

CHAPTER

3

Functions, Inequalities and Rigour

Knowing that something is true is not the same as knowing why, and the difference is where mathematics lives.

IN THIS CHAPTER

- | | | |
|-----|------------------------------------|---------------|
| 3.1 | Spot the Flaw | STEP I 1991 |
| 3.2 | AM–GM, Backwards | STEP II 1993 |
| 3.3 | The Power of Convexity | STEP II 2007 |
| 3.4 | Areas That Prove Inequalities | STEP III 2011 |
| 3.5 | π^e or e^π ? | STEP II 2013 |
| 3.6 | A Bound on Slopes | STEP I 1988 |
| 3.7 | $\pi^2 \leq 10$ | STEP II 2014 |
| 3.8 | Functions from Their Addition Laws | STEP III 2019 |
| 3.9 | Inventing a New Function | STEP II 2021 |

Getting It Right

In mathematics you don't understand things. You just get used to them.

— John von Neumann

Here is a “proof” that $1 = 2$. Let $a = b$. Then $a^2 = ab$, so $a^2 - b^2 = ab - b^2$, which factorises as $(a + b)(a - b) = b(a - b)$. Cancelling $a - b$ gives $a + b = b$, and since $a = b$, that says $2b = b$, so $2 = 1$.

You probably spotted the problem: since $a = b$, “cancelling $a - b$ ” means dividing by zero. But notice how plausible every step looked, and how easily the error hid inside a routine move. Real mistakes in real mathematics are rarely so obvious, and they have fooled some of the greatest mathematicians who ever lived. This chapter is about the habits of mind that catch them: precise definitions, careful use of theorems, and a healthy suspicion of arguments that seem too easy.

Why rigour matters

At school, mathematics is mostly about getting answers. At university it is mostly about *knowing that your answers are right*, and being able to convince someone else. That shift is the biggest single change between school and university mathematics, and it can be unsettling. Why spend a page proving something that “obviously” works?

The answer is that “obvious” is unreliable. The history of analysis is full of results that seemed obvious and were false: that a continuous function must be differentiable except perhaps at a few isolated points (André-Marie Ampère published a “proof” in 1806; in 1872 Weierstrass found a continuous function differentiable *nowhere*), that a convergent sum of continuous functions is continuous (Cauchy believed it; Abel found a counterexample), that you can rearrange the terms of a convergent series without changing its sum (you can't, in general). Rigour is not pedantry. It is the only reliable way to tell the true from the plausible.

Inequalities: the working tools of analysis

The second theme of this chapter is *inequalities*. In school mathematics most results are equations: this equals that. In analysis, most results are inequalities: this is at most that. To prove a sequence converges you show the error is *less than* any tolerance; to prove an integral exists you show some sum is *bounded*; to estimate how good an approximation is, you bound the error. A professional analyst spends much more time proving inequalities than equations, and the classical inequalities you will meet here (AM–GM, Jensen, Young, Hölder, Cauchy–Schwarz) are the tools of the trade.

They are also remarkable in their own right. A single inequality about functions, with a well-chosen example, proves $\pi^2 \leq 10$ (Section 3.7). A fact about the shape of the logarithm settles whether π^e or e^π is bigger (Section 3.5). A chemist's question about measurement errors leads to some of the most important polynomials in mathematics (Section 3.6).

What you will meet at university

Rigour is the theme of first-year *Analysis*, and also of a course often called *Numbers and Sets* or *Foundations*, where you learn to write proofs precisely. You will meet the formal definitions of continuity and differentiability, and theorems with precise hypotheses (the intermediate value theorem, the mean value theorem, the extreme value theorem) whose hypotheses really matter. Later courses develop the inequalities in this chapter into whole subjects: *functional analysis* studies spaces of functions measured by integrals like those in Section 3.4 and Section 3.7; *convex analysis* and *optimisation* grow out of Section 3.3; *approximation theory* grows out of Section 3.6; and *special functions* like the one in Section 3.9 appear throughout applied mathematics and physics.

The questions that run through this chapter

Three big questions

1. **How do you know an argument is correct?** What are the common ways for plausible arguments to fail, and how do you check?
2. **How do you prove an inequality?** Where do the classical inequalities come from, and how do you choose which one to use?
3. **What is a function, really?** Can a function be defined by a property it has, rather than by a formula? And can we invent new functions when the old ones are not enough?

The chapter opens with error-spotting (Section 3.1), then works through the classical inequalities, each from a different angle: an unusual induction, convexity, areas, the shape of a graph, polynomial interpolation and test functions. It closes with two sections on functions defined by their properties: e^x and $\tanh x$ from their addition laws, and a brand-new function defined as an inverse.

Ideas in this chapter

§	Topic	Idea
3.1	Plausible wrong arguments	Checking calculations, logic and hidden hypotheses
3.2	AM–GM	Forward–backward induction
3.3	Jensen’s inequality	Concave curves lie above their chords
3.4	Young’s inequality	Comparing areas in a picture
3.5	π^e or e^π ?	Reducing to one function and studying its graph
3.6	Slopes of quadratics	Expressing the unknown in terms of the known
3.7	$\pi^2 \leq 10$	Test functions in a general inequality
3.8	e^x and $\tanh x$	Functions defined by functional equations
3.9	The W function	Inventing a function as an inverse

How to read this chapter

As before, try each question before reading the solution. In this chapter, also try something else: after reading each solution, go back and find the one or two steps that a careless student would get wrong. Where would a hidden assumption slip in? Which hypothesis of which theorem needs checking? Reading in this way, critically and actively, is the single most useful habit you can build before university.

3.1 Spot the Flaw

The question

Criticise each step of the following arguments. You should correct the arguments where necessary and possible, and say (with justification) whether you think the conclusions are true even though the argument is incorrect.

- (i) The function g defined by

$$g(x) = \frac{2x^3 + 3}{x^4 + 4}$$

satisfies $g'(x) = 0$ only for $x = 0$ or $x = \pm 1$. Hence the stationary values are given by $x = 0$, $g(x) = \frac{3}{4}$ and $x = \pm 1$, $g(x) = 1$. Since $\frac{3}{4} < 1$, there is a minimum at $x = 0$ and maxima at $x = \pm 1$. Thus we must have $\frac{3}{4} \leq g(x) \leq 1$ for all x .

- (ii) $\int (1-x)^{-3} dx = -3(1-x)^{-4}$ and so $\int_{-1}^3 (1-x)^{-3} dx = 0$.

The hook

Most exam questions ask you to produce an argument. This one hands you two arguments and asks you to take them apart. That is a very different skill, and it is the one you will use most at university: checking your own work, reading a proof in a book and deciding whether you believe it, marking a friend's attempt, refereeing a paper.

The arguments here are bad in an instructive way. Every step *looks* like something you have seen a hundred times. Some steps contain arithmetic slips, but the more serious errors are errors of *logic*: conclusions that do not follow even when the calculations are right. Those are the errors that survive into professional mathematics, and spotting them is what this chapter is about.

Before you start



Ask yourself When you are asked to “criticise each step”, what should you check at each step? Make a list before you start.

A good checklist has at least three questions. *Is the calculation correct? Does this step follow from the previous ones? Are there hidden assumptions (that a function is continuous,*

that a limit exists, that a stationary point must be a maximum or minimum)? The question also asks whether the conclusions might be true anyway. A wrong argument can have a true conclusion, so that needs a separate check.

Working it through

STAGE 1 Is the derivative right?

? **Ask yourself** Differentiate g with the quotient rule. Where is $g'(x) = 0$? Check $x = 1$ and $x = -1$ directly.

By the quotient rule,

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

So $g'(x) = 0$ when $x = 0$ or $x^4 + 6x - 12 = 0$. At $x = 1$, $x^4 + 6x - 12 = -5$; at $x = -1$ it is -17 . Neither is a stationary point. The first claim is **false**.

The quartic $x^4 + 6x - 12$ is -12 at $x = 0$ and tends to $+\infty$ in both directions, so it has (at least) one positive root and one negative root; it has exactly two real roots, because its second derivative $12x^2$ is non-negative, so it is convex and can cross zero at most twice. Numerically they are $x \approx 1.386$ and $x \approx -2.247$.

STAGE 2 Are the values right?

$g(0) = \frac{3}{4}$ is correct, and $g(1) = \frac{5}{5} = 1$. But $g(-1) = \frac{-2+3}{1+4} = \frac{1}{5}$, not 1. Even on its own terms, the argument has an arithmetic error: it seems to have assumed that g is even.

STAGE 3 Does comparing values tell you which is a maximum?

? **Ask yourself** Suppose a function has exactly two stationary points, with values $\frac{3}{4}$ and 1. Does it follow that the first is a minimum and the second a maximum?

No. The *nature* of a stationary point is a local property: it depends on how the function behaves near that point, not on how its value compares with the value somewhere else. A function can have a local maximum lower than a local minimum (think of $x + 2 \sin x$), or a stationary point that is neither. Here, near $x = 0$, the factor $x^4 + 6x - 12$ is close to -12 , so

$$g'(x) = \frac{-2x^2(x^4 + 6x - 12)}{(x^4 + 4)^2} > 0 \quad \text{for small } x \neq 0.$$

The derivative is positive on both sides of 0: $x = 0$ is a stationary point of *inflection*, not a minimum.

STAGE 4 Do stationary values bound a function?

Pause and try Even with correct stationary points, why can't you conclude that every value of g lies between the smallest and largest stationary values? Find a value of x with $g(x) < \frac{3}{4}$.

Stationary points only describe *local* behaviour. The global behaviour also depends on what happens at the ends of the domain, here as $x \rightarrow \pm\infty$. Since $g(x) \rightarrow 0$ as $x \rightarrow \pm\infty$, and $g(x) < 0$ whenever $2x^3 + 3 < 0$, that is, $x < -\sqrt[3]{3/2} \approx -1.14$, the function takes negative values. For example $g(-2) = \frac{-13}{20}$. So the conclusion $\frac{3}{4} \leq g(x) \leq 1$ is **false** (Figure 3.1.1).

The correct statement: g increases to a global maximum of about 1.083 at $x \approx 1.386$, then decreases towards 0; to the left, it decreases from 0 to a global minimum of about -0.668 at $x \approx -2.247$, then increases through $x = 0$ (where it has a horizontal inflection) up to the maximum. So $-0.668 \lesssim g(x) \lesssim 1.083$ for all x .

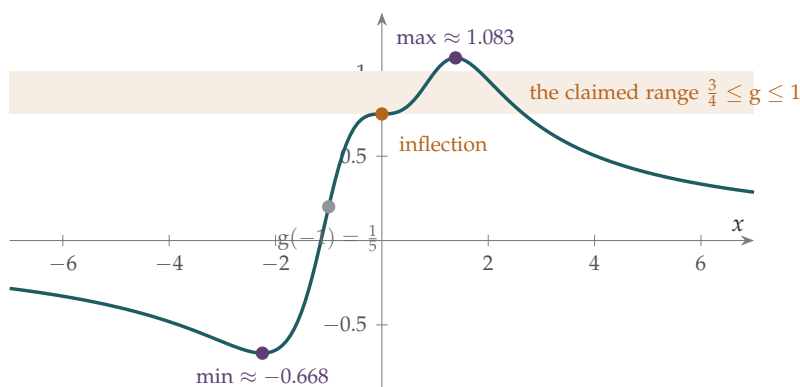


Figure 3.1.1 The true graph of $g(x) = \frac{2x^3+3}{x^4+4}$. The argument's claimed range (amber band) is wrong at every turn: $x = -1$ is not stationary and $g(-1) = \frac{1}{5}$; $x = 0$ is an inflection, not a minimum; and the function goes well outside the band, to a maximum above 1 and a minimum below 0.

STAGE 5 The integral

Ask yourself Differentiate $-3(1-x)^{-4}$. Do you get $(1-x)^{-3}$?

$\frac{d}{dx}[-3(1-x)^{-4}] = -3 \cdot (-4)(1-x)^{-5} \cdot (-1) = -12(1-x)^{-5}$. Wrong in both the power and the coefficient. The correct antiderivative is

$$\int (1-x)^{-3} dx = \frac{1}{2}(1-x)^{-2} + C,$$

as you can check by differentiating.



Ask yourself With the correct antiderivative, $\left[\frac{1}{2}(1-x)^{-2}\right]_{-1}^3 = \frac{1}{2} \cdot \frac{1}{4} - \frac{1}{2} \cdot \frac{1}{4} = 0$. So the answer was right after all. Or was it?

It was not. The fundamental theorem of calculus, $\int_a^b F'(x) dx = F(b) - F(a)$, requires the integrand to be continuous on the whole interval $[a, b]$. Here $(1-x)^{-3}$ is undefined at $x = 1$, which lies inside $[-1, 3]$, and near $x = 1$ it is enormous. The integral must be treated as an *improper* integral, split at the bad point:

$$\int_{-1}^1 (1-x)^{-3} dx = \lim_{\varepsilon \rightarrow 0^+} \left[\frac{1}{2}(1-x)^{-2} \right]_{-1}^{1-\varepsilon} = \lim_{\varepsilon \rightarrow 0^+} \left(\frac{1}{2\varepsilon^2} - \frac{1}{8} \right) = +\infty.$$

The integral from -1 to 1 is infinite (and the one from 1 to 3 is $-\infty$), so $\int_{-1}^3 (1-x)^{-3} dx$ **does not exist**. The conclusion “ $= 0$ ” is false, not because the true value is something else, but because there is no value. ■

Looking back

- **Three kinds of error.** Calculation errors (the derivative, $g(-1)$, the antiderivative), logical errors (comparing values to classify stationary points; stationary values as global bounds), and hidden assumptions (continuity in the fundamental theorem of calculus).
- **Right answer, wrong reason.** Correcting the antiderivative in part (ii) reproduces the answer 0 , but the argument is still invalid. A correct-looking answer is no evidence that the reasoning is sound.
- **Local versus global.** Derivatives tell you about the behaviour near a point. Global conclusions need more: what happens at the ends of the domain, or at points where the function misbehaves.

What's really going on

When great mathematicians got it wrong

Errors of exactly this kind were made by the founders of analysis. In his *Cours d'analyse* of 1821, Augustin-Louis Cauchy “proved” that a convergent sum of continuous functions is always continuous. In 1826 Niels Henrik Abel pointed out a counterexample: the Fourier series

$$\sin x - \frac{1}{2} \sin 2x + \frac{1}{3} \sin 3x - \dots$$

converges for every x , and every term is continuous, but the sum jumps at $x = \pm\pi$ (it equals $x/2$ for $-\pi < x < \pi$). Working out exactly what had gone wrong in Cauchy's argument took decades, and led to the concept of *uniform convergence*, which every analysis student now learns. Hidden assumptions, often about swapping the order of two limits, are the most common source of error in all of analysis.

Theorems with hypotheses

Two of the errors in this question come from ignoring the hypotheses of a theorem.

- The *extreme value theorem* says that a continuous function on a closed, bounded interval $[a, b]$ attains a maximum and a minimum, and they occur either at stationary points or at the endpoints. On the whole real line there are no endpoints to check, and the function might not have a maximum at all; you must examine the behaviour as $x \rightarrow \pm\infty$ instead.
- The *fundamental theorem of calculus* needs the integrand to be continuous on $[a, b]$. If it is not, the integral may still exist (try $\int_0^1 x^{-1/2} dx$) or may not (as here), and you must check using limits.

At university, every theorem you meet will come with a precise list of hypotheses. Part of the training is to check them, every time, before using the theorem.

A value for a divergent integral

There is a curious footnote to part (ii). The interval $[-1, 3]$ is symmetric about the bad point $x = 1$, and $(1 - x)^{-3}$ is “odd” about that point. If you cut out a symmetric gap $(1 - \varepsilon, 1 + \varepsilon)$ and let $\varepsilon \rightarrow 0$, the two infinite pieces cancel and you do get 0. This is called the *Cauchy principal value*, and it is genuinely useful in physics and complex analysis. But it is a *different definition*, chosen deliberately, and an argument that arrives at it by accident has not proved anything.

The thinking skill

Reading arguments critically

Check every step for three things: is the calculation right, does the step follow logically, and what has been assumed without being said?

Then check the conclusion separately: an incorrect argument can have a true conclusion, and a “correct” answer can come from an invalid argument. Apply the same scrutiny to your own work; the arguments you are most likely to believe are your own.

Try this next

- 1 **A substitution that lies.** (From STEP I, 1987.) Explain why the substitution $x = 1/t$ does not show that

$$\int_{-1}^1 \frac{dx}{(1+x^2)^2} \quad \text{and} \quad \int_{-1}^1 \frac{-t^2}{(1+t^2)^2} dt$$

are equal. Evaluate both integrals correctly. (One is positive and one is negative, so they certainly differ.)

2 True or false? (From STEP I, 1998.) Justify your answers.

- (a) $a^{\ln b} = b^{\ln a}$ for all $a, b > 0$.
- (b) $\cos(\sin \theta) = \sin(\cos \theta)$ for all real θ .
- (c) There is a polynomial P such that $|P(\theta) - \cos \theta| \leq 10^{-6}$ for all real θ .
- (d) $x^4 + 3 + x^{-4} \geq 5$ for all $x > 0$.

3 All horses are the same colour. Find the flaw in this “proof” by induction: any one horse is the same colour as itself; given $n + 1$ horses, the first n are the same colour by the inductive hypothesis, and so are the last n ; the two groups overlap, so all $n + 1$ are the same colour.

4 An integral that does exist. Show that $\int_{-1}^1 |x|^{-1/2} dx$ exists, although the integrand is undefined at 0, and find its value.

Further reading

- **E. A. Maxwell, *Fallacies in Mathematics*** (Cambridge University Press). A classic collection of plausible wrong arguments, organised by topic, written for sixth-formers.
- **Bernard Gelbaum & John Olmsted, *Counterexamples in Analysis*** (Dover). Hundreds of examples showing why the hypotheses of theorems matter. Dip in once you have started an analysis course.
- **Imre Lakatos, *Proofs and Refutations*** (Cambridge University Press). A brilliant dialogue about how mathematicians find, and fix, errors in proofs.

3.2 AM–GM, Backwards

The question

Suppose that $a_i > 0$ for all $i > 0$. Show that

$$a_1 a_2 \leq \left(\frac{a_1 + a_2}{2} \right)^2.$$

Prove by induction that for all positive integers m

$$a_1 \cdots a_{2^m} \leq \left(\frac{a_1 + \cdots + a_{2^m}}{2^m} \right)^{2^m}. \quad (*)$$

If $n < 2^m$, put $b_1 = a_1, b_2 = a_2, \dots, b_n = a_n$ and $b_{n+1} = \cdots = b_{2^m} = A$, where

$$A = \frac{a_1 + \cdots + a_n}{n}.$$

By applying $(*)$ to the b_i , show that

$$a_1 \cdots a_n A^{(2^m - n)} \leq A^{2^m}$$

(notice that $b_1 + \cdots + b_n = nA$). Deduce the (arithmetic mean)/(geometric mean) inequality

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

The hook

The arithmetic mean–geometric mean inequality (AM–GM) is perhaps the single most useful inequality in mathematics. It says that for positive numbers, the ordinary average is always at least the “multiplicative average”, with equality only when all the numbers are equal. It solves optimisation problems without calculus, it proves other inequalities, and it appears in every branch of analysis.

This question proves it by a strange and beautiful induction, due to Augustin-Louis Cauchy in 1821. Instead of going from n to $n + 1$, it leaps from 2 to 4 to 8 to 16, covering only the powers of 2, and then comes *back down* to fill in the numbers in between. It is a reminder that “induction” is a much more flexible idea than the textbook version suggests.

Before you start

The version of this paper in the online archive has a misprint: it reads “ $b_1 = a_2$ ”. It should of course be $b_1 = a_1$, as printed above; the idea is that the b s are the a s, padded out with copies of their average.

Ask yourself Why might it be easier to go from 2^m numbers to 2^{m+1} numbers than from n to $n + 1$?

Because 2^{m+1} numbers split naturally into two groups of 2^m , and you can apply the inductive hypothesis to each half and then use the two-number case to combine the halves. Going from n to $n + 1$ gives you nothing so symmetric to work with.

Working it through

STAGE 1 Two numbers

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

STAGE 2 Doubling

The case $m = 1$ is Stage 1. Suppose $(*)$ holds for 2^m numbers, and take 2^{m+1} positive numbers.

Ask yourself Split the 2^{m+1} numbers into a first half and a second half, and let P and Q be the arithmetic means of the two halves. What does the inductive hypothesis say about each half?

Applying $(*)$ to each half,

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

Multiplying (everything is positive), the product of all 2^{m+1} numbers is at most $(PQ)^{2^m}$. Now the two-number case gives $PQ \leq \left(\frac{P+Q}{2}\right)^2$, and $\frac{P+Q}{2}$ is exactly the mean of all 2^{m+1} numbers. So

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

which is $(*)$ for $m + 1$. By induction, AM–GM holds whenever the number of terms is a power of 2. ■

STAGE 3 Filling the gaps

Now take any n , and choose m with $2^m > n$. We pad the list $a_1, \dots, a_n$ with $2^m - n$ copies of its own average A .

? **Ask yourself** What is the average of the padded list $b_1, \dots, b_{2^m}$?

The sum is $(a_1 + \dots + a_n) + (2^m - n)A = nA + (2^m - n)A = 2^m A$, so the average is A itself. Adding copies of the average does not change the average. So $(*)$ applied to the b s gives

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

/ **Pause and try** Finish the proof.

Divide both sides by $A^{2^m - n}$ (which is positive): $a_1 \cdots a_n \leq A^n$. Taking n th roots,

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

Looking back

- **Induction along a different path.** Stage 2 is an induction on m , covering $n = 2, 4, 8, \dots$. Stage 3 is a way of stepping *down* from any power of 2 to any smaller n (Figure 3.2.1).
- **Padding with the average.** The trick of extending a list without changing its average is worth remembering; it is a way of turning a problem about any n into a problem about a convenient n .
- **When is there equality?** Tracing back through the proof, equality needs equality in every use of the two-number case, which happens only when all the numbers are equal.

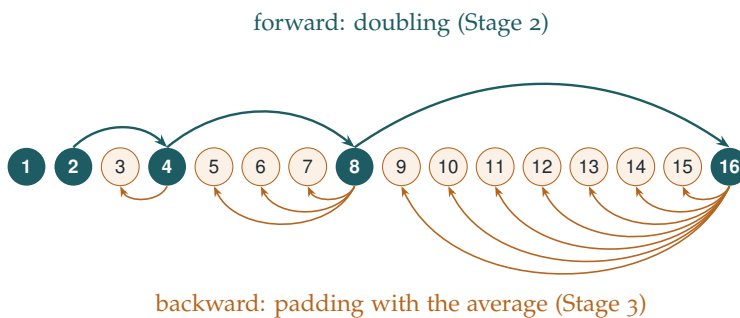


Figure 3.2.1 Cauchy's forward-backward induction. The teal arrows prove AM–GM for $2, 4, 8, 16, \dots$ numbers by doubling; the amber arrows fill in every other n by padding the list up to the next power of 2.

What's really going on

Cauchy's induction

Cauchy published this proof in his *Cours d'analyse* (1821), the book that began the rigorous treatment of analysis we discussed in Chapter 2. The technique is now called *forward-backward induction*: prove a statement for an infinite sequence of cases (here the powers of 2), and then show that the case n implies the case $n - 1$. Together, these reach every n . Our Stage 3 jumps straight down from 2^m to n , but you can also do it one step at a time: padding $n - 1$ numbers with one copy of their average shows that the case n implies the case $n - 1$.

Many proofs of one inequality

AM–GM has dozens of proofs, and comparing them is a good way to see different styles of mathematics.

- **By convexity.** The logarithm is concave, so the average of the logarithms is at most the logarithm of the average. That is AM–GM after taking exponentials, and it is the subject of Section 3.3.
- **George Pólya's proof.** Use $e^{x-1} \geq x$ (the tangent line to e^{x-1} at $x = 1$ lies below the curve). With A the mean, apply it to each a_i/A and multiply: $\prod \frac{a_i}{A} \leq e^{\sum(a_i/A-1)} = e^0 = 1$. Pólya said this proof came to him in a dream.
- **By smoothing.** If two of the numbers are unequal, replace them both by their average. The arithmetic mean is unchanged and, by the two-number case, the product increases. Repeating pushes all the numbers towards the average.

Optimisation without calculus

AM–GM turns many optimisation problems into one-liners. Of all boxes with a given surface area, which has the largest volume? Of all rectangles with a given perimeter, which has the largest area? (A square, by the two-number case: $xy \leq (\frac{x+y}{2})^2$, with equality when $x = y$.) The equality condition tells you where the optimum is, and the inequality tells you it really is the optimum, not just a stationary point. Compare the warnings of Section 3.1.

The thinking skill

Flexible induction

Induction does not have to go $1, 2, 3, \dots$. You can induct along any path that reaches every case: by doubling, by strong induction, or forwards on a subsequence and then

backwards to fill the gaps.

When the obvious step from n to $n + 1$ is hard, ask which step would be easy, and whether you can reach every case using it.

Try this next

- 1 **One step back.** Show directly that if AM–GM holds for n numbers, it holds for $n - 1$ numbers, by padding with one copy of the average.
- 2 **An increasing sequence.** Apply AM–GM to the $n + 1$ numbers $1 + \frac{1}{n}, \dots, 1 + \frac{1}{n}, 1$ (with n copies of $1 + \frac{1}{n}$) to show that $\left(1 + \frac{1}{n}\right)^n < \left(1 + \frac{1}{n+1}\right)^{n+1}$.
- 3 **The best box.** A closed rectangular box has surface area S . Use AM–GM on the areas of three different faces to show that its volume is at most $(S/6)^{3/2}$, with equality for a cube.
- 4 **Harmonic mean.** By applying AM–GM to the numbers $1/a_i$, show that the harmonic mean $\frac{n}{1/a_1 + \dots + 1/a_n}$ is at most the geometric mean.

Further reading

- **J. Michael Steele, *The Cauchy–Schwarz Master Class*** (Cambridge University Press). A wonderful, problem-driven book on inequalities, starting from exactly this level. Chapter 2 is about AM–GM.
- **Hardy, Littlewood & Pólya, *Inequalities*** (Cambridge University Press). The classic reference, written by three of the great analysts of the twentieth century.
- **Arthur Engel, *Problem-Solving Strategies*** (Springer). Chapter 7 is a large collection of inequality problems from olympiads.

3.3 The Power of Convexity

The question

A function $f(x)$ is said to be concave on some interval if $f''(x) < 0$ in that interval. Show that $\sin x$ is concave for $0 < x < \pi$ and that $\ln x$ is concave for $x > 0$.

Let $f(x)$ be concave on a given interval and let $x_1, x_2, \dots, x_n$ lie in the interval. *Jensen's inequality* states that

$$\frac{1}{n} \sum_{k=1}^n f(x_k) \leq f\left(\frac{1}{n} \sum_{k=1}^n x_k\right)$$

and that equality holds if and only if $x_1 = x_2 = \dots = x_n$. You may use this result without proving it.

(i) Given that A, B and C are angles of a triangle, show that

$$\sin A + \sin B + \sin C \leq \frac{3\sqrt{3}}{2}.$$

(ii) By choosing a suitable function f , prove that

$$\sqrt[n]{t_1 t_2 \cdots t_n} \leq \frac{t_1 + t_2 + \cdots + t_n}{n}$$

for any positive integer n and for any positive numbers $t_1, t_2, \dots, t_n$.

Hence:

- (a) show that $x^4 + y^4 + z^4 + 16 \geq 8xyz$, where x, y and z are any positive numbers;
- (b) find the minimum value of $x^5 + y^5 + z^5 - 5xyz$, where x, y and z are any positive numbers.

The hook

In Section 3.2 we proved AM–GM with a clever induction. Here it falls out in one line, as a special case of something much bigger. The bigger thing is *concavity*: the shape of a curve that bends downwards, like the top of a hill. A single geometric fact about such curves, that the curve lies above its chords, turns out to prove a startling range of inequalities, from a fact about triangles to AM–GM to inequalities in probability, information theory and economics.

At university this idea becomes *convexity*, and it is one of the great unifying themes of modern mathematics. Whole subjects, such as convex optimisation, are built on it.

Before you start



Ask yourself Draw a curve that bends downwards, such as $y = \sin x$ on $(0, \pi)$. Mark two points on it and join them with a straight line. Is the chord above or below the curve?

Below. For a concave curve, every chord lies below the curve between its endpoints. The average of two function values is the height of the midpoint of the chord; the function value at the average is the height of the curve above that midpoint. So “chord below curve” is exactly Jensen’s inequality for $n = 2$:

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

For n points, the average of the points $(x_k, f(x_k))$ is their centre of mass, which lies inside the region under the curve, so it is below the curve too (Figure 3.3.1).

Working it through

STAGE 1 Checking concavity

$\frac{d^2}{dx^2} \sin x = -\sin x < 0$ for $0 < x < \pi$, and $\frac{d^2}{dx^2} \ln x = -\frac{1}{x^2} < 0$ for $x > 0$. ■

STAGE 2 Triangles



Ask yourself Which function, which points, and what is their average? Do the angles of a triangle lie in an interval where the function is concave?

Take $f(x) = \sin x$ and the three points A, B, C . Each is strictly between 0 and π , where $\sin$ is concave, and their average is $\frac{A+B+C}{3} = \frac{\pi}{3}$. Jensen gives

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

so $\sin A + \sin B + \sin C \leq \frac{3\sqrt{3}}{2}$, with equality exactly for the equilateral triangle. ■

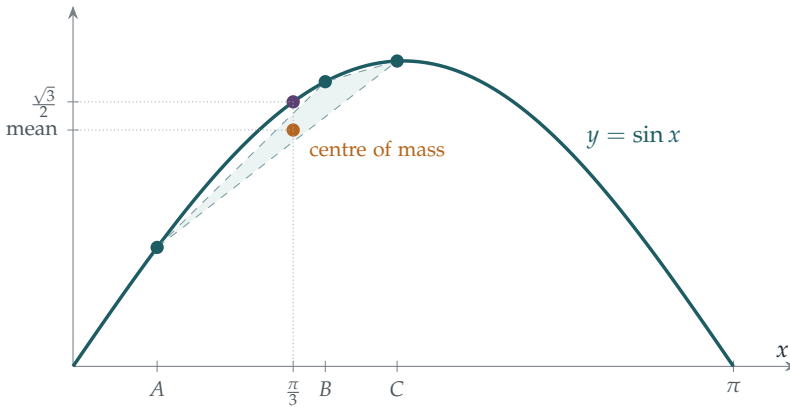


Figure 3.3.1 Jensen's inequality for $\sin x$ with three angles $A = 0.4$, $B = 1.2$, $C = \pi - 1.6$ of a triangle. The centre of mass of the three points on the curve (amber) lies inside the shaded triangle, and so below the curve. Its height is the average of $\sin A$, $\sin B$, $\sin C$, and the curve above it is at height $\sin \frac{\pi}{3}$.

STAGE 3 AM–GM in one line



Ask yourself AM–GM compares a product with a sum. Which concave function turns products into sums?

The logarithm. Apply Jensen to $f = \ln$ and the positive numbers $t_1, \dots, t_n$:

$$\frac{\ln t_1 + \dots + \ln t_n}{n} \leq \ln \left(\frac{t_1 + \dots + t_n}{n} \right).$$

The left-hand side is $\ln(\sqrt[n]{t_1 \cdots t_n})$. Since $\ln$ is increasing, we can remove it from both sides, and that is AM–GM. Equality holds exactly when all the t_k are equal. ■

STAGE 4 Using AM–GM



Ask yourself For (a), the left-hand side is a sum of four terms and the right-hand side contains a product of three variables. How many terms should you apply AM–GM to, and why does the 16 help?

Apply AM–GM to the four positive numbers $x^4, y^4, z^4, 16$:

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

Multiplying by 4 gives $x^4 + y^4 + z^4 + 16 \geq 8xyz$. ■



Pause and try Now find the minimum of $x^5 + y^5 + z^5 - 5xyz$. (Hint: you want a fifth root of $x^5y^5z^5$, so you need five numbers. What should the other two

be?)

Apply AM–GM to $x^5, y^5, z^5, 1, 1$:

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

so $x^5 + y^5 + z^5 - 5xyz \geq -2$. Equality holds when $x^5 = y^5 = z^5 = 1$, that is, $x = y = z = 1$, where the expression is $3 - 5 = -2$. So the minimum value is -2 . ■

Notice how the equality condition does double duty: it shows the bound is actually attained, so -2 is the minimum and not just a lower bound.

Looking back

- ▶ **One idea, many inequalities.** A geometric fact about concave curves gave a result about triangles and AM–GM, with no calculation beyond checking a second derivative.
- ▶ **Choose the function and the points.** Using Jensen means deciding which function to apply it to and at which points. The logarithm, which turns products into sums, is a favourite.
- ▶ **Pad to fit.** In (b), adding two 1s made the AM–GM fit the problem, just as padding with the average did in Section 3.2.

What's really going on

Convex functions

At university the standard definition of a *convex* function (the upside-down version of concave) does not mention derivatives at all: f is convex if

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2) \quad \text{for all } x_1, x_2 \text{ and } 0 \leq t \leq 1,$$

which says exactly that chords lie above the curve. This definition makes sense for functions that are not differentiable, such as $|x|$, and for functions of many variables. The Danish engineer and mathematician Johan Jensen made convex functions a subject in their own right in 1906, with the inequality that bears his name.

Convexity everywhere

Convexity appears throughout mathematics and its applications.

- ▶ **Probability.** Jensen's inequality says that for a concave f and any random variable X , $\mathbb{E}[f(X)] \leq f(\mathbb{E}[X])$. In economics, with f a concave "utility" function, this explains why people prefer a certain £100 to a fifty-fifty gamble on £0 or £200, and why insurance exists.

- **Information theory.** The proof that entropy is maximised by the uniform distribution, which underlies data compression, is Jensen’s inequality for $\ln$.
- **Optimisation.** A convex function has no local minima except the global one, so “walk downhill” always finds the best answer. This is why convex optimisation problems can be solved efficiently, and why so much effort goes into recasting problems in convex form.

In Section 3.4 we will see a different side of convexity: a geometric duality between a function and its inverse.

The thinking skill

Find the general principle

When several results have similar proofs, look for the single principle behind them. Then prove the principle once and apply it many times.

AM–GM, the triangle inequality for sines, and many others are all instances of “a concave curve lies above its chords”. Recognising the principle turns a collection of tricks into a method.

Try this next

- 1 The definition.** (From STEP II, 2018.) Draw a sketch illustrating the chord definition of concavity in the box above. Explain why a function with $f''(x) < 0$ satisfies it. Then, by choosing $t = \frac{1}{3}$, $x_1 = u$ and $x_2 = \frac{v+w}{2}$ suitably, show that a concave function satisfies Jensen’s inequality for three points.
- 2 A product of sines.** (From STEP II, 2018.) By considering $\ln(\sin x)$, show that if A, B, C are the angles of a triangle then $\sin A \sin B \sin C \leq \frac{3\sqrt{3}}{8}$.
- 3 Acute triangles.** Show that $\tan x$ is convex on $(0, \frac{\pi}{2})$, and deduce that in an acute-angled triangle, $\tan A + \tan B + \tan C \geq 3\sqrt{3}$.
- 4 Power means.** Show that x^2 is convex, and deduce that $\left(\frac{a_1 + \dots + a_n}{n}\right)^2 \leq \frac{a_1^2 + \dots + a_n^2}{n}$. Interpret this as a statement about the variance of a set of numbers.

Further reading

- **J. Michael Steele, *The Cauchy–Schwarz Master Class*** (Cambridge University Press). Chapter 6 is about convexity and Jensen’s inequality.
- **Claudi Alsina & Roger Nelsen, *When Less Is More: Visualizing Basic Inequalities*** (Mathematical Association of America). Inequalities proved by pictures, many of them

using convexity.

- **Stephen Boyd & Lieven Vandenberghe, *Convex Optimization*** (Cambridge University Press). A modern classic, free online; the first chapters show why convexity matters in practice.

3.4 Areas That Prove Inequalities

The question

The following result applies to any function f which is continuous, has positive gradient and satisfies $f(0) = 0$:

$$ab \leq \int_0^a f(x) \, dx + \int_0^b f^{-1}(y) \, dy, \quad (*)$$

where f^{-1} denotes the inverse function of f , and $a \geq 0$ and $b \geq 0$.

- (i) By considering the graph of $y = f(x)$, explain briefly why the inequality $(*)$ holds.

In the case $a > 0$ and $b > 0$, state a condition on a and b under which equality holds.

- (ii) By taking $f(x) = x^{p-1}$ in $(*)$, where $p > 1$, show that if $\frac{1}{p} + \frac{1}{q} = 1$ then

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

Verify that equality holds under the condition you stated above.

- (iii) Show that, for $0 \leq a \leq \frac{1}{2}\pi$ and $0 \leq b \leq 1$,

$$ab \leq b \arcsin b + \sqrt{1-b^2} - \cos a.$$

Deduce that, for $t \geq 1$,

$$\arcsin(t^{-1}) \geq t - \sqrt{t^2 - 1}.$$

The hook

Here is an inequality proved by a picture so simple that it fits on a postage stamp. Draw an increasing curve through the origin. The area under it, plus the area to its left, always covers a rectangle. From that one observation comes *Young's inequality*, $ab \leq \frac{a^p}{p} + \frac{b^q}{q}$, which in turn is the key to Hölder's inequality and the Cauchy-Schwarz inequality, two of the most important tools in analysis.

The picture also contains a deeper idea. It pairs a function with its inverse in a perfectly symmetric way, and that pairing, generalised, is the *Legendre transform*, which connects

the two great formulations of classical mechanics and the two main descriptions of thermodynamics.

Before you start

Ask yourself $\int_0^a f(x) dx$ is the area between the curve and the x -axis. What area does $\int_0^b f^{-1}(y) dy$ represent on the same picture?

The graph of $x = f^{-1}(y)$ is the same curve as $y = f(x)$, viewed sideways. So $\int_0^b f^{-1}(y) dy$ is the area between the curve and the y -axis, from height 0 to height b . With that, part (i) is a matter of drawing.

Working it through

STAGE 1 The picture

Ask yourself Draw the rectangle with corners $(0,0)$ and (a,b) . Shade the region under the curve from 0 to a in one colour and the region to the left of the curve from 0 to b in another. Do the two regions cover the rectangle? When is there no overlap and no excess?

The two regions together cover the whole rectangle $[0, a] \times [0, b]$, and usually some area outside it as well (Figure 3.4.1). Every point (x, y) of the rectangle is either below the curve (so in the first region, since $x \leq a$) or to the left of it (so in the second, since $y \leq b$). The total shaded area is therefore at least the area of the rectangle, ab . That is (*).

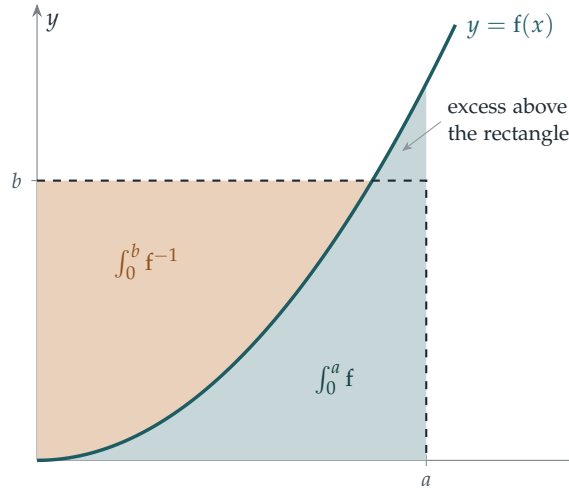


Figure 3.4.1 Young's inequality as a picture. The area under the curve from 0 to a (teal) and the area to the left of the curve from 0 to b (amber) together cover the dashed rectangle of area ab , with some area to spare. There is no excess exactly when the corner (a, b) lies on the curve.

The excess is the small curved region between the curve and the corner of the rectangle. It disappears exactly when the curve passes through the corner (a, b) , that is, when

$$b = f(a).$$

STAGE 2 Young's inequality

Ask yourself With $f(x) = x^{p-1}$, what is $f^{-1}(y)$? Compute both integrals, and use $\frac{1}{p} + \frac{1}{q} = 1$ to tidy the second.

Since $p > 1$, $f(x) = x^{p-1}$ is continuous and increasing with $f(0) = 0$, and $f^{-1}(y) = y^{1/(p-1)}$. Then

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

From $\frac{1}{p} + \frac{1}{q} = 1$ we get $q = \frac{p}{p-1}$, so $1 + \frac{1}{p-1} = \frac{p}{p-1} = q$, and the second integral is $\frac{b^q}{q}$. So (*) becomes

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}. \quad \blacksquare$$

Equality should hold when $b = f(a) = a^{p-1}$. Check: then $ab = a^p$, and $b^q = a^{(p-1)q} = a^p$, so the right-hand side is $a^p \left(\frac{1}{p} + \frac{1}{q} \right) = a^p$. $\blacksquare$

With $p = q = 2$ this is $ab \leq \frac{a^2+b^2}{2}$, the two-number AM–GM inequality of Section 3.2 in disguise.

STAGE 3 An inverse sine

Now take $f(x) = \sin x$ on $[0, \frac{\pi}{2}]$, which is continuous and increasing with $f(0) = 0$, and $f^{-1} = \arcsin$ on $[0, 1]$.

? **Ask yourself** $\int_0^a \sin x \, dx$ is easy. How would you integrate $\arcsin y$?

$\int_0^a \sin x \, dx = 1 - \cos a$. For the inverse sine, integrate by parts with $\arcsin y \times 1$:

$$\int_0^b \arcsin y \, dy = \left[y \arcsin y \right]_0^b - \int_0^b \frac{y}{\sqrt{1-y^2}} \, dy = b \arcsin b + \sqrt{1-b^2} - 1.$$

Adding, (*) gives $ab \leq b \arcsin b + \sqrt{1-b^2} - \cos a$. ■

Pause and try Now deduce the inequality for $\arcsin(t^{-1})$. Which value of a should you choose? (Try the simplest possible one.)

Take $a = 0$: then $0 \leq b \arcsin b + \sqrt{1-b^2} - 1$, that is, $b \arcsin b \geq 1 - \sqrt{1-b^2}$. (This is just the statement that $\int_0^b \arcsin y \, dy \geq 0$.) Put $b = 1/t$ with $t \geq 1$, so $0 < b \leq 1$, and divide by b :

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

Looking back

- **A picture as a proof.** Part (i) is a complete proof once you see the right picture. Translating a geometric fact into an inequality between integrals is a powerful technique.
- **Specialise a general result.** One general inequality, with different choices of f , gave Young's inequality and an inequality for $\arcsin$.
- **The simplest choice can be the right one.** The final deduction used $a = 0$, where the inequality is almost trivial. Always try the easy cases.

What's really going on**From Young to Hölder**

William Henry Young published his inequality in 1912. Its importance comes from what it leads to. Suppose $\sum a_i^p = 1$ and $\sum b_i^q = 1$. Applying Young to each pair and adding,

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

Scaling the a s and b s to make the sums 1 gives *Hölder's inequality*: for positive numbers,

$$\sum a_i b_i \leq \left(\sum a_i^p \right)^{1/p} \left(\sum b_i^q \right)^{1/q}.$$

With $p = q = 2$ this is the *Cauchy–Schwarz inequality*, $\sum a_i b_i \leq \sqrt{\sum a_i^2} \sqrt{\sum b_i^2}$, which says that the cosine of the angle between two vectors is at most 1. Hölder's inequality is the foundation of the theory of L^p spaces, spaces of functions in which “distance” is measured by $(\int |f|^p)^{1/p}$, and which you will meet in courses on functional analysis and probability.

The Legendre transform

The picture in Figure 3.4.1 treats f and f^{-1} completely symmetrically. Write $F(a) = \int_0^a f$ and $G(b) = \int_0^b f^{-1}$. The inequality says $F(a) + G(b) \geq ab$, with equality on the curve, so

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

This operation, turning a convex function F into a new convex function G , is the *Legendre transform*, and applying it twice gets you back to F . In mechanics it converts the *Lagrangian*, a function of position and velocity, into the *Hamiltonian*, a function of position and momentum: two equivalent formulations of Newton's laws that you will meet in a second-year course. In thermodynamics it converts between energy, enthalpy and free energy. For $F(a) = a^p/p$ the transform is $G(b) = b^q/q$, which is why the two exponents in Young's inequality are linked by $\frac{1}{p} + \frac{1}{q} = 1$.

The thinking skill

Inequalities from pictures

An integral is an area. Comparing areas in a well-chosen picture can prove an inequality that is hard to see algebraically.

Look for pictures in which one region obviously contains another, or several regions obviously cover a third. Then translate each area into an integral.

Try this next

- 1 The 1987 version.** (From STEP I, 1987.) Using the same picture with $y = \sin x$, and choosing s so that equality holds in (*), show that $\int_0^t \arcsin y \, dy = t \arcsin t - (1 - \cos(\arcsin t))$ for $0 \leq t \leq 1$. Check the result by differentiating both sides with respect to t .

2 Cauchy–Schwarz. Use Young’s inequality with $p = q = 2$ to prove $a_1b_1 + a_2b_2 + a_3b_3 \leq \sqrt{a_1^2 + a_2^2 + a_3^2} \sqrt{b_1^2 + b_2^2 + b_3^2}$ for positive numbers. (*Hint: first deal with the case where both square roots equal 1.*)

3 An exponential version. Take $f(x) = e^x - 1$ in (*) to show that, for $a, b \geq 0$,

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

When is there equality?

4 How good is the bound? HARDER Use a calculator to compare $\arcsin(1/t)$ with $t - \sqrt{t^2 - 1}$ for $t = 1, 2, 10$. Show that for large t , $t - \sqrt{t^2 - 1} \approx \frac{1}{2t}$ while $\arcsin(1/t) \approx \frac{1}{t}$. Why is the bound off by a factor of 2? (Look at the picture with $a = 0$: how much of the excess is wasted?)

Further reading

- **J. Michael Steele, *The Cauchy–Schwarz Master Class*** (Cambridge University Press). Chapters 1 and 9 cover Cauchy–Schwarz and Hölder, with Young’s inequality along the way.
- **Roger Nelsen, *Proofs Without Words*** (Mathematical Association of America). Includes Young’s inequality and many other picture proofs.
- **V. I. Arnold, *Mathematical Methods of Classical Mechanics*** (Springer). Section 14 explains the Legendre transform with exactly this picture. Advanced, but the relevant section is short.

3.5 π^e or e^π ?

The question

- (i) Find the value of m for which the line $y = mx$ touches the curve $y = \ln x$.
If instead the line intersects the curve when $x = a$ and $x = b$, where $a < b$, show that $a^b = b^a$. Show by means of a sketch that $a < e < b$.
- (ii) The line $y = mx + c$, where $c > 0$, intersects the curve $y = \ln x$ when $x = p$ and $x = q$, where $p < q$. Show by means of a sketch, or otherwise, that $p^q > q^p$.
- (iii) Show by means of a sketch that the straight line through the points $(p, \ln p)$ and $(q, \ln q)$, where $e \leq p < q$, intersects the y -axis at a positive value of y . Which is greater, π^e or e^π ?
- (iv) Show, using a sketch or otherwise, that if $0 < p < q$ and $\frac{\ln q - \ln p}{q - p} = e^{-1}$, then $q^p > p^q$.

The hook

Which is bigger, π^e or e^π ? Your calculator will tell you in a second: $e^\pi \approx 23.14$ and $\pi^e \approx 22.46$. But a calculator only tells you *that*; it does not tell you *why*, and it cannot help with infinitely many similar questions at once. Is 3^π bigger than π^3 ? Are there any whole numbers with $n^m = m^n$ apart from the obvious $n = m$?

All these questions turn out to be about a single curve, $y = \ln x$, and the straight lines that cross it. This question shows how a sketch, used carefully, can settle comparisons that look hopeless by algebra, and how the number e emerges as the natural dividing point.

Before you start



Ask yourself Taking logarithms, $a^b = b^a$ is the same as $b \ln a = a \ln b$. Rearrange this so that a appears only on one side and b only on the other.

$b \ln a = a \ln b$ is the same as

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

So every comparison of a^b with b^a is a comparison of the values of the single function $h(x) = \frac{\ln x}{x}$ at a and b . And $\frac{\ln a}{a}$ is the gradient of the line from the origin to the point $(a, \ln a)$ on the curve $y = \ln x$. That is why the question is about lines.

Working it through

STAGE 1 The tangent through the origin

? **Ask yourself** If $y = mx$ touches $y = \ln x$ at $x = x_0$, what two equations must hold?

The line and curve must meet, $mx_0 = \ln x_0$, and have the same gradient, $m = 1/x_0$. Substituting the second into the first, $1 = \ln x_0$, so $x_0 = e$ and $m = 1/e$. The tangent from the origin touches the curve at $(e, 1)$.

If instead $y = mx$ crosses the curve at $x = a$ and $x = b$, then $\ln a = ma$ and $\ln b = mb$, so $\frac{\ln a}{a} = m = \frac{\ln b}{b}$, which gives $b \ln a = a \ln b$ and hence $a^b = b^a$. ■

For the sketch: a line through the origin that is slightly less steep than the tangent cuts the curve twice, once on each side of the point of tangency, so $a < e < b$. Lines steeper than $1/e$ miss the curve altogether, because $\ln x$ is concave and lies entirely below its tangent $y = x/e$.

STAGE 2 Lines that cut the y -axis above the origin

? **Ask yourself** Now the line is $y = mx + c$ with $c > 0$. What are $\frac{\ln p}{p}$ and $\frac{\ln q}{q}$ in terms of m and c ?

From $\ln p = mp + c$ and $\ln q = mq + c$,

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

Since $c > 0$ and $p < q$, $\frac{c}{p} > \frac{c}{q}$, so $\frac{\ln p}{p} > \frac{\ln q}{q}$. Multiplying by $pq > 0$ gives $q \ln p > p \ln q$, so $p^q > q^p$. ■

Geometrically: from the origin, the point $(p, \ln p)$ is seen at a steeper angle than $(q, \ln q)$ when the line through them crosses the y -axis above the origin.

STAGE 3 Chords to the right of e

? **Ask yourself** The chord from $(p, \ln p)$ to $(q, \ln q)$, with $e \leq p < q$, has gradient less than the gradient of the curve at p . Why? And why does that make its y -intercept positive?

Because $\ln x$ is concave, its gradient $1/x$ decreases, so the chord from p to q is less steep

than the tangent at p : its gradient s satisfies $s < \frac{1}{p}$. The chord's y -intercept is

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

using $p \geq e$. So the chord crosses the y -axis above the origin. ■

 **Pause and try** Use this, and Stage 2, to decide which is greater, π^e or e^π .

Take $p = e$ and $q = \pi$. The chord between them has a positive intercept, so by Stage 2, $p^q > q^p$: that is,

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

STAGE 4 Chords parallel to the tangent

Now suppose the chord from p to q has gradient exactly $\frac{1}{e}$, the same as the tangent $y = \frac{x}{e}$.

 **Ask yourself** The chord is parallel to the tangent at $(e, 1)$. Since $\ln x$ is concave, the curve lies below that tangent. Does the chord cut the y -axis above or below the origin?

The chord is the line $y = \frac{x}{e} + c$ for some c . Its endpoints lie on the curve, which lies on or below the tangent $y = \frac{x}{e}$, and touches it only at $x = e$. The endpoints cannot both be at $x = e$, so at least one lies strictly below the tangent, and therefore the whole parallel chord lies below the tangent: $c < 0$. Now repeat Stage 2 with $c < 0$: the inequality reverses, $\frac{\ln p}{p} < \frac{\ln q}{q}$, and so $q^p > p^q$. ■

Looking back

- **Reduce to one function.** Comparing a^b with b^a became comparing $h(a)$ with $h(b)$, where $h(x) = \frac{\ln x}{x}$.
- **A sketch with reasons.** Each “by means of a sketch” was backed up by a precise fact: concavity of $\ln$ (chords below the curve, tangents above it) and the sign of an intercept.
- **The special role of e .** The tangent from the origin touches at $x = e$, and that is the dividing point for all of these comparisons.

What's really going on

The graph of $\frac{\ln x}{x}$

Everything in this question is visible in the graph of $h(x) = \frac{\ln x}{x}$ (Figure 3.5.1). Its derivative is $h'(x) = \frac{1-\ln x}{x^2}$, so h increases on $(0, e)$, reaches its maximum $\frac{1}{e}$ at $x = e$, and then decreases towards 0. A horizontal line at height m between 0 and $\frac{1}{e}$ cuts the graph twice, at points $a < e < b$ with $h(a) = h(b)$, which is exactly $a^b = b^a$. To the right of e , h is decreasing, so for $e \leq p < q$ we have $h(p) > h(q)$, that is, $p^q > q^p$. In particular $e^\pi > \pi^e$, and more generally $e^x > x^e$ for every $x \neq e$: no other number beats e at this game.

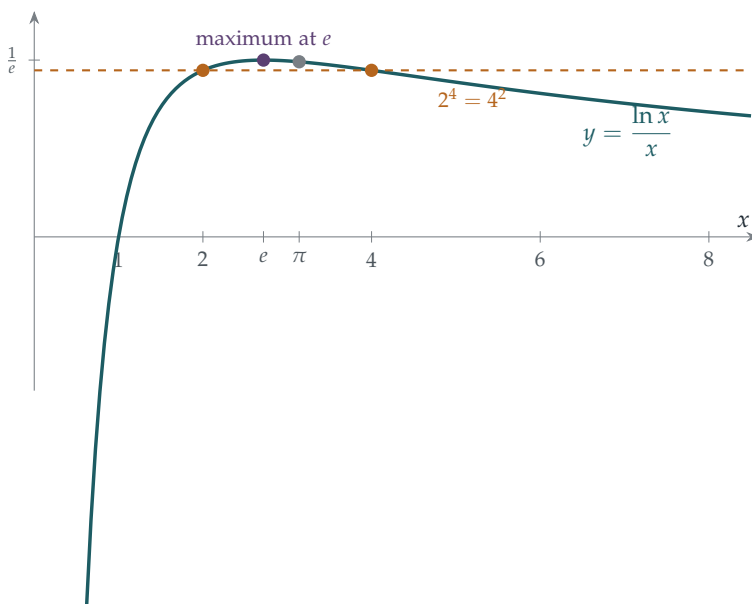


Figure 3.5.1 The graph of $h(x) = \frac{\ln x}{x}$. It rises to a maximum of $\frac{1}{e}$ at $x = e$ and then falls. Because $h(e) > h(\pi)$, we get $e^\pi > \pi^e$. The horizontal line at height $\frac{\ln 2}{2}$ (amber) meets the graph at $x = 2$ and $x = 4$: the only pair of distinct whole numbers with $n^m = m^n$.

All solutions of $x^y = y^x$

Which pairs of distinct positive reals satisfy $x^y = y^x$? A neat trick finds them all. Put $y = tx$ with $t > 1$; then $x^{tx} = (tx)^x$ gives $x^t = tx$, so $x^{t-1} = t$ and

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

Writing $t = 1 + \frac{1}{n}$ gives $x = \left(1 + \frac{1}{n}\right)^n$ and $y = \left(1 + \frac{1}{n}\right)^{n+1}$. With $n = 1$ this is $(2, 4)$; with $n = 2$ it is $(\frac{9}{4}, \frac{27}{8})$, and indeed $(\frac{9}{4})^{27/8} = (\frac{27}{8})^{9/4}$. As $n \rightarrow \infty$ both tend to e , the

point where the two branches meet. The problem was discussed in letters between Daniel Bernoulli and Christian Goldbach in 1728, and later by Euler.

The most economical base

The function $x^{1/x} = e^{h(x)}$ is also maximised at $x = e$. This answers a curious question: if the “cost” of a number system is the base times the number of digits needed, the most economical base is e , and among whole numbers, 3 (narrowly beating 2). A few experimental computers, such as the Soviet Setun of 1958, have used base 3.

The thinking skill

Rearrange to separate the variables, then study one function

An inequality or equation involving two quantities in a tangled way often becomes a statement about one function evaluated at two points.

Once you have that function, its graph (where it increases, decreases and peaks) answers every question at once. Here a^b versus b^a became $h(a)$ versus $h(b)$.

Try this next

- 1 Whole numbers.** (From STEP I, 1988.) Use the graph of $h(x) = \frac{\ln x}{x}$ to find all pairs of distinct positive integers m and n with $n^m = m^n$.
- 2 More comparisons.** (From STEP II, 2025.) Use the graph to show that $3^\pi > \pi^3$, and that $(\frac{9}{4})^{\sqrt{5}} > \sqrt{5}^{9/4}$. Given that $1 < x < 2$, decide which is larger, x^{x+2} or $(x+2)^x$. (Hint: $h(2) = h(4)$.)
- 3 Using $e^2 < 8$.** (From STEP II, 2025.) Show that $9^{\sqrt{2}} > \sqrt{2}^9$ and $3^{2\sqrt{2}} > (2\sqrt{2})^3$ are equivalent. Given that $e^2 < 8$, decide which of $9^{\sqrt{2}}$ and $\sqrt{2}^9$ is larger.
- 4 A one-line proof.** Use $e^x > 1 + x$ for $x \neq 0$, with $x = \frac{\pi}{e} - 1$, to prove $e^\pi > \pi^e$ directly.

Further reading

- **Eli Maor, *e: The Story of a Number*** (Princeton University Press). The history of e , including the equation $x^y = y^x$.
- **Paul Nahin, *An Imaginary Tale*** (Princeton University Press). Not only about $\sqrt{-1}$: full of curious calculations of exactly this kind.
- **Michael Spivak, *Calculus*** (Cambridge University Press). Chapter 11 covers the use of derivatives to understand the shape of graphs with complete rigour.

3.6 A Bound on Slopes

The question

The function f is defined by

$$f(x) = ax^2 + bx + c.$$

Show that

$$f'(x) = f(1) \left(x + \frac{1}{2}\right) + f(-1) \left(x - \frac{1}{2}\right) - 2f(0)x.$$

If a, b and c are real and such that $|f(x)| \leq 1$ for $|x| \leq 1$, show that $|f'(x)| \leq 4$ for $|x| \leq 1$.

Find particular values of a, b and c such that, for the corresponding function f of the above form, $|f(x)| \leq 1$ for all x with $|x| \leq 1$ and $|f'(x)| = 4$ for some x satisfying $|x| \leq 1$.

The hook

If you know a function stays between -1 and 1 on an interval, can you say anything about how steep it is? For a general function, no: $\sin(1000x)$ stays between -1 and 1 but has slopes as large as 1000 . But a quadratic cannot wiggle like that. This question shows that a quadratic trapped in the box $|x| \leq 1, |y| \leq 1$ can never have slope bigger than 4 inside the box, and that 4 is the best possible.

The problem has a surprising origin. In 1887 the chemist Dmitri Mendeleev, famous for the periodic table, was studying how the density of a solution depends on its concentration, and fitting quadratics to his data. He needed to know how badly errors in the data could affect the slope of his curves, and he proved exactly the result in this question. Two years later the young Andrei Markov solved the general problem for polynomials of any degree, and the answer led straight to some of the most important polynomials in mathematics.

Before you start



Ask yourself The formula for $f'(x)$ uses only three values of f : at $-1, 0$ and 1 . Why should three values be enough to determine the derivative of a quadratic?

A quadratic has three coefficients, so it is determined by its values at any three points. So its derivative, which is another function of the coefficients, must be expressible in terms of those three values. The formula tells you how. And it is exactly what you need, because the hypothesis $|f(x)| \leq 1$ tells you about the *values* of f , while the conclusion is about its *derivative*.

Working it through

STAGE 1 The formula

$f(1) = a + b + c$, $f(-1) = a - b + c$ and $f(0) = c$. So the right-hand side is

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

The coefficient of x is $(a + b + c) + (a - b + c) - 2c = 2a$, and the constant term is $\frac{1}{2}(a + b + c) - \frac{1}{2}(a - b + c) = b$. So the right-hand side is $2ax + b = f'(x)$. ■

STAGE 2 The bound



Ask yourself Take absolute values in the formula and use the triangle inequality, together with $|f(1)|, |f(-1)|, |f(0)| \leq 1$. What bound on $|f'(x)|$ do you get, as a function of x ?

By the triangle inequality $|u + v + w| \leq |u| + |v| + |w|$,

$$|f'(x)| \leq |f(1)| \left|x + \frac{1}{2}\right| + |f(-1)| \left|x - \frac{1}{2}\right| + 2|f(0)| |x| \leq \left|x + \frac{1}{2}\right| + \left|x - \frac{1}{2}\right| + 2|x|.$$



Pause and try Show that $\left|x + \frac{1}{2}\right| + \left|x - \frac{1}{2}\right| + 2|x| \leq 4$ for $|x| \leq 1$. (Split into cases according to the signs.)

If $|x| \leq \frac{1}{2}$, then $x + \frac{1}{2} \geq 0$ and $x - \frac{1}{2} \leq 0$, so $\left|x + \frac{1}{2}\right| + \left|x - \frac{1}{2}\right| = (x + \frac{1}{2}) - (x - \frac{1}{2}) = 1$, and the total is $1 + 2|x| \leq 2$.

If $\frac{1}{2} \leq |x| \leq 1$, both brackets have the same sign as x , so $\left|x + \frac{1}{2}\right| + \left|x - \frac{1}{2}\right| = 2|x|$, and the total is $4|x| \leq 4$.

Either way $|f'(x)| \leq 4$. ■

The bound is only achieved at $x = \pm 1$; in the middle half of the interval the slope is at most 2.

STAGE 3 An extreme example



Ask yourself For the bound to be attained at $x = 1$, every inequality in Stage 2 must be an equality. What values must $f(1)$, $f(-1)$ and $f(0)$ take, and with

? what signs?

At $x = 1$, the formula is $f'(1) = \frac{3}{2}f(1) + \frac{1}{2}f(-1) - 2f(0)$. To make this equal to 4 we need $f(1) = 1$, $f(-1) = 1$ and $f(0) = -1$. The quadratic through $(-1, 1)$, $(0, -1)$ and $(1, 1)$ is

$$f(x) = 2x^2 - 1,$$

so $a = 2$, $b = 0$, $c = -1$. On $|x| \leq 1$, $x^2 \in [0, 1]$, so $2x^2 - 1 \in [-1, 1]$, and $f'(1) = 4$. ■

Looking back

- **Express what you want in terms of what you know.** The formula rewrote the derivative in terms of values of f , which is where the hypothesis gave information.
- **The triangle inequality.** Bounding a sum by the sum of the absolute values is one of the most common moves in analysis.
- **Find the extremal case from the equality conditions.** Asking when every inequality is tight produced the example $2x^2 - 1$ directly.

What's really going on

Markov's inequality and Chebyshev polynomials

Andrei Markov proved in 1889 that if p is a polynomial of degree n with $|p(x)| \leq 1$ on $[-1, 1]$, then

$$|p'(x)| \leq n^2 \quad \text{for } |x| \leq 1.$$

For $n = 2$ this is our bound of 4. The extreme cases are the *Chebyshev polynomials* T_n , defined by

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

For example $\cos 2\theta = 2\cos^2 \theta - 1$ gives $T_2(x) = 2x^2 - 1$, our extreme quadratic, and $\cos 3\theta = 4\cos^3 \theta - 3\cos \theta$ gives $T_3(x) = 4x^3 - 3x$. Because $T_n(x) = \cos n\theta$, these polynomials oscillate between -1 and 1 as often as a polynomial of degree n possibly can, and they are as steep as possible at the ends (Figure 3.6.1). Markov's brother Vladimir extended the result to higher derivatives in 1892.

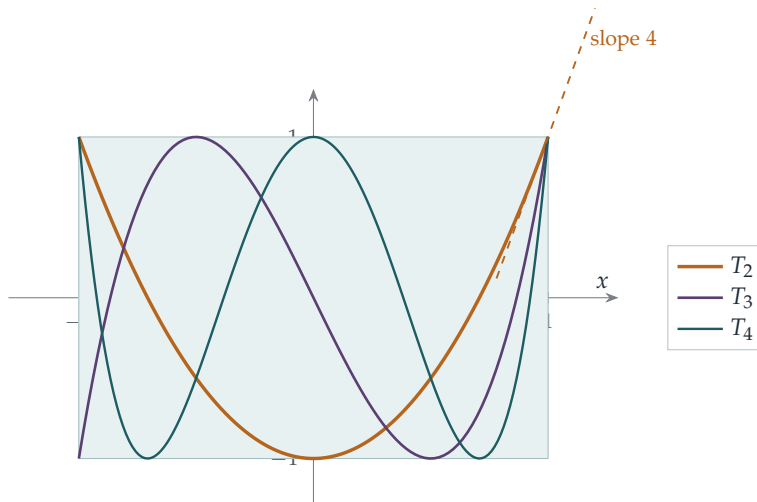


Figure 3.6.1 The Chebyshev polynomials $T_2 = 2x^2 - 1$, $T_3 = 4x^3 - 3x$ and $T_4 = 8x^4 - 8x^2 + 1$ stay inside the box $|x| \leq 1$, $|y| \leq 1$, touching its top and bottom as often as possible. They are the steepest polynomials of their degree at the ends: $T'_n(1) = n^2$. The dashed line is the tangent to T_2 at $x = 1$, with slope 4.

Interpolation

The formula in the question is a piece of *Lagrange interpolation*: given the values of a function at a few points, write down the unique polynomial of lowest degree through them, and use it to approximate the function or its derivative. Here the three points are $-1, 0, 1$. Formulas like this are used everywhere in numerical analysis, for example to estimate derivatives from measured data. Markov's inequality is the warning that comes with them: a small error in the values can be magnified by a factor of up to n^2 in the derivative, which is exactly the issue Mendelev was worried about.

Chebyshev polynomials in practice

Chebyshev polynomials are the workhorses of approximation theory. Interpolating a function at the roots of T_n , rather than at equally spaced points, keeps the error under control; with equally spaced points, high-degree interpolation can go wildly wrong near the ends of the interval (Runge's phenomenon). Software that evaluates functions to full machine precision, or solves differential equations very accurately, often works with Chebyshev expansions internally.

The thinking skill

Express the unknown in terms of the known

If the hypothesis tells you about one kind of quantity and the conclusion is about another, look for an identity linking them, then bound term by term.

And when you have a bound, ask whether it is best possible. Following the equality cases backwards usually leads you to the extreme example, and the extreme example is often interesting in its own right.

Try this next

- 1 **The general idea.** Show that if f is a quadratic then $f(x) = f(-1)\frac{x(x-1)}{2} + f(0)(1-x^2) + f(1)\frac{x(x+1)}{2}$, and differentiate to recover the formula in the question.
- 2 **Uniqueness.** HARDER Show that $\pm(2x^2 - 1)$ are the only quadratics with $|f(x)| \leq 1$ on $[-1, 1]$ and $|f'(x)| = 4$ somewhere in $[-1, 1]$.
- 3 **Chebyshev polynomials.** Use $\cos 3\theta = 4\cos^3 \theta - 3\cos \theta$ to check that $T_3(x) = 4x^3 - 3x$, and verify that $T'_3(1) = 9$. Show that $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$, and find T_4 .
- 4 **Slopes of T_n .** By differentiating $T_n(\cos \theta) = \cos n\theta$ with respect to θ , show that $T'_n(\cos \theta) = \frac{n \sin n\theta}{\sin \theta}$, and deduce that $T'_n(1) = n^2$ by letting $\theta \rightarrow 0$.

Further reading

- **Lloyd N. Trefethen, *Approximation Theory and Approximation Practice*** (SIAM). A beautiful modern book about Chebyshev polynomials and interpolation, readable by a strong school student, with computer experiments throughout.
- **Peter Borwein & Tamás Erdélyi, *Polynomials and Polynomial Inequalities*** (Springer). For the full story of Markov's inequality and its relatives. Advanced.
- **George M. Phillips, *Interpolation and Approximation by Polynomials*** (Springer). An undergraduate text covering Lagrange interpolation, Chebyshev polynomials and Markov-type inequalities.

3.7 $\pi^2 \leq 10$

The question

This question concerns the inequality

$$\int_0^\pi (f(x))^2 dx \leq \int_0^\pi (f'(x))^2 dx. \quad (*)$$

- (i) Show that $(*)$ is satisfied in the case $f(x) = \sin nx$, where n is a positive integer.

Show by means of counterexamples that $(*)$ is not necessarily satisfied if either $f(0) \neq 0$ or $f(\pi) \neq 0$.

- (ii) You may now assume that $(*)$ is satisfied for any (differentiable) function f for which $f(0) = f(\pi) = 0$.

By setting $f(x) = ax^2 + bx + c$, where a , b and c are suitably chosen, show that $\pi^2 \leq 10$.

By setting $f(x) = p \sin \frac{1}{2}x + q \cos \frac{1}{2}x + r$, where p , q and r are suitably chosen, obtain another inequality for π .

Which of these inequalities leads to a better estimate for π^2 ?

The hook

Every school student knows that π is a little more than 3. How would you *prove* an upper bound for it? Archimedes did it with polygons. This question does it in a completely unexpected way: by taking a general inequality about functions, plugging in a simple parabola, and reading off $\pi^2 \leq 10$. A different choice of function gives the famous approximation $\pi \leq \frac{22}{7}$.

The inequality $(*)$ is *Wirtinger's inequality*. It says that a function pinned to zero at both ends cannot be “big” without also being “steep”. Behind it lies one of the central ideas of mathematical physics: the vibrations of a string, the energy levels of a quantum particle, and the theory of eigenvalues.

Before you start



Ask yourself Why might a function that is zero at both ends of $[0, \pi]$ have to be steep if it is large in the middle? Why does the condition $f(0) = f(\pi) = 0$ matter?

To get from 0 up to some height H and back down to 0 within a distance π , a function must climb at an average rate of at least $H/(\pi/2)$. So its derivative cannot be small everywhere. Without the condition at the ends there is no such constraint: a constant function can be as large as you like with zero derivative. The inequality $(*)$ makes this precise, measuring “size” and “steepness” by integrals of squares.

Working it through

STAGE 1 Sine waves



Ask yourself What is $\int_0^\pi \sin^2 nx \, dx$? And $\int_0^\pi \cos^2 nx \, dx$? (Use $\sin^2 \theta = \frac{1}{2}(1 - \cos 2\theta)$.)

Both are $\frac{\pi}{2}$, since $\int_0^\pi \cos 2nx \, dx = 0$. With $f(x) = \sin nx$ we have $f'(x) = n \cos nx$, so

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

since $n \geq 1$. ■ Notice that $n = 1$ gives equality, so the inequality is as tight as it could be.

STAGE 2 The end conditions matter



Ask yourself Find a simple function with $f(0) \neq 0$ for which $(*)$ fails. Then find one with $f(0) = 0$ but $f(\pi) \neq 0$.

The constant function $f(x) = 1$ has $f(0) = 1 \neq 0$, and $\int_0^\pi 1 \, dx = \pi > 0 = \int_0^\pi 0 \, dx$.

For the second, $f(x) = x$ has $f(0) = 0$ but $f(\pi) = \pi$, and

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

So $(*)$ fails in both cases. ■

STAGE 3 A parabola



Ask yourself Which quadratic vanishes at $x = 0$ and $x = \pi$? (Up to a constant multiple, there is only one. Why doesn't the multiple matter?)

The quadratic must be a multiple of $x(\pi - x)$, and multiplying f by a constant k multiplies both sides of (*) by k^2 , so we may take $f(x) = x(\pi - x)$, that is, $a = -1$, $b = \pi$, $c = 0$. Then $f'(x) = \pi - 2x$, and with the substitution $u = x - \frac{\pi}{2}$ (which makes both integrands even),

$$\begin{aligned}\int_0^\pi x^2(\pi - x)^2 dx &= \int_{-\pi/2}^{\pi/2} \left(\frac{\pi^2}{4} - u^2\right)^2 du = \frac{\pi^5}{30}, \\ \int_0^\pi (\pi - 2x)^2 dx &= \int_{-\pi/2}^{\pi/2} 4u^2 du = \frac{\pi^3}{3}.\end{aligned}$$

(For the first: $\int_{-h}^h (h^2 - u^2)^2 du = 2(h^5 - \frac{2}{3}h^5 + \frac{1}{5}h^5) = \frac{16}{15}h^5$ with $h = \frac{\pi}{2}$.) So (*) says $\frac{\pi^5}{30} \leq \frac{\pi^3}{3}$, that is,

$$\pi^2 \leq 10. \quad \blacksquare$$

STAGE 4 A trigonometric choice

? **Ask yourself** For $f(x) = p \sin \frac{1}{2}x + q \cos \frac{1}{2}x + r$, what conditions do $f(0) = 0$ and $f(\pi) = 0$ impose on p, q, r ?

$f(0) = q + r = 0$ and $f(\pi) = p + r = 0$. So $p = q = -r$, and scaling, we take $f(x) = \sin \frac{x}{2} + \cos \frac{x}{2} - 1$.

Pause and try Compute $\int_0^\pi f^2$ and $\int_0^\pi (f')^2$ for this function. What inequality for π results?

Write $s = \sin \frac{x}{2}$ and $c = \cos \frac{x}{2}$. Then $(s + c)^2 = 1 + 2sc = 1 + \sin x$, so

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

Using $\int_0^\pi \sin x = 2$, $\int_0^\pi \sin \frac{x}{2} = 2$ and $\int_0^\pi \cos \frac{x}{2} = 2$,

$$\int_0^\pi f^2 dx = 2\pi + 2 - 4 - 4 = 2\pi - 6.$$

And $f' = \frac{1}{2}(c - s)$, so $(f')^2 = \frac{1}{4}(1 - \sin x)$ and $\int_0^\pi (f')^2 dx = \frac{1}{4}(\pi - 2)$. Now (*) gives

$$2\pi - 6 \leq \frac{\pi - 2}{4} \implies 8\pi - 24 \leq \pi - 2 \implies \pi \leq \frac{22}{7}. \quad \blacksquare$$

STAGE 5 Which is better?

From $\pi \leq \frac{22}{7}$ we get $\pi^2 \leq \frac{484}{49} \approx 9.878$, which is better than $\pi^2 \leq 10$. (The true value is $\pi^2 \approx 9.870$.) Notice that this proves $\frac{22}{7}$ is an *upper* bound for π ; a quite different integral, in Section 4.3, shows that it is not equal to π .

Looking back

- **A general inequality, a specific function.** Choosing a test function turned a statement about all functions into a statement about a number.
- **Scaling.** Both sides of (*) scale by the same factor, so we could fix the free constant in each family.
- **Better test functions give better bounds.** The closer the test function is to the extreme case $\sin x$, the closer the inequality is to equality, and the better the estimate (Figure 3.7.1).

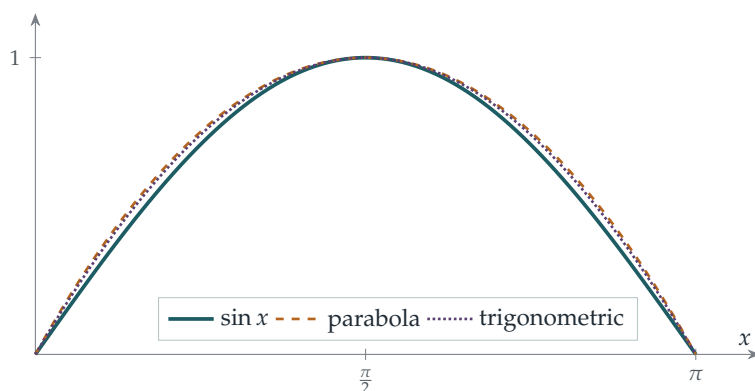


Figure 3.7.1 The three functions in this section, each scaled to have maximum 1: the extreme case $\sin x$ (solid), which gives equality in (*); the parabola $x(\pi - x)$ (dashed), which gives $\pi^2 \leq 10$; and $\sin \frac{x}{2} + \cos \frac{x}{2} - 1$ (dotted), which gives $\pi \leq \frac{22}{7}$. Both test functions are very close to $\sin x$, which is why the bounds are so good.

What's really going on

A proof with Fourier series

Why is (*) true? At university you will learn that any reasonable function on $[0, \pi]$ that vanishes at both ends can be written as a *Fourier sine series*,

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

Because $\int_0^\pi \sin mx \sin nx \, dx = 0$ for $m \neq n$, all the cross terms vanish when you square and integrate, giving

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

Each term on the right is at least the corresponding term on the left, which proves (*), and shows that equality holds only when f is a multiple of $\sin x$. Part (i) of the question is the single-term case.

Vibrating strings and quantum particles

The ratio $\int (f')^2 / \int f^2$ is called a *Rayleigh quotient*, and (*) says its smallest possible value, for functions pinned at 0 and π , is 1. This number has physical meaning. For a string fixed at both ends, $\int (f')^2$ measures the stretching energy and $\int f^2$ the size of the displacement, and the minimum value of the ratio determines the lowest note the string can play. In quantum mechanics, the same minimum is the lowest energy of a particle trapped in a box: it cannot be zero, because a wave pinned at both ends must bend. Lord Rayleigh used exactly this trick, guessing a test function and computing the ratio, to estimate the frequencies of vibrating systems; because the extreme case gives the minimum, any guess gives an upper bound on the true lowest frequency, and good guesses give very accurate ones.

The calculus of variations

Finding the function that minimises a ratio of integrals is a problem in the *calculus of variations*, where the unknown is a whole function rather than a number. The minimiser here, $\sin x$, satisfies the differential equation $f'' + f = 0$, and this is no coincidence: minimising an integral always leads to a differential equation for the minimiser. We return to this idea in Section 4.7.

The thinking skill

Test functions

An inequality that holds for every function in a class can be applied to a cleverly chosen function to give information about a specific number.

The art is in the choice: a test function close to the extreme case gives a nearly sharp result. This idea runs through analysis, physics and engineering, from estimating eigenvalues to the finite element method.

Try this next

- 1 **A quartic.** Use $f(x) = x^2(\pi - x)^2$ in (*). Show that it gives only $\pi^2 \leq 12$. Why is this test function worse than the parabola? (Compare its shape with $\sin x$.)
- 2 **Two sine waves.** Check (*) directly for $f(x) = \sin x + \sin 2x$, computing both sides.
- 3 **A different interval.** Show by a substitution that if $g(0) = g(L) = 0$ then $\int_0^L g^2 \leq \frac{L^2}{\pi^2} \int_0^L (g')^2$. What does this say about strings of different lengths?
- 4 **Rayleigh's method.** HARDER Take $f(x) = x(\pi - x) + \lambda x^2(\pi - x)^2$ and choose λ to make the bound on π^2 as good as possible. (The algebra is heavy; a computer algebra system helps.)

Further reading

- **T. W. Körner, *Fourier Analysis*** (Cambridge University Press). A wonderful, discursive book on Fourier series and their applications, written by a Cambridge analyst.
- **Hardy, Littlewood & Pólya, *Inequalities*** (Cambridge University Press). Chapter 7 discusses Wirtinger's inequality and its relatives.
- **Richard Feynman, *The Feynman Lectures on Physics***, Volume II, Chapter 19 ("The principle of least action"). Feynman's lecture on minimising integrals, with physical intuition throughout. Free online.

3.8 Functions from Their Addition Laws

The question

The definition of the derivative f' of a (differentiable) function f is

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}. \quad (*)$$

- (i) The function f has derivative f' and satisfies

$$f(x+y) = f(x)f(y)$$

for all x and y , and $f'(0) = k$ where $k \neq 0$. Show that $f(0) = 1$.

Using $(*)$, show that $f'(x) = kf(x)$ and find $f(x)$ in terms of x and k .

- (ii) The function g has derivative g' and satisfies

$$g(x+y) = \frac{g(x) + g(y)}{1 + g(x)g(y)}$$

for all x and y , $|g(x)| < 1$ for all x , and $g'(0) = k$ where $k \neq 0$.

Find $g'(x)$ in terms of $g(x)$ and k , and hence find $g(x)$ in terms of x and k .

The hook

How do you know what e^x is? At school it is introduced through compound interest, or as the function that is its own derivative, or as $\lim(1 + x/n)^n$. This question offers another answer: e^{kx} is the *only* differentiable function that turns addition into multiplication, $f(x+y) = f(x)f(y)$. The rule alone, plus one number $f'(0)$, pins the function down completely.

The second part does the same for a stranger rule, $g(x+y) = \frac{g(x)+g(y)}{1+g(x)g(y)}$. If that formula looks familiar, it should: it is Einstein's rule for adding velocities in special relativity, with the speed of light set to 1. The function it forces is $\tanh$, and that explains why nothing can be accelerated past the speed of light.

Before you start

An equation like $f(x+y) = f(x)f(y)$, where the unknown is a function, is called a *functional equation*.

Ask yourself The equation holds for *all* x and y . What is the most useful thing to try first?

Special values. Setting $y = 0$ (or $x = y$, or $y = -x$) often gives a lot of information cheaply. That is how both parts start.

Working it through

STAGE 1 The value at zero

Put $x = y = 0$: $f(0) = f(0)^2$, so $f(0) = 0$ or 1 .

Ask yourself What goes wrong if $f(0) = 0$?

If $f(0) = 0$, then putting $y = 0$ gives $f(x) = f(x)f(0) = 0$ for every x . Then f is identically zero and $f'(0) = 0$, contradicting $f'(0) = k \neq 0$. So $f(0) = 1$. ■

STAGE 2 The derivative everywhere from the derivative at zero

Ask yourself In (*), write $f(x+h) = f(x)f(h)$. What can you take outside the limit?

$$\begin{aligned} f'(x) &= \lim_{h \rightarrow 0} \frac{f(x)f(h) - f(x)}{h} = f(x) \lim_{h \rightarrow 0} \frac{f(h) - 1}{h} \\ &= f(x) \lim_{h \rightarrow 0} \frac{f(0+h) - f(0)}{h} = f(x) f'(0) = k f(x). \quad \blacksquare \end{aligned}$$

The functional equation lets the behaviour near 0 be transported to every other point.

STAGE 3 Solving $f' = kf$

Consider $u(x) = f(x)e^{-kx}$. Then $u'(x) = (f'(x) - kf(x))e^{-kx} = 0$, so u is constant, and $u(0) = f(0) = 1$. Therefore

$$f(x) = e^{kx}. \quad \blacksquare$$

(Separating the variables, $\int \frac{df}{f} = \int k dx$, gives the same answer, but needs $f(x) \neq 0$; the trick with u avoids that worry.)

STAGE 4 The second rule

Ask yourself Put $y = 0$ in the rule for g . What does that tell you about $g(0)$? Where does $|g(x)| < 1$ come in?

With $y = 0$: $g(x)(1 + g(x)g(0)) = g(x) + g(0)$, which simplifies to $g(0)(g(x)^2 - 1) = 0$. Since $|g(x)| < 1$, the bracket is never zero, so $g(0) = 0$.



Pause and try Now compute $g(x+h) - g(x)$ using the rule, simplify, divide by h and let $h \rightarrow 0$.

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

Divide by h . As $h \rightarrow 0$, $\frac{g(h)}{h} = \frac{g(h)-g(0)}{h} \rightarrow g'(0) = k$, and $g(h) \rightarrow g(0) = 0$ (a differentiable function is continuous), so the denominator tends to 1. Therefore

$$g'(x) = k(1 - g(x)^2). \quad \blacksquare$$

STAGE 5 Solving it



Ask yourself Separate the variables. Which function has derivative $\frac{1}{1-g^2}$ for $|g| < 1$?

Since $|g| < 1$, we can divide: $\frac{g'}{1-g^2} = k$. The function $\operatorname{artanh} g = \frac{1}{2} \ln \frac{1+g}{1-g}$ has derivative $\frac{1}{1-g^2}$, so $\frac{d}{dx} \operatorname{artanh} g(x) = k$ and $\operatorname{artanh} g(x) = kx + C$. From $g(0) = 0$, $C = 0$, so

$$g(x) = \tanh kx. \quad \blacksquare$$

Check: the addition formula $\tanh(A+B) = \frac{\tanh A + \tanh B}{1 + \tanh A \tanh B}$ is exactly the given rule.

Looking back

- **Special values first.** Putting $y = 0$ found $f(0)$ and $g(0)$.
- **From one point to everywhere.** The addition law moved the derivative at 0 to every x , turning a functional equation into a differential equation.
- **Uniqueness.** The argument does not just check that e^{kx} and $\tanh kx$ work; it shows they are the *only* possibilities. That is what makes a functional equation a definition.

What's really going on

Velocities in relativity

In everyday life, if a train moves at speed u and you walk forwards inside it at speed v , your speed relative to the ground is $u + v$. Einstein showed in 1905 that this is only an approximation. The correct rule, with c the speed of light, is

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

Measuring speeds as fractions of c , this is exactly the rule for g . Our result says that if we write each velocity as $v = \tanh \phi$, then velocities combine by *adding* the ϕ s. The quantity $\phi = \operatorname{artanh} v$ is called the *rapidity*. Each boost from a rocket engine adds a fixed amount of rapidity, but since $\tanh \phi < 1$ for every ϕ , the velocity creeps closer and closer to the speed of light without ever reaching it (Figure 3.8.1). Combining $0.9c$ with $0.9c$ gives not $1.8c$ but $\frac{1.8}{1.81}c \approx 0.9945c$.

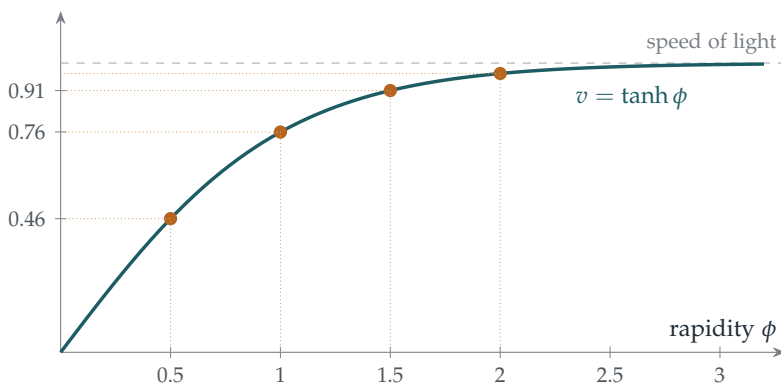


Figure 3.8.1 Velocity (as a fraction of the speed of light) against rapidity. Equal steps in rapidity, which is what equal boosts from an engine provide, give smaller and smaller increases in velocity: 0.46, 0.76, 0.91, 0.96, $\dots$, never reaching 1.

Cauchy's functional equation

The simplest functional equation is $f(x + y) = f(x) + f(y)$, studied by Cauchy in 1821. If f is differentiable, the method of this question shows $f(x) = kx$. In fact continuity is enough. But without any regularity assumption at all, there are other, utterly wild solutions, whose graphs are scattered densely over the whole plane. Georg Hamel constructed them in 1905; the construction needs the *axiom of choice*, and it is impossible to write any of them down explicitly. This is one reason the question insists that f is differentiable: the hypothesis is not a formality.

Homomorphisms

$f(x + y) = f(x)f(y)$ says that f turns the addition of real numbers into the multiplication of positive real numbers, respecting the structure of both. In the language of first-year algebra, f is a *group homomorphism* from $(\mathbb{R}, +)$ to $(\mathbb{R}_{>0}, \times)$, and the logarithm is its inverse. Many of the most important functions in mathematics are characterised in this way, by the structure they preserve. In Section 7.6 we meet the geometric meaning of $\tanh$: it measures “hyperbolic angle”, and its addition law comes from composing hyperbolic rotations.

The thinking skill

Define a function by its properties

Instead of a formula, describe a function by a rule it obeys, and prove that only one function obeys it.

This is how mathematicians often think about $\exp$, $\sin$, $\cos$ and many other functions: not as formulas, but as the unique solutions of functional or differential equations. It explains *why* they matter, not just how to calculate them.

Try this next

- 1 **A tangent rule.** Suppose g is differentiable, $g'(0) = k \neq 0$, and $g(x+y) = \frac{g(x) + g(y)}{1 - g(x)g(y)}$ whenever both sides are defined. Show that $g'(x) = k(1 + g(x)^2)$ and deduce that $g(x) = \tan kx$ near $x = 0$.
- 2 **Cauchy's equation.** If f is differentiable and $f(x+y) = f(x) + f(y)$ for all x, y , show that $f(x) = kx$.
- 3 **Without derivatives.** HARDER Suppose only that $f(x+y) = f(x)f(y)$ and $f(1) = a > 0$. Show that $f(n) = a^n$ for positive integers n , then that $f(1/n) = a^{1/n}$, and then that $f(q) = a^q$ for every positive rational q . (Why is continuity needed to go further?)
- 4 **Rocket science.** A rocket can increase its speed by $0.5c$ relative to its current frame each time it fires. Starting from rest, what is its speed (relative to the start) after 2, 3 and 4 firings? Use rapidities.

Further reading

- **Christopher Small, *Functional Equations and How to Solve Them*** (Springer). An accessible introduction aimed at problem-solvers.
- **Edwin Taylor & John Wheeler, *Spacetime Physics*** (W. H. Freeman). The classic introduction to special relativity, using rapidity throughout.
- **Tom Körner, *Calculus for the Ambitious*** (Cambridge University Press). Defines the exponential and trigonometric functions from their properties, exactly in the spirit of this section.

3.9 Inventing a New Function

The question

- (i) Sketch the curve $y = xe^x$, giving the coordinates of any stationary points.
- (ii) The function f is defined by $f(x) = xe^x$ for $x \geq a$, where a is the minimum possible value such that f has an inverse function. What is the value of a ?
Let g be the inverse of f . Sketch the curve $y = g(x)$.
- (iii) For each of the following equations, find a real root in terms of a value of the function g , or demonstrate that the equation has no real root. If the equation has two real roots, determine whether the root you have found is greater than or less than the other root.
 - (a) $e^{-x} = 5x$ (b) $2x \ln x + 1 = 0$ (c) $3x \ln x + 1 = 0$ (d) $x = 3 \ln x$
- (iv) Given that the equation $x^x = 10$ has a unique positive root, find this root in terms of a value of the function g .

The hook

Can you solve $e^{-x} = 5x$? Not with any function on your calculator. Rearranging gets nowhere: the x is stuck both inside and outside the exponential. You could find a numerical answer by trial and improvement, but there is no “formula”.

Or is there? This question takes a bold step: if a family of equations has no solution in terms of known functions, *invent a new function* that solves them. That is exactly what mathematicians did with $\sqrt{x}$ (the inverse of x^2), with $\ln x$ (the inverse of e^x) and with $\arcsin x$. The new function here is the inverse of xe^x , now called the *Lambert W function*. Once you have it, a surprising range of equations, from $x^x = 10$ to a law of physics, can be solved exactly.

Before you start



Ask yourself What does it actually mean to “solve” an equation? Is $x = \sqrt{2}$ a solution of $x^2 = 2$, or just a name for the solution?

Both. $\sqrt{2}$ is a name for “the positive number whose square is 2”; we accept it as an answer because the function $\sqrt{}$ is well understood, easy to compute and has known

properties. Whether an answer counts as a “formula” depends on which functions we have agreed to name. This question adds one more to the list.

Working it through

STAGE 1 The curve $y = xe^x$

$\frac{dy}{dx} = e^x + xe^x = (1+x)e^x$, which is zero only at $x = -1$, where $y = -e^{-1}$. The derivative is negative for $x < -1$ and positive for $x > -1$, so this is a minimum.



Ask yourself What happens to xe^x as $x \rightarrow +\infty$ and as $x \rightarrow -\infty$? Where does the curve cross the axes?

As $x \rightarrow \infty$, $xe^x \rightarrow \infty$. As $x \rightarrow -\infty$, $xe^x \rightarrow 0$ from below, because the exponential decay beats the linear growth. The curve passes through the origin. So it comes in from just below the negative x -axis, dips to its minimum $(-1, -\frac{1}{e})$, then rises through $(0, 0)$ and shoots upwards (Figure 3.9.1).

STAGE 2 The inverse

To have an inverse, f must be one-to-one, so it must be restricted to a region where it is strictly increasing or strictly decreasing. The largest such interval containing the right-hand part is $[-1, \infty)$, so $a = -1$. On $[-1, \infty)$, f increases from $-\frac{1}{e}$ to ∞ , so the inverse g is defined on $[-\frac{1}{e}, \infty)$, increases from $g(-\frac{1}{e}) = -1$, passes through $g(0) = 0$, and grows slowly (for large x , even more slowly than $\ln x$). Its graph is the reflection of the right-hand part of $y = xe^x$ in the line $y = x$, with a vertical tangent at $(-\frac{1}{e}, -1)$.

The defining property of g is

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

So the strategy for every part of (iii) is the same: rearrange the equation into the form (something) $e^{(\text{the same something})} = (\text{a number})$.

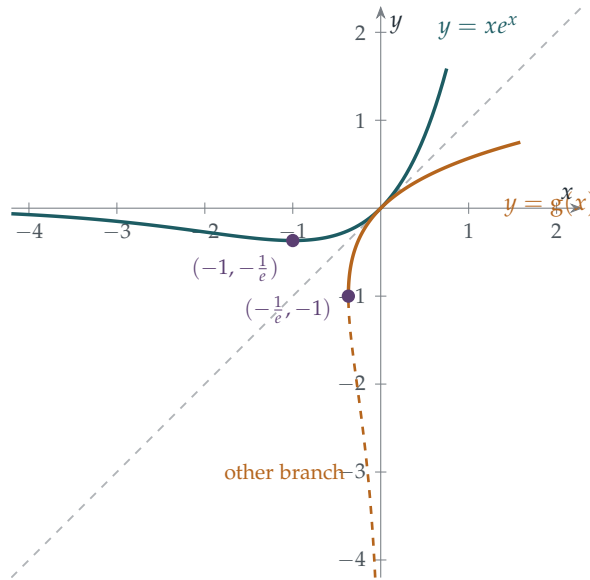


Figure 3.9.1 The curve $y = xe^x$ (teal) has a minimum at $(-1, -\frac{1}{e})$. Restricting to $x \geq -1$ and reflecting in $y = x$ gives the inverse g (amber, solid), the Lambert W function. The dashed amber curve is the reflection of the left-hand part of $y = xe^x$: a second branch, needed for the second roots in part (iii).

STAGE 3 Four equations

(a) $e^{-x} = 5x$. Multiply by e^x and divide by 5: $xe^x = \frac{1}{5}$. So $x = g(\frac{1}{5}) \approx 0.169$. Since xe^x is negative for $x < 0$, any root is positive, and xe^x is increasing for $x > 0$, so this is the only root.



Ask yourself For (b) and (c), the unknown appears as $x \ln x$. What substitution turns $x \ln x$ into something of the form ue^u ?

Put $x = e^u$, so $\ln x = u$ and $x \ln x = ue^u$.

(b) $2x \ln x + 1 = 0$ becomes $ue^u = -\frac{1}{2}$. But $ue^u \geq -\frac{1}{e} \approx -0.368$ for all real u , and $-\frac{1}{2} < -\frac{1}{e}$. So there is **no real root**.

(c) $3x \ln x + 1 = 0$ becomes $ue^u = -\frac{1}{3}$, and $-\frac{1}{e} < -\frac{1}{3} < 0$. A horizontal line at height $-\frac{1}{3}$ cuts $y = ue^u$ twice, once on each side of the minimum at $u = -1$. The root on the right is $u = g(-\frac{1}{3})$, giving

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

The other root has $u < -1 < g(-\frac{1}{3})$, and $x = e^u$ is increasing in u , so the root we have found is the **greater** of the two. (The other is about 0.220.)



Pause and try Now try (d), $x = 3 \ln x$. (Hint: exponentiate and aim for (something) $e^{\text{something}}$ with the something being a multiple of x .)

(d) $x = 3 \ln x$ means $e^{x/3} = x$, so $xe^{-x/3} = 1$, and multiplying by $-\frac{1}{3}$,

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

So $-\frac{x}{3}$ is one of the two solutions of $ue^u = -\frac{1}{3}$ found in (c). Taking the one given by g , $x = -3g(-\frac{1}{3}) \approx 1.857$. The other root comes from the other solution $u < g(-\frac{1}{3})$, which gives a larger $x = -3u$. So this time the root we have found is the **smaller** of the two. (The other is about 4.536.) ■

STAGE 4 $x^x = 10$

Taking logarithms, $x \ln x = \ln 10$. With $x = e^u$ this is $ue^u = \ln 10$, so $u = g(\ln 10)$ and

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

Since $g(z)e^{g(z)} = z$, we can also write this as $x = \frac{\ln 10}{g(\ln 10)}$. ■

Looking back

- **Name the missing function.** When an equation cannot be solved with known functions, defining a new function as an inverse is a legitimate, and often illuminating, move.
- **Rearrange to a standard form.** Every part of (iii) and (iv) came down to $ue^u = \text{constant}$, using a substitution such as $x = e^u$ or $u = -\frac{x}{3}$.
- **Count roots with the graph.** Whether $ue^u = c$ has zero, one or two solutions depends only on where c sits relative to the minimum value $-\frac{1}{e}$.

What's really going on

A function with a long history and a recent name

Johann Heinrich Lambert studied equations of this type in 1758, and Euler took them further in 1783. But the function had no standard name until the 1990s, when the developers of the computer algebra system Maple needed one. In 1996 Robert Corless, Gaston Gonnet, David Hare, David Jeffrey and Donald Knuth published a paper, "On the Lambert W function", that collected its properties and applications. It has been cited thousands of times, and W is now built into most mathematical software. The function g in the question is the *principal branch* W_0 ; the dashed curve in Figure 3.9.1, which gave the second roots in (c) and (d), is the other real branch W_{-1} .

Where W appears

- **Wien's displacement law.** The wavelength at which a hot body radiates most strongly is inversely proportional to its temperature. The constant comes from solving

$5(1 - e^{-x}) = x$, whose positive solution is $x = 5 + W(-5e^{-5}) \approx 4.965$. That is why the Sun, at about 5800 K, looks yellow-white, and why a cooler red dwarf looks red.

- **Power towers.** The infinite tower $x^{x^{x^{\cdots}}}$ converges for $e^{-e} \leq x \leq e^{1/e}$, and its value is $\frac{W(-\ln x)}{-\ln x}$.
- **Population and epidemic models.** The final size of an epidemic in the simplest model, and the time to reach a given population in some growth models, are both given by W .

Elementary functions

Is W really new, or could it be written using $\exp$, $\ln$, roots and the trigonometric functions? It is genuinely new: a theorem going back to Joseph Liouville in the 1830s shows that W is not an *elementary function*. The same theory shows that some innocent-looking integrals, such as $\int e^{-x^2} dx$, cannot be done in elementary terms, which is why the normal distribution needs its own special function.

The thinking skill

Inverse functions as solutions

An equation $f(x) = c$ is solved by $x = f^{-1}(c)$, provided f is restricted to make it one-to-one. The restriction determines which solution you get.

Many equations that look different are the same equation in disguise; finding the substitution that reveals the common form is the key step.

Try this next

- 1 **Two, four, and one more.** The equation $x^2 = 2^x$ has the obvious solutions $x = 2$ and $x = 4$ (compare Section 3.5). Show that it also has a negative solution, and that this solution is

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

- 2 **A derivative.** By differentiating $g(x)e^{g(x)} = x$ implicitly, show that, for $x \neq 0$,

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

What happens at $x = -\frac{1}{e}$?

- 3 **An integral.** Use the substitution $x = we^w$ to show that $\int_0^e g(x) dx = e - 1$.

- 4 The tower.** Assuming the tower $T = \sqrt{2}^{\sqrt{2}^{\sqrt{2}^{\cdots}}}$ converges, show that $T = \sqrt{2}^T$, and deduce that $T = 2$ or $T = 4$. Which is it? (Compute a few terms of the sequence $t_1 = \sqrt{2}$, $t_{n+1} = \sqrt{2}^{t_n}$, and think about the cobweb diagrams of Section 2.2.)

Further reading

- **Robert Corless et al.**, “On the Lambert W function”, *Advances in Computational Mathematics* 5 (1996). The paper that named the function. The first sections are readable with this section behind you.
- **Brian Hayes**, “Why W?”, *American Scientist* 93 (2005). A lively essay on the function’s history and uses. Free online.
- **Brian Conrad**, “Impossibility theorems for elementary integration”. Lecture notes, available online, explaining Liouville’s theory. Advanced.

What We Have Learned

We began with a “proof” that $1 = 2$ and three questions: how to know an argument is correct, how to prove inequalities, and what a function really is. Here is what the chapter’s nine questions have taught us.

1. How do you know an argument is correct?

Section 3.1 gave us a checklist. At every step, ask three things: *is the calculation right? does the step follow logically? what has been assumed without being said?* The errors there were typical. Comparing the values at two stationary points says nothing about which is a maximum. Stationary values do not bound a function on an infinite domain. The fundamental theorem of calculus needs a continuous integrand. And an argument can arrive at a “correct-looking” answer by accident.

The same vigilance ran through the rest of the chapter. Section 3.7 showed by counterexample that dropping a hypothesis ($f(0) = f(\pi) = 0$) breaks an inequality. In Section 3.8, the assumption that the function is differentiable was essential: without it, Cauchy’s functional equation has wild solutions that nobody can write down. And in Section 3.6 and Section 3.3, the equality conditions told us whether a bound was best possible, which is the kind of question a careful mathematician always asks.

2. How do you prove an inequality?

We met the classical inequalities from several directions, and saw how they are related.

- ▶ **Squares are non-negative.** The two-number AM–GM is just $(a - b)^2 \geq 0$. Cauchy’s forward–backward induction built the general case from it (Section 3.2).
- ▶ **Convexity.** A concave curve lies above its chords. That one fact gives Jensen’s inequality, and with the logarithm it gives AM–GM in a line; with $\sin x$ it gives a sharp inequality for triangles (Section 3.3).
- ▶ **Areas.** The area under a curve plus the area beside it covers a rectangle. That gives Young’s inequality, which leads to Hölder and Cauchy–Schwarz (Section 3.4).
- ▶ **Shape of a graph.** Rewriting a^b versus b^a as a comparison of $\frac{\ln x}{x}$ at two points settled π^e versus e^π (Section 3.5).
- ▶ **Interpolation and the triangle inequality.** Writing the derivative in terms of values, then bounding term by term, gave Markov’s inequality for quadratics (Section 3.6).
- ▶ **Test functions.** Applying a general inequality to a cleverly chosen function gave $\pi^2 \leq 10$ and $\pi \leq \frac{22}{7}$ (Section 3.7).

Choosing among these is a skill that comes with practice. A rough guide: products and sums suggest AM–GM; averages of function values suggest Jensen; a function and its inverse suggest Young; a comparison of two expressions of the same form suggests finding the single function behind them.

3. What is a function, really?

The last two sections answered this in two ways. In Section 3.8, e^{kx} and $\tanh kx$ were not given by formulas at all; they were the *only* differentiable functions obeying certain addition laws. That is a deeper kind of understanding than a formula: it explains why these functions keep appearing, in compound interest and in Einstein’s velocity addition. In Section 3.9 we went further and *invented* a function, the inverse of xe^x , because the functions we had were not enough to solve a natural family of equations. Every function on your calculator was once new in the same way.

4. A toolkit for rigour and inequalities

Method	Where you used it
Checking every step and every hypothesis	Plausible wrong arguments (3.1)
Local versus global behaviour	Stationary points and limits at infinity (3.1)
Forward–backward induction	AM–GM (3.2)
Padding with extra terms	AM–GM (3.2, 3.3)
Jensen’s inequality	Triangles; AM–GM (3.3)
Comparing areas	Young’s inequality (3.4)
Separating variables into one function	π^e versus e^π (3.5)
Triangle inequality, equality cases	Markov’s inequality (3.6)
Test functions	$\pi^2 \leq 10$ (3.7)
Functional equations	e^x and $\tanh x$ (3.8)
Inverse functions and substitutions	The W function (3.9)

Where next

- The calculus of variations, glimpsed in Section 3.7, is the subject of Section 4.7.
- The geometry behind $\tanh$ and its addition law appears in Section 7.6.
- Chapter 4 takes the careful attitude of this chapter to integration itself, starting from what an integral actually is.

If you want to go further before university

- **J. Michael Steele, *The Cauchy–Schwarz Master Class*:** the best introduction to inequalities, built entirely from problems.

- ▶ **Daniel Velleman, *How to Prove It*:** a systematic guide to writing correct proofs.
- ▶ **Michael Spivak, *Calculus*:** rigorous calculus from the ground up; many Cambridge students read it the summer before they arrive.
- ▶ **E. A. Maxwell, *Fallacies in Mathematics*:** for practice in spotting what has gone wrong.

Von Neumann joked that in mathematics you don't understand things, you just get used to them. There is some truth in it: the precise definitions and careful arguments of this chapter can feel strange at first. But getting used to them is exactly what turns a student who can calculate into a mathematician who can be sure. The proof that $1 = 2$ will never fool you again; the aim of a mathematics degree is to make every subtler error just as easy to see.