phi = \sqrt{b^2 - 4a^2\sin^2\tfrac{1}{2}\theta}.$$

由于 $b > 2a$，这个值是实数。由于 $0 < \theta < \pi$，我们有 $0 < \sin\frac{1}{2}\theta < 1$，所以 $0 < \phi < \frac{1}{2}\pi$。

力偶

对圆盘作铅直方向的分解，n 个相等的张力承托着重量：

$$nT\cos\phi = mg.$$

力偶 C 作用于圆盘所在的平面内，所以对这个平衡没有影响。

每个张力都有一个铅直分量，以及一个沿弦由 P' 指向 A 的水平分量 $T\sin\phi$。铅直分量与轴平行，所以对轴没有力矩。水平分量沿弦作用，而弦与 O 的垂直距离为 $OM = a\cos\frac{1}{2}\theta$。每条绳的水平拉力都倾向把圆盘转回 $\theta = 0$，所以 n 个力矩的转向全部相同。对轴取力矩，

$$\begin{aligned} C &= n\,T\sin\phi\cdot a\cos\tfrac{1}{2}\theta = mg\tan\phi\; a\cos\tfrac{1}{2}\theta \\ &= mg\,a\cos\tfrac{1}{2}\theta\cdot\frac{2a\sin\frac{1}{2}\theta}{\sqrt{b^2-4a^2\sin^2\frac{1}{2}\theta}} = \frac{mga^2\sin\theta}{\sqrt{b^2-4a^2\sin^2\frac{1}{2}\theta}}, \end{aligned}$$

其中先用了 $nT = mg/\cos\phi$，再用 $\tan\phi = b\sin\phi/(b\cos\phi)$，最后用 $2\sin\frac{1}{2}\theta\cos\frac{1}{2}\theta = \sin\theta$。这就是所求的大小。

由静止释放

移去力偶后，圆盘会反向扭回并下降。绳是不可伸长的，且一直保持拉紧，所以不做功：每条绳的张力都与其系点相对于天花板固定点的速度垂直。重力是唯一做功的力，所以机械能守恒。

由上面的几何关系，当圆盘由绳铅直的位置转过 α 时，其深度为

$$z(\alpha) = \sqrt{b^2 - 4a^2 \sin^2 \tfrac{1}{2}\alpha} = \sqrt{b^2 - 2a^2(1 - \cos\alpha)}.$$

圆盘的动能（kinetic energy）为 $\frac{1}{2}m\dot{z}^2 + \frac{1}{2}I\omega^2$，其中对轴的转动惯量（moment of inertia）为 $I = \frac{1}{2}ma^2$。

想一想 当绳处于铅直时，圆盘的圆心是否仍有铅直方向的速度？

没有。求导得

$$\dot{z} = \frac{\mathrm{d}z}{\mathrm{d}\alpha}\,\dot{\alpha} = -\frac{a^2 \sin\alpha}{z}\,\dot{\alpha},$$

在 $\alpha = 0$ 时这等于零。那一刻绳是铅直的，圆盘处于最低点，所以此时全部动能都是转动动能。

圆盘由深度 $\sqrt{b^2 - 4a^2 \sin^2 \frac{1}{2}\theta}$ 下降至深度 b。能量守恒给出

$$\tfrac{1}{2} \cdot \tfrac{1}{2}ma^2\,\omega^2 = mg\left(b - \sqrt{b^2 - 4a^2 \sin^2 \tfrac{1}{2}\theta}\right),$$

即

$$\frac{a^2\omega^2}{4g} = b - \sqrt{b^2 - 4a^2 \sin^2 \tfrac{1}{2}\theta},$$

此即所求。

验算

- *虚功（virtual work）*。逆着力偶把圆盘再转 $\delta\theta$，圆盘会升高 $-\delta z$，所以 $C = mg\,|\mathrm{d}z/\mathrm{d}\theta| = mg\,a^2 \sin\theta / z$。这与我们取力矩所得的力偶完全一致。
- *小扭转*。当 θ 很小时，$C \approx mga^2\theta/b$，这正是大家熟悉的双线悬挂（bifilar suspension，这里是 n 线悬挂）的回复力偶。
- *数值验证*。我们在三维空间中明确放置 n 条绳，并对随机的 a, b, θ, m, n 把各张力的力矩加起来，结果与公式吻合。我们亦取 $a = 1$、$b = 3$、$\theta = 2$、$g = 9.81$，以数值方法积分释放后的运动，得出绳变成铅直时 $a^2\omega^2/(4g) = 0.51651$，与 $3 - \sqrt{9 - 4\sin^2 1} = 0.51651$ 相符。

常见错误

- 把水平张力的力臂（lever arm）当作 a 而不是 $a\cos\frac{1}{2}\theta$。这个力沿弦作用，而不是沿切线作用。
- 忘记圆盘在扭转时会上升。这个高度变化正是最后一部分中能量的来源。
- 把平移动能 $\frac{1}{2}m\dot{z}^2$ 计算在内，却没有留意到它在绳铅直时等于零。

STEP 之外

这是一个运动受约束、只有一个自由度（degree of freedom）的系统。深度是扭转角的固定函数（function），所以能量只取决于 α 和 $\dot{\alpha}$。拉格朗日力学（Lagrangian mechanics）把这个观察转化为运动方程。它也解释了为何力偶恰好是 $mg\,\mathrm{d}z/\mathrm{d}\alpha$，这就是虚功原理。在 $\alpha = 0$ 附近线性化，便得到简谐（simple harmonic）扭转振动（torsional oscillations），而双线悬挂正是借此用来量度转动惯量。Kleppner 和 Kolenkow 的 *An Introduction to Mechanics*（《力学导论》）以这个程度讲解刚体转动和能量方法。

详细解答

STEP III 2011，第 12 题

概率与统计 · 以概率生成函数处理复合分布

> **STEP III 2011，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/11-s3-q12

关键思路

关于一个分布的平均值（mean）和方差（variance），我们所需要的一切都藏在它的概率母函数（probability generating function, pgf）在 $t = 1$ 处的首两阶导数之中。因此，我们用连锁法则（chain rule）对复合函数 $G(H(t))$ 求导（两次），代入 $t = 1$，并利用 $H(1) = 1$。至于掷硬币的部分，N 服从几何分布，而每个 X_i 都是单次掷公平硬币的指示变量（indicator），所以两个概率母函数都很简单，而 $G(H(t))$ 是一个有理函数（rational function），可以展开成几何级数（geometric series）。

解答

一般结果

对于任何取非负整数值、概率母函数为 $P(t) = \mathrm{E}(t^Z)$ 的随机变量 Z，我们有 $P(1) = 1$，

$$P'(1) = \mathrm{E}(Z), \qquad P''(1) = \mathrm{E}\big(Z(Z-1)\big),$$

因此

$$\mathrm{Var}(Z) = P''(1) + P'(1) - \big(P'(1)\big)^2.$$

记 $\mu = \mathrm{E}(X_i) = H'(1)$、$\sigma^2 = \mathrm{Var}(X_i)$，并以 $K(t) = G(H(t))$ 表示 Y 的概率母函数。

由连锁法则，$K'(t) = G'(H(t))\,H'(t)$。由于 $H(1) = 1$，

$$\mathrm{E}(Y) = K'(1) = G'(1)H'(1) = \mathrm{E}(N)\,\mathrm{E}(X_i),$$

这就是第一个结果。

再求导一次（先用乘积法则（product rule），再用连锁法则），

$$K''(t) = G''(H(t))\big(H'(t)\big)^2 + G'(H(t))\,H''(t),$$

所以，再次利用 $H(1) = 1$，

$$K''(1) = G''(1)\mu^2 + \mathrm{E}(N)\,H''(1).$$

把上面的方差公式分别应用于 N 和 X_i，得

$$G''(1) = \mathrm{Var}(N) - \mathrm{E}(N) + \big(\mathrm{E}(N)\big)^2, \qquad H''(1) = \sigma^2 - \mu + \mu^2.$$

? **想一想** 为甚么不直接由 $K''(1)$ 计算 $\mathrm{Var}(Y)$，而要先改写 $G''(1)$ 和 $H''(1)$?

因为我们想要的答案是以方差表示，而不是以阶乘矩（factorial moments）表示；通过 $\mathrm{Var}(N)$ 和 σ^2 来表达 $G''(1)$ 和 $H''(1)$，可以让多余的项互相抵消。记 $m = \mathrm{E}(N)$，

$$\begin{aligned}\mathrm{Var}(Y) &= K''(1) + K'(1) - \big(K'(1)\big)^2 \\ &= \big(\mathrm{Var}(N) - m + m^2\big)\mu^2 + m\big(\sigma^2 - \mu + \mu^2\big) + m\mu - m^2\mu^2 \\ &= \mathrm{Var}(N)\mu^2 + m\sigma^2 + \big(-m\mu^2 + m\mu^2\big) + \big(m^2\mu^2 - m^2\mu^2\big) + \big(-m\mu + m\mu\big) \\ &= \mathrm{Var}(N)\big(\mathrm{E}(X_i)\big)^2 + \mathrm{E}(N)\,\mathrm{Var}(X_i),\end{aligned}$$

即为所求。

掷硬币

*N 的概率母函数。*第一次掷出正面发生在第 n 次，当且仅当首 $n-1$ 次都是反面而第 n 次是正面，所以对 $n = 1, 2, 3, \ldots$，$\mathrm{P}(N = n) = \left(\frac{1}{2}\right)^n$（这是几何分布（geometric distribution））。因此，当 $|t| < 2$ 时，

$$G(t) = \sum_{n=1}^{\infty} \left(\frac{t}{2}\right)^n = \frac{t/2}{1 - t/2} = \frac{t}{2 - t}.$$

*每个 X_i 的概率母函数。*若其后的第 i 次掷出正面，令 $X_i = 1$，否则令 $X_i = 0$。于是 $Y = X_1 + \cdots + X_N$，而 $H(t) = \frac{1}{2} + \frac{1}{2}t = \frac{1}{2}(1 + t)$。其后各次掷硬币彼此独立，亦与决定 N 的各次掷硬币独立，所以一般结果的条件成立。

Y 的概率母函数。

$$G(H(t)) = \frac{\frac{1}{2}(1 + t)}{2 - \frac{1}{2}(1 + t)} = \frac{1 + t}{3 - t}.$$

*平均值与方差。*对于 $p = \frac{1}{2}$ 的几何分布，$\mathrm{E}(N) = 1/p = 2$，$\mathrm{Var}(N) = (1 - p)/p^2 = 2$（或者对 G 求导：$G'(t) = 2/(2 - t)^2$、$G''(t) = 4/(2 - t)^3$，得 $G'(1) = 2$、$G''(1) = 4$，以及

$\mathrm{Var}(N) = 4 + 2 - 4 = 2$）。每个 X_i 都服从伯努利（Bernoulli）分布 $\left(\frac{1}{2}\right)$，所以 $\mathrm{E}(X_i) = \frac{1}{2}$，$\mathrm{Var}(X_i) = \frac{1}{4}$。因此

$$\mathrm{E}(Y) = 2 \cdot \tfrac{1}{2} = 1, \qquad \mathrm{Var}(Y) = 2 \cdot \tfrac{1}{4} + 2 \cdot \tfrac{1}{4} = 1.$$

Y 的分布。当 $|t| < 3$ 时，我们展开

$$\frac{1+t}{3-t} = \frac{1+t}{3} \cdot \frac{1}{1 - t/3} = \frac{1+t}{3} \sum_{k=0}^{\infty} \frac{t^k}{3^k} = \sum_{k=0}^{\infty} \frac{t^k}{3^{k+1}} + \sum_{k=0}^{\infty} \frac{t^{k+1}}{3^{k+1}}.$$

t^0 的系数只来自第一个和。当 $r \geq 1$ 时，t^r 的系数是 $\dfrac{1}{3^{r+1}} + \dfrac{1}{3^r} = \dfrac{4}{3^{r+1}}$。所以

$$\mathrm{P}(Y = 0) = \frac{1}{3}, \qquad \mathrm{P}(Y = r) = \frac{4}{3^{r+1}} \quad (r = 1, 2, 3, \ldots).$$

验算

- 各概率之和为 1：$\frac{1}{3} + 4\sum_{r\geq 1} 3^{-(r+1)} = \frac{1}{3} + 4 \cdot \frac{1}{6} = 1$。
- 直接计算 $\mathrm{P}(Y = 0)$：$\sum_{n\geq 1} \left(\frac{1}{2}\right)^n \left(\frac{1}{2}\right)^n = \sum_{n\geq 1} 4^{-n} = \frac{1}{3}$。同理，$\mathrm{P}(Y = 1) = \sum_{n\geq 1} n\, 4^{-n} = \frac{4}{9}$，与 $4/3^2$ 吻合。
- 对 $n \leq 200$ 精确求和 $\mathrm{P}(N = n)\binom{n}{r}2^{-n}$，可重现 $r \leq 7$ 时的 $\mathrm{P}(Y = r)$，而所得平均值和方差均为 1.000；以 200 000 次模拟，得到 0.996 和 0.994。

常见错误

- 把 N 当作从 0 开始（概率母函数为 $1/(2-t)$）：这里 N 计算的是掷出正面的那一次，所以 $N \geq 1$。
- 忘记 $P''(1)$ 是 $\mathrm{E}(Z^2) - \mathrm{E}(Z)$ 而不是 $\mathrm{E}(Z^2)$，以致在方差公式中漏掉 $+P'(1)$。
- 对所有 $r \geq 0$ 都给出 $4/3^{r+1}$：$r = 0$ 的一项是不同的。

STEP 之外

Y 是一个随机和（random sum）（即复合分布（compound distribution）），而上述方差公式是全方差定律（law of total variance）$\mathrm{Var}(Y) = \mathrm{E}\big(\mathrm{Var}(Y \mid N)\big) + \mathrm{Var}\big(\mathrm{E}(Y \mid N)\big)$ 的一个特例。概率母函数的复合也是分支过程（branching process）的核心：在分支过程中，$G(G(\cdots G(t)))$ 给出第 n 代的规模，而灭绝的概率是 G 的一个不动点（fixed point）。Grimmett 和 Welsh 的 *Probability: An Introduction*（《概率导论》）正是在这个程度上讲解母函数、随机和及分支过程。

详细解答

STEP III 2011，第 13 题

概率与统计 • 离散分布的众数

STEP III 2011，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/11-s3-q13

关键思路

要找出离散随机变量（discrete random variable）最可能的取值，我们不用微分，而是比较相邻的概率（probability）。比值 $\mathrm{P}(X = r+1)/\mathrm{P}(X = r)$ 是 r 的一个简单有理函数（rational function），而问它何时大于 1，就变成一个关于 r 的线性不等式（linear inequality）。比值大于 1 时概率上升，一旦跌到 1 以下概率便下降，所以峰值就在转折前的最后一个 r。同一方法无论放回抽样（with replacement）还是不放回抽样（without replacement）都适用，改变的只是那个比值。

解答

第 (i) 部分

在放回抽样下，各次抽取互相独立，每次抽到黑球的概率都是 $p = b/n$，所以 $X \sim \mathrm{B}(k, p)$：

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

先假设 $0 < b < n$，使得这些概率全部为正。对于 $0 \le r \le k-1$，

$$\frac{\mathrm{P}(X = r+1)}{\mathrm{P}(X = r)} = \frac{k-r}{r+1} \cdot \frac{p}{1-p} = \frac{(k-r)\,b}{(r+1)(n-b)}.$$

由于分母为正，这个比值大于 1 当且仅当

$$\begin{aligned} & (k-r)b > (r+1)(n-b) \\ \iff \quad & kb - rb > rn - rb + n - b \\ \iff \quad & (k+1)b > (r+1)n \\ \iff \quad & r+1 < \frac{(k+1)b}{n}. \end{aligned}$$

同理，比值等于 1 当且仅当 $r+1=m$，小于 1 当且仅当 $r+1>m$，其中

$$m = \frac{(k+1)b}{n}.$$

因此，随着 r 增大，每一步若终点 $r+1<m$，$\mathrm{P}(X=r)$ 便严格递增；若终点 $r+1>m$，便严格递减。注意由于 $0<b<n$，有 $0<m<k+1$，所以转折发生在取值范围 $0,\ldots,k$ 之内。

> **想一想** 如果 m 不是整数，最大的概率究竟在哪里？

若 m 不是整数，记 $M=\lfloor m\rfloor$，则 $M<m<M+1$ 且 $0\le M\le k$。概率严格递增直至 $\mathrm{P}(X=M)$（每一步的终点都 $\le M<m$），之后严格递减（之后每一步的终点都 $\ge M+1>m$）。因此黑球数目唯一的最可能取值是

$$\left\lfloor \frac{(k+1)b}{n} \right\rfloor,$$

正如所求。

不唯一的情况。若 m 是整数，由 $r=m-1$ 到 $r=m$ 这一步的比值恰好是 1，所以 $\mathrm{P}(X=m-1)=\mathrm{P}(X=m)$，而这两个相等的值比其余所有值都大（概率严格上升至 $m-1$，并在 m 之后严格下降）。由于 $1\le m\le k$，两个取值都是可能的。所以答案不唯一，当且仅当 $(k+1)b/n$ 是整数（且 $0<b<n$），此时 $m-1$ 和 m 同样是最可能的取值。例如 $n=3$、$b=1$、$k=5$ 给出 $m=2$，而确实 $\mathrm{P}(X=1)=\mathrm{P}(X=2)=\frac{80}{243}$。

平凡情况不在上述讨论之内：若 $b=0$，则必然 $X=0$，与公式相符；若 $b=n$，则必然 $X=k$，而公式（会给出 $k+1$）并不适用，因为上面的比值涉及除以 $n-b$。

第 (ii) 部分

在不放回抽样下，X 服从超几何分布（hypergeometric distribution）：把样本视为 k 个球组成的无序集合，所有 $\binom{n}{k}$ 个集合机会均等，

$$\mathrm{P}(X=r) = \frac{\binom{b}{r}\binom{n-b}{k-r}}{\binom{n}{k}}.$$

由于 $k \le b$，最大可能取值是 $r = k$；最小的是 $r_0 = \max(0, k-(n-b))$，因为非黑球最多只能抽出 $n-b$ 个。对于 $r_0 \le r \le k$，概率为正；而对于 $r_0 \le r \le k-1$，利用 $\binom{a}{r+1} = \binom{a}{r}\frac{a-r}{r+1}$ 和 $\binom{a}{s-1} = \binom{a}{s}\frac{s}{a-s+1}$，

$$\frac{\mathrm{P}(X=r+1)}{\mathrm{P}(X=r)} = \frac{b-r}{r+1}\cdot\frac{k-r}{n-b-k+r+1}.$$

在此范围内两个分母都为正，所以比值大于 1 当且仅当

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

展开后，左边是 $bk-r(b+k)+r^2$，右边是 $(r+1)(n-b-k+1)+r^2+r$。r^2 项互相抵消，不等式变为

$$\begin{aligned} & bk-r(b+k) > r(n-b-k+1)+(n-b-k+1)+r \\ \iff\quad & bk+b+k+1 > r(n+2)+(n+2) \\ \iff\quad & r+1 < \frac{(b+1)(k+1)}{n+2}. \end{aligned}$$

想一想 转折点是否真的落在可能取值的范围 $r_0,\dots,k$ 之内?

设 $m = \dfrac{(b+1)(k+1)}{n+2}$。由于 $b+1 < n+2$，有 $m < k+1$。至于下端，$m > 0$ 是显然的；而若 $r_0 = k+b-n > 0$，则需要 $(b+1)(k+1) > (n+2)(k+b-n)$。记 $x = n-b$ 及 $y = n-k$，左边减右边等于

$$(n+1-x)(n+1-y) - \left[(n+1)^2 - 1 - (n+2)(x+y)\right] = 1+x+y+xy > 0.$$

所以在所有情况下都有 $r_0 < m < k+1$，第 (i) 部分的论证可以原封不动地套用：当 $r+1 < m$ 时概率严格上升，一旦 $r+1 > m$ 便严格下降。因此，在唯一的情况下，黑球数目最可能的取值是

$$\left\lfloor \frac{(b+1)(k+1)}{n+2} \right\rfloor.$$

它不唯一，当且仅当 $(b+1)(k+1)/(n+2)$ 是整数；此时 $m-1$ 和 m（两者都是可能的取值，因为 $r_0 \le m-1$ 且 $m \le k$）机会相等，并且比其他任何取值都更可能。例如 $n=10$、$b=7$、$k=5$ 给出 $m=4$，而 $X=3$ 与 $X=4$ 打和，概率各为 $\frac{105}{252} = \frac{5}{12}$。

验算

- 我们用电脑以分数精确计算了所有 $n \le 24$、$0 \le b \le n$ 及 $1 \le k \le 19$（第 (ii) 部分另要求 $k \le b$）的两个分布，以穷举法找出众数，并在所有情况下确认了两条公式及两个打和条件（第 (ii) 部分共 2565 个情况）。

- 对于 $n = 10$、$b = 6$、$k = 4$：$m = \frac{35}{12}$，公式给出 2，而 $X = 2$ 确实是唯一的众数。
- 若 n 远大于 k，取走几个球几乎不改变黑球所占的比例，所以第 (ii) 部分应该与第 (i) 部分相若。事实正是如此：$(b+1)(k+1)/(n+2) \approx (k+1)b/n$。

常见错误

- 断定众数（mode）是 $\lfloor kb/n \rfloor$，即平均值（mean）的整数部分。众数接近平均值，但不一定等于其整数部分。
- 忘记当比值恰好等于 1 时，众数有两个，而不是一个。
- 在第 (ii) 部分展开 $(r+1)(n-b-k+r+1)$ 时弄错正负号；r^2 项互相抵消是一个有用的检查。

STEP 之外

比较相邻概率是找出任何离散分布众数的标准方法，而同一计算作为第 (i) 部分的极限情况，给出泊松分布（Poisson distribution）的众数 $\lfloor \lambda \rfloor$。超几何分布是抽样理论的核心，也是统计学中费雪（Fisher）精确检验（Fisher's exact test）的基础。概率像这样先升后降的分布称为单峰分布（unimodal）；对于二项分布和超几何分布，这源自一个更强的性质——对数凹性（log-concavity）。如欲深入了解，可参阅 Grimmett 和 Welsh 的 *Probability: An Introduction*（《概率导论》），或 Blitzstein 和 Hwang 的 *Introduction to Probability*（《概率导论》）。

Beyond STEP

STEP 是大多数中学生一生中应考过最艰深的数学考试，也是最出色的考试之一：题目简短、陌生而深刻，当中最精彩的，其实是乔装了的大学数学小品。

本书精选最有趣的 STEP 试题，由 1987 年首次开考之前的样本试卷（specimen papers），到 2025 年的试卷，并以每一道题目作为一扇门。每道题目都会仔细解答，沿途提出优秀学生会问的问题，然后带你穿过这扇门：这个想法从何而来，到了大学会演变成甚么，以及接下来该读些甚么。

本书为正在准备 STEP、有天分又有志投身数学的学生而写。他们想要的不止是模拟试卷：他们想开始像数学系本科生那样思考，并一探前方有甚么正等着自己。

- 共 13 章，涵盖纯数学（pure mathematics）、力学（mechanics）、概率（probability）与统计（statistics）
- 136 道 STEP 历届试题，每题均附引导式解答
- 每道题目背后的大学数学
- 附答案的延伸练习，以及延伸阅读

2011 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。

