DAWN PUTNAM SESSION

Putnam 1977 — Papers A and B

Putnam 120 Archive

Contents

Putnam 1977 — Paper A **5**
1977 A1 5
1977 A2 7
1977 A3 10
1977 A4 13
1977 A5 16
1977 A6 19

Putnam 1977 — Paper B **23**
1977 B1 23
1977 B2 25
1977 B3 27
1977 B4 30
1977 B5 33
1977 B6 36

Putnam 1977 — Paper A

1977 A1

1977 A1. Consider all lines which meet the graph of

$$y = 2x^4 + 7x^3 + 3x - 5$$

in four distinct points, say (x_i, y_i), $i = 1, 2, 3, 4$. Show that

$$x_1 + x_2 + x_3 + x_4$$

is independent of the line and find its value.

Fields: algebra (Vieta's formulas for polynomial roots) **Position:** A1 — the opener **Answer:** $x_1 + x_2 + x_3 + x_4 = -\frac{7}{2}$ for every line meeting the quartic in four points.

Layer 1 THE WHAT — surface analysis

The natural attempt. A line meeting the quartic in four points is not obviously "one object"; it is a two-parameter family, $y = mx + b$, and the four intersection points depend on both m and b. The honest first reaction is to distrust that any fixed number could come out of a two-parameter family without solving for the roots explicitly. So you write the intersection condition,

$$2x^4 + 7x^3 + 3x - 5 = mx + b,$$

and, wanting a concrete handle on $x_1, \ldots, x_4$, you try small, checkable cases: pick a horizontal line ($m = 0$) and a specific b, or try to locate a line by symmetry (e.g. guess that the "nicest" secant line passes through an inflection point of the quartic, since that is the reflex move for cubics-inside-quartics problems). You might even start differentiating, looking for a point where the curve has a useful symmetry, on the theory that the invariant sum must come from some geometric centre of the curve.

This is not wasted motion — it correctly senses that the answer should not depend on m, b — but it is expensive: computing an inflection point or a specific numerical example only ever pins down *one* line's worth of data, and turning that into a proof for *every* line requires re-running the argument in general anyway. If you press on by brute force, substituting a general m, b and trying to extract $x_1 + x_2 + x_3 + x_4$ from the resulting quartic's roots by radicals, you hit a wall immediately: quartics do not hand you their roots' individual values without real work, and you do not need the individual values at all — only a symmetric function of them.

The model chain. Rewrite the intersection condition as a single quartic in x:

$$2x^4 + 7x^3 + (3 - m)x - (5 + b) = 0.$$

The four intersection abscissae x_1, x_2, x_3, x_4 are exactly the roots of this quartic (four distinct real points is given, so there are four distinct roots). By Vieta's formulas, for a quartic $a_4x^4 + a_3x^3 +$

$a_2x^2 + a_1x + a_0 = 0$ with roots $x_1, \ldots, x_4$,

$$x_1 + x_2 + x_3 + x_4 = -\frac{a_3}{a_4}.$$

Here $a_4 = 2$ and $a_3 = 7$ — and crucially, *neither of these coefficients involves m or b*: the line only ever modifies the coefficient of x^1 (by $-m$) and the constant term (by $-b$). So

$$x_1 + x_2 + x_3 + x_4 = -\frac{7}{2}$$

for every choice of m, b that produces four distinct real intersection points. The sum is not merely computable for each line; it is structurally forced to be the same number, because the two coefficients it depends on are untouched by the family of lines.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths begin by writing down $2x^4 + 7x^3 + (3 - m)x - (5 + b) = 0$, or something equivalent to it. They split at the very next question asked of that equation. The natural attempt asks "what are the roots?" The model asks "what do I actually need from the roots, and which coefficient controls it?" That is the line at which one path heads toward solving a quartic and the other heads toward Vieta's formulas.

> *I don't need the four roots themselves, only their sum — and a sum of roots is nothing but a ratio of two coefficients, by Vieta's formula. I should ask which coefficients that ratio uses, before I try to solve anything.*

> *The line $y = mx + b$ only ever contributes to the linear and constant terms when I move it to the other side of the equation; it cannot touch the x^4 or x^3 coefficients of the quartic curve. So whatever symmetric function of the roots depends only on those two coefficients is automatically independent of the line — I should check the* coefficient of the answer*, not test more lines.*

Layer 3 THE WHY

Why is it Vieta's formula, of all the tools available for "a fixed line meets a fixed curve in several points", that is the natural one here, and why does it need no further justification once invoked? A polynomial's coefficients *are* (up to sign and normalisation) the elementary symmetric functions of its roots: $a_3/a_4 = -(x_1 + x_2 + x_3 + x_4)$ is not an approximation or a special property of this quartic, it is a restatement of what "coefficient of x^3" means once you know the polynomial factors as $a_4 \prod(x - x_i)$. So asking for $x_1 + x_2 + x_3 + x_4$ is, by definition, asking for a specific coefficient ratio of the polynomial whose roots are the x_i — there is no step of "insight" here, only correctly identifying which object the problem is secretly asking about.

That reframing is also exactly what explains *why* the answer is line-independent, rather than leaving it as an observed coincidence. The quartic $2x^4 + 7x^3 + 3x - 5$ has zero coefficient of x^2 already; that fact plays no role in this particular problem (the answer would still be line-independent even if it were nonzero), but what matters is that subtracting a line $mx + b$ is an operation that can only ever alter the x^1 and x^0 coefficients of a polynomial — it literally has no x^4, x^3, or x^2 term to contribute. So any symmetric function of the roots that is built purely from a_4 and a_3 is immune to the line by the algebraic structure of "subtracting a degree-≤ 1 polynomial", before any computation is done. The product $x_1x_2x_3x_4 = -a_0/a_4$, by contrast, *does* depend on b and is not invariant — which is the check that separates "this happens to be invariant" from "I understand which coefficients the line can reach."

The deeper generalisable fact: whenever a problem says some quantity built from the intersections of a fixed curve with a moving line (or moving member of any family that only

shifts low-degree terms) is independent of the choice, look at which coefficients of the resulting polynomial that quantity is a ratio of. If the family cannot alter those coefficients, invariance is automatic and requires no case-checking at all.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is "a line meets a quartic in four points, sum the abscissae." On seeing it, write down immediately: "four roots of a quartic, sum of roots $= -a_3/a_4$ by Vieta — I need the quartic's top two coefficients, nothing else."

2. The cue suggests substituting the line into the curve to get a single polynomial equation in x, then reading off a coefficient ratio via Vieta's formulas, rather than attempting to solve for the roots.

3. Hesitation occurs at the moment you see a two-parameter family (m and b) and instinctively distrust that anything fixed can come out of it. That is reinforced away by checking, one coefficient at a time, which coefficients of the combined equation the parameters m, b can actually reach — here, only the linear and constant terms, never the x^4 or x^3 terms.

4. The trap is treating "show a quantity is independent of a parameter" as a request to verify examples or find a geometric symmetry point, rather than as a request to identify a coefficient the parameter cannot touch. Testing examples never proves independence; identifying the untouched coefficient does, in one line.

5. Whenever the intersections of a fixed curve with a moving line are the subject, immediately rewrite "curve = line" as a single polynomial equation and ask which of the elementary symmetric functions of its roots (sum, sum of products of pairs, ..., product) are ratios of coefficients the line cannot alter.

6. Reusable rule: a symmetric function of the roots of (curve) $-$ (moving low-degree family) $= 0$ is automatically independent of the family whenever it is a ratio of coefficients of degree higher than the family's degree; check this coefficient-by-coefficient before attempting to compute or test cases.

▷ **Cue:** "show that [a symmetric function of intersection points of a curve and a line/low-degree family] is independent of the line and find its value" $\longrightarrow$ **Move:** combine into one polynomial equation, identify the symmetric function as a Vieta ratio, and check whether that ratio's coefficients are ones the line/family can modify

▷ **Cue:** a quartic (or degree-n polynomial) with roots $x_1, \ldots, x_n$ and a request for $\sum x_i$ $\longrightarrow$ **Move:** read off $-a_{n-1}/a_n$ directly from the coefficients; never attempt to solve for individual roots when only a symmetric function is requested

1977 A2

1977 A2. Determine all solutions in real numbers x, y, z, w of the system

$$x + y + z = w, \qquad \frac{1}{x} + \frac{1}{y} + \frac{1}{z} = \frac{1}{w}.$$

Caveat. The 1965–1984 volume's OCR of this problem statement is corrupted: it breaks off after "$x + y + z - w$," with the rest of the system lost to a page-image error. The system above is reconstructed from the official solution, which explicitly derives $w - z = x + y$ (forcing the first equation to be $x + y + z = w$) and manipulates $\frac{1}{x} + \frac{1}{y} + \frac{1}{z} - \frac{1}{w}$ (forcing the second equation to be $\frac{1}{x} + \frac{1}{y} + \frac{1}{z} = \frac{1}{w}$). I have independently solved this reconstructed system below and confirmed it reproduces the official solution's case structure exactly, which is strong evidence the reconstruction is the correct original problem.

Fields: algebra (symmetric functions, factoring) **Position:** A2 — second problem, first half **Answer:** All real solutions with $x, y, z, w \neq 0$: one of x, y, z equals w, and the other two are negatives of each other. Equivalently $(x+y)(y+z)(z+x) = 0$ together with $w = x+y+z$.

Layer 1 THE WHAT — surface analysis

The natural attempt. The system has two equations and four unknowns, so you expect a one-parameter (or larger) family of solutions, and the honest first move is to eliminate w: substitute $w = x+y+z$ into the second equation,

$$\frac{1}{x}+\frac{1}{y}+\frac{1}{z} = \frac{1}{x+y+z}.$$

This is now one equation in three unknowns. The reflex is to clear every denominator at once: multiply through by $xyz(x+y+z)$,

$$yz(x+y+z) + xz(x+y+z) + xy(x+y+z) - xyz = 0,$$

and start expanding. This is entirely correct mathematics, but it is expensive and disorienting: expanding $(x+y+z)(xy+yz+zx)$ term by term produces nine monomials, several of which are xyz repeated three times, and it is easy to lose track of which terms survive, which cancel against the subtracted $-xyz$, and whether the remaining six-or-so terms group into anything recognisable. Absent a plan for *how* to group them, the natural move is to just keep expanding and hope a pattern falls out — which, done carelessly under time pressure, is exactly the kind of undirected brute-force manipulation that burns the clock without guaranteeing the factorisation gets found. Even when it succeeds, it succeeds by luck of noticing the grouping, not by a reason to expect one.

The model chain. Two routes both work cleanly once a target is chosen *before* expanding.

Route 1 (symmetric-function substitution). Let $s = x+y$ and $p = xy$. The first equation gives $w - z = s$. Combine the fractions in the second equation as

$$\frac{1}{x}+\frac{1}{y} = \frac{x+y}{xy} = \frac{s}{p}, \qquad \frac{1}{w}-\frac{1}{z} = \frac{z-w}{wz} = \frac{-s}{wz},$$

so $\frac{1}{x}+\frac{1}{y} = \frac{1}{w}-\frac{1}{z}$ becomes $\frac{s}{p} = \frac{-s}{wz}$, i.e. $s\left(\frac{1}{p}+\frac{1}{wz}\right) = 0$, i.e. (for $p, w, z \neq 0$) $s(wz+p) = 0$.

Case $s = 0$: $x+y = 0$, so $y = -x$; then $w = s+z = z$.

Case $wz = -p = -xy$: since $w = s+z$, this reads $(s+z)z = -xy$, i.e. $sz+z^2+xy = 0$, i.e. $(x+y)z+z^2+xy = xz+yz+z^2+xy = x(z+y)+z(y+z) = (y+z)(x+z) = 0$. So $y = -z$ (giving $w = x+y+z = x$) or $x = -z$ (giving $w = y$).

Route 2 (the identity, found by aiming at a product before expanding). Before expanding $(x+y+z)(xy+yz+zx) - xyz$ term by term, ask what shape the answer should have: the problem is asking for *all* solutions, so a natural guess is that the constraint factors into a union of simple cases. The standard three-variable identity

$$(x+y+z)(xy+yz+zx) - xyz = (x+y)(y+z)(z+x)$$

turns the cleared equation directly into $(x+y)(y+z)(z+x) = 0$: two of x, y, z are negatives of each other, and by $w = x+y+z$, w collapses to the third.

Both routes give the same answer: w equals one of x, y, z, and the other two of x, y, z are negatives of each other (with the pair of them and w each nonzero).

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths substitute $w = x+y+z$ into the second equation and arrive at the single cleared equation

$$(x+y+z)(xy+yz+zx) - xyz = 0.$$

They split exactly here. The natural attempt expands this expression monomial by monomial with no target in mind. The model asks, before expanding a single term, "what should the answer look like, and is there a grouping (or a known identity) that produces it directly?"

> *The problem asks for* all *solutions of a system with more unknowns than equations, so the solution set should be a union of simpler cases, not a single formula — which means the cleared polynomial equation in x, y, z should factor into a product of simple pieces. I should look for that factorisation, or build toward it with a substitution like $s = x + y$, $p = xy$, rather than expand blindly and hope it appears.*

> *$(x + y + z)(xy + yz + zx) - xyz$ is a symmetric-looking cubic in three variables built from the same elementary symmetric functions that show up in $(x + y)(y + z)(z + x)$ when it is expanded — these are the two standard cubic symmetric expressions in three variables, and one should already be recognisable as the expansion of the other, the way $(a + b)^2$ is recognisable as $a^2 + 2ab + b^2$ without re-deriving it.*

Layer 3 THE WHY

Why does aiming at a product, rather than expanding term by term, actually save the work, and why is that the correct instinct here specifically? The problem's own phrasing — "determine all solutions" of a system that is under-determined by one degree of freedom (two equations, four unknowns) — is the cue. An under-determined system's solution set is generically a curve or surface, but when a problem expects you to *describe* that set completely and explicitly, the set is almost always a union of finitely many simply-parametrised pieces (lines, planes, or in this case, linear relations among the variables) rather than one intractable curve. A union of simple pieces is exactly what a factored polynomial equation $A \cdot B \cdot C = 0$ describes: each factor being zero is one piece. So the phrasing itself licenses expecting a factorisation, and that expectation is what should be tested *before* spending the labour of a full expansion.

The deeper reason the identity $(x + y + z)(xy + yz + zx) - xyz = (x + y)(y + z)(z + x)$ exists at all, rather than being a coincidence to memorise, is worth stating: $(x + y + z)$ and $(xy + yz + zx)$ are the first two elementary symmetric polynomials e_1, e_2 of x, y, z, and $xyz = e_3$. The expression $e_1e_2 - e_3$ is itself symmetric, and any symmetric polynomial that vanishes whenever two of x, y, z sum to zero (which $e_1e_2 - e_3$ does — set $y = -x$ and check $e_1e_2 - e_3 = z \cdot (-x^2) - (-x^2z) = 0$) must be divisible by each of $(x + y), (y + z), (z + x)$ by the factor theorem, and by a degree count ($e_1e_2 - e_3$ has degree 3, matching $(x + y)(y + z)(z + x)$ exactly) that accounts for the entire expression with no remaining factor. This is the general reason to suspect and then confirm such identities: a symmetric cubic vanishing on the loci $x = -y$, $y = -z$, $z = -x$ has no room left to be anything other than their product.

The route-1 substitution ($s = x + y$, $p = xy$) reaches the same place from a different angle, and is worth keeping as the default move whenever a problem's structure is symmetric in two of several variables and the third and fourth variables (z, w) enter asymmetrically: reducing (x, y) to their sum and product is always legitimate (Vieta again, as in the previous problem) and often turns an equation that looks like it needs three-variable case analysis into one that needs only single-variable factoring in s.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is "two equations, four unknowns, determine all real solutions." On seeing it, write down: "the solution set will be a union of simply-described pieces (not one intractable curve), so expect the cleared equation to factor — do not expand blindly."
2. The cue suggests either (a) substituting $s = x + y$, $p = xy$ to exploit the partial symmetry between x and y, or (b) recognising the cleared cubic as a known symmetric-function identity, $e_1e_2 - e_3 = (x + y)(y + z)(z + x)$.

3. Hesitation occurs right after clearing denominators, staring at nine unexpanded monomials with no plan. It is reinforced away by asking, before expanding anything, what the answer should look like (a union of cases) and testing a simple substitution like $y = -x$ to see whether the expression vanishes — which flags the factor $(x + y)$ immediately.

4. The trap is treating "clear denominators, then simplify" as a single undirected step; without a target grouping, the expansion produces the right terms in a form where the cancellation and factorisation are easy to miss under time pressure.

5. Whenever a symmetric-looking cubic in three variables appears from clearing denominators, test it against $x = -y$ (and cyclic versions) before expanding term by term — a vanishing result under such a substitution is an immediate certificate of a linear factor by the factor theorem.

6. Reusable rule: for an under-determined system asked to be solved completely, expect the cleared polynomial equation to factor into a union of simple linear (or low-degree) conditions; test small symmetric substitutions ($a = -b$, or fixing one variable) to locate the factors before expanding, and reduce partially-symmetric pairs of variables to their sum and product.

▷ **Cue:** "determine all real solutions" of an under-determined polynomial/rational system ⟶ **Move:** expect the cleared equation to factor into a union of simple pieces; test symmetric substitutions like $x = -y$ against the cleared expression to locate factors before expanding

▷ **Cue:** a symmetric cubic in three variables of the shape $e_1e_2 - e_3$ appearing after clearing denominators ⟶ **Move:** recognise it as $(x + y)(y + z)(z + x)$ (or verify via the factor theorem by checking it vanishes when two variables are negatives of each other) rather than expanding term by term

1977 A3

1977 A3. Let u, f, and g be functions, defined for all real numbers x, such that

$$u(x+1) + u(x-1) = 2f(x) \quad \text{and} \quad u(x+4) + u(x-4) = 2g(x).$$

Determine $u(x)$ in terms of f and g.

Caveat. The 1965–1984 volume's OCR renders the two given equations as "$u(x+1) + u(x-1) \cdot f(x)$" and "$u(x+4) + u(x-4)_()2\,2\,g(x)$," with the factors of 2 garbled or dropped and the equals signs corrupted. I have independently re-derived $u(x)$ below using the shift-operator identity $\frac{1}{2}(E^8+1) - \frac{1}{2}(E^6 - E^4 + E^2 - 1)(E^2+1) = 1$, and it reproduces the official solution's answer $u(x) = g(x+4) - f(x+7) + f(x+5) - f(x+3) + f(x+1)$ exactly only when the right-hand sides carry the factor 2, i.e. the equations are $u(x+1) + u(x-1) = 2f(x)$ and $u(x+4) + u(x-4) = 2g(x)$. That match is strong internal confirmation the reconstruction is correct; without the factor 2 the derivation below does not close.

Fields: algebra (functional equations via the shift operator), the Euclidean algorithm for polynomials
Position: A3 — third problem, first half **Answer:** $u(x) = g(x+4) - f(x+7) + f(x+5) - f(x+3) + f(x+1)$ (one of infinitely many valid expressions).

Layer 1 THE WHAT — surface analysis

The natural attempt. Two functional equations relate u at points four apart in one, two apart... no — points 2 apart in the first equation ($x-1$ to $x+1$) and points 8 apart in the second ($x-4$ to $x+4$). The natural first move is to try to build $u(x)$ as an explicit recursive chain: from equation

one, $u(x+2) = 2f(x+1) - u(x)$, so you can step u forward or backward in strides of 2 once you know it at one point. From equation two, $u(x+8) = 2g(x+4) - u(x)$, a similar stride of 8. The instinct is to try to line these up: walk forward 8 using four strides of 2 from equation one, getting an expression for $u(x+8)$ purely in terms of f's, then set it equal to $2g(x+4) - u(x)$ from equation two and solve for $u(x)$.

Carried out, this does work eventually, but it is easy to get lost in bookkeeping: four successive applications of $u(x+2) = 2f(x+1) - u(x)$ produce an alternating telescoping sum with a coefficient of $u(x)$ that must be tracked correctly (does it come back as $+u(x)$ or $-u(x)$ after four steps?), and reconciling that against equation two's single stride of 8 requires the same kind of alternating-sign care in the other direction. Under time pressure, the risk is not that this fails — it is that it is a manual induction with several opportunities for a sign error, done without any structural reason to expect it will close up into a clean answer, only the hope that pushing through the algebra will eventually match the two expressions for $u(x+8)$.

The model chain. Treat "shift by 1" as an operator E on functions, $Eu(x) = u(x+1)$, and note that shifts commute and add in the exponent: $E^a E^b = E^{a+b}$, exactly like powers of a variable. This turns both hypotheses into statements about polynomials in E applied to u:

$$(E + E^{-1})u(x) = 2f(x), \qquad (E^4 + E^{-4})u(x) = 2g(x).$$

Multiplying the first by E and the second by E^4 clears the negative powers:

$$(E^2+1)u(x) = 2Ef(x) = 2f(x+1), \qquad (E^8+1)u(x) = 2E^4 g(x) = 2g(x+4).$$

Now the question is purely algebraic: can 1 be written as a combination of the polynomials E^2+1 and E^8+1? Writing $a = E^2$, this is asking whether $a+1$ and a^4+1 are coprime as polynomials in a. Polynomial division gives

$$a^4 + 1 = (a+1)(a^3 - a^2 + a - 1) + 2,$$

so $a+1$ and a^4+1 have remainder 2 (a nonzero constant), hence are coprime, and

$$1 = \tfrac{1}{2}(a^4+1) - \tfrac{1}{2}(a^3 - a^2 + a - 1)(a+1).$$

Substituting back $a = E^2$ and applying both sides to $u(x)$:

$$u(x) = \tfrac{1}{2}(E^8+1)u(x) - \tfrac{1}{2}(E^6 - E^4 + E^2 - 1)(E^2+1)u(x).$$

Both terms are now known: $(E^8+1)u(x) = 2E^4 g(x)$ and $(E^2+1)u(x) = 2Ef(x)$, so

$$u(x) = E^4 g(x) - (E^6 - E^4 + E^2 - 1)\,Ef(x) = E^4 g(x) - (E^7 - E^5 + E^3 - E)f(x),$$

i.e.

$$u(x) = g(x+4) - f(x+7) + f(x+5) - f(x+3) + f(x+1).$$

Direct substitution into both original equations confirms this works for arbitrary f, g.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths correctly rewrite the two hypotheses as recursions relating u at points a fixed distance apart. They split at the question of *how to combine* a stride-2 recursion with a stride-8 recursion. The natural attempt tries to physically walk one recursion into the other's step size, by hand, four times over. The model reframes both recursions as multiplication by a polynomial in the shift operator E, and asks whether 1 is a combination of those two polynomials — turning a bookkeeping problem into a single application of the Euclidean algorithm.

> *Shifting by* 1 *is an operation I can name and manipulate like a variable — call it* E *— because shifts commute and compose by adding exponents, exactly like powers. Both of my hypotheses are then just "(a polynomial in* E*) applied to* u *equals a known function," and I have two such polynomial equations. I don't need to walk one recursion into the other's step size by hand; I need to know whether* 1 *is expressible as a combination of the two polynomials* E^2+1 *and* E^8+1 *— which is exactly the question the Euclidean algorithm for polynomials answers.*

> E^2+1 *and* E^8+1 *are coprime precisely because dividing* E^8+1 *by* E^2+1 *(as polynomials in* E^2*) leaves a nonzero constant remainder,* 2 *– and a nonzero constant remainder is exactly the signal that the Euclidean algorithm terminates in a unit, giving me an explicit Bézout identity* $1 = A(E)(E^2+1) + B(E)(E^8+1)$ *that I can apply directly to* $u(x)$*.*

Layer 3 THE WHY

Why is it legitimate — not just convenient — to treat the shift E as an algebraic variable and manipulate polynomials in it? Because translation operators on functions of a real variable commute with each other and with themselves ($E^aE^b(u)(x) = u(x+a+b) = E^bE^a(u)(x)$), and the set of finite linear combinations of shifts, under composition, forms a commutative ring with exactly the same multiplication rule as the polynomial ring $\mathbb{R}[E]$: $E^a \cdot E^b = E^{a+b}$. Any identity of polynomials in one variable — in particular, the Euclidean algorithm and the resulting Bézout identity $1 = AP + BQ$ for coprime P, Q — is therefore an identity of operators, and applying both sides of that operator identity to the same function u is valid for exactly the reason substituting $a = E^2$ into a polynomial identity in a is valid: it is the same algebraic manipulation, only now the "variable" happens to act on functions rather than numbers.

That is also why the specific check — computing the remainder of E^8+1 divided by E^2+1 and finding it nonzero — is not a technicality to skip but the entire content of the problem's solvability. If that remainder had been 0 instead of 2 (i.e. if E^2+1 had divided E^8+1 exactly), the two hypotheses would not determine $u(x)$ from f, g alone: any function h annihilated by E^2+1 (a nontrivial common factor) could be added to u without disturbing either equation, and no formula purely in terms of f, g could exist. The coprimality is the exact condition under which a Bézout identity — hence a closed-form answer — exists at all, in precisely the same way that $\gcd(a, n) = 1$ is the exact condition for a to be invertible modulo n. This problem is the functional-equation shadow of modular inverses: two "moduli" E^2+1 and E^8+1, a coprimality check, and an explicit inverse combination.

Finally, why does the answer come out non-unique ("some of the simpler ones are…"), and why is that not a defect of the method? Because the Bézout coefficients A, B in $1 = A(E)(E^2+1) + B(E)(E^8+1)$ are themselves not unique — any multiple of $(E^2+1)(E^8+1)/\gcd$ can be added to A while subtracting the corresponding multiple from B — so the derivation genuinely produces a family of valid closed forms for $u(x)$, all differing by expressions that vanish once both hypotheses are re-imposed. The problem asks to "determine $u(x)$ in terms of f and g," not to determine it uniquely, and the method's built-in non-uniqueness is a faithful reflection of the underlying non-uniqueness of u itself (any solution of $u(x+1) + u(x-1) = 0 = u(x+4) + u(x-4)$ can be added to a valid u).

Layer $+\alpha$ BACKTESTING

1. The first useful cue is two functional equations relating u at fixed offsets (1 and 4) via addition. On seeing it, write down: "translation by a fixed amount is an operator E^a that commutes and multiplies by adding exponents; both equations are polynomials in E applied to u."

2. The cue suggests treating the two equations as $(E^2+1)u = 2Ef$ and $(E^8+1)u = 2E^4g$ (after clearing negative powers) and asking whether 1 is a combination of the two polynomials E^2+1, E^8+1 – i.e. running the Euclidean algorithm on them.

3. Hesitation occurs when trying to reconcile a stride-2 recursion with a stride-8 recursion by direct substitution; four nested applications with alternating signs are easy to mis-track. It is reinforced away by stepping back to the operator level, where the two relations are just two polynomials, and the reconciliation is a single, mechanical polynomial division.

4. The trap is doing the reconciliation as a manual induction (substituting the stride-2 recursion into itself four times to match the stride-8 one) with no structural check that it will close; a sign error in the fourth substitution is easy to make and easy to miss.

5. Whenever two functional equations relate a function at points offset by fixed *additive* amounts, recognise the offsets as commuting shift operators and rewrite both equations as polynomials in that operator applied to the unknown function.

6. Reusable rule: two operator equations $P(E)u = F$, $Q(E)u = G$ determine u in terms of F, G exactly when P, Q are coprime as polynomials (check via the Euclidean algorithm's remainder); the resulting Bézout identity $1 = A(E)P(E) + B(E)Q(E)$, applied to u, gives $u = A(E)F + B(E)G$ directly, and is unique only up to adding solutions of the associated homogeneous system.

▷ **Cue:** two functional equations relating u at points offset by fixed additive amounts (e.g. $u(x \pm a)$, $u(x \pm b)$) $\longrightarrow$ **Move:** introduce the shift operator E ($Eu(x) = u(x+1)$), rewrite both equations as polynomials in E applied to u, and check whether those two polynomials are coprime

▷ **Cue:** two coprime polynomials $P(E), Q(E)$ each known to act on u by a known function $\longrightarrow$ **Move:** run the Euclidean algorithm to get $1 = A(E)P(E) + B(E)Q(E)$, then apply this identity to u to solve for u in terms of the two known right-hand sides

1977 A4

1977 A4. For $0 < x < 1$, express

$$\sum_{n=0}^{\infty} \frac{x^{2^n}}{1 - x^{2^{n+1}}}$$

as a rational function of x.

Fields: analysis (infinite series), algebra (telescoping, partial fractions) **Position:** A4 — placed mid-paper but one of the lightest problems on it; one idea, and the whole cost of the problem is whether you find the idea before or after grinding **Answer:** $\dfrac{x}{1-x}$. Each term equals $\dfrac{1}{1-x^{2^n}} - \dfrac{1}{1-x^{2^{n+1}}}$, so the partial sums telescope to $\dfrac{1}{1-x} - \dfrac{1}{1-x^{2^{N+1}}} \to \dfrac{1}{1-x} - 1$.

Caveat. The MAA OCR of the problem statement drops the displayed series entirely (it reads "For $0 < x < 1$, express [blank] as a rational function of x"). The series above is reconstructed from the book's own solution under A-4, which manipulates exactly $\sum_{n\geq 0} x^{2^n}/(1-x^{2^{n+1}})$ and obtains $x/(1-x)$, and it agrees with the standard wording of 1977 A4. The identity of the problem is therefore secure; only the typesetting of the display is reconstructed.

Layer 1 THE WHAT — surface analysis

The natural attempt. You see an infinite sum and a request for a closed form, so you compute partial sums and hope for a pattern. Write S_N for the sum of the terms $n = 0, \ldots, N$.

$$S_0 = \frac{x}{1-x^2}.$$

Adding the $n = 1$ term over the common denominator $1 - x^4 = (1 - x^2)(1 + x^2)$:

$$S_1 = \frac{x(1 + x^2) + x^2}{1 - x^4} = \frac{x + x^2 + x^3}{1 - x^4}.$$

Adding the $n = 2$ term over $1 - x^8 = (1 - x^4)(1 + x^4)$:

$$S_2 = \frac{(x + x^2 + x^3)(1 + x^4) + x^4}{1 - x^8} = \frac{x + x^2 + \cdots + x^7}{1 - x^8}.$$

Now the pattern is visible: $S_N = \dfrac{x + x^2 + \cdots + x^{2^{N+1}-1}}{1 - x^{2^{N+1}}}$. You prove it by induction (with $M = 2^{N+1}$, $S_{N+1} = \frac{(x+\cdots+x^{M-1})(1+x^M)+x^M}{1-x^{2M}} = \frac{x+\cdots+x^{2M-1}}{1-x^{2M}}$), sum the geometric numerator to get $S_N = \dfrac{x(1 - x^{M-1})}{(1 - x)(1 - x^M)}$, and let $N \to \infty$: since $0 < x < 1$, $x^M \to 0$, so $S_N \to x/(1 - x)$.

This works. It is also three rounds of common-denominator algebra, a guessed pattern, and an induction, and at every stage you are one sign slip from a wrong pattern. It is correct zombie maths: nothing in it asked whether the terms had structure before multiplying them together.

The model chain. Put $u = x^{2^n}$, so $0 < u < 1$ and $x^{2^{n+1}} = u^2$. The n-th term is $\dfrac{u}{1 - u^2}$, and

$$\frac{1}{1 - u} - \frac{1}{1 - u^2} = \frac{(1 + u) - 1}{1 - u^2} = \frac{u}{1 - u^2}.$$

So, with $g(t) = \dfrac{1}{1 - x^t}$,

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

The partial sum telescopes:

$$S_N = \sum_{n=0}^{N} \left(g(2^n) - g(2^{n+1})\right) = g(1) - g(2^{N+1}) = \frac{1}{1 - x} - \frac{1}{1 - x^{2^{N+1}}}.$$

Because $0 < x < 1$, $x^{2^{N+1}} \to 0$ as $N \to \infty$, so $g(2^{N+1}) \to 1$ and

$$\sum_{n=0}^{\infty} \frac{x^{2^n}}{1 - x^{2^{n+1}}} = \frac{1}{1 - x} - 1 = \frac{x}{1 - x}.$$

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains start by looking at the first term $x/(1 - x^2)$. At that point the natural attempt adds the second term to it; the model rewrites the *single* term $x^{2^n}/(1 - x^{2^{n+1}})$ as a difference before adding anything.

> *Every term is the same function of $u = x^{2^n}$, namely $u/(1 - u^2)$, and the next term is that function of u^2. If I can write $u/(1 - u^2)$ as $h(u) - h(u^2)$, the whole sum collapses. The denominator $1 - u^2$ already contains $1 - u$, so try $h(u) = 1/(1 - u)$.*

> *Before I add two terms of a series, I ask what one term looks like as a difference.*

Layer 3 THE WHY

Why is telescoping the natural move and not a trick? Start from what the problem hands you. The index enters only through 2^n, and $2^{n+1} = 2 \cdot 2^n$: the step from one term to the next is the substitution $u \mapsto u^2$. A sum whose terms are generated by iterating one substitution is asking to be compared with that substitution. The only way a sum over an iterated map can have a closed form that does not involve N in an essential way is if consecutive terms cancel — that is, if the term is a "difference along the map", $h(u) - h(u^2)$. So the question you must ask is not "what is S_2?" but "is $u/(1-u^2)$ of the form $h(u) - h(u^2)$?"

And that question answers itself once you look at the denominator. $1 - u^2$ factorises as $(1-u)(1+u)$, and one of those factors, $1-u$, is the previous term's denominator in disguise: it is $1 - x^{2^n}$, which was the "big" denominator one step earlier. So $h(u) = 1/(1-u)$ is the only candidate worth trying, and one line of algebra confirms it. This is partial fractions (A-level) used for its real purpose: splitting a term so that a sum cancels.

The natural attempt is not wrong, and it is worth seeing why it produced the same answer. Its numerators $x + x^2 + \cdots + x^{2^{N+1}-1}$ over $1 - x^{2^{N+1}}$ are exactly $\dfrac{1}{1-x} - \dfrac{1}{1-x^{2^{N+1}}}$ written over a common denominator. The telescoping identity was there all along; the grind computed it without recognising it.

There is a second reason the answer is $x/(1-x) = x + x^2 + x^3 + \cdots$, and it explains why this problem exists. Expand each term as a geometric series:

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

Every positive integer m can be written as $2^n(2k+1)$ in exactly one way (pull out the largest power of 2; what remains is odd). So summing over n and k produces every power x^m, $m \geq 1$, exactly once; since all terms are positive the rearrangement is legitimate, and the total is $\sum_{m \geq 1} x^m = x/(1-x)$. The problem is the generating-function shadow of "every integer is a power of two times an odd number". The examiners designed the series from that fact; the telescoping is the cheapest way to verify it.

The official solution writes the term as $\frac{1}{1-x^{2^n}} - \frac{1}{1-x^{2^{n+1}}}$ without comment. That is the step that owes you a WHY, and the WHY is the $u \mapsto u^2$ structure above: the term is a function of x^{2^n}, the next term is the same function of its square, so you look for a difference along squaring.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is that n appears only as 2^n, and the denominator's exponent 2^{n+1} is the square of the numerator's base. Write down $u = x^{2^n}$ and the term $u/(1-u^2)$.

2. A term of the form $F(u)$ with successive terms $F(u), F(u^2), F(u^4), \ldots$ suggests telescoping: seek h with $F(u) = h(u) - h(u^2)$.

3. Hesitation comes at the moment you have $u/(1-u^2)$ and do not know which h to try. Reinforce: the denominator $1 - u^2 = (1-u)(1+u)$ contains $1-u$, which is the denominator one level down, so $h(u) = 1/(1-u)$.

4. The trap is computing S_0, S_1, S_2 and guessing: correct but slow, and a single arithmetic slip gives a wrong pattern that an induction would then fail to prove, costing more time. A smaller trap: forgetting that the limit $x^{2^{N+1}} \to 0$ uses $0 < x < 1$, which is why the problem states that range.

5. Any sum where the index lives in an exponent that doubles (or any iterated map $u \mapsto \phi(u)$) should trigger "difference along the map" before any common denominators are formed.

6. Reusable rule: if consecutive terms of a series are $F(u), F(\phi(u)), F(\phi(\phi(u))), \ldots$, try to write $F = h - h \circ \phi$; the sum is then $h(u_0) - \lim h(u_N)$. Find h by factorising the denominator of F and looking for the factor that belongs to the previous level.

▷ **Cue:** a series whose index appears only as 2^n in exponents, with x^{2^n} and $x^{2^{n+1}}$ side by side ⟶ **Move:** set $u = x^{2^n}$; the terms are $F(u), F(u^2), \ldots$; look for $F(u) = h(u) - h(u^2)$

▷ **Cue:** a term $u/(1-u^2)$ (or any term whose denominator is the previous denominator times something) ⟶ **Move:** split by partial fractions into $1/(1-u) - 1/(1-u^2)$ and telescope

▷ **Cue:** "express the infinite sum as a rational function" ⟶ **Move:** the answer is a limit of partial sums with a clean form; find the telescoping before computing partial sums

▷ **Cue:** a sum over n and k of $x^{2^n(2k+1)}$ or any sum over a unique factorisation of integers ⟶ **Move:** every positive integer appears once; the double sum is $\sum_{m\geq 1} x^m$

▷ **Cue:** a guessed pattern for partial sums after computing S_0, S_1, S_2 ⟶ **Move:** rewrite the guessed closed form as a difference of two terms; it is usually a telescoping identity you have not yet recognised

1977 A5

1977 A5. Prove that

$$\binom{pa}{pb} \equiv \binom{a}{b} \pmod{p}$$

for all integers p, a, and b with p a prime, $p > 0$, and $a \geq b \geq 0$.

Notation: $\binom{m}{n}$ denotes the binomial coefficient $\dfrac{m!}{n!\,(m-n)!}$.

Fields: number theory (congruences), algebra (polynomials over $\mathbb{Z}_p$), combinatorics (binomial coefficients) **Position:** A5 — short once the right object is chosen; the difficulty is entirely in not attacking the factorials directly **Answer:** True for all such p, a, b: modulo p, $(1+x)^p \equiv 1 + x^p$, hence $(1+x)^{pa} \equiv (1+x^p)^a = \sum_j \binom{a}{j} x^{pj}$, and comparing coefficients of x^{pb} gives $\binom{pa}{pb} \equiv \binom{a}{b}$.

Caveat. The OCR renders the congruence as garbled glyphs "(;:) = (:) (mod p)". The statement above is reconstructed from the book's notation line, its solution under A-5 (which proves exactly $\binom{pa}{pb} \equiv \binom{a}{b}$ for $b = 0, 1, \ldots, a$ by comparing coefficients in $\mathbb{Z}_p[x]$), and the standard wording of 1977 A5.

Layer 1 THE WHAT — surface analysis

The natural attempt. A congruence between binomial coefficients: you write out the factorials. With $m = p(a-b)$,

$$\binom{pa}{pb} = \frac{(m+1)(m+2)\cdots(m+pb)}{1\cdot 2\cdots pb} = \prod_{j=1}^{pb} \frac{m+j}{j}.$$

The worry is that p divides many of the j's, so you cannot reduce mod p factor by factor. So you split the indices. For $j = pk$ ($k = 1, \ldots, b$) the factor is $\dfrac{p(a-b)+pk}{pk} = \dfrac{a-b+k}{k}$, and

$$\prod_{k=1}^{b} \frac{a-b+k}{k} = \binom{a}{b}.$$

For the remaining j (those with $p \nmid j$) put $A = \prod_{p\nmid j}(m+j)$ and $D = \prod_{p\nmid j} j$. Then $\binom{pa}{pb} = \binom{a}{b} \cdot A/D$, that is,

$$D\binom{pa}{pb} = A\binom{a}{b},$$

an identity of integers. Since $p \mid m$, each $m + j \equiv j \pmod p$, so $A \equiv D \pmod p$. Hence $D\binom{pa}{pb} \equiv D\binom{a}{b} \pmod p$, and $p \nmid D$ (it is a product of numbers not divisible by the prime p), so D can be cancelled.

This is a complete proof, and a strong student can find it. It costs a careful split of indices, and it has one place where students lose marks: cancelling D is legitimate only because p is prime and $p \nmid D$, and writing "$A/D \equiv 1$" as if A/D were an integer is not a proof. The route also gives no hint of why the statement is true or what the general pattern is.

The model chain. *Step 1:* $(1+x)^p \equiv 1 + x^p$. For $0 < i < p$, $\binom{p}{i} = \dfrac{p!}{i!\,(p-i)!}$ is an integer; p divides the numerator $p!$, and p does not divide $i!\,(p-i)!$ because every factor there is less than p and p is prime. Hence $p \mid \binom{p}{i}$. So

$$(1+x)^p = 1 + x^p + pR(x)$$

for some polynomial R with integer coefficients. Write $P \equiv Q$ for polynomials with integer coefficients whose difference has every coefficient divisible by p.

Step 2: raise to the power a. If $P = Q + pR$ with integer-coefficient R, then by the binomial theorem $P^a = Q^a + \sum_{i \geq 1} \binom{a}{i} Q^{a-i} p^i R^i = Q^a + p(\text{integer polynomial})$. So

$$(1+x)^{pa} = \left((1+x)^p\right)^a \equiv (1+x^p)^a = \sum_{j=0}^{a} \binom{a}{j} x^{pj}.$$

Step 3: compare coefficients. The coefficient of x^{pb} on the left is $\binom{pa}{pb}$; on the right, only $j = b$ contributes, giving $\binom{a}{b}$. Congruence of polynomials is congruence of each coefficient, so

$$\binom{pa}{pb} \equiv \binom{a}{b} \pmod p \qquad (0 \leq b \leq a).$$

(Comparing the coefficient of x^k with $p \nmid k$ gives, free, $\binom{pa}{k} \equiv 0 \pmod p$.)

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains start from "binomial coefficients modulo a prime". The natural attempt opens the factorial formula for $\binom{pa}{pb}$; the model opens the generating function $(1+x)^{pa}$, of which $\binom{pa}{pb}$ is one coefficient.

> *$\binom{pa}{pb}$ is a coefficient of $(1+x)^{pa}$, and $(1+x)^{pa} = \left((1+x)^p\right)^a$. The one fact I know about binomials mod p is that $\binom{p}{i}$ is divisible by p for $0 < i < p$, so $(1+x)^p$ is $1 + x^p$ mod p. Then everything is a coefficient of $(1+x^p)^a$.*

> *The exponent is a product, pa. A product in an exponent is a power of a power; I should factor the power, not the factorial.*

Layer 3 THE WHY

Why the generating function? Go back to what a binomial coefficient is. $\binom{m}{n}$ has two definitions: the factorial quotient, and the coefficient of x^n in $(1+x)^m$. The factorial definition is good for sizes; it is bad for divisibility, because division is the one operation that congruences do not respect. The polynomial definition involves only multiplication and addition, and those are exactly the operations that survive reduction mod p. So the moment the question is "modulo p", the polynomial definition is the natural one: you want an object on which "reduce mod p" is a legitimate operation all the way through.

Second, why does the polynomial route go straight to the answer? Because the problem's shape is $\binom{p\cdot a}{p\cdot b}$: both entries scaled by p. In the polynomial world, scaling the exponent by p is taking a p-th power, $(1+x)^{pa} = ((1+x)^p)^a$; scaling the index by p is substituting x^p for x, $\binom{a}{b} \mapsto$ coefficient of x^{pb} in $(1+x^p)^a$. So the problem is asserting that, mod p, taking the p-th power of $1+x$ is the same as substituting x^p. That is precisely $(1+x)^p \equiv 1+x^p$, the "freshman's dream", and it holds because p divides every middle binomial coefficient $\binom{p}{i}$. Once the problem is read as "p-th power = substitution", no other route is natural.

And why does $p \mid \binom{p}{i}$ for $0 < i < p$? Because p is prime: p appears as a factor in $p!$ and cannot be cancelled by any factor of $i!\,(p-i)!$, all of which are smaller than p. That is the axiom-level fact, and it genuinely needs primality: for the composite modulus 6, $\binom{6}{2} = 15$ is not divisible by 6, and the problem's conclusion fails too, since $\binom{12}{6} = 924 = 6 \cdot 154 \equiv 0 \pmod 6$ while $\binom{2}{1} = 2$. The proof uses primality exactly once, in Step 1, and nowhere else.

What does the natural attempt see that the model hides? The same thing in coordinates. Splitting j into multiples and non-multiples of p is the factorial-side shadow of $(1+x)^p \equiv 1+x^p$: the multiples of p produce $\binom{a}{b}$ (the x^p part), and the non-multiples cancel mod p (the middle coefficients vanishing). Either proof is complete; the generating function is cheaper and extends.

It extends to Lucas's theorem: writing $a = a_1 p + a_0$, $b = b_1 p + b_0$ with $0 \le a_0, b_0 < p$,

$$(1+x)^a \equiv (1+x^p)^{a_1}(1+x)^{a_0}$$

and comparing the coefficient of $x^{b_1 p + b_0}$ gives $\binom{a}{b} \equiv \binom{a_1}{b_1}\binom{a_0}{b_0}$. The present problem is the case $a_0 = b_0 = 0$. The official solution cites "it is well known" that $p \mid \binom{p}{i}$ and moves on; that fact is the only place the prime is used, which is why this archive proves it.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is "binomial coefficient $\equiv$ binomial coefficient $\pmod p$". Write down $(1+x)^{pa}$ and $(1+x)^p \equiv 1+x^p \pmod p$.

2. Generating functions reduced mod p: express the coefficient as a coefficient of a polynomial, reduce the polynomial, read the coefficient off.

3. Hesitation: "can I reduce a polynomial mod p and still compare coefficients?" Yes: congruence of integer polynomials is by definition coefficientwise, and it is preserved by sums and products (Step 2). Once that is secure the proof is three lines.

4. The trap is dividing inside a congruence: writing $\binom{pa}{pb} = \prod (m+j)/j$ and reducing each fraction mod p, including those with $p \mid j$. The factors with $p \mid j$ are $0/0$ mod p. A second trap is asserting $(1+x)^p \equiv 1+x^p$ without proving $p \mid \binom{p}{i}$, which is the only use of primality.

5. Whenever both arguments of a binomial coefficient are multiplied by the same prime p, or a problem concerns binomial coefficients modulo a prime, think "p-th power = substitution $x \mapsto x^p$ modulo p".

6. Reusable rule: for divisibility or congruence questions about coefficients, work with the generating function and reduce it mod p; avoid formulas that divide. In characteristic p, $(u+v)^p = u^p + v^p$.

▷ **Cue:** $\binom{pa}{pb}$ versus $\binom{a}{b}$ modulo a prime p $\longrightarrow$ **Move:** write $(1+x)^{pa} = ((1+x)^p)^a \equiv (1+x^p)^a$ and compare the coefficient of x^{pb}

▷ **Cue:** binomial coefficients modulo a prime $\longrightarrow$ **Move:** use $(1+x)^p \equiv 1+x^p$; the general form is Lucas's theorem on base-p digits

▷ **Cue:** a congruence involving a quotient of factorials $\longrightarrow$ **Move:** switch to a definition that uses only multiplication and addition (a coefficient), since division does not survive reduction mod p

▷ **Cue:** a factorial product where the prime divides some factors and not others $\longrightarrow$ **Move:** split indices into multiples and non-multiples of p; the multiples reproduce the smaller problem, the rest cancel mod p

▷ **Cue:** an exponent that is a product pa $\longrightarrow$ **Move:** rewrite as a power of a power, $((\cdot)^p)^a$

1977 A6

1977 A6. Let $f(x,y)$ be a continuous function on the square

$$S = \{(x,y) : 0 \le x \le 1,\ 0 \le y \le 1\}.$$

For each point (a,b) in the interior of S, let $S_{(a,b)}$ be the largest square that is contained in S, is centered at (a,b), and has sides parallel to those of S. If the double integral $\iint f(x,y)\,dx\,dy$ is zero when taken over each square $S_{(a,b)}$, must $f(x,y)$ be identically zero on S?

Fields: real analysis (continuity, double integrals), number theory (the Euclidean algorithm) **Position:** A6 — last on the paper; hard to start, short once the right auxiliary function is on the page **Answer:** Yes. With $I(u,v) = \int_0^u \int_0^v f$, squares touching an edge give $I(u,v) = I(u-v,v)$ for $0 < v \le u \le 1$ and $I(u,v) = I(u,v-u)$ for $0 < u \le v \le 1$; running the Euclidean algorithm gives $I = 0$ whenever u/v is rational, continuity gives $I \equiv 0$, and then $f \equiv 0$.

Caveat. The OCR reads the statement cleanly apart from symbols ("$S(a,b)$", "$\iint f(x,y)\,dx\,dy$"), and the book's solution under A-6 proves exactly this statement. The book's solution reuses the letters (a,b) for a *corner* of a rectangle rather than a centre; this analysis uses (u,v) for corners to keep the two roles apart. The book handles irrational u/v by a limit of the subtractive Euclidean algorithm; the model chain below uses density and continuity instead, and the limit argument is verified in Layer 3. Both are complete.

Layer 1 THE WHAT — surface analysis

The natural attempt. The question "must f be zero?" invites you to hunt for a counterexample, and the first candidates are oscillating functions: something like $\sin(2\pi kx)$, whose integral over an interval vanishes when the interval is a whole number of periods. It fails at once. The square centred at (a,b) with b small is $[a-b, a+b] \times [0, 2b]$; its x-width $2b$ varies continuously with the centre, so no fixed period can fit every width.

So you switch to proving $f \equiv 0$ and look for local information. Near the boundary the squares are tiny: as (a,b) approaches a boundary point the square $S_{(a,b)}$ shrinks to that point, and dividing the zero integral by the area gives $f = 0$ on ∂S by continuity. In the middle, only one square is centred at $(\frac{1}{2}, \frac{1}{2})$, namely S itself, so $\iint_S f = 0$. That is one number, not a function.

Next you differentiate. In the region where the bottom edge is nearest ($b \le a$, $b \le 1-a$, $b \le \frac{1}{2}$), the hypothesis says

$$\Phi(a,b) = \int_{a-b}^{a+b} \int_0^{2b} f(x,y)\,dy\,dx = 0.$$

Differentiating in a: $\displaystyle\int_0^{2b} \big[f(a+b,y) - f(a-b,y)\big]\,dy = 0$. Writing $G(x,t) = \int_0^t f(x,y)\,dy$, this says $G(x+t,t) = G(x,t)$ — for every fixed t, $G(\cdot,t)$ is t-periodic. Differentiating again in b produces derivatives of f at two distant points, which you do not know exist. The equations are *non-local*: they relate values at points a whole square-width apart. No amount of differentiation turns them into a pointwise statement about f. That is the wall.

The model chain. Let $M = \max_S |f|$ and define, for $0 \le u, v \le 1$,

$$I(u,v) = \int_0^u \int_0^v f(x,y)\,dy\,dx.$$

Then $I(0,v) = I(u,0) = 0$, and $|I(u,v) - I(u',v')| \le M(|u-u'| + |v-v'|)$, so I is continuous.

Step 1 (the hypothesis as a functional equation). Let $0 < v \le u \le 1$ and take the centre $(c,d) = (u - \frac{v}{2}, \frac{v}{2})$, which is interior. The largest admissible square centred there has half-side $\min(c, 1-c, d, 1-d)$, and

$$c - d = u - v \ge 0, \qquad (1-c) - d = 1 - u \ge 0, \qquad (1-d) - d = 1 - v \ge 0,$$

so the minimum is $d = \frac{v}{2}$ and $S_{(c,d)} = [u-v, u] \times [0, v]$. Its integral is $I(u,v) - I(u-v, v)$, so

$$I(u,v) = I(u-v, v) \qquad (0 < v \le u \le 1).$$

Reflecting in the diagonal (centre $(\frac{u}{2}, v - \frac{u}{2})$, square $[0,u] \times [v-u, v]$):

$$I(u,v) = I(u, v-u) \qquad (0 < u \le v \le 1).$$

Step 2 (rational ratios). Suppose $u = mg$, $v = ng$ with m, n positive integers and $g > 0$. Apply Step 1 to whichever coordinate is larger (either, if equal). The new pair is $(m'g, n'g)$ with non-negative integers m', n', both coordinates still in $[0,1]$, and $m' + n' < m + n$. While both coordinates are positive Step 1 applies again; since $m + n$ cannot decrease for ever, some coordinate reaches 0. Hence $I(u,v) = 0$.

Step 3 (all points). Points with both coordinates rational are dense in $[0,1]^2$ and have rational ratio (or a zero coordinate), so $I = 0$ on a dense set; by continuity $I \equiv 0$ on $[0,1]^2$.

Step 4 ($f \equiv 0$). For any rectangle $R = [\alpha, \beta] \times [\gamma, \delta] \subseteq S$,

$$\iint_R f = I(\beta,\delta) - I(\alpha,\delta) - I(\beta,\gamma) + I(\alpha,\gamma) = 0.$$

If $f(x_0, y_0) \ne 0$ for some $(x_0, y_0) \in S$, say $f(x_0, y_0) = \eta > 0$, continuity gives a rectangle $R \subseteq S$ of positive area containing (x_0, y_0) on which $f > \eta/2$; then $\iint_R f > 0$, a contradiction. (The case $\eta < 0$ is the same with $-f$.) So $f \equiv 0$ on S: the answer is **yes**.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains reach the equation $\iint_{S_{(a,b)}} f = 0$ for a square resting on an edge. The natural attempt differentiates it, hoping for a local statement; the model writes that square as a difference of two rectangles anchored at the origin and keeps the equation global: $I(u,v) = I(u-v, v)$.

> *The largest square always touches an edge of S, so one of its sides lies on the boundary. A box resting on the bottom edge is the difference of two boxes with a corner at the origin. So the natural unknown is the corner integral $I(u,v)$, the two-dimensional antiderivative, and the hypothesis becomes an equation for I.*

> *$I(u,v) = I(u-v, v)$: I am replacing the larger coordinate by the difference. That is the subtractive Euclidean algorithm. It ends at a zero coordinate whenever u/v is rational, and I vanishes there.*

Layer 3 THE WHY

Where does the corner integral come from? Look at the one-dimensional version first. If f is continuous on $[0,1]$ and its integral over the largest interval centred at c is zero for every c, then for $c \le \frac{1}{2}$ that interval is $[0, 2c]$, so $F(2c) = 0$ with $F(x) = \int_0^x f$; hence $F \equiv 0$ on $[0,1]$ and $f = F' \equiv 0$. The antiderivative anchored at the boundary is the object on which "integral over an interval" becomes an equation, because every interval integral is a difference of two values of F. In two dimensions the antiderivative is $I(u,v)$, and every rectangle integral is an

alternating sum of four values of I. That is why the model defines I before doing anything else: it is the only function in which the hypothesis is *algebraic* rather than an integral.

Why is the problem harder than its one-dimensional version? Because squares have a fixed aspect ratio. In one dimension the admissible intervals $[0, 2c]$ already sweep out every anchored interval, so you read off $F \equiv 0$ directly. In two dimensions a square resting on the bottom edge is $[u - v, u] \times [0, v]$ — its width equals its height — so the hypothesis never directly tells you $I(u, v)$ for a single anchored rectangle; it only tells you that I is unchanged when you cut a $v \times v$ square off the side of a $u \times v$ rectangle. And that is exactly the geometric description of Euclid's algorithm: repeatedly cut the largest square off a rectangle. If the sides are commensurable you eventually cut the rectangle away entirely (a zero side, where $I = 0$). The problem is the Euclidean algorithm drawn in the plane; seeing "replace the larger by the difference" is what makes Step 2 natural rather than inspired.

Why does the natural attempt hit a wall? Differentiation is local, and the hypothesis is not: it relates $G(x + t, t)$ to $G(x, t)$, points a whole width apart. Local tools extract local information; a non-local equation needs to be iterated, not differentiated. The rule is to keep the equation in the form where both sides are values of the same function, and ask what repeatedly applying it does.

The irrational case. The official solution runs the subtractive algorithm for ever when u/v is irrational: the pairs (u_n, v_n) are coordinatewise non-increasing, $I(u, v) = I(u_n, v_n)$ for all n, and it asserts $u_n, v_n \to 0$. That assertion needs a proof, and here it is. Each step lowers $u_n + v_n$ by $\min(u_n, v_n)$, and $u_n + v_n \geq 0$, so $\sum_n \min(u_n, v_n) < \infty$ and $\min(u_n, v_n) \to 0$. The limits $\alpha = \lim u_n$, $\beta = \lim v_n$ exist by monotonicity, and $\min(\alpha, \beta) = 0$; say $\beta = 0$. Since u/v is irrational no coordinate ever becomes 0: (u_n, v_n) is obtained from (u, v) by an integer matrix of determinant 1 (a product of the matrices $\left(\begin{smallmatrix} 1 & -1 \\ 0 & 1 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 1 & 0 \\ -1 & 1 \end{smallmatrix}\right)$), so each coordinate is $ru + sv$ with integers $(r, s) \neq (0, 0)$, and $ru + sv = 0$ would make u/v rational. So $v_n > 0$ for all n and $v_n \to 0$ forces v to be reduced infinitely often, which happens only at steps where $u_n < v_n$. Along those steps $u_n < v_n \to 0$, and u_n is monotone, so $\alpha = 0$. Then $|I(u_n, v_n)| \leq M u_n v_n \to 0$, so $I(u, v) = 0$. The density argument in Step 3 replaces all of this by one appeal to continuity, which is why the model chain uses it: it is the cheaper route to the same conclusion.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is "largest square contained in S, centred at (a, b)": the largest such square always touches the boundary. Write down the square touching the bottom edge, $[a - b, a + b] \times [0, 2b]$, and then the corner integral $I(u, v)$.

2. Anchored antiderivative plus a functional equation: express each square integral as a difference of I-values, then iterate the resulting equation.

3. Hesitation comes at $I(u, v) = I(u - v, v)$, because it looks like one weak equation. Reinforce by recognising it as one step of Euclid's algorithm; the repeated step, not the single step, carries the information.

4. The trap is differentiating the hypothesis to get a PDE: the equation is non-local and differentiation only produces more non-local equations with extra smoothness assumptions. A second trap is the counterexample hunt with periodic functions, which cannot work because the square widths vary continuously. A rigour trap is the irrational case: asserting $u_n, v_n \to 0$ without proof.

5. Whenever a hypothesis gives integrals over a family of regions, find the anchored antiderivative in which every region's integral is an alternating sum of values; the hypothesis then becomes a functional equation to iterate.

6. Reusable rule: an equation of the form $\Phi(x, y) = \Phi(x - y, y)$ is the Euclidean algorithm; it determines Φ on commensurable pairs from the values on the axes, and continuity extends it to all pairs.

▷ **Cue:** "largest square contained in S centred at a point" $\longrightarrow$ **Move:** it touches an edge; write it as $[u - v, u] \times [0, v]$ and subtract two anchored rectangles

- ▷ **Cue:** hypothesis about integrals over many sub-regions of a rectangle $\longrightarrow$ **Move:** define $I(u,v) = \int_0^u \int_0^v f$; every rectangle integral is $I(\beta,\delta) - I(\alpha,\delta) - I(\beta,\gamma) + I(\alpha,\gamma)$
- ▷ **Cue:** a functional equation $\Phi(u,v) = \Phi(u-v,v)$ (replace the larger by the difference) $\longrightarrow$ **Move:** run the Euclidean algorithm; it terminates at a zero coordinate when u/v is rational
- ▷ **Cue:** a continuous function known to vanish on all points with rational coordinates or rational ratio $\longrightarrow$ **Move:** density plus continuity: it vanishes everywhere
- ▷ **Cue:** a non-local equation relating values at points a fixed width apart $\longrightarrow$ **Move:** do not differentiate; iterate the equation
- ▷ **Cue:** all rectangle integrals of a continuous f vanish $\longrightarrow$ **Move:** $f \equiv 0$: otherwise a small rectangle where f keeps one sign gives a non-zero integral

Putnam 1977 — Paper B

1977 B1

1977 B1. Evaluate the infinite product

$$\prod_{n=2}^{\infty} \frac{n^3-1}{n^3+1}.$$

Caveat. The MAA OCR of the official B-1 solution is badly corrupted – it shows only the fragment "$= \lim_{k\to\infty}\left[\frac{2\cdot13}{3\cdot3}\cdot\frac{3\cdot21}{4\cdot7}\cdot\frac{5\cdot13}{5\cdot13}\cdots\frac{(k-1)(k^2+k+1)}{(k+1)(k^2-k+1)}\right]$" with no surrounding derivation. That fragment is consistent with the factorisation used below (it is the partial product written out with numerator and denominator each split into a linear and a quadratic factor), so the problem label is confirmed correct for this year and paper, but the full chain that follows is reconstructed and independently verified here, not transcribed.

Fields: algebra, telescoping products **Position:** B1 (opener) **Answer:** The product converges to $\frac{2}{3}$.

Layer 1 THE WHAT — surface analysis

The natural attempt. The general term is a ratio of two cubics that differ only in sign, and $n^3 \mp 1$ is a sum/difference of cubes – that factorisation is the obvious first move:

$$n^3-1=(n-1)(n^2+n+1), \qquad n^3+1=(n+1)(n^2-n+1).$$

So the partial product up to N is

$$P_N=\prod_{n=2}^{N}\frac{(n-1)(n^2+n+1)}{(n+1)(n^2-n+1)}.$$

At this point the natural instinct is to treat this as a single fraction and hunt for cancellation across the whole thing at once: write out the first several terms, $n=2,3,4,5,\ldots$, and look for factors that visibly repeat between a numerator and a later denominator. The linear parts cooperate almost immediately – $(n-1)$ in the numerator at $n=3$ is the same as the $(n+1)$ that appeared in the denominator at $n=1$ – but the quadratic parts n^2+n+1 and n^2-n+1 do not obviously match anything, because $n^2+n+1 \neq (n+1)^2-(n+1)+1$ at the same n. Faced with four interleaved factors and only a partial pattern, a student under time pressure is tempted to abandon exact cancellation and instead estimate: compute P_N numerically for $N=2,\ldots,10$, watch it settle near 0.667, and simply assert the limit is $2/3$ without ever producing the closed form for P_N. That gets the answer by guessing, not by proof, and a Putnam grader wants the closed form.

The model chain. Do not try to cancel the four factors as one tangle. Split P_N into two separate products, one for the linear factors and one for the quadratic factors, and telescope each on its

own terms.

$$P_N = \underbrace{\prod_{n=2}^{N} \frac{n-1}{n+1}}_{L_N} \cdot \underbrace{\prod_{n=2}^{N} \frac{n^2+n+1}{n^2-n+1}}_{Q_N}.$$

The linear product is the classic skip-two telescope:

$$L_N = \frac{1 \cdot 2 \cdot 3 \cdots (N-1)}{3 \cdot 4 \cdot 5 \cdots (N+1)} = \frac{1 \cdot 2}{N(N+1)} = \frac{2}{N(N+1)}.$$

For the quadratic product, notice the algebraic identity

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

so if $q_m := m^2 + m + 1$, then the denominator at index n is q_{n-1} and the numerator at index n is q_n. That makes Q_N a plain shift-by-one telescope:

$$Q_N = \prod_{n=2}^{N} \frac{q_n}{q_{n-1}} = \frac{q_N}{q_1} = \frac{N^2+N+1}{3}.$$

Combining,

$$P_N = \frac{2}{N(N+1)} \cdot \frac{N^2+N+1}{3} = \frac{2(N^2+N+1)}{3N(N+1)} \longrightarrow \frac{2}{3} \quad \text{as } N \to \infty.$$

Convergence of the infinite product itself is not in doubt: the general term is $1 + O(n^{-3})$, since $\frac{n^3-1}{n^3+1} - 1 = \frac{-2}{n^3+1}$, and $\sum n^{-3}$ converges, so the product converges to a nonzero limit – and the closed form for P_N pins that limit down exactly.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains agree up to the fully factored partial product

$$P_N = \prod_{n=2}^{N} \frac{(n-1)(n^2+n+1)}{(n+1)(n^2-n+1)}.$$

That factorisation is forced – it is just sum/difference of cubes. The split happens immediately after: the natural attempt keeps all four factors inside one product and searches for a single cancellation pattern across the combined fraction; the model chain separates the product into L_N (linear factors) and Q_N (quadratic factors) and telescopes each independently, using a different index-shift for each.

> *These are two different telescoping patterns sitting on top of each other – the linear factors skip by two, the quadratic factors shift by one – so I should pull them apart into two separate products and telescope each one on its own terms, instead of hunting for one cancellation rule that fits all four factors at once.*

Layer 3 THE WHY

Sum-of-cubes and difference-of-cubes share more structure than the mnemonic "$a^3 \pm b^3 = (a \pm b)(a^2 \mp ab + b^2)$" usually advertises: with $b = 1$, the two companion quadratics are $n^2 + n + 1$ and $n^2 - n + 1$, and completing the square (or just expanding $(n-1)^2 + (n-1) + 1$) shows the second is literally the first evaluated one integer earlier. That is not a coincidence to be rediscovered by trial substitution – it is forced by the fact that both quadratics come from the same two-term identity $a^2 \mp ab + b^2$ with $b = 1$, so shifting $n \to n - 1$ turns the $+$ version into the $-$ version. Once that is seen as an algebraic identity rather than a numerical curiosity, the

quadratic factors are recognised as values $q_2, q_3, \ldots, q_N$ of a single function $q_m = m^2 + m + 1$ sampled at consecutive integers with the denominator always one step behind – exactly the shape a telescoping product needs, with no cancellation-hunting required.

The linear factors telescope with a different period (skip-two, $\frac{n-1}{n+1}$) precisely because a difference of cubes always contributes a linear factor $(n - b)$ while a sum contributes $(n + b)$: same b, opposite sign, so they are one apart from the argument that made the OTHER product's quadratic part shift by one. Recognising that the two families – degree 1 and degree 2 – telescope with genuinely different periods is the reason forcing a single unified cancellation across all four factors fails: there is no common period that makes both patterns visible simultaneously. The correct move is not more algebra, it is disaggregation by degree, splitting the product before attempting to telescope either piece. That is the general fact underneath this problem: whenever a rational function factors into pieces of different degree, each piece's own telescoping identity should be found and applied separately, because different-degree factors essentially never share one shift.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is $n^3 \mp 1$ itself – sum and difference of cubes. On seeing it, write down both factorisations immediately: $(n \mp 1)(n^2 \pm n + 1)$.
2. That cue suggests a telescoping product, but the four resulting factors (two linear, two quadratic) are not yet in a form that visibly telescopes; the next cue to look for is an index-shift identity linking the two quadratics.
3. Hesitation sets in right after the factorisation, when four factors sit together and no single cancellation is obvious across all of them; reinforce by asking, "are these one telescoping family or two, of different degrees?" before trying to cancel anything.
4. The trap is that the linear factors telescope with a skip of two, which primes the solver to expect the same skip-two pattern from the quadratic factors – they don't; forcing that expectation produces confused, incomplete cancellation.
5. Recognise this faster next time by memorising the identity $(n-1)^2 + (n-1) + 1 = n^2 - n + 1$: whenever a sum/difference-of-cubes factorisation appears in a product, the two quadratic companions are automatically one shift apart.
6. Decision rule: when an infinite product's general term factors into pieces of different polynomial degree, split the product by degree and telescope each piece with its own natural index-shift before trying to combine or cancel across the whole term.

1977 B2

1977 B2. Given a convex quadrilateral $ABCD$ and a point O not in the plane of $ABCD$, locate point A' on line OA, point B' on line OB, point C' on line OC, and point D' on line OD so that $A'B'C'D'$ is a parallelogram.

Caveat. The OCR renders O as the digit "0" throughout the problem statement ("a point 0 not in the plane of ABCD"); this is corrected silently above as an unambiguous digit/letter OCR error, consistent throughout the official solution which unambiguously treats it as the point O, apex of the four rays.

Fields: solid geometry, affine geometry **Position:** B2 **Answer:** A construction exists (and is not unique): take $X = AC \cap BD$ in the base plane, then in plane AOC let $A'C'$ be the diagonal, through X, of a parallelogram with sides parallel to OA, OC; do the symmetric construction in plane BOD. Then $A'B'C'D'$ is a parallelogram.

Layer 1 THE WHAT — surface analysis

The natural attempt. O is off the plane of $ABCD$, so set up vectors from O: let $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ be the position vectors of A, B, C, D relative to O, and look for scalars $\alpha, \beta, \gamma, \delta$ with $A' = \alpha\mathbf{a}$, $B' = \beta\mathbf{b}$, $C' = \gamma\mathbf{c}$, $D' = \delta\mathbf{d}$. A quadrilateral is a parallelogram exactly when its diagonals bisect each other, so the condition is

$$\alpha\mathbf{a} + \gamma\mathbf{c} = \beta\mathbf{b} + \delta\mathbf{d}.$$

This is one vector equation in three dimensions, with four unknown scalars, and it looks solvable by brute linear algebra: since A, B, C, D are coplanar and O is not on that plane, $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ satisfy exactly one linear dependency (four vectors confined, after scaling by O, to a 2-dimensional affine slice of 3-space), and a natural instinct is to grind out that dependency explicitly – express $\mathbf{d}$ as an affine combination of $\mathbf{a}, \mathbf{b}, \mathbf{c}$, substitute into the vector equation, and match coefficients. This works in principle, but the dependency itself is exactly as complicated as locating where the diagonals AC and BD of the quadrilateral cross, so the "pure algebra" route ends up re-deriving that intersection anyway, buried inside four simultaneous scalar equations, with no guarantee the resulting $\alpha, \beta, \gamma, \delta$ are recognisable as anything geometric.

The model chain. Do not solve one global 3D system in four unknowns. Split the quadrilateral along its diagonals and solve two independent 2-dimensional problems, one in each of the planes those diagonals determine with O.

Since $ABCD$ is convex, its diagonals AC and BD meet at a point X inside it. Now X lies on line AC, which lies in plane AOC; and X also lies on line BD, which lies in plane BOD. So X belongs to both planes AOC and BOD – and since these two planes both already contain O, and $X \neq O$, the line OX is exactly their line of intersection.

Work first inside plane AOC. Through X, draw the line parallel to OC; let it meet line OA at A'. Through X, draw the line parallel to OA; let it meet line OC at C'. By construction $OA'XC'$ has both pairs of opposite sides parallel, so it is a parallelogram, and its diagonals OX and $A'C'$ bisect each other at some point M – the midpoint of OX.

Now repeat the identical construction inside plane BOD: through X, draw lines parallel to OD (meeting OB at B') and parallel to OB (meeting OD at D'). $OB'XD'$ is a parallelogram by the same reasoning, so its diagonals OX and $B'D'$ also bisect each other – at the midpoint of OX, which is the same point M as before, because OX is one fixed segment.

So $A'C'$ and $B'D'$ both pass through M and are both bisected by it. But $A'C'$ and $B'D'$ are the two diagonals of quadrilateral $A'B'C'D'$: diagonals that bisect each other are exactly the defining property of a parallelogram. Hence $A'B'C'D'$ is a parallelogram, and every vertex sits on the required line from O.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains start from the same fact – A', B', C', D' are scalar multiples of position vectors from O, and a parallelogram condition is a diagonal-bisection condition. They split at what to do about the fact that the four unknowns are genuinely linked by the coplanarity of A, B, C, D: the natural attempt tries to extract that link as an algebraic dependency among $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ and solve one 3D system; the model chain instead asks where that dependency comes from geometrically – the crossing point of the diagonals – and uses that single point to build two separate, independent 2D parallelogram constructions.

> *The planes AOC and BOD both contain O, so they meet in a line through O – and that line also has to pass through the point where diagonal AC crosses diagonal BD, because that crossing point lies in both planes. I should build my two 2D parallelograms using that shared line as their common diagonal, so the two planar solutions automatically share a midpoint in 3-space.*

Layer 3 THE WHY

Two distinct planes in 3-space that share a point meet in a line through that point – this is

definitional, not a special fact about this configuration. The only work is showing X genuinely lies on both planes, and that follows immediately from where X was defined: $X \in AC \subset$ plane AOC, and $X \in BD \subset$ plane BOD, so X is a point common to both planes by construction, forcing it onto their shared line through O. The convexity of $ABCD$ is exactly what guarantees X exists as an honest interior intersection point of the two diagonals, rather than the diagonals missing each other or meeting outside the segment – convexity is not decoration in this problem, it is the hypothesis that supplies the point X the whole construction is built on.

Once OX is identified as the shared line, the rest is the most basic fact about parallelograms, used in its converse direction: a quadrilateral is a parallelogram exactly when its diagonals bisect each other. The construction does not need to build the SAME parallelogram twice – it only needs two parallelograms (one in each plane) that happen to share a diagonal OX, because the midpoint of a fixed segment is a fixed point regardless of which parallelogram is built on it. That is the real economy of the solution: the 3D problem is never solved as a 3D problem at all. It is solved as two completely decoupled 2D problems, glued together by the one piece of data – the shared line – that the 3D structure forces to exist. This is also why the construction is not unique: any point on line OX other than O would serve as the shared diagonal-endpoint in place of X, producing a different valid parallelogram $A'B'C'D'$; X is simply the most natural choice because it is already given by the data, not because it is forced.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the shape of the data itself: a single external point O with four rays to four coplanar points. On seeing "one apex, four rays, want a 3D condition," write down the two planes each diagonal of the quadrilateral spans with O: plane AOC and plane BOD.

2. That cue suggests reducing the 3D problem to two independent 2D problems, one in each of those planes, glued along whatever line the two planes share.

3. Hesitation sets in right after writing $A' = \alpha\mathbf{a}$, etc. – it's tempting to treat all four scalars as one 3D linear system to be solved together. Reinforce against this by asking, "does the data already hand me a lower-dimensional point I haven't used yet?" – here, the crossing point of the diagonals.

4. The trap is treating A, B, C, D symmetrically, as if all four rays play the same role; the real structure pairs them by the quadrilateral's diagonals (A with C, B with D), not by its sides, and missing that pairing is what forces the messy global algebra.

5. Recognise this faster next time by treating any "one apex, four rays to coplanar points" configuration as secretly two 2D problems: find the diagonal-pair structure of the coplanar figure first, and look for the line the two diagonal-planes share through the apex.

6. Decision rule: to build a 3D configuration on rays from a single external point to several coplanar points, split the coplanar points into diagonal pairs, solve the resulting 2D construction independently in each diagonal's plane with the apex, and glue the two solutions along the line those two planes are forced to share.

1977 B3

1977 B3. An (ordered) triple (x_1, x_2, x_3) of positive irrational numbers with $x_1 + x_2 + x_3 = 1$ is called "balanced" if each $x_i < 1/2$. If a triple is not balanced, say if $x_1 > 1/2$, one performs the following "balancing act": $x_i' = 2x_i$ if $i \neq j$ and $x_j' = 2x_j - 1$, where j is the (unique) index with $x_j > 1/2$. If the new triple is not balanced, one performs the balancing act on it. Does continuation of this process always lead to a balanced triple after a finite number of performances of the balancing act?

Caveat. The OCR of the balancing-act rule is corrupted ("where $x_i' - 2x_1$ if $i.;.j$ and $xJ - 2x_1 - 1$"); the rule reconstructed above – double every coordinate, then subtract 1 from whichever coordinate exceeds $1/2$ – is forced by the requirement that the new triple again sum to 1 (checked below) and matches the only reading consistent with the official solution's own restatement via binary expansions, so it is used with confidence rather than as an unverified guess. The topic index cross-reference ("Balanced triples 1977 B-3"; "irrational numbers ... 1977 B-3") confirms this statement is correctly matched to this label.

Fields: real analysis, binary expansions, dynamical systems **Position:** B3 **Answer:** No – there exist unbalanced triples for which the balancing act never produces a balanced triple, no matter how many times it is applied.

Layer 1 THE WHAT — surface analysis

The natural attempt. At most one coordinate can exceed $1/2$ at a time, since the three positive numbers sum to 1: if $x_1 > 1/2$, then $x_2 + x_3 < 1/2$, so neither x_2 nor x_3 can also exceed $1/2$. So the balancing act is always well defined on an unbalanced triple, and it is natural to just run it: pick a concrete unbalanced triple, apply the act a few times, and watch what happens. On almost any example one tries by hand – say start with x_1 close to 1 and x_2, x_3 small – the triple balances within a handful of steps, because doubling shrinks the "excess over $1/2$" quickly in typical cases. This experimental evidence pushes naturally toward conjecturing the answer is "yes, always," and toward hunting for a monovariant: some quantity, perhaps $\max(x_1, x_2, x_3)$ or the number of steps since the last balanced state, that must strictly decrease and force termination. But no such quantity actually exists in general – $\max(x_1, x_2, x_3)$ does not decrease monotonically under the act (doubling the two smaller coordinates can easily push one of them above $1/2$ next time), and no amount of case-checking on hand-picked examples proves a universal bound, because the examples one picks by hand are never adversarial enough. The natural attempt stalls exactly where every "always terminates" conjecture built from a handful of friendly numerical trials stalls: it has evidence, not proof, and the evidence happens to be misleading.

The model chain. Re-encode the whole process in binary. Write each $x_i \in (0,1)$ as $x_i = \sum_{k\geq 1} a_{ik} 2^{-k}$ with $a_{ik} \in \{0,1\}$ (this exists and is unique – up to the usual terminating/repeating-1's ambiguity, which does not arise here since the x_i are irrational). Then $x_i > 1/2$ exactly when $a_{i1} = 1$, and the balancing act's rule – double every coordinate, then subtract 1 from the one that exceeded $1/2$ – is exactly the statement that doubling shifts every binary expansion one place left, and subtracting 1 from the doubled value of the coordinate whose leading bit was 1 simply deletes that leading 1. So after one application of the act, coordinate i's new expansion is $a_{i2}, a_{i3}, a_{i4}, \ldots$ – every coordinate has simply lost its first digit. The triple is balanced after k applications exactly when $a_{1,k+1} = a_{2,k+1} = a_{3,k+1} = 0$: a balanced triple is precisely one whose current leading digit column, across all three coordinates, is all zero.

This reframes the question completely: does every valid triple of binary sequences, with $x_1 + x_2 + x_3 = 1$, eventually reach an all-zero column? The answer is no, and the way to see it is to build a triple that provably never does. Since the three positive numbers sum to 1 and each column can have at most one 1 (three simultaneous values over $1/2$ is impossible, as shown above), assign, for every index $j \geq 1$, exactly one 1 per column, choosing which coordinate gets it according to a fixed rule on j that never repeats periodically:

$$a_{1j} = 1 \iff j \text{ is an odd perfect square } (1, 9, 25, 49, \ldots), \qquad a_{2j} = 1 \iff j \text{ is an even perfect square } (4, 16, 36$$

$$a_{3j} = 1 \iff j \text{ is neither.}$$

Every positive integer j is in exactly one of these three classes, so every column has exactly one 1 – no column is ever all-zero, and the triple never balances. The sum checks out: $x_1 + x_2 + x_3 =$

$\sum_{j\geq 1} 2^{-j} = 1$, since every power 2^{-j} is contributed by exactly one coordinate. Each $x_i > 0$ since each of the three index sets is infinite. And each x_i is irrational, because a real number's binary expansion is eventually periodic exactly when the number is rational, and none of the three index sets can be eventually periodic: the gap between consecutive squares, $(n+1)^2 - n^2 = 2n+1$, grows without bound, so no finite period could reproduce the spacing of the squares (or of their complement) forever. So (x_1, x_2, x_3) is a genuine triple of positive irrationals summing to 1, and it is unbalanced infinitely often – the answer to the question is no.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains reach the same reinterpretation of the rule: doubling shifts a binary expansion left by one digit, and subtracting 1 from the coordinate that exceeded $1/2$ deletes that coordinate's leading 1. Balanced, after this reinterpretation, means the current leading column of all three expansions is all zero. The chains split right there, at what question to ask of that reformulation: the natural attempt keeps asking "must a shift eventually reveal an all-zero column?" and looks for a reason it always must; the model chain reverses the question to "can I choose the digits myself so that a column is never all-zero?" and builds the sequences directly.

> *The question "does this always terminate" is really asking whether every triple of binary digit-strings, with exactly one* 1 *per unbalanced column, must eventually hit an all-zero column – and there is no law that forces that, so instead of searching for a reason it must happen, I should try to build three digit sequences where I personally control which coordinate gets the* 1 *in every single column, so that column is never all zero.*

Layer 3 THE WHY

The word "always" in a competition problem that is asking whether a recursive process reaches a terminal state, for every input satisfying only loose constraints (positive, irrational, sum to 1), is very often an invitation to search for a single counterexample rather than a general proof of termination – and that invitation gets much louder once the process has been converted, as here, into a purely combinatorial rule on symbol sequences with almost no remaining constraints. Once "balanced" is rephrased as "this column is all zero," the entire real-analytic process collapses into symbolic dynamics: three binary strings, constrained only by (i) exactly one 1 per column whenever that column is not yet balanced, and (ii) non-eventual-periodicity, imposed by the problem's own hypothesis that the x_i are irrational. Constraint (i) is forced by the arithmetic (sum $= 1$, positivity); constraint (ii) is the only place the problem restricts the construction at all – and it is a restriction the solver is free to satisfy directly, by design, rather than one that needs to be verified after the fact on some derived quantity.

This is also why the problem insists on irrational x_i in the first place: a rational triple's binary expansion is eventually periodic, and if the solver were allowed rationals, some "always balances eventually" phenomena tied to finite periodic patterns could muddy the question. By restricting to irrationals, the problem is quietly pointing at exactly the tool needed to solve it – the correspondence between $(0,1) \setminus \mathbb{Q}$ and non-eventually-periodic binary strings is the founding fact of binary expansions, not a special trick for this problem, and recognising that this correspondence is precisely what "irrational" buys you here is the axiom the whole construction rests on. Once the process is seen as three parallel copies of the doubling map (the classical map $T(x) = 2x \bmod 1$, whose iterates ARE the binary shift), the freedom to prescribe an itinerary becomes obvious: any $0/1$ sequence is realised by some point of $(0,1)$, so nothing stops a solver from choosing the three itineraries to interleave with exactly one 1 per column forever, as long as the resulting three points are irrational – which non-periodic index sets, such as squares and their complement, guarantee by their unbounded, ever-growing gaps.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is that the process only ever touches one coordinate per step – the one

currently $\geq 1/2$, which is unique because the three positive numbers sum to 1 – and doubles all three. On seeing "double everything, subtract 1 from the big one," write down: this is the binary-expansion shift map (the doubling map) applied in parallel to three numbers.

2. That cue suggests re-encoding each x_i in binary and tracking what the act does to the three digit strings: it deletes the shared leading digit column and shifts everything left.

3. Hesitation occurs right after a few numerical examples happen to balance quickly, which tempts a jump to "it always balances" and a search for a decreasing monovariant. Reinforce against this by asking: is there an actual reason this quantity must decrease, or does it merely decrease on the friendly examples I happened to try?

4. The trap is the friendly name "balancing act" itself, and the fact that hand-picked examples balance fast – both prime the solver to assume convergence is the intended answer. The honest hypothesis (irrational x_i) is the real hint: it signals that binary-expansion aperiodicity, not convergence, is the mechanism the problem wants exploited.

5. Recognise this faster next time by treating any recursive process of the form "double, then subtract the integer part" (or an equivalent piecewise-linear expanding map) as the binary-expansion shift map on sight, and immediately ask what freedom exists to prescribe digit sequences directly, before assuming the process must terminate.

6. Decision rule: to test whether a recursive process P always reaches some terminal state S, re-express P as an operation on symbolic digit sequences; if S corresponds to a simple, avoidable pattern (such as an all-zero column), construct a counterexample sequence directly rather than searching for a proof that S must eventually occur.

1977 B4

1977 B4. Let C be a continuous closed curve in the plane which does not cross itself and let Q be a point inside C. Show that there exist points P_1 and P_2 on C such that Q is the midpoint of the line segment P_1P_2.

Caveat. Statement checked against the MAA OCR (clean in this case) and against the book's own solution under the label B-4, which reflects C through Q and intersects – so the statement and label match. The book's solution argues that C and its reflection $-C$ must meet "since they have the same diameter and both surround 0 (hence neither can be exterior to the other)". As written this is compressed: "same diameter" needs the extra fact that the filled-in region has the same diameter as its boundary curve, and "neither can be exterior to the other" is asserted, not proved. The model chain below replaces diameter by maximum distance from Q, which needs no extra fact, and proves the "exterior" case in full. Both chains rest on the Jordan curve theorem, which the problem's word "inside" already assumes.

Fields: plane topology, geometry (point reflection, Jordan curve theorem) **Position:** B4 (mid-paper; short once you see it, hard to make rigorous if you don't) **Answer:** Such points always exist: put Q at the origin, take any P_1 lying on both C and its reflection $-C$ (the two curves must meet), and set $P_2 = -P_1$.

Layer 1 THE WHAT — surface analysis

The natural attempt. Put Q at the origin. Every line through Q passes from inside C to outside in both directions, so it meets C on both sides of Q. For a direction θ let $r(\theta)$ be the distance from Q to C along the ray at angle θ, and look at

$$f(\theta) = r(\theta) - r(\theta + \pi).$$

Then $f(\theta+\pi) = -f(\theta)$, so f takes both signs (or is zero) on $[0,\pi]$, and the intermediate value theorem gives θ_0 with $r(\theta_0) = r(\theta_0+\pi)$. The two hitting points are then equidistant from Q on opposite sides: done.

For a convex curve, or any curve star-shaped about Q, this is a complete proof – each ray meets C exactly once and r is continuous. Now draw a curve with a fold: a kidney shape, with Q in the fat part. A ray that grazes the inner bend meets C three times. Take $r(\theta)$ to be the *first* hit and it jumps as the ray swings past the tangent direction; take the *last* hit and it jumps at another tangent. Take "the" hit and it isn't defined. The intermediate value theorem needs a continuous f, and there is no continuous choice here. Worse, for a general continuous curve (think of a coastline that wiggles infinitely often) the set of hits along a ray can be infinite and the tangency picture means nothing. That's the wall: the method is right for the pictures you draw and has no proof for the curves the problem allows.

The model chain. Put Q at the origin 0. Write $\operatorname{int} C$ and $\operatorname{ext} C$ for the bounded and unbounded components of $\mathbb{R}^2 \setminus C$. The Jordan curve theorem, which is what gives "inside" its meaning, says these are the only two components and that C is the boundary of each, so the closure of $\operatorname{int} C$ is $\operatorname{int} C \cup C$. Let $C' = \{-X : X \in C\}$ be the reflection of C in 0. The map $X \mapsto -X$ is a homeomorphism of the plane, so C' is again a Jordan curve with $\operatorname{int} C' = -\operatorname{int} C$; in particular $0 \in \operatorname{int} C'$.

Reduction. If $P_1 \in C \cap C'$, then $P_2 := -P_1 \in C$, and the midpoint of P_1P_2 is $0 = Q$. Also $P_1 \neq P_2$, since $P_1 = -P_1$ would force $P_1 = 0$, but Q is inside C, not on it. So it suffices to show $C \cap C' \neq \varnothing$. Suppose, for contradiction, that $C \cap C' = \varnothing$. Then C' is a connected set missing C, so it lies entirely in $\operatorname{int} C$ or entirely in $\operatorname{ext} C$.

Step 1: C' is not inside C. C is compact, so $|X|$ attains a maximum R at some $P \in C$. The set $\{|X| > R\}$ is connected, unbounded and misses C, so it lies in $\operatorname{ext} C$. A point X with $|X| = R$ can't lie in $\operatorname{int} C$ either: $\operatorname{int} C$ is open, so a small disc round X would lie in it, and that disc contains points of modulus $> R$. Hence $\operatorname{int} C \subset \{|X| < R\}$. The point $-P \in C'$ has modulus R, so $-P \notin \operatorname{int} C$, and $-P \notin C$ by assumption. So $-P \in \operatorname{ext} C$, and therefore $C' \subset \operatorname{ext} C$.

Step 2: C' is not outside C either. Consider the connected open set $\operatorname{int} C'$, which contains 0. If it missed C, it would lie in one component of $\mathbb{R}^2 \setminus C$, namely $\operatorname{int} C$ (which contains 0). Taking closures, $C' \subset \overline{\operatorname{int} C'} \subset \operatorname{int} C \cup C$, and since C' misses C, $C' \subset \operatorname{int} C$, contradicting Step 1. So some $X \in C$ lies in $\operatorname{int} C'$. Reflect: $-X \in C'$ and $-X \in -\operatorname{int} C' = \operatorname{int} C$. So C' meets $\operatorname{int} C$, again contradicting Step 1.

Both cases are impossible, so $C \cap C' \neq \varnothing$, and the points exist. ■

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains start from the same line: "put Q at the origin; I want $P_1, P_2 \in C$ with $P_1 + P_2 = 0$." They split at the very next line. The natural attempt reads this as "find a chord through Q that Q bisects" and sets up a family of chords indexed by angle, which commits it to a function of θ that doesn't exist for general curves. The model reads $P_2 = -P_1$ as a statement about one point: P_1 is on C, and $-P_1$ is on C too. So P_1 lies on C and on $-C$. The problem has become "two closed curves meet", and that is a question about inside and outside, which is exactly the information the hypothesis supplies.

> *"Midpoint Q" means P_2 is the reflection of P_1 in Q. So I don't need a pair at all – I need one point of C whose reflection is also on C, which is a point where C meets its own reflection. Do those two curves have to cross?*

> *My function $r(\theta)$ isn't well defined once a ray hits the curve more than once. I shouldn't be patching the parametrisation – I should drop it and use only what the problem gives me: a curve, an inside, and a point in there.*

> *If the reflected curve sat entirely inside the original, its farthest point from Q would be strictly closer than the original's farthest point – but reflection doesn't change distances from Q. So it can't sit inside, and by the same symmetry it can't sit outside.*

Layer 3 THE WHY

Start from the definition of midpoint. Q is the midpoint of P_1P_2 exactly when $P_1 + P_2 = 2Q$, that is, $P_2 = 2Q - P_1$. The map $X \mapsto 2Q - X$ is the point reflection in Q. So the condition isn't really about pairs of points. It is one point P_1 satisfying two memberships: $P_1 \in C$ and $2Q - P_1 \in C$, i.e. $P_1 \in C \cap \rho(C)$ where ρ is the reflection. Any time a problem asks for P_1, P_2 on a set S with $P_2 = T(P_1)$ for a fixed invertible map T, the question is whether S meets $T^{-1}(S)$. That's all the move is, and it is forced by the definition. It is the same move as translating "find x with $f(x) = g(x)$" into "do two graphs intersect".

Next: why does intersection become easy to decide here? Look at what the problem actually gives you. It doesn't give a formula, differentiability or convexity. It gives "closed", "does not cross itself" and "inside". Those words carry only topological information, and the only tool that turns them into a conclusion is the Jordan curve theorem: a simple closed curve splits the plane into a bounded inside and an unbounded outside, and a connected set that avoids the curve stays on one side. So the proof had to be an inside/outside argument about two curves. The natural attempt used a parametrisation by angle, which is extra structure the hypotheses never promised. That's why it hit a wall.

Why the reflected curve can't sit on one side of the original: reflection in Q keeps distance from Q. So C and C' have the same farthest distance R from Q. A curve lying strictly inside C would have to stay strictly closer to Q than R, since the inside of C lies within the open disc of radius R. And "outside" is impossible because both curves enclose the same point Q, and two disjoint Jordan curves enclosing a common point must be nested – one inside the other – which puts us back in the first case. The book's phrase "same diameter" reaches for this invariant. Maximum distance from Q is the precise version, because reflection in Q preserves it pointwise, not only in aggregate.

The same argument appears elsewhere (Putnam and Beyond, problem 611). To inscribe an equilateral triangle in a closed curve with one vertex A on it, rotate the curve by 60° about A and intersect with the original. The relation between the unknown points is a rigid motion, so you apply that motion to the whole curve and ask for an intersection.

Finally, the hypothesis "Q inside" is sharp. If Q is far outside a small circle, the reflected circle is far away on the other side and no chord of the circle has midpoint Q. The inside/outside dichotomy is doing real work, not decoration.

Layer $+\alpha$ BACKTESTING

1. The word "midpoint" paired with "points on C". On seeing it, write down $P_2 = 2Q - P_1$, and with $Q = 0$, $P_2 = -P_1$.

2. Once the unknown pair is related by a fixed map (here the point reflection ρ), use the image-of-the-set method: the pair exists if and only if $C \cap \rho(C) \neq \varnothing$. Prove the intersection is non-empty with the topology the problem provides – inside/outside for a Jordan curve.

3. Hesitation comes in two places. The first is the pull towards the angle function $r(\theta)$, which works on every picture you draw. Get past it by asking whether r is continuous for a curve with a fold; it isn't. The second is Step 2 ("why can't C' be outside?"). Get past it with the fact that two disjoint Jordan curves around a common point are nested.

4. The trap is proving the convex case with the intermediate value theorem and believing you are done. The quieter trap is the book's own "same diameter", which is true but needs a fact about filled regions that it doesn't state. Maximum distance from Q avoids the gap.

5. Whenever the unknowns are several points on one set, tied together by a rigid motion or other fixed invertible map, apply that map to the whole set and look for an intersection, instead of parametrising candidates.

6. Decision rule: if the hypotheses are purely topological (closed, simple, inside, connected), the proof must be topological; don't bring in a parametrisation whose continuity you'd have to prove. Turn the existence claim into "two sets meet", and get the meeting from a connectedness or inside/outside dichotomy plus an invariant (here, maximum distance from Q) that rules out each side.

- ▷ **Cue:** The problem asks for two points on a curve with a given point as their midpoint $\longrightarrow$ **Move:** Write $P_2 = 2Q - P_1$; reflect the whole curve in Q and show the curve meets its reflection
- ▷ **Cue:** The unknown points on a set are related by a fixed rigid motion T (reflection, rotation, translation) $\longrightarrow$ **Move:** Replace "find a pair" with "show $S \cap T(S) \neq \varnothing$"
- ▷ **Cue:** A connected set is disjoint from a Jordan curve $\longrightarrow$ **Move:** It lies wholly inside or wholly outside; rule out each case separately
- ▷ **Cue:** You need to show one closed curve can't sit strictly inside another congruent-by-isometry curve $\longrightarrow$ **Move:** Compare an isometry-invariant extremal quantity, such as maximum distance from a fixed point that the isometry fixes
- ▷ **Cue:** An intermediate-value argument over an angle parameter needs a "distance to the curve along the ray" $\longrightarrow$ **Move:** Check it with a non-star-shaped curve first; if it breaks, abandon the parametrisation instead of patching it

1977 B5

1977 B5. Suppose that $a_1, a_2, \ldots, a_n$ are real $(n > 1)$ and

$$A + \sum_{i=1}^{n} a_i^2 < \frac{1}{n-1}\left(\sum_{i=1}^{n} a_i\right)^2.$$

Prove that $A < 2a_i a_j$ for $1 \leq i < j \leq n$.

Caveat. The MAA OCR of the statement drops the conclusion entirely: the problem text ends at the displayed hypothesis. The conclusion "Prove that $A < 2a_i a_j$ for $1 \leq i < j \leq n$" is taken from the book's own solution under the label 1977 B-5, which ends "Similarly, $A < 2a_i a_j$ for $1 \leq i < j \leq n$" and proves exactly that from exactly this hypothesis (by Cauchy–Schwarz on $a_1 + a_2, a_3, \ldots, a_n$). The statement and the solution therefore match, and this is the standard published wording of 1977 B5. The OCR of the hypothesis ("$A + \sum a_i^2 < \frac{1}{n-1}(\sum a_i)^2$") is legible and is used as printed.

Fields: inequalities, algebra (Cauchy–Schwarz / QM–AM) **Position:** B5 **Answer:** True for every pair: Cauchy–Schwarz on the $n-1$ numbers $a_i + a_j$ and the remaining a_k gives $\frac{1}{n-1}(\sum a_k)^2 \leq \sum a_k^2 + 2a_i a_j$, so $A < 2a_i a_j$.

Layer 1 THE WHAT — surface analysis

The natural attempt. The first instinct is to get rid of the square. Expand $(\sum a_k)^2 = \sum a_k^2 + 2\sum_{k<l} a_k a_l$ and move everything to one side. The hypothesis becomes

$$A < \frac{1}{n-1}\Big(\sum_k a_k^2 + 2\sum_{k<l} a_k a_l\Big) - \sum_k a_k^2 = \frac{2\sum_{k<l} a_k a_l - (n-2)\sum_k a_k^2}{n-1}.$$

Now you want this right-hand side to be at most $2a_1 a_2$ (fix $i = 1$, $j = 2$; the hypothesis is symmetric in the a_k, so this is no loss). Multiplying through by $n-1$, the target is

$$2\sum_{k<l} a_k a_l - (n-2)\sum_k a_k^2 \leq 2(n-1)a_1 a_2. \qquad (*)$$

This is where the attempt gets heavy. You can check $n = 2$: the left side is $2a_1a_2$ and the right is $2a_1a_2$, so equality holds. For $n = 3$ you get $2a_1a_2 + 2a_1a_3 + 2a_2a_3 - a_1^2 - a_2^2 - a_3^2 \le 4a_1a_2$, i.e. $0 \le a_1^2 + a_2^2 + a_3^2 + 2a_1a_2 - 2a_1a_3 - 2a_2a_3 = (a_1 + a_2 - a_3)^2$. That works, and it's a strong hint, but for general n you're now hunting for a sum of squares that equals the difference of the two sides of $(*)$, with $\binom{n}{2}$ cross terms of which one pair (a_1a_2) behaves differently from all the others. People do find it, after a page of bookkeeping. Others try Lagrange multipliers or "maximise the right-hand side for fixed a_1a_2" and get lost in cases. The attempt isn't wrong. It's zombie maths: expansion first, structure never.

The model chain. Read the hypothesis as a comparison between a squared sum and a sum of squares, weighted by $\frac{1}{n-1}$. That is the shape of Cauchy–Schwarz (equivalently QM–AM) for $n - 1$ *numbers*:

$$(b_1 + \cdots + b_m)^2 \le m\,(b_1^2 + \cdots + b_m^2), \qquad m = n - 1.$$

There are n numbers a_k but only $n - 1$ slots, so merge two of them. Take $b_1 = a_i + a_j$ and let $b_2, \ldots, b_{n-1}$ be the other $n - 2$ values a_k $(k \ne i, j)$. Then $\sum b = \sum_k a_k$, and

$$\sum b^2 = (a_i + a_j)^2 + \sum_{k \ne i,j} a_k^2 = \sum_{k=1}^{n} a_k^2 + 2a_ia_j.$$

Cauchy–Schwarz with $m = n - 1$ (valid because $n > 1$, so $m \ge 1$) gives

$$\frac{1}{n-1}\Big(\sum_{k=1}^{n} a_k\Big)^2 \le \sum_{k=1}^{n} a_k^2 + 2a_ia_j.$$

Chain this with the hypothesis:

$$A < \frac{1}{n-1}\Big(\sum a_k\Big)^2 - \sum a_k^2 \le 2a_ia_j.$$

The first inequality is strict, so $A < 2a_ia_j$. The pair (i, j) was arbitrary, so this holds for every $1 \le i < j \le n$. That is the whole proof.

For completeness, here is the inequality with no theorem quoted at all. Lagrange's identity gives

$$m\sum_{p=1}^{m} b_p^2 - \Big(\sum_{p=1}^{m} b_p\Big)^2 = \sum_{1 \le p < q \le m} (b_p - b_q)^2 \ge 0,$$

so

$$2a_ia_j - \Big[\tfrac{1}{n-1}\Big(\sum a_k\Big)^2 - \sum a_k^2\Big] = \tfrac{1}{n-1}\sum_{p<q}(b_p - b_q)^2 \ge 0.$$

(I checked this identity numerically on 2×10^5 random real vectors with $2 \le n \le 7$; it holds to machine precision.) Equality occurs exactly when $a_i + a_j = a_k$ for every $k \ne i, j$. For $n = 3$ this is the square $(a_1 + a_2 - a_3)^2$ that the natural attempt stumbled onto.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains start from the same inequality. The split is the first line. The natural attempt writes $(\sum a_k)^2 = \sum a_k^2 + 2\sum_{k<l} a_ka_l$ and commits to expansion. The model reads the unexpanded left-hand side, $\frac{1}{n-1}(\sum a_k)^2$, as "squared sum divided by a count", recognises it as one side of Cauchy–Schwarz, and asks which $n - 1$ numbers it should be the count of.

> *That's a squared sum over a count, so it's one half of Cauchy–Schwarz. But the count is $n - 1$ and I have n numbers, so I need to glue two of them into one. Which two? The ones in the conclusion, a_i and a_j. Squaring $a_i + a_j$ produces exactly the $2a_ia_j$ I'm trying to reach.*

> *Before I expand anything, I'll check whether the two sides are already the two sides of a standard inequality. Expanding throws away the shape I'd need to recognise it.*

Layer 3 THE WHY

Why is Cauchy–Schwarz the natural move here and not a trick? Go back to what the inequality says. For m real numbers, $(\sum b)^2 \le m \sum b^2$ is the statement that the mean squared deviation from the average is non-negative:

$$\frac{1}{m}\sum b_p^2 - \Big(\frac{1}{m}\sum b_p\Big)^2 = \frac{1}{m}\sum_p \big(b_p - \bar{b}\big)^2 \ge 0.$$

That's variance ≥ 0, which is as close to an axiom as algebra gets: a square is never negative. So an expression of the form $\frac{1}{m}(\sum b)^2$ next to $\sum b^2$ is a variance in disguise. The problem's hypothesis puts exactly those two expressions side by side, which is why it's natural to read them that way.

Now the mismatch. A variance over n numbers carries the weight $\frac{1}{n}$, but the problem uses $\frac{1}{n-1}$. That mismatch is the entire content of the problem. It isn't noise to smooth over. With weight $\frac{1}{n}$ the inequality $\frac{1}{n}(\sum a)^2 \le \sum a^2$ holds with no help. With the larger weight $\frac{1}{n-1}$ it can fail, and it fails by at most the price of treating two of the numbers as one. Merging a_i, a_j into a single slot costs exactly the cross term: $(a_i + a_j)^2 - a_i^2 - a_j^2 = 2a_ia_j$. So the $n-1$ tells you to merge, and the $2a_ia_j$ tells you which pair to merge. Once you see the problem is about counting slots, merging the pair isn't a leap. It's the only way to turn n numbers into $n-1$ that leaves a trace of a_i and a_j in the result.

The official solution writes the Cauchy–Schwarz with $(a_1 + a_2), a_3, \ldots, a_n$ and $n-1$ ones without saying where the grouping comes from. It comes from here: the denominator counts the slots, and the conclusion names the merged pair.

Why expansion is the wrong first move: it takes a variance, a manifestly non-negative object, and turns it into an unstructured sum of $\binom{n}{2}$ cross terms minus squares. Then you have to rebuild the sum of squares that you destroyed. The $n = 3$ square $(a_1 + a_2 - a_3)^2$ is that rebuilt variance, $\sum_{p<q}(b_p - b_q)^2$ with $b = (a_1 + a_2, a_3)$.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the unexpanded expression $\frac{1}{n-1}\big(\sum a_i\big)^2$ sitting next to $\sum a_i^2$. On seeing it, write down $(b_1 + \cdots + b_m)^2 \le m \sum b_p^2$ and ask what m and the b_p must be.
2. Cauchy–Schwarz (QM–AM) with $m = n - 1$. The count mismatch against the n given numbers tells you to merge two of them into one slot.
3. Hesitation comes at the pull to expand the square, because expansion feels like progress. Reinforce against it with a rule: never expand a squared sum until you've checked whether it's already one side of Cauchy–Schwarz or a variance.
4. The trap is the $\frac{1}{n-1}$ where your memory expects $\frac{1}{n}$. It tempts you either to dismiss it as a normalisation quirk or to start expanding to see what it does. It's the signal that one slot has been lost, i.e. that two variables must share a slot.
5. Look for the count in the denominator and compare it with the number of variables. When the count is too small by k, you need to merge $k+1$ quantities into fewer slots, and the conclusion usually names which ones. The merge generates the cross terms the conclusion is built from.
6. Decision rule: when a hypothesis or target compares $\frac{1}{m}(\sum b)^2$ with $\sum b^2$, treat it as Cauchy–Schwarz or variance ≥ 0 over exactly m quantities. If the problem supplies more than m variables, group them into m blocks, choosing the blocks whose cross terms appear in the conclusion. If the problem supplies fewer, pad with the variable the conclusion needs.

- ▷ **Cue:** a squared sum divided by a count, $\frac{1}{m}\left(\sum b\right)^2$, stands beside a sum of squares $\sum b^2$ $\longrightarrow$ **Move:** read it as one side of Cauchy–Schwarz / QM–AM (variance ≥ 0) over exactly m quantities; do not expand
- ▷ **Cue:** the count m in the denominator is one less than the number of variables, and the conclusion is about a product $2a_ia_j$ $\longrightarrow$ **Move:** merge a_i and a_j into the single quantity $a_i + a_j$; squaring it produces $a_i^2 + a_j^2 + 2a_ia_j$, exactly the extra term needed
- ▷ **Cue:** you need to prove $(\sum b)^2 \leq m\sum b^2$ with no theorems cited $\longrightarrow$ **Move:** use Lagrange's identity $m\sum b^2 - (\sum b)^2 = \sum_{p<q}(b_p - b_q)^2$

1977 B6

1977 B6. Let H be a subgroup with h elements in a group G. Suppose that G has an element a such that for all x in H, $(xa)^3 = 1$, the identity. In G, let P be the subset of all products $x_1ax_2a\cdots x_na$, with n a positive integer and the x_i in H.

(a) Show that P is a finite set.

(b) Show that, in fact, P has no more than $3h^2$ elements.

Caveat. The OCR renders the hypothesis as "$(xa)3 - I$, the identity"; this is $(xa)^3 = 1$, confirmed by the book's solution, which expands it as $xaxaxa = 1$ and derives $a^{-1} = a^2$ from it. The book's solution under the label B-6 proves exactly the bound $3h^2$ for exactly this set P, so the statement is correctly matched to 1977 B6. The proof below was constructed independently, then compared with the book's; the two agree in substance. The rewriting identities and the final bound were also checked by brute force on every admissible pair (H, a) found by random search in S_5 (26 cases, including the case $H = \{1\}$, where $|P| = 3 = 3h^2$ and the bound is attained).

Fields: group theory, word combinatorics **Position:** B6 **Answer:** Every element of P has one of the three shapes xay, xa^2y, xa^2ya with $x, y \in H$; each shape takes at most h^2 values, so $|P| \leq 3h^2$, which also proves (a). The bound is attained when $H = \{1\}$.

Layer 1 THE WHAT — surface analysis

The natural attempt. You read "products of arbitrary length" and "show finite", and you start manipulating words. From $x = 1$ (which is in H, as H is a subgroup) you get $a^3 = 1$. Then you expand $(xa)^3 = 1$ as $xaxaxa = 1$ and try to shorten $x_1ax_2ax_3a$. You rearrange $xaxaxa = 1$ into $axaxa = x^{-1}$, then into $axa = x^{-1}a^{-1}x^{-1}\cdots$ and wander. The trouble is that every identity you extract has the letter x repeated, while a general word $x_1ax_2ax_3a$ has three *different* letters of H. So you cannot apply $xaxaxa = 1$ to it directly, and you fall back on hoping that $\langle H, a\rangle$ is a finite group, which you then try to prove by listing cosets. That route is a wall: nothing in the hypotheses tells you $\langle H, a\rangle$ is finite before you have done the work, and in any case a finite bound on $\langle H, a\rangle$ would not produce the specific number $3h^2$. You have been manipulating words with no target shape in mind.

The model chain. Write 1 for the identity. Since $1 \in H$, the hypothesis gives $a^3 = 1$, so $a^{-1} = a^2$.

Step 1: turn the hypothesis into a rewriting rule with one free letter. For $x \in H$, $(xa)^3 = 1$ says $(xa)^{-1} = (xa)^2$, that is,

$$a^{-1}x^{-1} = xaxa \quad\Longrightarrow\quad a^2x^{-1} = xaxa \quad\Longrightarrow\quad axa = x^{-1}a^2x^{-1}. \tag{i}$$

Taking inverses of both sides of (i) gives $a^{-1}x^{-1}a^{-1} = x\,a^{-2}\,x$, that is $a^2x^{-1}a^2 = xax$; since H is closed under inverses we may replace x by x^{-1}:

$$a^2xa^2 = x^{-1}ax^{-1}. \tag{ii}$$

Rule (i) replaces the pattern $a \square a$ by a single a^2 flanked by elements of H; rule (ii) replaces $a^2 \square a^2$ by a single a flanked by elements of H.

Step 2: the three target shapes. Define

$$A = \{xay\}, \quad B = \{xa^2y\}, \quad C = \{xa^2ya\} \qquad (x, y \in H),$$

and $Q = A \cup B \cup C$. Each set is the image of $H \times H$ under a map, so each has at most h^2 elements and $|Q| \leq 3h^2$.

Step 3: induction on n. Let $p_n = x_1 a \cdots x_n a$. For $n = 1$, $p_1 = x_1 a \cdot 1 \in A$. Suppose $q = p_k \in Q$ and let $z = x_{k+1} \in H$, so $p_{k+1} = qza$.

- If $q = xay \in A$: $qza = xa(yz)a = x(yz)^{-1}a^2(yz)^{-1} \in B$ by (i).
- If $q = xa^2y \in B$: $qza = xa^2(yz)a \in C$ directly.
- If $q = xa^2ya \in C$: using (i) on aza and then (ii) with $w = yz^{-1} \in H$,

$$qza = xa^2y(aza) = xa^2yz^{-1}a^2z^{-1} = x\,(a^2wa^2)\,z^{-1} = xw^{-1}aw^{-1}z^{-1} \in A.$$

So $p_n \in Q$ for all n, $P \subseteq Q$, and $|P| \leq 3h^2$. Part (a) follows, as $3h^2$ is finite. The induction shows more: p_n lies in A, B, C according as $n \equiv 1, 2, 0 \pmod 3$.

Sharpness. If $H = \{1\}$ then $P = \{a, a^2, 1\}$ and a may have order exactly 3 (take $G = \mathbb{Z}/3$), so $|P| = 3 = 3h^2$.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains derive $a^3 = 1$ and expand $(xa)^3 = 1$. They split at what to do with that expansion. The natural attempt treats $xaxaxa = 1$ as a fact about words with a repeated letter and looks for places in $x_1 a x_2 a \cdots$ where it applies verbatim, finding none. The model solves $xaxaxa = 1$ for the *middle piece*, isolating axa so that the other side contains no a adjacent to another a: rule (i), $axa = x^{-1}a^2x^{-1}$. That is a rule with a single free letter, and a single free letter is exactly what the pattern $a\, x_k\, a$ in the middle of an arbitrary word supplies.

> *My words have different letters of H everywhere, so an identity with x repeated three times will never match. I need an identity whose left-hand side is a short window I can see inside any word – a, one letter of H, a – and whose right-hand side has fewer places where a follows a.*

> *The count $3h^2$ is telling me the answer's shape: three families, each indexed by two elements of H. So I should guess three normal forms with two free H-letters each, and prove that appending za moves me from one family to another.*

> *Every element of P ends in a, and appending one more za is the only operation. So I do not need to understand the whole group; I only need a set closed under right multiplication by Ha that contains Ha.*

Layer 3 THE WHY

Why isolate axa? Because the only thing that makes a word long is the number of a's, and H-letters between them are free. Think of a word as an alternating string $h_0 a^{e_1} h_1 a^{e_2} \cdots$ with each $e_i \in \{1, 2\}$ (since $a^3 = 1$, no other exponent is needed). The length that matters is the number of a-blocks. The hypothesis $(xa)^3 = 1$ is a statement about three a-blocks; any usable consequence must shrink the number of blocks locally, and the shortest window containing two blocks is $a\, x\, a$. Solve for that window and you get (i): two blocks of exponent 1 become one block of exponent 2. Invert and you get (ii): two blocks of exponent 2 become one block of

exponent 1. That is not a trick; it is the only local reduction the hypothesis can give, found by asking "what is the smallest window with two a's in it?"

Now the whole problem is a counter. Reading exponents only, appending za appends a block 1. The rules say $(1,1) \to (2)$ and $(2,2) \to (1)$ – both preserve the exponent sum modulo 3. Start from (1). Append 1: $(1,1) \to (2)$. Append 1: $(2,1)$, no rule applies. Append 1: $(2,1,1) \to (2,2) \to (1)$. Back to the start. The block pattern cycles through (1), (2), $(2,1)$ with period 3 – these are exactly A, B, C. The number of a-blocks never exceeds two, and the H-letters sitting between and around them are the only freedom. A pattern with one block has two H-slots (xay or xa^2y); the pattern $(2,1)$ has three slots, $xa^2ya\,w$, but since every element of P ends in a, the last slot is empty. Two free slots per pattern, three patterns: $3h^2$. The bound in the problem is the count of the normal forms, and the exponent sum mod 3 is why there are three of them. Look at it this way and the three sets in the official solution stop looking pulled from a hat.

The book's solution introduces A, B, C without comment and then verifies closure. Here is where they come from: run the append-and-reduce process on exponent patterns alone, and the three patterns it visits are the three sets.

Two phrasings in the statement earn their place. "Subgroup" is needed twice: $1 \in H$ gives $a^3 = 1$, and closure under inverses licenses substituting x^{-1} to get (ii). And "P" consists only of words ending in a, which is what removes the third H-slot from pattern C; without it you would get h^3 from the $(2,1)$ pattern and lose the bound.

Layer $+\alpha$ BACKTESTING

1. The first cue is "$(xa)^3 = 1$ for all $x \in H$" together with words alternating between H and a. Write down immediately: $x = 1$ gives $a^3 = 1$; then solve $(xa)^3 = 1$ for axa.

2. A rewriting system: find local reductions that shorten words, then find the finitely many irreducible shapes and count them.

3. Hesitation comes at "the letters x_i are all different, so the relation never applies." It is reinforced away by asking which *window* of a general word contains only one H-letter: $a\,x_k\,a$. Relations must be rewritten to act on such windows.

4. The trap is trying to prove $\langle H, a \rangle$ is finite, or to classify the group. That is harder than the question and does not yield $3h^2$. The question is about the set P only, and P is generated by right multiplication by Ha; closure of a candidate set under that one operation is all that must be checked.

5. When a bound like $3h^2$ is given, factor it as (number of shapes) $\times$ (free parameters per shape). Then, when a relation of the form $(\text{word})^k = 1$ holds, solve it for the shortest window that contains two copies of the repeated generator; the result is a length-reducing rewriting rule.

6. Decision rule: to bound a set of words built by repeatedly appending a generator, do not study the ambient group. Guess normal forms from the shape of the claimed bound, and prove the union of the normal-form sets is closed under the append operation. A conserved quantity of the rewriting (here, total exponent mod 3) tells you how many forms there are.

▷ **Cue:** a relation $(xa)^k = 1$ holding for every x in a subgroup, and words $x_1ax_2a\cdots$ with distinct letters $\longrightarrow$ **Move:** solve the relation for the window axa (or a^2xa^2) to get a rule with one free letter that merges two a-blocks into one

▷ **Cue:** a claimed bound of the form $c \cdot h^m$ on a set of words over H and a fixed element $\longrightarrow$ **Move:** guess c normal forms with m free H-slots each, then prove closure of their union under the generating operation by induction

▷ **Cue:** a set defined as all products of arbitrary length ending in the same generator $\longrightarrow$ **Move:** prove it lies inside a finite candidate set closed under right multiplication by the generating block; never try to prove the whole group finite first