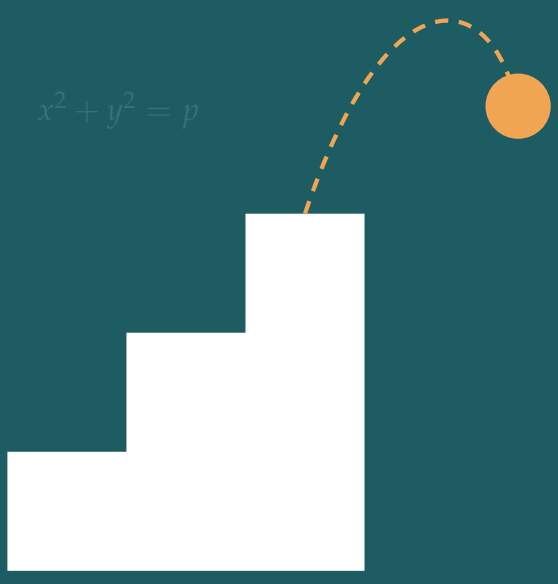
STEP 备试 · 大学数学

Beyond STEP

第 3 章

函数、不等式与严谨性



KEITH McNULTY

Beyond STEP

以历届 STEP 试题为门
走进大学数学

第 3 章

函数、不等式与严谨性

Keith McNulty

Beyond STEP：以历届 STEP 试题为门，走进大学数学

第 3 章：函数、不等式与严谨性

本册是 *Beyond STEP* 全书 13 章中的第 3 章。全书、各章的单行本、每一节的可列印工作纸，以及每一道 STEP 试题的详细解答，均可在 `beyondstep.keithmcnulty.org` 免费阅读及下载。

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。

章

3

函数、不等式与严谨

知道一件事是对的，不等于知道它为甚么对；数学就活在这两者的差别之中。

本章内容

3.1	找出错处	STEP I 1991
3.2	倒过来的 AM–GM	STEP II 1993
3.3	凸性的威力	STEP II 2007
3.4	以面积证明不等式	STEP III 2011
3.5	π^e 还是 e^π?	STEP II 2013
3.6	斜率的界	STEP I 1988
3.7	$\pi^2 \leq 10$	STEP II 2014
3.8	由加法公式决定的函数	STEP III 2019
3.9	创造一个新函数	STEP II 2021
A	练习答案	第 65 页

把事情做对

> 在数学里，你并不是理解了甚么，你只是习惯了它们。
>
> — 冯·诺伊曼（John von Neumann）

以下是 $1 = 2$ 的一个“证明”（proof）。设 $a = b$。那么 $a^2 = ab$，所以 $a^2 - b^2 = ab - b^2$，因式分解后得 $(a+b)(a-b) = b(a-b)$。约去 $a-b$ 得 $a+b = b$，而由于 $a = b$，这就是说 $2b = b$，所以 $2 = 1$。

你大概已经看出问题所在：由于 $a = b$，“约去 $a-b$”其实就是除以零。但请留意，每一步看起来都多么合理，而错误又多么容易藏在一个例行的步骤之中。真正的数学里，真正的错误很少这样明显，有史以来最伟大的一些数学家也曾被它们骗倒。本章要谈的，正是能把这些错误揪出来的思考习惯：精确的定义、小心运用定理，以及对看来太过容易的论证抱持适度的怀疑。

为甚么严谨很重要

在中学，数学主要是求出答案。到了大学，数学主要是知道自己的答案是对的，并且能够说服别人。这个转变是中学数学与大学数学之间最大的一个变化，也可能令人不安。一件“显然”成立的事，为甚么要花一整页去证明？

答案是：“显然”并不可靠。分析（analysis）的历史上，满是看似显然、实则错误的结果：连续（continuous）函数除了可能在少数几个孤立点之外必定可导（differentiable）（安培（André-Marie Ampère）在 1806 年发表过一个“证明”；到了 1872 年，魏尔斯特拉斯（Weierstrass）找到一个处处不可导的连续函数）；连续函数的收敛（convergent）和仍然连续（柯西（Cauchy）相信这一点；阿贝尔（Abel）找到了反例（counterexample））；收敛级数（series）的各项可以重新排列而不改变其和（一般而言，不可以）。严谨（rigour）并不是吹毛求疵，而是分辨真确与似是而非的唯一可靠方法。

不等式：分析的基本工具

本章的第二个主题是不等式（inequality）。在中学数学里，大部分结果都是等式：这个等于那个。在分析里，大部分结果都是不等式：这个至多是那个。要证明一个数列收敛，你要证明误差小于任何容许的限度；要证明一个积分存在，你要证明某个和有界；要估计一个近似值有多好，你要给误差定出一个界。专业的分析学家花在证明不等式上的时间，远多于证明等式，而你将在这里遇到的经典不等式（AM–GM、詹森（Jensen）、杨

（Young）、赫尔德（Hölder）、柯西—施瓦兹（Cauchy–Schwarz））正是这一行的看家工具。

这些不等式本身也十分精彩。一条关于函数的不等式，配上一个精心挑选的例子，就能证明 $\pi^2 \leq 10$（Section 3.7）。关于对数图像形状的一个事实，就能判定 π^e 与 e^π 哪个较大（Section 3.5）。一位化学家关于量度误差的提问，引出了数学中最重要的一些多项式（Section 3.6）。

你在大学会遇到甚么

严谨是大学一年级分析（Analysis）课程的主题，也是另一门常称为数与集合（Numbers and Sets）或数学基础（Foundations）的课程的主题，你会在那里学习怎样精确地书写证明。你会遇到连续性和可导性的正式定义，以及带有精确假设（hypothesis）的定理（介值定理（intermediate value theorem）、中值定理（mean value theorem）、极值定理（extreme value theorem）），而这些假设确实缺一不可。往后的课程会把本章的不等式发展成一门门完整的学科：泛函分析（functional analysis）研究函数空间，并以 Section 3.4 和 Section 3.7 中那类积分来量度其中的函数；凸分析（convex analysis）和最优化（optimisation）由 Section 3.3 发展而来；逼近论（approximation theory）由 Section 3.6 发展而来；而像 Section 3.9 中那样的特殊函数（special functions），则在应用数学和物理学中随处可见。

贯穿本章的问题

三个大问题

1. **怎样知道一个论证是正确的？** 看似合理的论证通常会在哪些地方出错？又应该怎样检查？
2. **怎样证明一条不等式？** 经典不等式从何而来？又怎样选择用哪一条？
3. **函数究竟是甚么？** 一个函数可否由它具有的某种性质来定义，而不是由公式来定义？当旧有的函数不够用时，我们可否创造新的函数？

本章以找错开始（Section 3.1），然后逐一探讨各条经典不等式，每条都从不同的角度切入：一种不寻常的归纳法（induction）、凸性（convexity）、面积、图像的形状、多项式插值（polynomial interpolation）和试验函数（test function）。最后两节讨论由性质定义的函数：由加法定律（addition law）得出的 e^x 和 $\tanh x$，以及一个以反函数（inverse）形式定义的全新函数。

本章的数学思想

§	课题	思想
3.1	似是而非的错误论证	检查计算、逻辑和隐藏的假设
3.2	AM–GM	前推—后推归纳法
3.3	詹森不等式	凹（concave）曲线位于其弦之上
3.4	杨氏不等式	在图中比较面积
3.5	π^e 还是 e^π?	化为单一函数并研究其图像
3.6	二次函数的斜率	以已知量表示未知量
3.7	$\pi^2 \leq 10$	把试验函数代入一般的不等式
3.8	e^x 与 $\tanh x$	由函数方程（functional equation）定义的函数
3.9	W 函数	以反函数的形式创造一个函数

怎样阅读本章

和之前一样，请先自己尝试每道题目，然后才阅读解答。在这一章，请再多做一件事：读完每个解答之后，回头找出粗心的学生会做错的一两个步骤。隐藏的假设会在哪里悄悄混进来？哪条定理的哪个假设需要验证？这样带著批判眼光、主动地阅读，是你在上大学之前所能养成的最有用的一个习惯。

3.1 找出错处

题目

STEP I 1991，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/91-s1-q6`

引子

大多数考试题目都要求你写出一个论证。这一题却给你两个论证，要你把它们逐一拆解。这是一种截然不同的技巧，也是你在大学里最常用到的技巧：检查自己的作业、阅读书中的证明（proof）并判断是否相信它、批改朋友的解答、审阅论文。

这里的论证错得很有启发性。每一步看起来都像你见过上百次的东西。有些步骤含有计算上的失误，但更严重的是逻辑上的错误：即使计算无误，结论也推不出来。这类错误会一直残留到专业数学之中，而本章要谈的正是如何把它们找出来。

动笔之前

想一想 当题目要求你“评论每一步”时，你在每一步应该检查甚么？动笔之前先列一张清单。

一张好的清单至少有三个问题。计算正确吗？这一步能从前面的步骤推出来吗？有没有隐藏的假设（例如函数是连续的、极限存在、驻点必定是极大点或极小点）？题目还问到结论会否无论如何都是对的。错误的论证也可以有正确的结论，所以这一点需要另行检查。

逐步解题

阶段 1 导数对吗？

想一想 用商法则（quotient rule）对 g 求导。$g'(x) = 0$ 在哪里成立？直接检查 $x = 1$ 和 $x = -1$。

由商法则，

$$g'(x) = \frac{6x^2(x^4+4) - (2x^3+3)\cdot 4x^3}{(x^4+4)^2} = \frac{-2x^6 - 12x^3 + 24x^2}{(x^4+4)^2} = \frac{-2x^2(x^4+6x-12)}{(x^4+4)^2}.$$

所以当 $x = 0$ 或 $x^4 + 6x - 12 = 0$ 时，$g'(x) = 0$。在 $x = 1$，$x^4 + 6x - 12 = -5$；在 $x = -1$，其值为 -17。两者都不是驻点（stationary point）。第一个断言是**错的**。

四次多项式 $x^4 + 6x - 12$ 在 $x = 0$ 的值是 -12，而且在两个方向上都趋向 $+\infty$，所以它有（至少）一个正根和一个负根；它恰好有两个实根，因为它的二阶导数 $12x^2$ 非负，所以它是凸（convex）的，最多只能穿过零两次。数值上，两个根是 $x \approx 1.386$ 和 $x \approx -2.247$。

阶段 2　函数值对吗？

$g(0) = \frac{3}{4}$ 是正确的，而 $g(1) = \frac{5}{5} = 1$。但 $g(-1) = \frac{-2+3}{1+4} = \frac{1}{5}$，不是 1。即使按它自己的思路，这个论证也有计算错误：它似乎假设了 g 是偶函数。

阶段 3　比较函数值能判断哪个是极大点吗？

想一想 假设一个函数恰好有两个驻点，驻点值分别是 $\frac{3}{4}$ 和 1。能否由此推出前者是极小点、后者是极大点？

不能。驻点的*性质*是局部（local）的性质：它取决于函数在该点附近的表现，而不是该点的函数值与别处的函数值孰大孰小。一个函数的局部极大值可以低于局部极小值（想想 $x + 2\sin x$），驻点也可以两者皆不是。在这里，在 $x = 0$ 附近，因式 $x^4 + 6x - 12$ 接近 -12，所以

$$g'(x) = \frac{-2x^2(x^4+6x-12)}{(x^4+4)^2} > 0 \qquad \text{对足够小的} x \neq 0.$$

导数在 0 的两侧都是正的：$x = 0$ 是一个*驻拐点*（stationary point of inflection），不是极小点。

阶段 4　驻点值能界定函数的范围吗？

停一停，试一试 即使驻点都求对了，为甚么仍不能断定 g 的每一个值都介乎最小与最大的驻点值之间？找出一个使 $g(x) < \frac{3}{4}$ 的 x 值。

驻点只描述*局部*的表现。全局（global）的表现还取决于定义域两端的情况，在这里即 $x \to \pm\infty$ 时的情况。由于当 $x \to \pm\infty$ 时 $g(x) \to 0$，而且每当 $2x^3 + 3 < 0$，即 $x <$

$-\sqrt[3]{3/2} \approx -1.14$ 时，$g(x) < 0$，所以这个函数会取负值。例如 $g(-2) = \frac{-13}{20}$。因此结论 $\frac{3}{4} \le g(x) \le 1$ 是**错的**（图 3.1.1）。

正确的说法是：g 递增至 $x \approx 1.386$ 处约为 1.083 的全局极大值，然后向 0 递减；在左边，它由 0 递减至 $x \approx -2.247$ 处约为 -0.668 的全局极小值，然后递增，经过 $x = 0$（在该处有一个水平拐点），一直升到极大值。所以对所有 x，$-0.668 \lesssim g(x) \lesssim 1.083$。

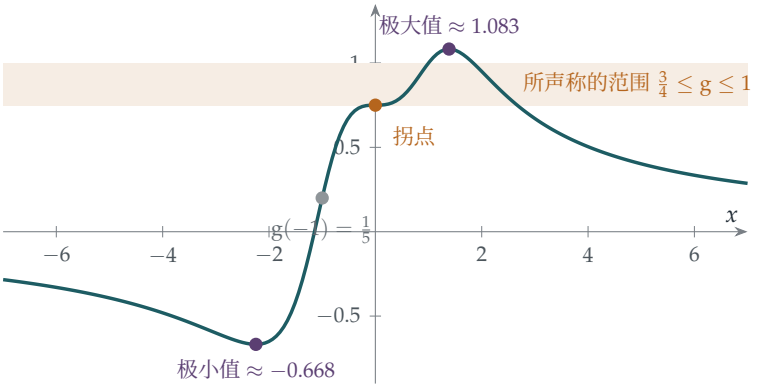


图 3.1.1 $g(x) = \frac{2x^3+3}{x^4+4}$ 的真实图像。该论证所声称的范围（琥珀色带）处处皆错：$x = -1$ 不是驻点，而且 $g(-1) = \frac{1}{5}$；$x = 0$ 是拐点，不是极小点；函数更远远超出色带之外，极大值高于 1，极小值低于 0。

阶段 5 积分

? **想一想** 对 $-3(1-x)^{-4}$ 求导。你会得到 $(1-x)^{-3}$ 吗？

$\frac{\mathrm{d}}{\mathrm{d}x}\left[-3(1-x)^{-4}\right] = -3 \cdot (-4)(1-x)^{-5} \cdot (-1) = -12(1-x)^{-5}$。幂和系数都错了。正确的原函数（antiderivative）是

$$\int (1-x)^{-3}\,\mathrm{d}x = \tfrac{1}{2}(1-x)^{-2} + C,$$

你可以求导来验证。

? **想一想** 用正确的原函数，$\left[\frac{1}{2}(1-x)^{-2}\right]_{-1}^{3} = \frac{1}{2}\cdot\frac{1}{4} - \frac{1}{2}\cdot\frac{1}{4} = 0$。那么答案原来还是对的。真的吗？

并不是。微积分基本定理（fundamental theorem of calculus）$\int_a^b F'(x)\,\mathrm{d}x = F(b) - F(a)$ 要求被积函数在整个区间 $[a,b]$ 上连续（continuous）。这里 $(1-x)^{-3}$ 在 $x = 1$ 处没有定义，而这一点位于 $[-1,3]$ 之内，而且在 $x = 1$ 附近函数值极其巨大。这个积分必须当作瑕积分（improper integral）来处理，在出问题的那一点拆开：

$$\int_{-1}^{1} (1-x)^{-3}\,\mathrm{d}x = \lim_{\varepsilon\to 0^+}\left[\tfrac{1}{2}(1-x)^{-2}\right]_{-1}^{1-\varepsilon} = \lim_{\varepsilon\to 0^+}\left(\frac{1}{2\varepsilon^2} - \frac{1}{8}\right) = +\infty.$$

由 -1 到 1 的积分是无穷大（而由 1 到 3 的积分是 $-\infty$），所以 $\int_{-1}^{3}(1-x)^{-3}\,\mathrm{d}x$ **不存在**。结论“$=0$”是错的，原因不在于真正的值是另一个数，而在于根本没有值。 ■

回顾

- **三类错误。**计算错误（导数、$g(-1)$、原函数）、逻辑错误（靠比较函数值来判别驻点的性质；把驻点值当作全局的界）以及隐藏的假设（微积分基本定理中的连续性）。
- **答案对，理由错。**在第 (ii) 部分改正原函数之后，仍然得出答案 0，但论证依然无效。答案看似正确，并不能证明推理站得住脚。
- **局部与全局。**导数告诉你的是函数在一点附近的表现。要得出全局的结论，还需要更多资料：定义域两端的情况，或者函数表现异常之处的情况。

背后的数学

大数学家也会出错

分析学的奠基者也犯过一模一样的错误。柯西（Augustin-Louis Cauchy）在 1821 年的 *Cours d'analyse* 中“证明”了连续函数的收敛（convergent）和必定连续。1826 年，阿贝尔（Niels Henrik Abel）指出了一个反例（counterexample）：傅里叶级数（Fourier series）

$$\sin x - \tfrac{1}{2}\sin 2x + \tfrac{1}{3}\sin 3x - \cdots$$

对每一个 x 都收敛，而且每一项都连续，但它的和在 $x = \pm\pi$ 处出现跳跃（当 $-\pi < x < \pi$ 时，它等于 $x/2$）。要弄清楚柯西的论证究竟错在哪里，花了数十年时间，并由此产生了一致收敛（uniform convergence）的概念，如今每个修读分析的学生都要学。隐藏的假设，尤其是关于交换两个极限次序的假设，是整个分析学中最常见的错误来源。

带有前提的定理

这一题中有两个错误，都源于忽视了定理的前提（hypotheses）。

- 极值定理（extreme value theorem）指出，在有界闭区间 $[a,b]$ 上的连续函数必定取得最大值和最小值。要找出它们，须检查端点、驻点，以及函数不可导的点：$|x|$ 在 $[-1,1]$ 上的最小值在 0 处取得，而这一点既非端点，亦非驻点。如果 f 在区间内部处处可导，则位于内部的最大点或最小点必定是驻点。在整条实数线上，没有端点可供检查，函数甚至可能根本没有最大值；你必须改为考察 $x \to \pm\infty$ 时的表现。
- 微积分基本定理要求被积函数在 $[a,b]$ 上连续。如果不连续，积分可能仍然存在（试试

$\int_0^1 x^{-1/2}\,\mathrm{d}x$），也可能不存在（如本题），你必须用极限来检查。

在大学里，你遇到的每一条定理都附有一张精确的前提清单。训练的一部分，就是每次使用定理之前，都要把这些前提检查一遍。

发散积分的一个值

第 (ii) 部分还有一个耐人寻味的注脚。区间 $[-1,3]$ 关于出问题的点 $x=1$ 对称，而 $(1-x)^{-3}$ 关于该点是“奇”的。如果你挖去一个对称的缺口 $(1-\varepsilon,1+\varepsilon)$，再令 $\varepsilon\to 0$，两块无穷大的部分便会互相抵消，你确实会得到 0。这称为*柯西主值*（Cauchy principal value），在物理学和复分析中确实很有用。但它是一个刻意选取的*不同定义*，一个误打误撞得出这个值的论证，甚么也没有证明。

思维技巧

以批判的眼光阅读论证

每一步都要检查三件事：计算对不对，这一步在逻辑上是否成立，以及有甚么是没有明说却已假设了的？

然后另行检查结论：不正确的论证可以有正确的结论，而“正确”的答案也可以出自无效的论证。对自己的作业也要同样严格审视；你最容易相信的论证，正是你自己的论证。

延伸练习

1 **会说谎的代换。**试做 STEP I 1987，第 8 题（`step.maths.org/questions/87-s1-q8`）。一个代换（substitution）似乎显示某个正的积分等于一个负的积分。准确找出论证在哪里出错，然后正确地计算两个积分。

2 **是对还是错？**试做 STEP I 1998，第 3 题（`step.maths.org/questions/98-s1-q3`）：有四个断言要你证明或否证。对每一个断言，动笔之前先判断你预期是哪一种。

3 **所有马都是同一种颜色。**找出以下用归纳法（induction）写成的“证明”的漏洞：任何一匹马都与自己颜色相同；给定 $n+1$ 匹马，由归纳假设，前 n 匹颜色相同，后 n 匹也是；两组有重叠，所以全部 $n+1$ 匹颜色相同。

4 **一个确实存在的积分。**证明 $\int_{-1}^{1}|x|^{-1/2}\,\mathrm{d}x$ 存在，尽管被积函数在 0 处没有定义，并求出它的值。

延伸阅读

- **E. A. Maxwell, *Fallacies in Mathematics***（Cambridge University Press）。一本经典的谬误论证集，收录各种似是而非的错误论证，按课题编排，为高中第六学级（sixth form）学生而写。
- **Bernard Gelbaum & John Olmsted, *Counterexamples in Analysis***（Dover）。数以百计的例子，说明定理的前提为何重要。开始修读分析课程之后，不妨随手翻阅。
- **Imre Lakatos, *Proofs and Refutations***（Cambridge University Press）。一篇精彩的对话，探讨数学家如何发现并修正证明中的错误。

3.2 倒过来的 AM–GM

题目

STEP II 1993，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/93-s2-q8`

引子

算术平均—几何平均不等式（arithmetic mean–geometric mean inequality，简称 AM–GM）也许是数学中最有用的一条不等式。它说的是：对于正数，普通的平均数总是不小于“乘法意义下的平均数”，而且只有当所有数都相等时，等号（equality）才成立。它不用微积分就能解决最优化（optimisation）问题，可以用来证明其他不等式，并且在分析学（analysis）的每一个分支中都会出现。

这道题用一种奇特而优美的归纳法（induction）来证明它，这个方法由柯西（Augustin-Louis Cauchy）于 1821 年提出。它不是由 n 走到 $n+1$，而是由 2 跃到 4，再到 8、16，只涵盖 2 的幂，然后掉头往下走，把中间的数补上。它提醒我们，“归纳法”这个想法远比课本上的版本灵活。

动笔之前

网上试题库所载的这份试卷有一处误植：它写的是“$b_1 = a_2$”。那当然应该是 $b_1 = a_1$；意思是 b 这一列数就是 a 那一列数，再用它们的平均数的若干个副本补足。

? **想一想** 为甚么由 2^m 个数走到 2^{m+1} 个数，可能比由 n 走到 $n+1$ 容易？

因为 2^{m+1} 个数可以自然地分成两组,每组 2^m 个;你可以对每一半应用归纳假设(inductive hypothesis)，然后用两个数的情形把两半合起来。由 n 走到 $n+1$，就没有这样对称的结构可供利用。

逐步解题

阶段 1　两个数

$\left(\frac{a_1+a_2}{2}\right)^2 - a_1a_2 = \frac{a_1^2+2a_1a_2+a_2^2-4a_1a_2}{4} = \frac{(a_1-a_2)^2}{4} \geq 0.$ ■

阶段 2　加倍

$m = 1$ 的情形就是第 1 阶段。假设 $(*)$ 对 2^m 个数成立，现取 2^{m+1} 个正数。

> **想一想**　把这 2^{m+1} 个数分成前半和后半，设 P 和 Q 分别为两半的算术平均。归纳假设对每一半说了甚么？

对每一半应用 $(*)$，

$$a_1 \cdots a_{2^m} \leq P^{2^m} \qquad \text{及} \qquad a_{2^m+1} \cdots a_{2^{m+1}} \leq Q^{2^m}.$$

两式相乘（所有数都是正数），全部 2^{m+1} 个数的积不大于 $(PQ)^{2^m}$。现在由两个数的情形得 $PQ \leq \left(\frac{P+Q}{2}\right)^2$，而 $\frac{P+Q}{2}$ 正好是全部 2^{m+1} 个数的平均数。所以

$$a_1 \cdots a_{2^{m+1}} \leq (PQ)^{2^m} \leq \left(\frac{P+Q}{2}\right)^{2^{m+1}} = \left(\frac{a_1 + \cdots + a_{2^{m+1}}}{2^{m+1}}\right)^{2^{m+1}},$$

这就是 $m+1$ 时的 $(*)$。由归纳法可知，只要项数是 2 的幂，AM–GM 就成立。 ■

阶段 3　填补空隙

现在取任意的 n，并选取 m 使得 $2^m > n$。我们在数列 $a_1, \ldots, a_n$ 之后补上 $2^m - n$ 个它自己的平均数 A。

> **想一想**　补足后的数列 $b_1, \ldots, b_{2^m}$ 的平均数是多少？

其总和为 $(a_1 + \cdots + a_n) + (2^m - n)A = nA + (2^m - n)A = 2^m A$，所以平均数就是 A 本身。加入平均数的副本并不会改变平均数。因此，对 b 这一列数应用 $(*)$，得

$$a_1 \cdots a_n \cdot A^{2^m - n} \leq A^{2^m}.$$

> **停一停，试一试**　完成这个证明。

两边同除以 A^{2^m-n}（这是正数）：$a_1 \cdots a_n \leq A^n$。取 n 次方根，

$$(a_1 \cdots a_n)^{1/n} \leq A = \frac{a_1 + \cdots + a_n}{n}.$$ ■

回顾

- **沿另一条路径作归纳。**第 2 阶段是对 m 作归纳，涵盖 $n = 2, 4, 8, \ldots$；第 3 阶段则是由任何一个 2 的幂往下走到任何较小的 n 的方法（图 3.2.1）。
- **用平均数补足。**把数列加长而不改变其平均数，这个技巧值得记住；它能把关于任意 n 的问题，化为关于一个方便的 n 的问题。
- **等号何时成立？**沿着证明往回追溯，等号成立需要每一次使用两个数的情形时等号都成立，而这只有在所有数都相等时才会发生。

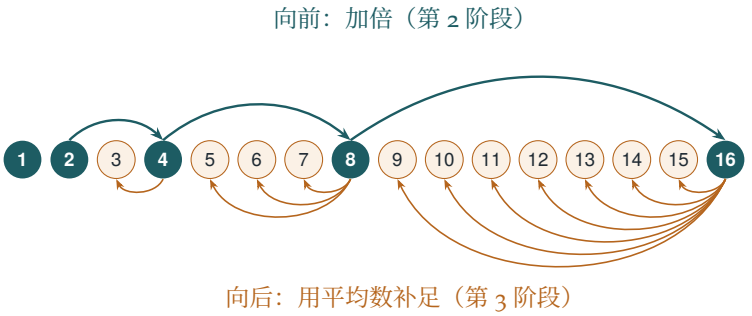


图 3.2.1 柯西的向前—向后归纳法。蓝绿色箭嘴以加倍的方式，证明 AM–GM 对 $2, 4, 8, 16, \ldots$ 个数成立；琥珀色箭嘴则把数列补足至下一个 2 的幂，从而填上其余每一个 n。

背后的数学

柯西的归纳法

柯西在他的《*Cours d'analyse*》(1821) 中发表了这个证明；我们在第 2 章谈过，正是这本书开创了分析学的严格处理。这种技巧现在称为*向前—向后归纳法* (forward–backward induction)：先对无穷多个情形组成的一个序列（这里是 2 的幂）证明命题成立，然后证明 n 的情形蕴含 $n-1$ 的情形。两者合起来，就能到达每一个 n。我们的第 3 阶段是由 2^m 直接跳到 n，但你也可以一步一步来：在 $n-1$ 个数之后补上一个它们的平均数，就能证明 n 的情形蕴含 $n-1$ 的情形。

一条不等式，多种证明

AM–GM 有数十种证明，把它们互相比较，是认识不同数学风格的好方法。

- **利用凸性（convexity）。**对数（logarithm）函数是凹（concave）函数，所以对数的平均不大于平均的对数。取指数之后，这就是 AM–GM；这也是 Section 3.3 的主题。
- **波利亚(George Pólya)的证明。**利用 $e^{x-1} \geq x$(e^{x-1} 在 $x=1$ 处的切线位于曲线下方)。设 A 为平均数，把这条不等式应用于每一个 a_i/A，然后相乘：$\prod \frac{a_i}{A} \leq e^{\sum(a_i/A-1)} = e^0 = 1$。波利亚说这个证明是他在梦中想到的。
- **利用磨光法（smoothing）。**如果其中两个数不相等，就把它们都换成两者的平均数。算术平均不变，而由两个数的情形可知，乘积会增大。重复这个步骤，所有数就会被推向平均数。

不用微积分的最优化

AM–GM 能把许多最优化问题变成一行就能解决的问题。在表面积固定的所有长方体盒子当中，哪一个体积最大？在周界固定的所有长方形当中，哪一个面积最大？（是正方形，由两个数的情形可知：$xy \leq \left(\frac{x+y}{2}\right)^2$，等号当 $x=y$ 时成立。）等号成立的条件告诉你最优值在哪里出现，而不等式则告诉你那确实是最优值，而不仅仅是一个驻点（stationary point）。可与 Section 3.1 中的告诫比较。

思维技巧

灵活的归纳法

归纳法不一定要按 $1, 2, 3, \ldots$ 的次序进行。只要一条路径能到达每一个情形，你就可以沿着它作归纳：可以加倍，可以用强归纳法（*strong induction*），也可以先在一个子数列（*subsequence*）上向前走，再向后走以填补空隙。

当由 n 到 $n+1$ 这显而易见的一步难以走通时，不妨问问哪一步会容易走，以及能否靠这一步到达每一个情形。

延伸练习

1. **后退一步。**补上一个平均数的副本，直接证明：若 AM–GM 对 n 个数成立，则它对 $n-1$ 个数也成立。
2. **一个递增数列。**把 AM–GM 应用于 $1+\frac{1}{n}, \ldots, 1+\frac{1}{n}, 1$ 这 $n+1$ 个数（其中 $1+\frac{1}{n}$ 有 n 个），证明 $\left(1+\frac{1}{n}\right)^n < \left(1+\frac{1}{n+1}\right)^{n+1}$。

3 **最好的盒子。**一个密封的长方体盒子的表面积为 S。把 AM–GM 用于三个不同的面的面积，证明其体积不大于 $(S/6)^{3/2}$，等号在盒子为正方体时成立。

4 **调和平均（harmonic mean）。**把 AM–GM 应用于 $1/a_i$ 这些数，证明调和平均 $\dfrac{n}{1/a_1 + \cdots + 1/a_n}$ 不大于几何平均。

延伸阅读

- **J. Michael Steele, *The Cauchy–Schwarz Master Class***（Cambridge University Press）。一本以问题带动、讲不等式的出色著作，正好由这个程度起步。第 2 章讲的就是 AM–GM。
- **Hardy, Littlewood & Pólya, *Inequalities***（Cambridge University Press）。经典参考书，由二十世纪三位伟大的分析学家撰写。
- **Arthur Engel, *Problem-Solving Strategies***（Springer）。第 7 章收录了大量来自数学奥林匹克的不等式题目。

3.3 凸性的威力

题目

STEP II 2007，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/07-s2-q7`

引子

在 Section 3.2 中，我们用一个巧妙的归纳法（induction）证明了 AM–GM 不等式。在这一节，它只需一行便水到渠成，因为它不过是某个更宏大的结果的特殊情形。这个更宏大的东西就是*凹性*（concavity）：向下弯的曲线所具有的形状，就像山顶一样。关于这类曲线，有一个简单的几何事实：曲线位于它的弦（chord）之上。单凭这一点，竟然可以证明范围广得惊人的各种不等式（inequality），从关于三角形的一个结果，到 AM–GM 不等式，以至概率（probability）、信息论（information theory）和经济学中的不等式。

到了大学，这个想法会发展成*凸性*（convexity），它是现代数学中最能贯通各个领域的重要主题之一。有些学科，例如凸优化（convex optimisation），整个就是建基于它之上。

动笔之前

想一想 画一条向下弯的曲线，例如 $(0, \pi)$ 上的 $y = \sin x$。在曲线上标出两点，并用直线把它们连起来。这条弦在曲线之上还是之下？

在曲线之下。对于凹曲线，每一条弦在两个端点之间都位于曲线之下。两个函数值的平均，就是弦的中点的高度；而函数在平均值处的取值，就是该中点正上方的曲线高度。因此，“弦在曲线之下”正正就是 $n = 2$ 时的琴生不等式（Jensen's inequality）：

$$\frac{f(x_1) + f(x_2)}{2} \leq f\left(\frac{x_1 + x_2}{2}\right).$$

若有 n 个点，则各点 $(x_k, f(x_k))$ 的平均就是它们的质心（centre of mass），它位于曲线下方的区域之内，所以同样在曲线之下（图 3.3.1）。

逐步解题

阶段 1　验证凹性

当 $0 < x < \pi$ 时，$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\sin x = -\sin x < 0$；当 $x > 0$ 时，$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\ln x = -\frac{1}{x^2} < 0$。 ■

阶段 2　三角形

想一想　应该用哪个函数、取哪些点？它们的平均是甚么？三角形的各个角是否都落在函数呈凹性的区间之内？

取 $\mathrm{f}(x) = \sin x$ 和三个点 A, B, C。每个角都严格介乎 0 与 π 之间，而 sin 在这个区间上是凹的，它们的平均是 $\frac{A+B+C}{3} = \frac{\pi}{3}$。由琴生不等式得

$$\frac{\sin A + \sin B + \sin C}{3} \le \sin\frac{\pi}{3} = \frac{\sqrt{3}}{2},$$

所以 $\sin A + \sin B + \sin C \le \frac{3\sqrt{3}}{2}$，而等号恰好在等边三角形（equilateral triangle）时成立。 ■

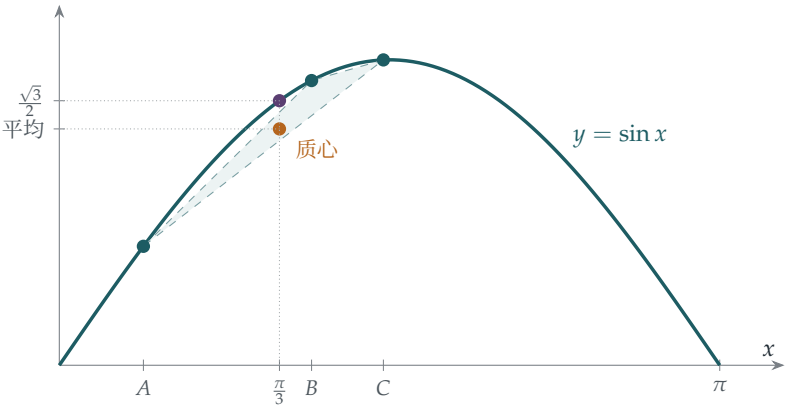


图 3.3.1　以三角形的三个角 $A = 0.4$、$B = 1.2$、$C = \pi - 1.6$ 说明 $\sin x$ 的琴生不等式。曲线上三点的质心（琥珀色）位于阴影三角形之内，因而在曲线之下。它的高度是 $\sin A$、$\sin B$、$\sin C$ 的平均，而它正上方的曲线高度是 $\sin\frac{\pi}{3}$。

阶段 3　一行证明 AM–GM

想一想　AM–GM 不等式比较的是一个积与一个和。哪个凹函数能把积变成和？

对数（logarithm）。把琴生不等式应用于 $\mathrm{f} = \ln$ 和正数 $t_1, \dots, t_n$：

$$\frac{\ln t_1 + \cdots + \ln t_n}{n} \le \ln\left(\frac{t_1 + \cdots + t_n}{n}\right).$$

左边是 $\ln\left(\sqrt[n]{t_1 \cdots t_n}\right)$。由于 ln 是递增函数，我们可以把两边的对数同时去掉，这就是 AM–GM 不等式。等号恰好在所有 t_k 都相等时成立。 ■

阶段 4 运用 AM–GM

> **想一想** 在 (a) 中，左边是四项之和，右边含有三个变量的积。你应该对多少项应用 AM–GM 不等式？那个 16 又为何有用？

对四个正数 x^4、y^4、z^4、16 应用 AM–GM 不等式：

$$\frac{x^4 + y^4 + z^4 + 16}{4} \geq \sqrt[4]{16\, x^4 y^4 z^4} = 2xyz.$$

两边乘以 4，得 $x^4 + y^4 + z^4 + 16 \geq 8xyz$。 ■

> **停一停，试一试** 现在求 $x^5 + y^5 + z^5 - 5xyz$ 的最小值。（提示：你想要的是 $x^5y^5z^5$ 的五次方根，所以需要五个数。另外两个数应该是甚么?）

对 $x^5, y^5, z^5, 1, 1$ 应用 AM–GM 不等式：

$$\frac{x^5 + y^5 + z^5 + 1 + 1}{5} \geq \sqrt[5]{x^5y^5z^5} = xyz,$$

所以 $x^5 + y^5 + z^5 - 5xyz \geq -2$。等号在 $x^5 = y^5 = z^5 = 1$ 时成立，即 $x = y = z = 1$，这时该式的值为 $3 - 5 = -2$。因此最小值是 -2。 ■

留意等号成立的条件在这里身兼两职：它说明这个界确实可以取到，所以 -2 是最小值，而不仅仅是一个下界（lower bound）。

回顾

- **一个想法，多个不等式。**关于凹曲线的一个几何事实，既给出了一个关于三角形的结果，又给出了 AM–GM 不等式，而所需的计算不过是验证一个二阶导数（second derivative）。
- **选择函数和点。**运用琴生不等式，就是要决定把它应用于哪个函数、取哪些点。对数能把积变成和，是常用之选。
- **补项以配合题目。**在 (b) 中，补上两个 1 令 AM–GM 不等式切合题目，正如在 Section 3.2 中补上平均值一样。

背后的数学

凸函数

在大学里，凸函数（convex function，即把凹函数上下倒转）的标准定义完全不涉及导数：若

$$f\big(tx_1 + (1-t)x_2\big) \le t\,f(x_1) + (1-t)\,f(x_2) \qquad \text{对所有} x_1, x_2 \text{及} 0 \le t \le 1,$$

则称 f 为凸函数，这正正是说弦位于曲线之上。这个定义对不可微（not differentiable）的函数（例如 $|x|$）以及多元函数同样适用。丹麦工程师兼数学家琴生（Johan Jensen）在 1906 年提出了以他命名的不等式，使凸函数成为一门独立的学问。

无处不在的凸性

凸性在数学及其应用中随处可见。

- **概率。**琴生不等式指出，对于凹函数 f 和任意随机变量（random variable）X，有 $\mathrm{E}[f(X)] \le f(\mathrm{E}[X])$。在经济学中，若以 f 为一个凹的“效用”（utility）函数，这就解释了为何人们宁愿稳拿 £100，也不愿参与一场各有一半机会得到 £0 或 £200 的赌局，亦解释了为何会有保险。
- **信息论。**均匀分布（uniform distribution）使熵（entropy）达到最大，这个结果是数据压缩的基础，而它的证明就是 ln 的琴生不等式。
- **优化。**凸函数除了全局最小值之外，再没有其他局部极小值（local minima），所以“一直往下坡走”总能找到最佳答案。这就是凸优化问题能够有效率地求解的原因，也是人们花那么多工夫把问题改写成凸形式的原因。

在 Section 3.4 中，我们会看到凸性的另一面：函数与其反函数（inverse）之间的一种几何对偶（duality）。

思维技巧

找出一般原理

当几个结果的证明相似时，要找出它们背后的同一个原理。然后把这个原理证明一次，再多次应用。

AM–GM 不等式、关于正弦的三角形不等式，以及许多其他结果，全都是“凹曲线位于它的弦之上”的实例。认出这个原理，就能把一堆零散的技巧变成一套方法。

延伸练习

1 **定义。**STEP II 2018，第 2 题（`step.maths.org/questions/18-s2-q2`）的第 (i) 及 (ii) 部分由上面方框中以弦刻划凹性的定义出发，逐步建立三个点的琴生不等式。

2 **正弦之积。**试做 STEP II 2018，第 2 题（`step.maths.org/questions/18-s2-q2`）的第 (iii) 部分。

3 **锐角三角形。**证明 $\tan x$ 在 $(0, \frac{\pi}{2})$ 上是凸的，并由此推出在锐角三角形（acute-angled triangle）中，$\tan A + \tan B + \tan C \geq 3\sqrt{3}$。

4 **幂平均。**证明 x^2 是凸的，并由此推出 $\left(\frac{a_1+\cdots+a_n}{n}\right)^2 \leq \frac{a_1^2+\cdots+a_n^2}{n}$。试把这个结果诠释为关于一组数的方差（variance）的命题。

延伸阅读

- **J. Michael Steele, *The Cauchy–Schwarz Master Class***（Cambridge University Press）。第 6 章讲述凸性和琴生不等式。
- **Claudi Alsina & Roger Nelsen, *When Less Is More: Visualizing Basic Inequalities***（Mathematical Association of America）。以图形证明不等式，当中不少用到凸性。
- **Stephen Boyd & Lieven Vandenberghe, *Convex Optimization***（Cambridge University Press）。一部现代经典，可在网上免费阅读；开首几章说明凸性在实际应用中为何重要。

3.4 以面积证明不等式

题目

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`

引子

这里有一个不等式（inequality），证明它的图简单得可以画在一枚邮票上。画一条通过原点的递增曲线。曲线下方的面积，加上曲线左方的面积，总能盖住一个矩形。单凭这一个观察，就得出杨氏不等式（Young's inequality）$ab \le \frac{a^p}{p} + \frac{b^q}{q}$，而它又是通往赫尔德不等式（Hölder's inequality）和柯西–施瓦兹不等式（Cauchy–Schwarz inequality）的钥匙，这两个不等式是分析学中最重要的工具之一。

这幅图还蕴藏着一个更深刻的想法。它以完全对称的方式把一个函数与它的反函数（inverse function）配成一对；把这种配对加以推广，就是勒让德变换（Legendre transform），它把古典力学的两大表述方式联系起来，也把热力学的两种主要描述联系起来。

动笔之前

想一想 $\int_0^a \mathrm{f}(x)\,\mathrm{d}x$ 是曲线与 x 轴之间的面积。在同一幅图上，$\int_0^b \mathrm{f}^{-1}(y)\,\mathrm{d}y$ 代表哪一块面积?

$x = \mathrm{f}^{-1}(y)$ 的图像与 $y = \mathrm{f}(x)$ 是同一条曲线，只是横着看而已。所以 $\int_0^b \mathrm{f}^{-1}(y)\,\mathrm{d}y$ 是曲线与 y 轴之间、由高度 0 到高度 b 的面积。明白了这一点，第 (i) 部分只需要画图。

逐步解题

阶段 1 图形

想一想　画出以 $(0,0)$ 和 (a,b) 为顶点的矩形。用一种颜色涂上曲线下方由 0 到 a 的区域，再用另一种颜色涂上曲线左方由 0 到 b 的区域。这两个区域能盖住整个矩形吗？甚么时候既没有重叠，也没有多出来的部分？

这两个区域合起来盖住整个矩形 $[0,a]\times[0,b]$，通常还盖住矩形以外的一些面积（图 3.4.1）。矩形内的每一点 (x,y)，不是在曲线下方（因为 $x\le a$，所以属于第一个区域），就是在曲线左方（因为 $y\le b$，所以属于第二个区域）。因此涂色部分的总面积至少等于矩形的面积 ab。这就是 $(*)$。

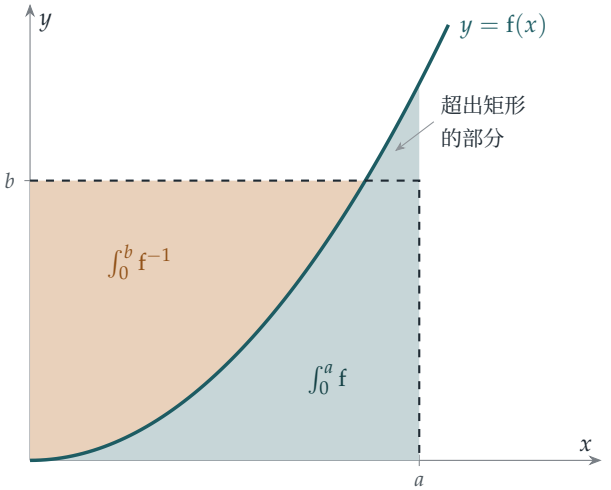


图 3.4.1　用一幅图看杨氏不等式。曲线下方由 0 到 a 的面积（蓝绿色）与曲线左方由 0 到 b 的面积（琥珀色）合起来盖住面积为 ab 的虚线矩形，而且还有剩余。恰好当顶点 (a,b) 位于曲线上时，才没有多出来的部分。

多出来的部分，就是曲线与矩形顶点之间那一小块曲边区域。它恰好在曲线通过顶点 (a,b) 时消失，也就是当

$$b=f(a).$$

阶段 2　杨氏不等式

想一想　取 $f(x)=x^{p-1}$，$f^{-1}(y)$ 是甚么？计算两个积分，并用 $\frac{1}{p}+\frac{1}{q}=1$ 化简第二个。

由于 $p>1$，$f(x)=x^{p-1}$ 连续而且递增，并且 $f(0)=0$，而 $f^{-1}(y)=y^{1/(p-1)}$。于是

$$\int_0^a x^{p-1}\,dx=\frac{a^p}{p},\qquad \int_0^b y^{1/(p-1)}\,dy=\frac{b^{1+1/(p-1)}}{1+1/(p-1)}.$$

由 $\frac{1}{p}+\frac{1}{q}=1$ 得 $q=\frac{p}{p-1}$，所以 $1+\frac{1}{p-1}=\frac{p}{p-1}=q$，第二个积分就是 $\frac{b^q}{q}$。因此 $(*)$ 变成

$$ab \le \frac{a^p}{p}+\frac{b^q}{q}. \qquad \blacksquare$$

等号应在 $b=\mathrm{f}(a)=a^{p-1}$ 时成立。验证：这时 $ab=a^p$，而 $b^q=a^{(p-1)q}=a^p$，所以右边是 $a^p\left(\frac{1}{p}+\frac{1}{q}\right)=a^p$。 $\blacksquare$

取 $p=q=2$，这就是 $ab\le\frac{a^2+b^2}{2}$，其实是 Section 3.2 中两个数的算术–几何平均不等式（AM–GM inequality）换了个样子。

阶段 3 反正弦

现在取 $[0,\frac{\pi}{2}]$ 上的 $\mathrm{f}(x)=\sin x$，它连续而且递增，并且 $\mathrm{f}(0)=0$，而在 $[0,1]$ 上 $\mathrm{f}^{-1}=\arcsin$。

> **想一想** $\int_0^a \sin x\,\mathrm{d}x$ 很容易。你会怎样求 $\arcsin y$ 的积分？

$\int_0^a \sin x\,\mathrm{d}x = 1-\cos a$。至于反正弦，把它看成 $\arcsin y\times 1$，用分部积分法（integration by parts）：

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

两式相加，由 $(*)$ 得 $ab\le b\arcsin b+\sqrt{1-b^2}-\cos a$。 $\blacksquare$

> **停一停，试一试** 现在推导关于 $\arcsin(t^{-1})$ 的不等式。应该选哪一个 a 的值？（试试最简单的那一个。）

取 $a=0$：这时 $0\le b\arcsin b+\sqrt{1-b^2}-1$，即 $b\arcsin b\ge 1-\sqrt{1-b^2}$。（这不过是说 $\int_0^b \arcsin y\,\mathrm{d}y\ge 0$。）设 $b=1/t$，其中 $t\ge 1$，于是 $0<b\le 1$，再除以 b：

$$\arcsin\left(\frac{1}{t}\right)\ge t\left(1-\sqrt{1-\frac{1}{t^2}}\right)=t-\sqrt{t^2-1}. \qquad \blacksquare$$

回顾

- **以图为证。**只要看出合适的图，第 (i) 部分就是一个完整的证明。把几何事实转化为积分之间的不等式，是一种威力强大的技巧。
- **把一般结果特殊化。**同一个一般的不等式，选取不同的 f，就得出杨氏不等式和一个关于 arcsin 的不等式。
- **最简单的选择可能正是对的选择。**最后的推导用了 $a=0$，这时不等式几乎是显然的。简单的情况一定要试试。

背后的数学

从杨氏到赫尔德

杨（William Henry Young）在 1912 年发表他的不等式。它之所以重要，在于它所引出的结果。假设 $\sum a_i^p = 1$ 及 $\sum b_i^q = 1$。对每一对数应用杨氏不等式，然后相加，

$$\sum a_i b_i \le \frac{\sum a_i^p}{p} + \frac{\sum b_i^q}{q} = \frac{1}{p} + \frac{1}{q} = 1.$$

把各个 a 和各个 b 按比例缩放，使两个和都等于 1，就得到赫尔德不等式：对正数，

$$\sum a_i b_i \le \Big(\sum a_i^p\Big)^{1/p} \Big(\sum b_i^q\Big)^{1/q}.$$

取 $p = q = 2$，这就是柯西—施瓦兹不等式 $\sum a_i b_i \le \sqrt{\sum a_i^2}\sqrt{\sum b_i^2}$，它说的是两个向量夹角的余弦最多为 1。赫尔德不等式是 L^p 空间理论的基础；在这些函数空间里，“距离”是用 $\left(\int |f|^p\right)^{1/p}$ 来量度的，你会在泛函分析（functional analysis）和概率的课程中遇到它们。

勒让德变换

图 3.4.1 对 f 和 f^{-1} 的处理完全对称。记 $F(a) = \int_0^a \mathrm{f}$ 及 $G(b) = \int_0^b \mathrm{f}^{-1}$。不等式说的是 $F(a) + G(b) \ge ab$，在曲线上等号成立，所以

$$G(b) = \max_a \big(ab - F(a)\big).$$

这个把凸函数（convex function）F 变成另一个凸函数 G 的运算，就是勒让德变换，而连续做两次就会回到 F。在力学中，它把拉格朗日量（Lagrangian，位置与速度的函数）转换成哈密顿量（Hamiltonian，位置与动量的函数）：这是牛顿（Newton）定律的两种等价表述，你会在大学二年级的课程中学到。在热力学中，它在能量、焓（enthalpy）和自由能（free energy）之间互相转换。对 $F(a) = a^p/p$，变换的结果是 $G(b) = b^q/q$，这就是杨氏不等式中两个指数由 $\frac{1}{p} + \frac{1}{q} = 1$ 联系起来的原因。

思维技巧

从图形得出不等式

积分就是面积。在一幅精心挑选的图中比较面积，可以证明一个从代数上难以看出的不等式。

留意这样的图：其中一个区域明显包含另一个区域，或者几个区域明显盖住第三个区域。然后把每块面积转化为积分。

延伸练习

1 **1987 年的版本。**试做 STEP I 1987, 第 6 题（`step.maths.org/questions/87-s1-q6`）。它用同一类图来求反函数的积分。用求导来验算你的答案。

2 **柯西–施瓦兹。**用 $p = q = 2$ 的杨氏不等式，证明对正数有 $a_1b_1 + a_2b_2 + a_3b_3 \le \sqrt{a_1^2 + a_2^2 + a_3^2}\,\sqrt{b_1^2 + b_2^2 + b_3^2}$。（提示：先处理两个平方根都等于 1 的情况。）

3 **指数版本。**在 $(*)$ 中取 $\mathrm{f}(x) = e^x - 1$，证明对 $a, b \ge 0$，

$$ab \le e^a - 1 - a + (1+b)\ln(1+b) - b.$$

等号何时成立？

4 **这个界有多好？** 较难 用计算机比较 $t = 1, 2, 10$ 时的 $\arcsin(1/t)$ 与 $t - \sqrt{t^2-1}$。证明当 t 很大时，$t - \sqrt{t^2-1} \approx \frac{1}{2t}$，而 $\arcsin(1/t) \approx \frac{1}{t}$。为甚么这个界相差了一个因子 2？（看看 $a = 0$ 时的图：多出来的部分有多少被浪费了？）

延伸阅读

- **J. Michael Steele, *The Cauchy–Schwarz Master Class***（Cambridge University Press）。第 1 章和第 9 章讲述柯西–施瓦兹不等式和赫尔德不等式，途中也谈到杨氏不等式。
- **Roger Nelsen, *Proofs Without Words***（Mathematical Association of America）。收录了杨氏不等式和许多其他的图形证明。
- **V. I. Arnold, *Mathematical Methods of Classical Mechanics***（Springer）。第 14 节正是用这幅图来解释勒让德变换。内容较深，但相关的一节很短。

3.5 π^e 还是 e^π?

题目

STEP II 2013，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/13-s2-q1`

引子

π^e 和 e^π，哪一个较大？计算机一秒钟就能告诉你：$e^\pi \approx 23.14$ 而 $\pi^e \approx 22.46$。但计算机只告诉你结果是这样，却没有告诉你为甚么，更无法一次过解答无穷多条同类的问题。3^π 比 π^3 大吗？除了显而易见的 $n = m$ 之外，还有没有整数满足 $n^m = m^n$？

原来这些问题全都关乎同一条曲线 $y = \ln x$，以及与它相交的直线。本题说明，只要运用得当，一幅草图就能解决一些单靠代数看似无从入手的大小比较，也说明 e 这个数怎样自然而然地成为分界点。

动笔之前

想一想 取对数（logarithm）后，$a^b = b^a$ 等同于 $b \ln a = a \ln b$。试把这条式重新整理，使 a 只出现在一边，b 只出现在另一边。

$b \ln a = a \ln b$ 等同于

$$\frac{\ln a}{a} = \frac{\ln b}{b}.$$

因此，每次比较 a^b 与 b^a，其实都是比较同一个函数 $h(x) = \frac{\ln x}{x}$ 在 a 和 b 两处的值。而 $\frac{\ln a}{a}$ 正是由原点（origin）连到曲线 $y = \ln x$ 上的点 $(a, \ln a)$ 的直线的斜率（gradient）。这就是本题与直线有关的原因。

逐步解题

阶段 1　通过原点的切线

想一想　若 $y = mx$ 与 $y = \ln x$ 在 $x = x_0$ 处相切，哪两条方程必须成立？

直线与曲线必须相交，即 $mx_0 = \ln x_0$，而且斜率相同，即 $m = 1/x_0$。把第二式代入第一式，得 $1 = \ln x_0$，所以 $x_0 = e$ 及 $m = 1/e$。由原点引出的切线（tangent）与曲线相切于 $(e, 1)$。

若 $y = mx$ 改为与曲线相交于 $x = a$ 和 $x = b$，则 $\ln a = ma$ 及 $\ln b = mb$，所以 $\frac{\ln a}{a} = m = \frac{\ln b}{b}$，由此得 $b \ln a = a \ln b$，因而 $a^b = b^a$。 ■

至于草图：一条通过原点、比切线稍为平坦的直线会与曲线相交两次，切点两侧各一次，所以 $a < e < b$。斜率大于 $1/e$ 的直线则完全碰不到曲线，因为 $\ln x$ 是凹（concave）函数，整条曲线都在其切线 $y = x/e$ 之下。

阶段 2　与 y 轴相交于原点上方的直线

想一想　现在直线是 $y = mx + c$，其中 $c > 0$。试以 m 和 c 表示 $\frac{\ln p}{p}$ 和 $\frac{\ln q}{q}$。

由 $\ln p = mp + c$ 及 $\ln q = mq + c$，得

$$\frac{\ln p}{p} = m + \frac{c}{p}, \qquad \frac{\ln q}{q} = m + \frac{c}{q}.$$

由于 $c > 0$ 且 $p < q$，可知 $\frac{c}{p} > \frac{c}{q}$，所以 $\frac{\ln p}{p} > \frac{\ln q}{q}$。两边乘以 $pq > 0$，得 $q \ln p > p \ln q$，因此 $p^q > q^p$。 ■

从几何角度看：若通过 $(p, \ln p)$ 和 $(q, \ln q)$ 两点的直线与 y 轴相交于原点上方，那么从原点望去，前一点的仰角比后一点的更陡。

阶段 3　e 右方的弦

想一想　设 $e \le p < q$。由 $(p, \ln p)$ 连到 $(q, \ln q)$ 的弦（chord），其斜率小于曲线在 p 处的斜率。为甚么？又为甚么这样会令它的 y 截距（intercept）为正数？

因为 $\ln x$ 是凹函数，它的斜率 $1/x$ 递减，所以由 p 到 q 的弦比 p 处的切线平坦：弦的斜率 s 满足 $s < \frac{1}{p}$。弦的 y 截距是

$$c = \ln p - sp > \ln p - \frac{1}{p} \cdot p = \ln p - 1 \ge 0,$$

当中用到 $p \ge e$。所以这条弦与 y 轴相交于原点上方。 ■

停一停，试一试 利用这个结果和第 2 阶段，判断 π^e 与 e^π 哪一个较大。

取 $p = e$ 及 $q = \pi$。连接这两点的弦的截距为正数，所以根据第 2 阶段，$p^q > q^p$，即

$$e^\pi > \pi^e.$$

阶段 4 与切线平行的弦

现在假设由 p 到 q 的弦的斜率刚好是 $\frac{1}{e}$，与切线 $y = \frac{x}{e}$ 的斜率相同。

想一想 这条弦与 $(e, 1)$ 处的切线平行。由于 $\ln x$ 是凹函数，曲线位于该切线之下。这条弦与 y 轴相交于原点的上方还是下方?

这条弦是直线 $y = \frac{x}{e} + c$，其中 c 为某个常数。它的两个端点都在曲线上，而曲线位于切线 $y = \frac{x}{e}$ 上或其下方，并且只在 $x = e$ 处与切线相接。两个端点不可能同时位于 $x = e$，所以最少有一个端点严格位于切线下方，于是整条平行的弦都在切线之下：$c < 0$。现在以 $c < 0$ 重做第 2 阶段：不等式的方向反转，$\frac{\ln p}{p} < \frac{\ln q}{q}$，因此 $q^p > p^q$。 ■

回顾

- **化为单一函数。**比较 a^b 与 b^a，变成了比较 $h(a)$ 与 $h(b)$，其中 $h(x) = \frac{\ln x}{x}$。
- **有理有据的草图。**每一处“藉草图说明”的背后，都有确切的事实支持：ln 的凹性（弦在曲线之下，切线在曲线之上），以及截距的正负。
- **e 的特殊角色。**由原点引出的切线在 $x = e$ 处与曲线相切，而这一点正是以上所有比较的分界点。

背后的数学

$\frac{\ln x}{x}$ 的图像

本题的一切，都可以在 $h(x) = \frac{\ln x}{x}$ 的图像中看到（图 3.5.1）。它的导数（derivative）是 $h'(x) = \frac{1-\ln x}{x^2}$，所以 h 在 $(0, e)$ 上递增，在 $x = e$ 处达到最大值（maximum）$\frac{1}{e}$，然后递减并趋向 0。一条高度为 m 的水平线，若高度介乎 0 与 $\frac{1}{e}$ 之间，便会与图像相交两次，交点为 $a < e < b$，且 $h(a) = h(b)$，这正正就是 $a^b = b^a$。在 e 的右方，h 递减，所以当 $e \le p < q$ 时，我们有 $h(p) > h(q)$，即 $p^q > q^p$。特别地，$e^\pi > \pi^e$；更一般而言，对每个 $x \neq e$ 都有 $e^x > x^e$：在这场较量中，没有其他数能胜过 e。

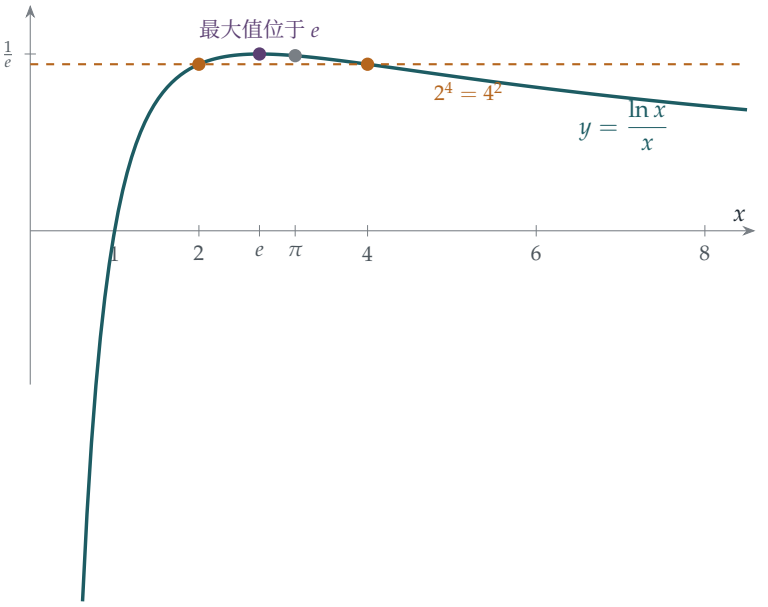


图 3.5.1　$h(x) = \frac{\ln x}{x}$ 的图像。它上升至 $x = e$ 处的最大值 $\frac{1}{e}$，然后下降。因为 $h(e) > h(\pi)$，所以 $e^\pi > \pi^e$。高度为 $\frac{\ln 2}{2}$ 的水平线（琥珀色）与图像相交于 $x = 2$ 和 $x = 4$：这是唯一一对满足 $n^m = m^n$ 的相异整数。

$x^y = y^x$ 的所有解

哪些相异的正实数对满足 $x^y = y^x$？有一个巧妙的方法可以把它们全部找出来。设 $y = tx$，其中 $t > 1$；那么 $x^{tx} = (tx)^x$ 给出 $x^t = tx$，所以 $x^{t-1} = t$，并且

$$x = t^{1/(t-1)}, \qquad y = t^{t/(t-1)}.$$

写成 $t = 1 + \frac{1}{n}$，便得 $x = \left(1 + \frac{1}{n}\right)^n$ 及 $y = \left(1 + \frac{1}{n}\right)^{n+1}$。取 $n = 1$ 时，这就是 $(2, 4)$；取 $n = 2$ 时则是 $\left(\frac{9}{4}, \frac{27}{8}\right)$，而的确有 $\left(\frac{9}{4}\right)^{27/8} = \left(\frac{27}{8}\right)^{9/4}$。当 $n \to \infty$ 时，两者都趋向 e，也就是两个分支会合之处。1728 年，丹尼尔·伯努利（Daniel Bernoulli）与哥德巴赫（Christian Goldbach）曾在书信中讨论这个问题，其后欧拉（Euler）亦有研究。

最经济的底

函数 $x^{1/x} = e^{h(x)}$ 同样在 $x = e$ 处取得最大值。这解答了一个有趣的问题：假如一个记数系统的“成本”是它的底（base）乘以所需的数位数目，那么最经济的底就是 e；若只考虑整数，则是 3（仅仅胜过 2）。有少数实验性质的电脑采用过以 3 为底的记数法，例如 1958 年苏联的 Setun。

思维技巧

重新整理以分离变量，然后研究单一函数

一条不等式或方程，若当中两个量纠缠在一起，往往可以化为关于同一个函数在两点取值的命题。

一旦找到这个函数，它的图像（在哪里递增、在哪里递减、在哪里达到顶点）就能一次过回答所有问题。在这里，a^b 与 b^a 的比较变成了 $h(a)$ 与 $h(b)$ 的比较。

延伸练习

1 **整数。**试做 STEP I 1988，第 1 题（`step.maths.org/questions/88-s1-q1`）。

2 **更多比较。**试做 STEP II 2025，第 3 题（载于 OCR 提供下载的 2025 年试卷，`www.ocr.org.uk/Images/750486-2025-past-papers.zip`）。该题利用本节的图像，解决更多 a^b 与 b^a 之间的比较，其中一个要靠 e 的数值界限才能完成。

3 **一行证明。**利用当 $x \neq 0$ 时 $e^x > 1 + x$，并取 $x = \frac{\pi}{e} - 1$，直接证明 $e^\pi > \pi^e$。

延伸阅读

- **Eli Maor, *e: The Story of a Number***（Princeton University Press）。讲述 e 的历史，包括方程 $x^y = y^x$。
- **Paul Nahin, *An Imaginary Tale***（Princeton University Press）。不只谈 $\sqrt{-1}$：书中满载正是这一类的有趣计算。
- **Michael Spivak, *Calculus***（Cambridge University Press）。第 11 章以完全严谨的方式，讲解如何运用导数了解图像的形状。

3.6 斜率的界

题目

STEP I 1988，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/88-s1-q7`

引子

假如你知道一个函数在某个区间上的值总是介乎 -1 与 1 之间，你能否说出它有多陡？对一般的函数来说，答案是不能：$\sin(1000x)$ 的值介乎 -1 与 1 之间，但它的斜率可以大至 1000。不过二次函数（quadratic）不可能这样摆动。这道题目要证明的是：一个困在方框 $|x| \leq 1$、$|y| \leq 1$ 之内的二次函数，在方框内的斜率永远不会大于 4，而且 4 已是最佳的界。

这个问题的来历出人意表。1887 年，以元素周期表闻名的化学家门捷列夫（Dmitri Mendeleev）正在研究溶液的密度如何随浓度变化，并用二次函数拟合他的数据。他需要知道数据中的误差对曲线斜率的影响最坏可以有多大，于是证明了与本题完全相同的结果。两年后，年轻的马可夫（Andrei Markov）解决了任意次数多项式（polynomial）的一般问题，而答案直接引出了数学中最重要的一些多项式。

动笔之前

想一想 $f'(x)$ 的公式只用到 f 的三个值：在 -1、0 和 1 的值。为甚么三个值就足以确定一个二次函数的导数（derivative）？

二次函数有三个系数，所以只要知道它在任意三点的值，就能把它确定下来。它的导数同样是由这些系数决定的函数，因此必定可以用这三个值表示出来。题中的公式正好告诉你怎样表示。而这恰恰是你所需要的，因为假设 $|f(x)| \leq 1$ 告诉你的是 f 的值，而结论谈的却是它的导数。

逐步解题

阶段 1　公式

$\mathrm{f}(1)=a+b+c$，$\mathrm{f}(-1)=a-b+c$，而 $\mathrm{f}(0)=c$。所以右方是

$$(a+b+c)\left(x+\tfrac{1}{2}\right)+(a-b+c)\left(x-\tfrac{1}{2}\right)-2cx.$$

x 的系数是 $(a+b+c)+(a-b+c)-2c=2a$，常数项是 $\tfrac{1}{2}(a+b+c)-\tfrac{1}{2}(a-b+c)=b$。所以右方等于 $2ax+b=\mathrm{f}'(x)$。 ■

阶段 2　上界

> **想一想**　对公式取绝对值，然后运用三角不等式（triangle inequality），再配合 $|\mathrm{f}(1)|,|\mathrm{f}(-1)|,|\mathrm{f}(0)|\le 1$。你会得到 $|\mathrm{f}'(x)|$ 的甚么上界（写成 x 的函数）？

根据三角不等式 $|u+v+w|\le|u|+|v|+|w|$，

$$|\mathrm{f}'(x)|\le|\mathrm{f}(1)|\,|x+\tfrac{1}{2}|+|\mathrm{f}(-1)|\,|x-\tfrac{1}{2}|+2|\mathrm{f}(0)|\,|x|\le|x+\tfrac{1}{2}|+|x-\tfrac{1}{2}|+2|x|.$$

> **停一停，试一试**　证明当 $|x|\le 1$ 时，$|x+\tfrac{1}{2}|+|x-\tfrac{1}{2}|+2|x|\le 4$。（按正负号分情况讨论。）

若 $|x|\le\tfrac{1}{2}$，则 $x+\tfrac{1}{2}\ge 0$ 而 $x-\tfrac{1}{2}\le 0$，所以 $|x+\tfrac{1}{2}|+|x-\tfrac{1}{2}|=(x+\tfrac{1}{2})-(x-\tfrac{1}{2})=1$，总和是 $1+2|x|\le 2$。

若 $\tfrac{1}{2}\le|x|\le 1$，两个括号的正负号都与 x 相同，所以 $|x+\tfrac{1}{2}|+|x-\tfrac{1}{2}|=2|x|$，总和是 $4|x|\le 4$。

无论哪种情况，都有 $|\mathrm{f}'(x)|\le 4$。 ■

这个上界只会在 $x=\pm1$ 处达到；在区间中间的一半，斜率最多只是 2。

阶段 3　一个极端例子

> **想一想**　要在 $x=1$ 处达到上界，第 2 阶段中的每个不等式都必须取等号。$\mathrm{f}(1)$、$\mathrm{f}(-1)$ 和 $\mathrm{f}(0)$ 必须取甚么值？正负号又是甚么？

在 $x=1$ 处，公式是 $\mathrm{f}'(1)=\tfrac{3}{2}\mathrm{f}(1)+\tfrac{1}{2}\mathrm{f}(-1)-2\mathrm{f}(0)$。要令它等于 4，我们需要 $\mathrm{f}(1)=1$、$\mathrm{f}(-1)=1$ 和 $\mathrm{f}(0)=-1$。通过 $(-1,1)$、$(0,-1)$ 和 $(1,1)$ 的二次函数是

$$\mathrm{f}(x)=2x^2-1,$$

所以 $a = 2$，$b = 0$，$c = -1$。在 $|x| \le 1$ 上，$x^2 \in [0,1]$，所以 $2x^2 - 1 \in [-1,1]$，而 $f'(1) = 4$。 ■

回顾

- **用已知的量表示想求的量。**公式把导数改写成 f 的值，而假设提供的资料正正在这些值上。
- **三角不等式。**用各项绝对值之和来估计一个和的上界，是分析学（analysis）中最常见的手法之一。
- **从等号成立的条件找出极值情况。**追问每个不等式何时取等号，就直接得出例子 $2x^2 - 1$。

背后的数学

马可夫不等式与切比雪夫多项式

马可夫在 1889 年证明：若 p 是 n 次多项式，且在 $[-1,1]$ 上 $|p(x)| \le 1$，则

$$|p'(x)| \le n^2 \qquad \text{当} |x| \le 1.$$

当 $n = 2$ 时，这就是我们的上界 4。极端情况是*切比雪夫多项式*（Chebyshev polynomials）T_n，其定义为

$$T_n(\cos\theta) = \cos n\theta.$$

例如 $\cos 2\theta = 2\cos^2\theta - 1$ 给出 $T_2(x) = 2x^2 - 1$，即我们的极端二次函数；而 $\cos 3\theta = 4\cos^3\theta - 3\cos\theta$ 给出 $T_3(x) = 4x^3 - 3x$。由于 $T_n(x) = \cos n\theta$，这些多项式在 -1 与 1 之间来回振荡的次数，已是 n 次多项式所能做到的最多，而且它们在两端也是最陡的（图 3.6.1）。马可夫的弟弟 Vladimir 在 1892 年把结果推广到高阶导数。

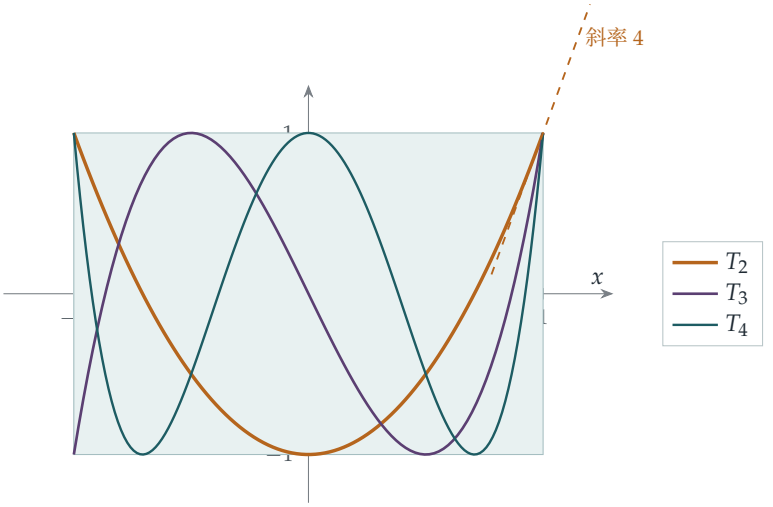


图 3.6.1　切比雪夫多项式 $T_2 = 2x^2 - 1$、$T_3 = 4x^3 - 3x$ 和 $T_4 = 8x^4 - 8x^2 + 1$ 都留在方框 $|x| \le 1$、$|y| \le 1$ 之内，并且尽可能多次触及方框的顶边和底边。在同次数的多项式中，它们在两端最陡：$T_n'(1) = n^2$。虚线是 T_2 在 $x = 1$ 处的切线，斜率为 4。

插值

题目中的公式是*拉格朗日插值*（Lagrange interpolation）的一个例子：已知函数在若干点的值，写出通过这些点而次数最低的唯一多项式，再用它来逼近该函数或其导数。这里的三个点是 $-1, 0, 1$。这类公式在数值分析（numerical analysis）中无处不在，例如用来从量度所得的数据估计导数。马可夫不等式就是随之而来的警告：函数值的小误差，在导数中最多可以被放大 n^2 倍，这正是门捷列夫当年担心的问题。

切比雪夫多项式的实际应用

切比雪夫多项式是逼近理论（approximation theory）的主力工具。在 T_n 的根处对函数插值，而不是在等距的点上插值，可以把误差控制住；若用等距的点，高次插值在区间两端附近可以错得非常离谱（龙格现象，Runge's phenomenon）。那些把函数计算至机器精度极限、或以极高准确度求解微分方程（differential equation）的软件，内部往往使用切比雪夫展开式。

思维技巧

用已知的量表示未知的量

如果假设告诉你的是某一类量，而结论谈的是另一类量，就去找一个把两者连系起来的恒等式，然后逐项估计上界。

得到一个界之后，还要问它是否最佳。沿着等号成立的情况倒推回去，通常会把你带到极端例子，而极端例子本身往往就很值得研究。

延伸练习

1 **一般的想法。**证明若 f 是二次函数，则 $\mathrm{f}(x) = \mathrm{f}(-1)\dfrac{x(x-1)}{2} + \mathrm{f}(0)(1-x^2) + \mathrm{f}(1)\dfrac{x(x+1)}{2}$，然后求导，重新得出题目中的公式。

2 **唯一性。**较难 证明在 $[-1,1]$ 上满足 $|\mathrm{f}(x)| \le 1$、且在 $[-1,1]$ 内某处有 $|\mathrm{f}'(x)| = 4$ 的二次函数，只有 $\pm(2x^2-1)$。

3 **切比雪夫多项式。**利用 $\cos 3\theta = 4\cos^3\theta - 3\cos\theta$ 验证 $T_3(x) = 4x^3 - 3x$，并验证 $T_3'(1) = 9$。证明 $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$，并求 T_4。

4 **T_n 的斜率。**把 $T_n(\cos\theta) = \cos n\theta$ 对 θ 求导，证明 $T_n'(\cos\theta) = \dfrac{n\sin n\theta}{\sin\theta}$，再令 $\theta \to 0$，推导出 $T_n'(1) = n^2$。

延伸阅读

- **Lloyd N. Trefethen,** ***Approximation Theory and Approximation Practice*** （SIAM）。一本出色的现代著作，讲述切比雪夫多项式与插值，能力强的中学生也读得懂，全书穿插电脑实验。
- **Peter Borwein & Tamás Erdélyi,** ***Polynomials and Polynomial Inequalities*** （Springer）。想了解马可夫不等式及其相关结果的来龙去脉，可读此书。程度较深。
- **George M. Phillips,** ***Interpolation and Approximation by Polynomials*** （Springer）。一本大学本科教科书，涵盖拉格朗日插值、切比雪夫多项式和马可夫型不等式。

3.7 $\pi^2 \leq 10$

题目

STEP II 2014，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/14-s2-q2`

引子

每个中学生都知道 π 比 3 大一点。但你会怎样证明它的一个上界（upper bound）呢？阿基米德（Archimedes）用的是多边形。这道题的做法却完全出人意表：取一条关于函数的一般不等式，代入一条简单的抛物线（parabola），便直接读出 $\pi^2 \leq 10$。换另一个函数，更可得出著名的近似值 $\pi \leq \frac{22}{7}$。

不等式 $(*)$ 称为 *Wirtinger* 不等式（Wirtinger's inequality）。它说的是：一个两端都固定为零的函数，不可能只“大”而不“陡”。它的背后是数学物理中最核心的思想之一：弦的振动、量子粒子的能级，以及特征值（eigenvalue）理论。

动笔之前

想一想 一个在 $[0, \pi]$ 两端都为零的函数，如果在中间取值很大，为甚么就必须很陡？条件 $f(0) = f(\pi) = 0$ 为甚么重要？

要在长度为 π 的距离之内，由 0 升到某个高度 H 再降回 0，函数上升的平均速率至少要有 $H/(\pi/2)$。因此它的导数（derivative）不可能处处都小。若没有两端的条件，就没有这种限制：常数函数可以任意大，而导数为零。不等式 $(*)$ 把这个想法精确化，以平方的积分（integral）来量度“大小”和“陡峭程度”。

逐步解题

阶段 1　正弦波

想一想　$\int_0^{\pi}\sin^2 nx\,\mathrm{d}x$ 等于甚么？$\int_0^{\pi}\cos^2 nx\,\mathrm{d}x$ 呢？（利用 $\sin^2\theta=\frac{1}{2}(1-\cos 2\theta)$。）

两者都等于 $\frac{\pi}{2}$，因为 $\int_0^{\pi}\cos 2nx\,\mathrm{d}x=0$。取 $\mathrm{f}(x)=\sin nx$，则 $\mathrm{f}'(x)=n\cos nx$，所以

$$\int_0^{\pi}\mathrm{f}^2=\frac{\pi}{2}\le\frac{n^2\pi}{2}=\int_0^{\pi}(\mathrm{f}')^2,$$

因为 $n\ge 1$。　■留意 $n=1$ 时等号成立，所以这条不等式已经紧得不能再紧。

阶段 2　端点条件不可或缺

想一想　找一个简单的函数，满足 $\mathrm{f}(0)\neq 0$，而 $(*)$ 对它不成立。然后再找一个满足 $\mathrm{f}(0)=0$ 但 $\mathrm{f}(\pi)\neq 0$ 的。

常数函数 $\mathrm{f}(x)=1$ 有 $\mathrm{f}(0)=1\neq 0$，而且 $\int_0^{\pi}1\,\mathrm{d}x=\pi>0=\int_0^{\pi}0\,\mathrm{d}x$。

至于第二个，$\mathrm{f}(x)=x$ 有 $\mathrm{f}(0)=0$ 但 $\mathrm{f}(\pi)=\pi$，而且

$$\int_0^{\pi}x^2\,\mathrm{d}x=\frac{\pi^3}{3}\approx 10.3>\pi=\int_0^{\pi}1^2\,\mathrm{d}x.$$

所以 $(*)$ 在两种情况下都不成立。　■

阶段 3　抛物线

想一想　哪个二次函数在 $x=0$ 和 $x=\pi$ 处都等于零？（若不计常数倍，就只有一个。为甚么这个倍数无关紧要？）

这个二次函数必定是 $x(\pi-x)$ 的倍数，而把 f 乘以常数 k，$(*)$ 的两边都会乘以 k^2，所以我们可以取 $\mathrm{f}(x)=x(\pi-x)$，即 $a=-1$、$b=\pi$、$c=0$。于是 $\mathrm{f}'(x)=\pi-2x$，再用代换（substitution）$u=x-\frac{\pi}{2}$（这样两个被积函数都变成偶函数），

$$\int_0^{\pi}x^2(\pi-x)^2\,\mathrm{d}x=\int_{-\pi/2}^{\pi/2}\left(\frac{\pi^2}{4}-u^2\right)^2\mathrm{d}u=\frac{\pi^5}{30},$$

$$\int_0^{\pi}(\pi-2x)^2\,\mathrm{d}x=\int_{-\pi/2}^{\pi/2}4u^2\,\mathrm{d}u=\frac{\pi^3}{3}.$$

（第一个积分：$\int_{-h}^{h}(h^2-u^2)^2\,\mathrm{d}u=2(h^5-\frac{2}{3}h^5+\frac{1}{5}h^5)=\frac{16}{15}h^5$，其中 $h=\frac{\pi}{2}$。）所以 $(*)$ 给出 $\frac{\pi^5}{30}\le\frac{\pi^3}{3}$，即

$$\pi^2\le 10.\qquad\blacksquare$$

阶段 4 三角函数的选择

想一想 对于 $\mathrm{f}(x) = p\sin\frac{1}{2}x + q\cos\frac{1}{2}x + r$，$\mathrm{f}(0) = 0$ 和 $\mathrm{f}(\pi) = 0$ 对 p、q、r 施加了甚么条件?

$\mathrm{f}(0) = q + r = 0$ 及 $\mathrm{f}(\pi) = p + r = 0$。所以 $p = q = -r$，经缩放后，我们取 $\mathrm{f}(x) = \sin\frac{x}{2} + \cos\frac{x}{2} - 1$。

停一停，试一试 对这个函数计算 $\int_0^\pi \mathrm{f}^2$ 和 $\int_0^\pi (\mathrm{f}')^2$。由此得出关于 π 的甚么不等式?

记 $s = \sin\frac{x}{2}$ 及 $c = \cos\frac{x}{2}$。则 $(s+c)^2 = 1 + 2sc = 1 + \sin x$，所以

$$\mathrm{f}^2 = (s+c)^2 - 2(s+c) + 1 = 2 + \sin x - 2\sin\tfrac{x}{2} - 2\cos\tfrac{x}{2}.$$

利用 $\int_0^\pi \sin x = 2$、$\int_0^\pi \sin\frac{x}{2} = 2$ 及 $\int_0^\pi \cos\frac{x}{2} = 2$，

$$\int_0^\pi \mathrm{f}^2\,\mathrm{d}x = 2\pi + 2 - 4 - 4 = 2\pi - 6.$$

另外 $\mathrm{f}' = \frac{1}{2}(c - s)$，所以 $(\mathrm{f}')^2 = \frac{1}{4}(1 - \sin x)$，而 $\int_0^\pi (\mathrm{f}')^2\,\mathrm{d}x = \frac{1}{4}(\pi - 2)$。于是 $(*)$ 给出

$$2\pi - 6 \le \frac{\pi - 2}{4} \quad\Longrightarrow\quad 8\pi - 24 \le \pi - 2 \quad\Longrightarrow\quad \pi \le \frac{22}{7}. \qquad \blacksquare$$

阶段 5 哪个较好?

由 $\pi \le \frac{22}{7}$ 可得 $\pi^2 \le \frac{484}{49} \approx 9.878$，比 $\pi^2 \le 10$ 更好。（真实值是 $\pi^2 \approx 9.870$。）留意这里证明的是 $\frac{22}{7}$ 为 π 的一个上界；而 Section 4.3 中一个截然不同的积分，则证明它并不等于 π。

回顾

- **一般的不等式，特定的函数。**选定一个测试函数（test function），就把一个关于所有函数的命题，变成一个关于某个数的命题。
- **缩放。**$(*)$ 的两边按同一个因子缩放，所以在每一族函数中，我们都可以把那个自由常数固定下来。
- **测试函数愈好，界愈好。**测试函数愈接近极端情况 $\sin x$，不等式就愈接近等式，估计也就愈准确（图 3.7.1）。

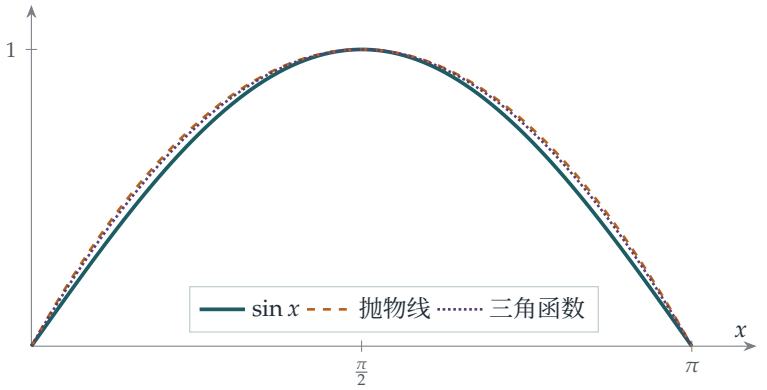


图 3.7.1　本节的三个函数，各自缩放至最大值为 1：极端情况 $\sin x$（实线），它令 $(*)$ 的等号成立；抛物线 $x(\pi - x)$（虚线），它给出 $\pi^2 \le 10$；以及 $\sin\frac{x}{2} + \cos\frac{x}{2} - 1$（点线），它给出 $\pi \le \frac{22}{7}$。两个测试函数都非常接近 $\sin x$，所以得出的界才这样好。

背后的数学

用傅里叶级数的证明

为甚么 $(*)$ 成立？到了大学你会学到，$[0, \pi]$ 上任何在两端都为零的合理函数，都可以写成傅里叶正弦级数（Fourier sine series）：

$$\mathrm{f}(x) = b_1 \sin x + b_2 \sin 2x + b_3 \sin 3x + \cdots .$$

由于当 $m \neq n$ 时 $\int_0^\pi \sin mx \sin nx \, \mathrm{d}x = 0$，平方后再积分，所有交叉项都会消失，得出

$$\int_0^\pi \mathrm{f}^2 = \frac{\pi}{2}\left(b_1^2 + b_2^2 + b_3^2 + \cdots\right), \qquad \int_0^\pi (\mathrm{f}')^2 = \frac{\pi}{2}\left(1^2 b_1^2 + 2^2 b_2^2 + 3^2 b_3^2 + \cdots\right).$$

右式的每一项都不小于左式的对应项，这就证明了 $(*)$，同时说明等号只在 f 是 $\sin x$ 的倍数时成立。题目的第 (i) 部分就是只有一项的情况。

振动的弦与量子粒子

比值 $\int (\mathrm{f}')^2 / \int \mathrm{f}^2$ 称为瑞利商（Rayleigh quotient），而 $(*)$ 说的是：对于固定在 0 和 π 的函数，它的最小可能值是 1。这个数有物理意义。对一条两端固定的弦来说，$\int (\mathrm{f}')^2$ 量度拉伸的能量，$\int \mathrm{f}^2$ 量度位移的大小，而这个比值的最小值决定了弦所能奏出的最低音。在量子力学（quantum mechanics）中，同一个最小值就是困在盒子里的粒子的最低能量：它不可能是零，因为两端固定的波必须弯曲。瑞利勋爵（Lord Rayleigh）正是用这一招

——猜一个测试函数，再计算比值——来估计振动系统的频率；由于极端情况给出最小值，任何猜测都会给出真正最低频率的一个上界，而好的猜测更会给出非常准确的上界。

变分法

找出令某个积分之比取得最小值的函数，是*变分法*（calculus of variations）的问题，当中的未知量是整个函数，而不是一个数。这里取得最小值的函数 $\sin x$ 满足微分方程（differential equation）$f'' + f = 0$，而这并非巧合：求积分的最小值，总会导出一条由取得最小值的函数所满足的微分方程。我们会在 Section 4.7 再谈这个想法。

思维技巧

测试函数

一条对某类函数中每一个都成立的不等式，可以套用在一个巧妙选取的函数上，从而得到关于某个特定数的信息。

个中艺术在于选择：接近极端情况的测试函数，会给出近乎最佳的结果。这个想法贯穿分析学、物理学和工程学，由估计特征值到有限元法（finite element method）皆然。

延伸练习

1 **四次函数。**在 $(*)$ 中取 $f(x) = x^2(\pi - x)^2$。证明它只能给出 $\pi^2 \le 12$。为甚么这个测试函数比抛物线差？（把它的形状与 $\sin x$ 比较。）

2 **两个正弦波。**对 $f(x) = \sin x + \sin 2x$ 直接验证 $(*)$，把两边都计算出来。

3 **另一个区间。**用代换证明：若 $g(0) = g(L) = 0$，则 $\displaystyle\int_0^L g^2 \le \frac{L^2}{\pi^2}\int_0^L (g')^2$。这对不同长度的弦说明了甚么？

4 **瑞利法。**较难 取 $f(x) = x(\pi - x) + \lambda x^2(\pi - x)^2$，并选取 λ 使 π^2 的界尽可能好。（代数运算相当繁重，用电脑代数系统会有帮助。）

延伸阅读

- **T. W. Körner, *Fourier Analysis***（Cambridge University Press）。一本精彩而旁征博引的书，讲述傅里叶级数及其应用，作者是剑桥的分析学家。

- **Hardy, Littlewood & Pólya, *Inequalities***（Cambridge University Press）。第 7 章讨论 Wirtinger 不等式及相关的不等式。
- **Richard Feynman, *The Feynman Lectures on Physics***，第 II 卷第 19 章（“The principle of least action”，最小作用量原理）。费曼（Feynman）关于求积分最小值的一讲，通篇贯穿物理直觉。可于网上免费阅读。

3.8 由加法公式决定的函数

题目

STEP III 2019，第 2 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2019 年 STEP 试卷（zip 文件）内的 STEP III 试卷：

`www.ocr.org.uk/Images/696590-2019-past-papers.zip`

引子

你怎样知道 e^x 是甚么？在学校里，它是通过复利息引入的，或者说成是导数（derivative）等于自身的函数，又或者写成 $\lim(1+x/n)^n$。这道题给出另一个答案：e^{kx} 是唯一能把加法变成乘法的可微（differentiable）函数，即 $\mathrm{f}(x+y)=\mathrm{f}(x)\mathrm{f}(y)$。单凭这条规则，再加上一个数 $\mathrm{f}'(0)$，就足以把这个函数完全确定下来。

第二部分对一条较陌生的规则做同样的事：$\mathrm{g}(x+y)=\frac{\mathrm{g}(x)+\mathrm{g}(y)}{1+\mathrm{g}(x)\mathrm{g}(y)}$。如果你觉得这条公式似曾相识，那就对了：它正是狭义相对论（special relativity）中爱因斯坦（Einstein）的速度相加法则，只是把光速设为 1。这条规则迫使函数必然是 tanh，而这正好解释了为何任何物体都不能被加速至超越光速。

动笔之前

像 $\mathrm{f}(x+y)=\mathrm{f}(x)\mathrm{f}(y)$ 这样以函数为未知量的方程，称为*函数方程*（functional equation）。

? **想一想** 这条方程对所有 x 和 y 都成立。最值得首先尝试的是甚么？

特殊值。设 $y=0$（或 $x=y$，或 $y=-x$），往往不费甚么工夫就能得到大量资料。两个部分都是这样起步的。

逐步解题

阶段 1 在零处的值

代入 $x = y = 0$：$\mathrm{f}(0) = \mathrm{f}(0)^2$，所以 $\mathrm{f}(0) = 0$ 或 1。

> **想一想**　如果 $\mathrm{f}(0) = 0$，会出甚么问题？

若 $\mathrm{f}(0) = 0$，则代入 $y = 0$ 可得对每个 x 都有 $\mathrm{f}(x) = \mathrm{f}(x)\mathrm{f}(0) = 0$。于是 f 恒等于零，且 $\mathrm{f}'(0) = 0$，与 $\mathrm{f}'(0) = k \neq 0$ 矛盾。所以 $\mathrm{f}(0) = 1$。 ■

阶段 2　由零处的导数得出处处的导数

> **想一想**　在 $(*)$ 中，写成 $\mathrm{f}(x+h) = \mathrm{f}(x)\mathrm{f}(h)$。有甚么可以抽到极限之外？

$$\begin{aligned}\mathrm{f}'(x) &= \lim_{h\to 0}\frac{\mathrm{f}(x)\mathrm{f}(h) - \mathrm{f}(x)}{h} = \mathrm{f}(x)\lim_{h\to 0}\frac{\mathrm{f}(h)-1}{h}\\ &= \mathrm{f}(x)\lim_{h\to 0}\frac{\mathrm{f}(0+h)-\mathrm{f}(0)}{h} = \mathrm{f}(x)\,\mathrm{f}'(0) = k\,\mathrm{f}(x).\end{aligned}$$ ■

函数方程让我们把 0 附近的性态搬到其他每一点去。

阶段 3　求解 $\mathrm{f}' = k\mathrm{f}$

考虑 $\mathrm{u}(x) = \mathrm{f}(x)e^{-kx}$。则 $\mathrm{u}'(x) = \big(\mathrm{f}'(x) - k\mathrm{f}(x)\big)e^{-kx} = 0$，所以 u 是常数，而 $\mathrm{u}(0) = \mathrm{f}(0) = 1$。因此

$$\mathrm{f}(x) = e^{kx}.$$ ■

（分离变量（separating the variables），$\int \frac{\mathrm{df}}{\mathrm{f}} = \int k\,\mathrm{d}x$，会得到相同的答案，但需要 $\mathrm{f}(x) \neq 0$；用 u 这个技巧就免却了这重顾虑。）

阶段 4　第二条规则

> **想一想**　在 g 的规则中代入 $y = 0$。这告诉你 $\mathrm{g}(0)$ 是甚么？$|\mathrm{g}(x)| < 1$ 在哪里派上用场？

取 $y = 0$：$\mathrm{g}(x)\big(1 + \mathrm{g}(x)\mathrm{g}(0)\big) = \mathrm{g}(x) + \mathrm{g}(0)$，化简后得 $\mathrm{g}(0)\big(\mathrm{g}(x)^2 - 1\big) = 0$。由于 $|\mathrm{g}(x)| < 1$，括号内的式子永不为零，所以 $\mathrm{g}(0) = 0$。

> **停一停，试一试**　现在利用这条规则计算 $\mathrm{g}(x+h) - \mathrm{g}(x)$，化简，除以 h，再令 $h \to 0$。

$$\mathrm{g}(x+h) - \mathrm{g}(x) = \frac{\mathrm{g}(x) + \mathrm{g}(h) - \mathrm{g}(x) - \mathrm{g}(x)^2\mathrm{g}(h)}{1 + \mathrm{g}(x)\mathrm{g}(h)} = \frac{\mathrm{g}(h)\big(1 - \mathrm{g}(x)^2\big)}{1 + \mathrm{g}(x)\mathrm{g}(h)}.$$

除以 h。当 $h \to 0$ 时，$\frac{\mathrm{g}(h)}{h} = \frac{\mathrm{g}(h)-\mathrm{g}(0)}{h} \to \mathrm{g}'(0) = k$，而 $\mathrm{g}(h) \to \mathrm{g}(0) = 0$（可微函数必

定连续（continuous）），所以分母趋向 1。因此

$$\mathrm{g}'(x) = k\bigl(1 - \mathrm{g}(x)^2\bigr). \qquad \blacksquare$$

阶段 5 求解

> ? **想一想** 分离变量。当 $|g| < 1$ 时，哪个函数的导数是 $\frac{1}{1-g^2}$？

由于 $|\mathrm{g}| < 1$，我们可以相除：$\dfrac{\mathrm{g}'}{1-\mathrm{g}^2} = k$。函数 $\operatorname{artanh} g = \frac{1}{2}\ln\frac{1+g}{1-g}$ 的导数是 $\frac{1}{1-g^2}$，所以 $\frac{\mathrm{d}}{\mathrm{d}x}\operatorname{artanh}\mathrm{g}(x) = k$，而 $\operatorname{artanh}\mathrm{g}(x) = kx + C$。由 $\mathrm{g}(0) = 0$ 得 $C = 0$，所以

$$\mathrm{g}(x) = \tanh kx. \qquad \blacksquare$$

验算：加法公式 $\tanh(A+B) = \dfrac{\tanh A + \tanh B}{1 + \tanh A \tanh B}$ 正正就是题目给出的规则。

回顾

- **先试特殊值。**代入 $y = 0$ 便求得 $\mathrm{f}(0)$ 和 $\mathrm{g}(0)$。
- **由一点到处处。**加法定律（addition law）把 0 处的导数搬到每个 x，把函数方程化为微分方程（differential equation）。
- **唯一性（uniqueness）。**这个论证不单是验证 e^{kx} 和 $\tanh kx$ 符合要求，更证明了它们是仅有的可能。函数方程之所以能成为定义，原因正在于此。

背后的数学

相对论中的速度

在日常生活中，如果一列火车以速率 u 行驶，而你在车厢内以速率 v 向前走，你相对于地面的速率就是 $u + v$。爱因斯坦在 1905 年证明这只是一个近似。正确的法则（其中 c 为光速）是

$$u \oplus v = \frac{u + v}{1 + uv/c^2}.$$

若以 c 的分数来量度速率，这正是 g 的规则。我们的结果说明，如果把每个速度写成 $v = \tanh\phi$，那么速度的合成就是把各个 ϕ 相加。$\phi = \operatorname{artanh} v$ 这个量称为快度（rapidity）。火箭引擎每次推进都会增加固定份量的快度，但由于对每个 ϕ 都有 $\tanh\phi < 1$，速度只会愈来愈接近光速，却永远达不到（图 3.8.1）。把 $0.9c$ 与 $0.9c$ 合成，得到的不是 $1.8c$，而是 $\frac{1.8}{1.81}c \approx 0.9945c$。

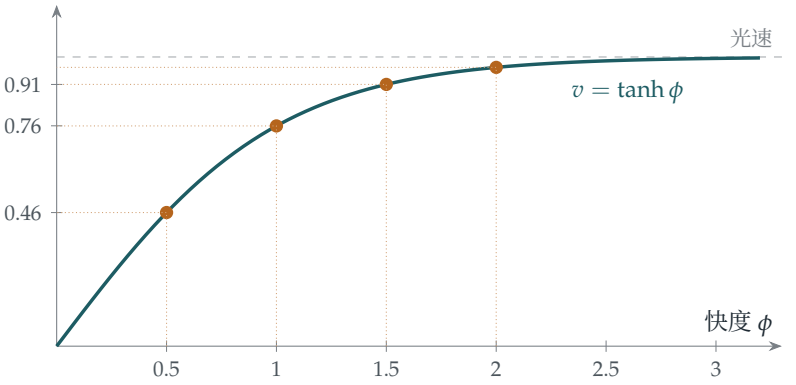


图 3.8.1　速度（以光速的分数表示）对快度的图像。快度每次增加相同的份量（引擎每次相同的推进所提供的正是这样），速度的增幅却愈来愈小：0.46、0.76、0.91、0.96、...，永远达不到 1。

柯西函数方程

最简单的函数方程是 $f(x+y) = f(x) + f(y)$，柯西（Cauchy）在 1821 年研究过它。如果 f 可微，用本题的方法可证明 $f(x) = kx$。事实上只要连续就够了。但若完全不作任何正则性（regularity）假设，就会有其他极其怪异的解，它们的图像稠密地散布在整个平面上。Georg Hamel 在 1905 年构造出这些解；这个构造需要用到*选择公理*（axiom of choice），而且不可能把其中任何一个明确地写出来。这就是题目坚持 f 可微的原因之一：这个假设并非例行公事。

同态

$f(x+y) = f(x)f(y)$ 的意思是，f 把实数的加法变成正实数的乘法，并且保持两者的结构。用大学一年级代数的语言来说，f 是由 $(\mathbb{R}, +)$ 到 $(\mathbb{R}_{>0}, \times)$ 的*群同态*（group homomorphism），而对数就是它的反函数。数学中许多最重要的函数，都是这样由它们所保持的结构来刻划的。在 Section 7.6 中，我们会看到 tanh 的几何意义：它量度的是“双曲角”（hyperbolic angle），而它的加法定律来自双曲旋转的复合。

思维技巧

用性质来定义函数

不用公式，而是用函数所遵守的规则来描述它，并证明只有一个函数遵守这条规则。

数学家往往就是这样看待 exp、sin、cos 和许多其他函数的：不把它们看成公式，而是看成函数方程或微分方程的唯一解。这解释了它们*为何*重要，而不只是怎样计算它们。

延伸练习

1 **正切的规则。**设 g 可微，$g'(0) = k \neq 0$，且只要等式两边都有定义，就有 $g(x+y) = \dfrac{g(x)+g(y)}{1-g(x)g(y)}$。证明 $g'(x) = k\big(1+g(x)^2\big)$，并由此推出在 $x=0$ 附近 $g(x) = \tan kx$。

2 **柯西方程。**若 f 可微，且对所有 x、y 都有 $f(x+y) = f(x)+f(y)$，证明 $f(x) = kx$。

3 **不用导数。**较难 只假设 $f(x+y) = f(x)f(y)$ 及 $f(1) = a > 0$。证明对正整数 n 有 $f(n) = a^n$，然后证明 $f(1/n) = a^{1/n}$，再证明对每个正有理数 q 都有 $f(q) = a^q$。（为甚么要再进一步就需要连续性?）

4 **火箭科学。**一枚火箭每次点火，都能相对于它当时所处的参考系把速率增加 $0.5c$。由静止开始，经过 2、3 和 4 次点火后，它（相对于起点）的速率是多少？请使用快度。

延伸阅读

- **Christopher Small, *Functional Equations and How to Solve Them***（Springer）。一本浅白易读的函数方程入门书，以喜欢解题的读者为对象。
- **Edwin Taylor & John Wheeler, *Spacetime Physics***（W. H. Freeman）。狭义相对论的经典入门书，全书都使用快度。
- **Tom Körner, *Calculus for the Ambitious***（Cambridge University Press）。由指数函数和三角函数的性质出发来定义它们，与本节的精神完全一致。

3.9 创造一个新函数

题目

STEP II 2021，第 4 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2021 年 STEP 试卷（zip 文件）内的 STEP II 试卷：

`www.ocr.org.uk/Images/696588-2021-past-papers.zip`

引子

你能解 $e^{-x} = 5x$ 吗？计算机上没有任何一个函数帮得上忙。移项整理也无济于事：x 同时困在指数函数的里面和外面。你可以用试误法（trial and improvement）求出数值答案，但并没有甚么“公式”。

真的没有吗？这道题踏出了大胆的一步：如果某一族方程无法用已知的函数解出来，那就*发明一个新函数*来解它们。数学家当年正是这样得到 $\sqrt{x}$（x^2 的反函数）、$\ln x$（e^x 的反函数）和 $\arcsin x$ 的。这里的新函数是 xe^x 的反函数（inverse function），现称为*兰伯特 W 函数*（Lambert W function）。有了它，从 $x^x = 10$ 到一条物理定律，范围广得出人意料的各种方程都可以精确地解出来。

动笔之前

想一想 “解”一条方程，究竟是甚么意思？$x = \sqrt{2}$ 是 $x^2 = 2$ 的解，还是只不过是那个解的名字？

两者皆是。$\sqrt{2}$ 是“平方等于 2 的那个正数”的名字；我们接受它作为答案，是因为函数 $\sqrt{\ }$ 已为人熟知、容易计算，而且性质清楚。一个答案算不算“公式”，取决于我们约定了为哪些函数命名。这道题为这份名单再添一个。

逐步解题

阶段 1　曲线 $y = xe^x$

$\frac{dy}{dx} = e^x + xe^x = (1+x)e^x$，它只在 $x = -1$ 处为零，此时 $y = -e^{-1}$。导数（derivative）在 $x < -1$ 时为负，在 $x > -1$ 时为正，所以这是一个极小值（minimum）。

> **想一想**　当 $x \to +\infty$ 和 $x \to -\infty$ 时，xe^x 会怎样？曲线在哪里与坐标轴相交？

当 $x \to \infty$ 时，$xe^x \to \infty$。当 $x \to -\infty$ 时，$xe^x \to 0$（从下方趋近），因为指数衰减胜过线性增长。曲线通过原点。所以它从负 x 轴的稍下方进来，下降至极小点 $(-1, -\frac{1}{e})$，然后上升穿过 $(0,0)$，再急速向上（图 3.9.1）。

阶段 2　反函数

要有反函数，f 必须是一对一（one-to-one）的，所以必须把它限制在一个严格递增（strictly increasing）或严格递减的区域上。包含右边部分的最大区间是 $[-1,\infty)$，所以 $a = -1$。在 $[-1,\infty)$ 上，f 由 $-\frac{1}{e}$ 递增至 ∞，所以反函数 g 定义在 $[-\frac{1}{e},\infty)$ 上，由 $g(-\frac{1}{e}) = -1$ 开始递增，经过 $g(0) = 0$，并且增长缓慢（当 x 很大时，比 $\ln x$ 还要慢）。它的图像是 $y = xe^x$ 右边部分关于直线 $y = x$ 的反射，在 $(-\frac{1}{e}, -1)$ 处有一条垂直切线。

g 的定义性质是

$$g(z)\, e^{g(z)} = z, \qquad g(z) \geq -1.$$

所以 (iii) 中每一部分的策略都一样：把方程整理成 (某个式子) $e^{(\text{同一个式子})}$ = (一个数) 的形式。

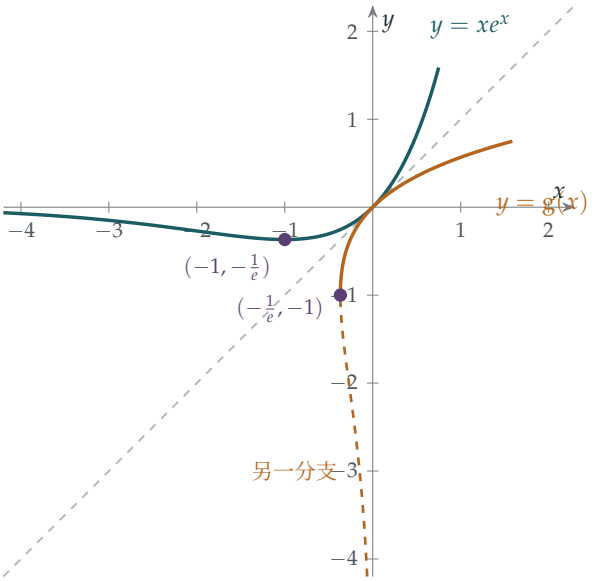


图 3.9.1 曲线 $y = xe^x$（蓝绿色）在 $(-1, -\frac{1}{e})$ 处有极小值。限制在 $x \geq -1$ 上，再关于 $y = x$ 反射，便得到反函数 g（琥珀色实线），即兰伯特 W 函数。琥珀色虚线是 $y = xe^x$ 左边部分的反射：这是第二个分支，第 (iii) 部分的第二个根要用到它。

阶段 3 四条方程

(a) $e^{-x} = 5x$。乘以 e^x 再除以 5：$xe^x = \frac{1}{5}$。所以 $x = g(\frac{1}{5}) \approx 0.169$。由于 xe^x 在 $x < 0$ 时为负，任何根都是正数，而 xe^x 在 $x > 0$ 时递增，所以这是唯一的根。

> **想一想** 在 (b) 和 (c) 中，未知数以 $x \ln x$ 的形式出现。用甚么代换（substitution）可以把 $x \ln x$ 变成 ue^u 的形式？

设 $x = e^u$，则 $\ln x = u$，而 $x \ln x = ue^u$。

(b) $2x \ln x + 1 = 0$ 变成 $ue^u = -\frac{1}{2}$。但对所有实数 u 都有 $ue^u \geq -\frac{1}{e} \approx -0.368$，而 $-\frac{1}{2} < -\frac{1}{e}$。所以**没有实根**。

(c) $3x \ln x + 1 = 0$ 变成 $ue^u = -\frac{1}{3}$，而 $-\frac{1}{e} < -\frac{1}{3} < 0$。高度为 $-\frac{1}{3}$ 的水平线与 $y = ue^u$ 相交**两次**，在极小点 $u = -1$ 的两侧各一次。右边的根是 $u = g(-\frac{1}{3})$，由此得

$$x = e^{g(-1/3)} \approx 0.538.$$

另一个根满足 $u < -1 < g(-\frac{1}{3})$，而 $x = e^u$ 随 u 递增，所以我们找到的根是两者之中**较大**的一个。（另一个约为 0.220。）

> **停一停，试一试** 现在试做 (d)，$x = 3\ln x$。（提示：两边取指数，设法凑成 (某个式子) $e^{\text{某个式子}}$ 的形式，其中“某个式子”是 x 的倍数。）

(d) $x = 3\ln x$ 即 $e^{x/3} = x$，所以 $xe^{-x/3} = 1$，再乘以 $-\frac{1}{3}$，

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

所以 $-\frac{x}{3}$ 是在 (c) 中找到的 $ue^u = -\frac{1}{3}$ 的两个解之一。取由 g 给出的那一个，$x = -3\,\mathrm{g}\left(-\frac{1}{3}\right) \approx 1.857$。另一个根来自另一个解 $u < \mathrm{g}(-\frac{1}{3})$，它给出较大的 $x = -3u$。所以这一次我们找到的根是两者之中**较小**的一个。（另一个约为 4.536。） ■

阶段 4 $x^x = 10$

取对数，$x\ln x = \ln 10$。设 $x = e^u$，这就是 $ue^u = \ln 10$，所以 $u = \mathrm{g}(\ln 10)$，而

$$x = e^{\mathrm{g}(\ln 10)} \approx 2.506.$$

由于 $\mathrm{g}(z)e^{\mathrm{g}(z)} = z$，我们也可以把它写成 $x = \dfrac{\ln 10}{\mathrm{g}(\ln 10)}$。 ■

回顾

- **为缺少的函数命名。**当一条方程无法用已知函数解出时，把一个新函数定义为反函数，是正当的做法，而且往往富有启发性。
- **整理成标准形式。**(iii) 和 (iv) 的每一部分最终都归结为 $ue^u =$ 常数，用的是 $x = e^u$ 或 $u = -\frac{x}{3}$ 之类的代换。
- **用图像数根的个数。**$ue^u = c$ 有零个、一个还是两个解，只取决于 c 相对于极小值 $-\frac{1}{e}$ 的位置。

背后的数学

历史悠久、名字却很新的函数

兰伯特（Johann Heinrich Lambert）在 1758 年研究过这类方程，欧拉（Euler）在 1783 年作了进一步的研究。但这个函数一直没有标准名称，直到 1990 年代，电脑代数系统（computer algebra system）Maple 的开发者需要为它起一个名字。1996 年，Robert Corless、Gaston Gonnet、David Hare、David Jeffrey 和 Donald Knuth 发表了论文“On the Lambert W function”，汇集了它的性质和应用。这篇论文已被引用数千次，而 W

现已内置于大多数数学软件之中。题目中的函数 g 是*主分支*（principal branch）W_0；图 3.9.1 中的虚线，即给出 (c) 和 (d) 中第二个根的那条曲线，是另一个实分支 W_{-1}。

W 出现在哪里

- **维恩位移定律（Wien's displacement law）。**热的物体辐射最强的波长与其温度成反比。当中的常数来自解方程 $5(1-e^{-x})=x$，其正数解为 $x=5+W\left(-5e^{-5}\right)\approx 4.965$。这就是为甚么约 5800 K 的太阳看起来是黄白色，而温度较低的红矮星看起来是红色。
- **幂塔（power tower）。**无穷幂塔 $x^{x^{x^{\cdots}}}$ 在 $e^{-e}\le x\le e^{1/e}$ 时收敛（converge），其值为 $\dfrac{W(-\ln x)}{-\ln x}$。
- **人口与传染病模型。**在最简单的模型中，一场传染病的最终规模，以及某些增长模型中达到指定人口所需的时间，都由 W 给出。

初等函数

W 真的是新函数吗？还是可以用 exp、ln、根式和三角函数写出来？它确实是新的：一条可追溯至 1830 年代刘维尔（Joseph Liouville）的定理表明，W 不是*初等函数*（elementary function）。同一套理论还表明，一些看似平平无奇的积分，例如 $\int e^{-x^2}\,dx$，无法用初等函数表示，这就是为甚么正态分布（normal distribution）需要自己专用的特殊函数。

思维技巧

以反函数作为解

方程 $f(x)=c$ 的解是 $x=f^{-1}(c)$，前提是把 f 限制成一对一的函数。这个限制决定了你得到的是哪一个解。

许多看似不同的方程，其实是同一条方程的不同伪装；找出能揭示共同形式的代换，就是关键的一步。

延伸练习

1 **二、四，还有一个。**方程 $x^2 = 2^x$ 有显而易见的解 $x = 2$ 和 $x = 4$（比较 Section 3.5）。证明它还有一个负数解，而这个解是

$$x = -\frac{2}{\ln 2}\,\mathrm{g}\left(\frac{\ln 2}{2}\right) \approx -0.767.$$

2 **一个导数。**对 $\mathrm{g}(x)e^{\mathrm{g}(x)} = x$ 作隐函数微分（implicit differentiation），证明当 $x \neq 0$ 时，

$$\mathrm{g}'(x) = \frac{\mathrm{g}(x)}{x\big(1 + \mathrm{g}(x)\big)}.$$

在 $x = -\frac{1}{e}$ 处会怎样？

3 **一个积分。**利用代换 $x = we^w$ 证明 $\displaystyle\int_0^e \mathrm{g}(x)\,\mathrm{d}x = e - 1$。

4 **幂塔。**假设幂塔 $T = \sqrt{2}^{\sqrt{2}^{\sqrt{2}^{\cdots}}}$ 收敛，证明 $T = \sqrt{2}^{T}$，并由此推出 $T = 2$ 或 $T = 4$。究竟是哪一个？（计算数列 $t_1 = \sqrt{2}$、$t_{n+1} = \sqrt{2}^{t_n}$ 的头几项，并想想 Section 2.2 中的蛛网图（cobweb diagram）。）

延伸阅读

- **Robert Corless et al., “On the Lambert W function”**，*Advances in Computational Mathematics* 5 (1996)。为这个函数命名的论文。读完本节之后，论文的头几节应可读懂。
- **Brian Hayes, “Why W?”**，*American Scientist* 93 (2005)。一篇生动的文章，讲述这个函数的历史和用途。可在网上免费阅读。
- **Brian Conrad, “Impossibility theorems for elementary integration”**。网上可找到的讲义，解释刘维尔的理论。程度较深。

我们学到了甚么

本章以一个 $1 = 2$ 的“证明”（proof）和三个问题开始：怎样知道一个论证是正确的？怎样证明不等式（inequality）？函数（function）究竟是甚么？以下是本章九道题目教晓我们的东西。

1. 怎样知道一个论证是正确的？

Section 3.1 给了我们一份检查清单。每走一步，都要问三件事：*计算对吗？这一步在逻辑上成立吗？有甚么假设是没有明说的？*那一节出现的错误都很典型。比较两个驻点（stationary point）上的函数值，并不能说明哪一个是极大值。在无限的定义域上，驻点的函数值不能作为函数的界。微积分基本定理（fundamental theorem of calculus）要求被积函数连续（continuous）。此外，一个论证也可能只是碰巧得出“看似正确”的答案。

同样的警觉贯穿本章其余部分。Section 3.7 用反例（counterexample）说明，去掉一个假设（$f(0) = f(\pi) = 0$），不等式便不再成立。在 Section 3.8 中，函数可微（differentiable）这个假设必不可少：没有它，柯西（Cauchy）函数方程（functional equation）便有一些谁也写不出来的怪异解。而在 Section 3.6 和 Section 3.3 中，等号成立的条件告诉我们某个界是否已是最佳，这正是严谨的数学家总会问的那类问题。

2. 怎样证明不等式？

我们从几个不同方向接触了各个经典不等式，也看到它们之间的关系。

- **平方非负。**两个数的算术－几何平均不等式（AM–GM）其实就是 $(a-b)^2 \geq 0$。柯西的前推－后推归纳法（forward–backward induction）由此建立起一般情形（Section 3.2）。
- **凸性（convexity）。**凹（concave）曲线位于它的弦之上。单凭这一点便得出琴生不等式（Jensen's inequality）；配合对数，一行便得出 AM–GM；配合 $\sin x$，则得出一个关于三角形的精确不等式（Section 3.3）。
- **面积。**曲线之下的面积加上曲线旁边的面积，足以盖过一个长方形。这给出杨氏不等式（Young's inequality），再由此推出赫尔德（Hölder）不等式和柯西－施瓦兹（Cauchy–Schwarz）不等式（Section 3.4）。
- **图像的形状。**把 a^b 与 b^a 的比较改写成 $\frac{\ln x}{x}$ 在两点上的比较，便解决了 π^e 与 e^π 孰大孰小的问题（Section 3.5）。

- **插值（interpolation）与三角不等式（triangle inequality）。**把导数用函数值表示，然后逐项估计其界，便得出二次多项式的马尔可夫不等式（Markov's inequality）（Section 3.6）。
- **试验函数（test function）。**把一个一般的不等式应用于巧妙选取的函数，便得出 $\pi^2 \leq 10$ 和 $\pi \leq \frac{22}{7}$（Section 3.7）。

在这些方法之中作出选择，是一种要靠练习才能养成的技巧。粗略的指引是：见到积与和，可考虑 AM–GM；见到函数值的平均，可考虑琴生不等式；见到一个函数和它的反函数（inverse function），可考虑杨氏不等式；要比较两个形式相同的式子，则可尝试找出它们背后的那一个函数。

3. 函数究竟是甚么？

最后两节从两方面回答了这个问题。在 Section 3.8 中，e^{kx} 和 $\tanh kx$ 根本不是由公式给出的；它们是满足某些加法定律（addition law）的仅有可微函数。这是比公式更深一层的理解：它解释了为何这些函数不断出现，既见于复利息，也见于爱因斯坦（Einstein）的速度相加公式。在 Section 3.9 中，我们更进一步，创造了一个函数，即 xe^x 的反函数，因为手上已有的函数不足以解一族很自然的方程。你计算机上的每一个函数，当初都是这样诞生的新事物。

4. 严谨论证与不等式的工具箱

方法	用于何处
检查每一步和每一个假设	似是而非的错误论证（3.1）
局部性质与整体性质	驻点与无穷远处的极限（3.1）
前推－后推归纳法	AM–GM（3.2）
补上额外的项	AM–GM（3.2、3.3）
琴生不等式	三角形；AM–GM（3.3）
比较面积	杨氏不等式（3.4）
分离变量，归结为单一函数	π^e 与 e^π 的比较（3.5）
三角不等式、等号成立的情况	马尔可夫不等式（3.6）
试验函数	$\pi^2 \leq 10$（3.7）
函数方程	e^x 和 $\tanh x$（3.8）
反函数与代换	W 函数（3.9）

下一步

- 在 Section 3.7 中略窥一二的变分法（calculus of variations），是 Section 4.7 的主题。
- tanh 及其加法定律背后的几何，见于 Section 7.6。

- 第 4 章会以本章这种审慎的态度看待积分（integration）本身，由积分究竟是甚么谈起。

如果你想在入读大学前再进一步

- **J. Michael Steele, *The Cauchy–Schwarz Master Class***：不等式的最佳入门书，全书由题目构成。
- **Daniel Velleman, *How to Prove It***：有系统地教你写出正确证明的指南。
- **Michael Spivak, *Calculus***：由根基建立起来的严谨微积分；不少剑桥学生在入学前的那个暑假读这本书。
- **E. A. Maxwell, *Fallacies in Mathematics***：用来练习找出哪里出了错。

冯纽曼（Von Neumann）曾打趣说，在数学里你不是理解事物，只是习惯了它们。这话不无道理：本章那些精确的定义和严谨的论证，起初或会令人觉得陌生。但正是习惯了它们，一个懂得计算的学生才会变成一个能够确信无误的数学家。$1 = 2$ 的证明再也骗不到你；数学学位的目标，就是让每一个更隐蔽的错误都同样一眼可见。

“延伸练习”答案

以下只是简短的答案和提示，并非完整解答：这里只列出结果和关键步骤，方便你核对自己的做法。请先认真做过每道练习，然后才看答案。

3.1 找出错处

1. 当 x 由 -1 变到 1 时会经过 0，而 $t = 1/x$ 在该处没有定义：这个代换（substitution）在区间上并不连续（continuous）。正确的结果是 $\int_{-1}^{1} \frac{\mathrm{d}x}{(1+x^2)^2} = \frac{1}{2} + \frac{\pi}{4}$（用 $x = \tan\theta$），而 $\int_{-1}^{1} \frac{-t^2}{(1+t^2)^2}\,\mathrm{d}t = \frac{1}{2} - \frac{\pi}{4}$。

2. (a) 正确：两边的对数（logarithm）都是 $\ln a \ln b$。(b) 错误：当 $\theta = 0$ 时，$\cos 0 = 1 \neq \sin 1$。（事实上两者永不相等，因为 $|\sin\theta \pm \cos\theta| \leq \sqrt{2} < \frac{\pi}{2}$。）(c) 错误：非常数的多项式（polynomial）是无界的，而一个常数不可能同时与 1 和 -1 相差不超过 10^{-6}。(d) 正确：由算术—几何平均不等式（AM–GM），$x^4 + x^{-4} \geq 2$，等号在 $x = 1$ 时成立。

3. 由 $n = 1$ 推到 $n = 2$ 的一步并不成立：两匹马当中的“前 1 匹”与“后 1 匹”并不重叠。

4. $\int_{\varepsilon}^{1} x^{-1/2}\,\mathrm{d}x = 2 - 2\sqrt{\varepsilon} \to 2$，所以由对称性，该积分（作为极限）存在，其值为 4。

3.2 倒过来的 AM–GM

1. 设 $a_1, \ldots, a_{n-1}$ 的平均值为 A，几何平均（geometric mean）为 G，再加入 A 作为第 n 个数。平均值仍然是 A，所以 $(G^{n-1}A)^{1/n} \leq A$，由此得 $G \leq A$。

2. 这 $n+1$ 个数的平均值为 $1 + \frac{1}{n+1}$，几何平均为 $(1 + \frac{1}{n})^{n/(n+1)}$，而且它们并非全部相等，所以不等式是严格的；把两边同时取 $n+1$ 次方即可。

3. 各个面的面积为 ab、bc、ca，而 $S = 2(ab + bc + ca)$。由 AM–GM 得 $\frac{S}{6} \geq (abc)^{2/3} = V^{2/3}$，等号在 $ab = bc = ca$ 时成立，即正方体。

4. 对 $1/a_i$ 应用 AM–GM：$\frac{1}{n}\sum \frac{1}{a_i} \geq \frac{1}{G}$。再取倒数。

3.3 凸性的威力

1. 若 $\mathrm{f}'' < 0$，斜率递减，所以曲线位于每条弦（chord）的上方。取 $t = \frac{1}{3}$：$\mathrm{f}\big(\frac{u+v+w}{3}\big) \geq \frac{1}{3}\mathrm{f}(u) + \frac{2}{3}\mathrm{f}\big(\frac{v+w}{2}\big) \geq \frac{1}{3}\big(\mathrm{f}(u) + \mathrm{f}(v) + \mathrm{f}(w)\big)$。

2. $(\ln\sin x)'' = -\operatorname{cosec}^2 x < 0$，所以 $\ln\sin$ 在 $(0,\pi)$ 上是凹（concave）函数。由延森不等式（Jensen's inequality）得 $\frac{1}{3}\ln(\sin A \sin B \sin C) \leq \ln\sin\frac{\pi}{3}$，所以该乘积最大为 $\big(\frac{\sqrt{3}}{2}\big)^3$。

3. 在 $(0, \frac{\pi}{2})$ 上 $(\tan x)'' = 2\sec^2 x \tan x > 0$；由延森不等式得 $\tan A + \tan B + \tan C \geq 3\tan\frac{\pi}{3} = 3\sqrt{3}$。

4. $(x^2)'' = 2 > 0$；由延森不等式即得所求的不等式，它说明 $\frac{1}{n}\sum a_i^2 - \bar{a}^2$（即方差（variance））是非负的。

3.4 以面积证明不等式

1. 取 $\mathrm{f} = \sin$，定义在 $[0, \frac{\pi}{2}]$ 上，等号在 $b = \mathrm{f}(a)$ 时成立；取 $a = \arcsin t$、$b = t$：$t\arcsin t = (1 - \cos(\arcsin t)) + \int_0^t \arcsin y \,\mathrm{d}y$。把所得结果的右边求导，得 $\arcsin t + \frac{t}{\sqrt{1-t^2}} - \frac{t}{\sqrt{1-t^2}} = \arcsin t$。

2. 在杨氏不等式（Young's inequality）中取 $p = q = 2$，得 $ab \leq \frac{1}{2}(a^2 + b^2)$。若 $\sum a_i^2 = \sum b_i^2 = 1$，逐项相加得 $\sum a_i b_i \leq 1$。一般情况下，把这个结果应用于 $a_i / \sqrt{\sum a_j^2}$ 和 $b_i / \sqrt{\sum b_j^2}$。

3. $\int_0^a (e^x - 1)\,\mathrm{d}x = e^a - 1 - a$，而 $\int_0^b \ln(1+y)\,\mathrm{d}y = (1+b)\ln(1+b) - b$。等号在 $b = e^a - 1$ 时成立。

4. $t = 1$：1.571 对 1；$t = 2$：0.524 对 0.268；$t = 10$：0.1002 对 0.0501。当 t 很大时，$t - \sqrt{t^2-1} = \frac{1}{t+\sqrt{t^2-1}} \approx \frac{1}{2t}$。这个界来自 $a = 0$ 的情况，此时 $(*)$ 只是说 $0 \leq \int_0^b \arcsin y \,\mathrm{d}y$；当 b 很小时，这块面积约为 $\frac{b^2}{2}$，是 $b\arcsin b \approx b^2$ 的一半，把它舍去便损失了因子 2。

3.5 π^e 还是 e^π？

1. $n^m = m^n$ 即 $h(n) = h(m)$。h 在 $(1,e)$ 上递增，其后递减，所以两数之中有一个小于 e：它必定是 2（因为 $h(1) = 0$ 这个值不会再次取到），于是另一个是 4。唯一的一对是 $\{2,4\}$。

2. $e < 3 < \pi$，而 h 在该处递减，所以 $h(3) > h(\pi)$，即 $3^\pi > \pi^3$。$\frac{9}{4}$ 和 $\sqrt{5}$ 都小于 e，而 $\frac{9}{4} > \sqrt{5}$，所以 $h(\frac{9}{4}) > h(\sqrt{5})$，由此得第二个结果。对于 $1 < x < 2$：由于 $3 < x+2 < 4$，有 $h(x) < h(2) = h(4) < h(x+2)$，所以 $x^{x+2} < (x+2)^x$。最后，

$9^{\sqrt{2}} = 3^{2\sqrt{2}}$，而 $\sqrt{2}^9 = 16\sqrt{2} = (2\sqrt{2})^3$。因为 $e^2 < 8$，所以 $e < 2\sqrt{2} < 3$，而 h 在该处递减，故 $h(2\sqrt{2}) > h(3)$，由此得 $(2\sqrt{2})^3 > 3^{2\sqrt{2}}$：$\sqrt{2}^9$ 较大。

3. $e^{\pi/e-1} > \frac{\pi}{e}$，所以 $e^{\pi/e} > \pi$；再把两边取 e 次方。

3.6 斜率的界

1. 两边都是二次多项式，并且在 $x = -1, 0, 1$ 处取值相同，所以两者相等（拉格朗日插值（Lagrange interpolation））。求导得 $\mathrm{f}(-1)(x - \frac{1}{2}) - 2\mathrm{f}(0)x + \mathrm{f}(1)(x + \frac{1}{2})$。

2. 在 $[-1, 1]$ 上 $|\mathrm{f}'(x)| \le |x - \frac{1}{2}| + 2|x| + |x + \frac{1}{2}| \le 4$，等号只在 $x = \pm 1$ 时成立，而且必须 $\mathrm{f}(1)$、$\mathrm{f}(-1)$、$-\mathrm{f}(0)$ 全部等于 1 或全部等于 -1。二次多项式由这三个值唯一确定，于是得 $\pm(2x^2 - 1)$。

3. 由 $T_3(\cos\theta) = \cos 3\theta$ 得 $4x^3 - 3x$，而 $T_3'(1) = 12 - 3 = 9$。由 $\cos(n+1)\theta + \cos(n-1)\theta = 2\cos\theta\cos n\theta$，得 $T_{n+1} = 2xT_n - T_{n-1}$，以及 $T_4 = 8x^4 - 8x^2 + 1$。

4. $-\sin\theta\, T_n'(\cos\theta) = -n\sin n\theta$。当 $\theta \to 0$ 时，$\frac{n\sin n\theta}{\sin\theta} \to n^2$。

3.7 $\pi^2 \le 10$

1. $\int_0^\pi \mathrm{f}^2 = \frac{\pi^9}{630}$，而 $\int_0^\pi (\mathrm{f}')^2 = \frac{2\pi^7}{105}$，所以由 $(*)$ 得 $\pi^2 \le 12$。与 $\sin x$ 相比，这个四次多项式在两端附近过于平坦，在中间又过于尖耸，所以它的斜率与大小之比离最优值较远。

2. $\int_0^\pi (\sin x + \sin 2x)^2 = \frac{\pi}{2} + \frac{\pi}{2} = \pi$（交叉项的积分为 0），而 $\int_0^\pi (\cos x + 2\cos 2x)^2 = \frac{\pi}{2} + 2\pi = \frac{5\pi}{2}$。

3. 设 $\mathrm{g}(x) = \mathrm{f}(\pi x/L)$。左边按 $\frac{L}{\pi}$ 的比例缩放，而 $\int(\mathrm{g}')^2$ 则按 $\frac{\pi}{L}$ 的比例缩放。对弦线而言，最低频率与 $\frac{1}{L}$ 成正比：长度减半，音高便加倍。

4. 界 $\pi^2 \le \int(\mathrm{f}')^2 / \int \mathrm{f}^2$（采用适当的单位）在 $\lambda \approx 1.133/\pi^2$ 时取得最小值，由此得 $\pi^2 \le 9.86975$，而真确值为 9.86960。

3.8 由加法公式决定的函数

1. 代入 $x = y = 0$ 得 $\mathrm{g}(0) = 0$。对 y 求导并取 $y = 0$，得 $\mathrm{g}'(x) = k(1 + \mathrm{g}(x)^2)$，所以 $\frac{\mathrm{d}}{\mathrm{d}x}\arctan\mathrm{g}(x) = k$，于是 $\mathrm{g}(x) = \tan kx$。

2. 对 y 求导得 $\mathrm{f}'(x+y) = \mathrm{f}'(y)$，所以 f' 是常数 k；又 $\mathrm{f}(0) = 0$（代入 $x = y = 0$）。

3. $\mathrm{f}(x) = \mathrm{f}(x/2)^2 \ge 0$，而且 f 永不为 0（否则它处处为 0）。用归纳法（induction）得 $\mathrm{f}(n) = a^n$；由 $\mathrm{f}(1/n)^n = \mathrm{f}(1) = a$ 及 $\mathrm{f}(1/n) > 0$ 得 $a^{1/n}$；于是 $\mathrm{f}(m/n) = \mathrm{f}(1/n)^m$。若没有连续性，函数在无理数处的值便不受限制：存在一些利用“哈默尔基（Hamel

basis)”构造出来的奇怪解，它们并不是 a^x。

4. 每次点火都使快度（rapidity）增加 $\phi = \operatorname{artanh}\frac{1}{2}$，所以点火 k 次之后，速率为 $\tanh k\phi = \frac{3^k-1}{3^k+1}$：$0.8c$、$\frac{13}{14}c \approx 0.929c$ 和 $\frac{40}{41}c \approx 0.976c$。

3.9 创造一个新函数

1. 当 $x = -1$ 时，$x^2 - 2^x = \frac{1}{2} > 0$；当 $x = 0$ 时其值为 -1：所以有一个负根。设 $x = -2u/\ln 2$，则 $x^2 = 2^x$（其中 $x < 0$，取 $x = -2^{x/2}$）化为 $ue^u = \frac{\ln 2}{2}$，所以 $u = \mathrm{g}\left(\frac{\ln 2}{2}\right)$，而 $x \approx -0.7667$。

2. $\mathrm{g}'e^{\mathrm{g}}(1+\mathrm{g}) = 1$，而 $e^{\mathrm{g}} = x/\mathrm{g}$。当 $x = -\frac{1}{e}$ 时，$\mathrm{g} = -1$，斜率为无限大（图像在该处是垂直的）。

3. 设 $x = we^w$，则 $\mathrm{d}x = (1+w)e^w\,\mathrm{d}w$，而 w 由 0 变到 1：$\int_0^1 w(1+w)e^w\,\mathrm{d}w = \left[(w^2-w+1)e^w\right]_0^1 = e-1$。

4. $T = \sqrt{2}^T$ 即 $\frac{\ln T}{T} = \frac{\ln 2}{2}$，所以 $T = 2$ 或 4。各项 $1.414, 1.633, 1.761, \ldots$ 递增，并且一直小于 2（若 $t_n < 2$，则 $t_{n+1} < \sqrt{2}^2 = 2$），所以 $T = 2$；不动点（fixed point）4 是排斥（repelling）的。

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

全书共 13 章，本册为第 3 章。 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。

