DAWN PUTNAM SESSION

Putnam 2003 — Papers A and B

Putnam 120 Archive

Contents

Putnam 2003 — Paper A **5**
2003 A1 5
2003 A2 7
2003 A3 9
2003 A4 12
2003 A5 15
2003 A6 17

Putnam 2003 — Paper B **21**
2003 B1 21
2003 B2 23
2003 B3 25
2003 B4 27
2003 B5 29
2003 B6 32

Putnam 2003 — Paper A

2003 A1

2003 A1. Let n be a fixed positive integer. How many ways are there to write n as a sum of positive integers, $n = a_1 + a_2 + \cdots + a_k$, with k an arbitrary positive integer and $a_1 \le a_2 \le \cdots \le a_k \le a_1 + 1$? For example, with $n = 4$ there are four ways: 4, 2+2, 1+1+2, 1+1+1+1.

Fields: combinatorics, elementary number theory **Position:** A1 — the opener **Answer:** Exactly n ways: for every $k = 1, \ldots, n$ there is precisely one such representation with k parts, and no other k works.

Layer 1 THE WHAT — surface analysis

The natural attempt. The worked example is organised by $n = 4$: it lists the four sums directly. The pull is to keep organising by n — to try to build the count for n out of the count for $n - 1$. One natural idea, visible in how the example reads (4, then 2+2, then 1+1+2, then 1+1+1+1), is that each representation of n looks like a representation of $n - 1$ with one part "augmented" — bump some 1 up to a 2, or a 2 up to a 3, and so on, whichever move keeps the near-equal condition intact.

Try to make this precise. Given a valid representation of $n - 1$, which single entry do you increment to reach a valid representation of n? Sometimes it's forced (the all-ones sum $1 + 1 + \cdots + 1$ of $n - 1$ ones must become n ones plus nothing — actually it has to gain a new part, not have an existing one incremented, since incrementing any single 1 in a run of $n - 1$ ones to a 2 breaks $a_1 \le \cdots \le a_k$ unless the 2 is at the very end). Sometimes there's a genuine choice: for $2 + 2$ ($n = 4$), incrementing either copy of 2 to 3 gives $2 + 3$ (invalid — $n = 5$, but $3 - 2 = 1$, wait $2 \le 3 \le 2 + 1 = 3$, so $2 + 3$ is actually valid for $n = 5$). So the "augment" idea does produce valid children, but multiple parents can plausibly reach the same child, and some children (a brand new smallest-part representation) need to be added by fiat rather than by incrementing anything. Pinning down a clean bijection this way means casework on where the run of a_1's ends, which is exactly the information the model chain below extracts directly — but reached by trial and error on an induction that was never guaranteed to be the efficient description in the first place.

The model chain. Organise not by n but by k, the number of parts — fix k first and ask how many valid representations of n have exactly k parts.

Suppose $n = a_1 + \cdots + a_k$ with $a_1 \le a_2 \le \cdots \le a_k \le a_1 + 1$. Every part equals either a_1 or $a_1 + 1$, so

$$ka_1 = a_1 + \cdots + a_1 \;\le\; n \;\le\; a_1 + (a_1 + 1) + \cdots + (a_1 + 1) = ka_1 + (k - 1).$$

This pins a_1 down to a unique integer: $a_1 = \lfloor n/k \rfloor$. (Two integers satisfying $ka_1 \le n \le ka_1 + k - 1$ would have to differ by less than 1.)

Once a_1 is fixed, the representation is determined by where the run of a_1's ends: if $a_i = a_1$ for $i \le m$ and $a_i = a_1 + 1$ for $i > m$, the sum is $ka_1 + (k - m)$. As m ranges over $0, 1, \ldots, k$ this hits $k + 1$ values, but $m = k$ and the "next" case coincide in total only once n is fixed —

concretely, exactly one value of $k - m \in \{0, \ldots, k-1\}$ satisfies $ka_1 + (k - m) = n$, since $n - ka_1$ is a fixed integer in $\{0, \ldots, k-1\}$ by the squeeze above. So for each k there is exactly one valid representation.

Finally, which k are achievable? Since $a_1 \geq 1$, $ka_1 \leq n$ forces $k \leq n$. And every $k \in \{1, \ldots, n\}$ is achieved by the construction above (take $a_1 = \lfloor n/k \rfloor \geq 1$ since $k \leq n$, and set the correct number of parts, forced by the remainder $n - ka_1$, to $a_1 + 1$). So k ranges over exactly $\{1, \ldots, n\}$, contributing one representation each, for a total of n.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains want a bijective or constructive handle on the count. They split on which variable to hold fixed first. The natural attempt holds n fixed and tries to induct on it, augmenting representations of $n - 1$ — an approach organised around the quantity being decomposed. The model chain holds k fixed instead, and asks how many representations of the fixed target n use exactly k parts — an approach organised around the constraint's own free parameter.

> *The condition $a_1 \leq \cdots \leq a_k \leq a_1 + 1$ doesn't bound the parts in absolute size — it bounds how much they can vary from each other. That means once I fix how many parts there are, only two values are even available for each part, so the whole representation should be forced by arithmetic (a floor and a remainder), not something I need to search for or induct into existence.*

Layer 3 THE WHY

Why does fixing k collapse the problem to a triviality while fixing n (organising by induction) does not? Because the constraint $a_k \leq a_1 + 1$ is a statement about the *spread* of the parts, not their total. A constraint on spread only becomes informative once the number of parts is known: with k parts each within 1 of each other, the total n satisfies $ka_1 \leq n \leq ka_1 + (k-1)$, a window of width exactly $k - 1$ that contains at most one integer per value of a_1 — the squeeze is tight enough to force uniqueness immediately. Fixing n instead and asking how a representation grows leaves the spread condition doing no comparable work: incrementing a part of a representation of $n - 1$ can violate $a_1 \leq \cdots \leq a_k$ depending on which part you touch, and nothing about the induction variable n interacts directly with the two-values-only structure the way k does.

Put differently: the phrase "arbitrary positive integer k" in the problem is not incidental bookkeeping to be summed over at the end — it is the natural coordinate the constraint is stated in terms of, since $a_1, \ldots, a_k \leq a_1 + 1$ is a condition on a k-tuple. Treating k as the object to range over, with n fixed throughout, matches the constraint to the variable it actually governs. That the final count is exactly n (one representation per $k = 1, \ldots, n$) is then not a coincidence to verify after the fact; it falls straight out of the range of achievable k, which is itself forced by $a_1 \geq 1$.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the shape of the constraint itself: $a_1 \leq a_2 \leq \cdots \leq a_k \leq a_1 + 1$ bounds the spread between parts, not their absolute size. On seeing it, write down: with the number of parts fixed, only two distinct values are possible for any part, so the whole tuple should be forced by a small amount of arithmetic.

2. That cue suggests fixing k and deriving $a_1 = \lfloor n/k \rfloor$ from the squeeze $ka_1 \leq n \leq ka_1 + (k-1)$, then locating the unique split point between the a_1's and the $(a_1 + 1)$'s from the remainder $n - ka_1$.

3. Hesitation occurs right after reading the worked example, which is presented as a list indexed by $n = 4$ and invites organising the whole argument by n via induction. This is reinforced away by asking what the constraint is actually a condition on — it is a condition on a k-tuple, so k, not n, is the variable that makes it informative.

4. The trap is trying to build a bijection between representations of $n-1$ and n by incrementing a single part, without a rule for which part to increment; some representations need a wholly new part added rather than an existing one bumped, and the rule for when each applies is exactly the casework the fixed-k approach avoids.

5. Recognise this family faster by checking, on any problem with a constraint on the spread (not the size) of a tuple's entries, whether fixing the tuple's length first turns the constraint into a two-valued forced arithmetic statement.

6. Reusable rule: when a counting constraint bounds how much a set of quantities can differ from one another, rather than their absolute magnitudes, organise the count by the number of quantities involved, not by the target total — the spread condition becomes a tight squeeze (often pinning everything down uniquely) exactly once that count is fixed.

▷ **Cue:** a decomposition constraint bounds the spread between parts (e.g. $a_k \le a_1 + 1$) rather than their absolute size $\longrightarrow$ **Move:** fix the number of parts k first; the spread condition then squeezes the smallest part to a unique value via floor/ceiling arithmetic

▷ **Cue:** a counting problem's worked example is presented indexed by the total n $\longrightarrow$ **Move:** do not default to inducting on n; check whether some other parameter appearing in the constraint (here the number of parts k) is the one the constraint actually governs

2003 A2

2003 A2. Let $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ be nonnegative real numbers. Show that

$$(a_1 a_2 \cdots a_n)^{1/n} + (b_1 b_2 \cdots b_n)^{1/n} \le [(a_1 + b_1)(a_2 + b_2) \cdots (a_n + b_n)]^{1/n}.$$

Fields: inequalities, AM-GM **Position:** A2 **Answer:** The inequality holds for all nonnegative a_i, b_i, with equality when some $a_i = b_i = 0$ or when $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_n)$ are proportional; proved by normalising each pair a_i, b_i by $a_i + b_i$ and applying AM-GM to the two resulting weight sequences.

Layer 1 THE WHAT — surface analysis

The natural attempt. The shape "geometric mean of a's plus geometric mean of b's $\le$ geometric mean of $(a+b)$'s" screams AM-GM, and the natural first move is to apply it to the raw quantities directly: AM-GM gives $(a_1 \cdots a_n)^{1/n} \le \frac{1}{n}\sum a_i$ and similarly for b. Adding, $(a_1 \cdots a_n)^{1/n} + (b_1 \cdots b_n)^{1/n} \le \frac{1}{n}\sum(a_i + b_i)$. This is a true statement, but it bounds the left side by the *arithmetic* mean of $a_i + b_i$, not the *geometric* mean $[(a_1 + b_1) \cdots (a_n + b_n)]^{1/n}$ that the problem actually wants — and the arithmetic mean is always at least the geometric mean, so this bound is in the wrong direction to be useful; it proves something weaker than what was asked, not the target itself.

A second instinct is to attack $n = 2$ concretely — $\sqrt{a_1 a_2} + \sqrt{b_1 b_2} \le \sqrt{(a_1 + b_1)(a_2 + b_2)}$ — where squaring both sides and simplifying reduces to $2\sqrt{a_1 a_2 b_1 b_2} \le a_1 b_2 + a_2 b_1$, itself an instance of AM-GM (or Cauchy–Schwarz applied to $(\sqrt{a_1}, \sqrt{a_2})$ and $(\sqrt{b_2}, \sqrt{b_1})$). This works and is a genuine base case, but there is no clean way to bootstrap it to general n by pairing up indices two at a time unless n is a power of 2 — for $n = 3$ or $n = 5$ the pairing runs out, and patching the gap requires either a separate argument or an induction whose inductive step is exactly the harder general-n problem in miniature. The $n = 2$ case, however satisfying, does not by itself supply a route to all n.

The model chain. Assume $a_i + b_i > 0$ for every i (if some $a_i = b_i = 0$, both sides of the target inequality vanish, so this case is immediate). Normalise: apply the AM-GM inequality (geometric mean $\le$ arithmetic mean) to the n nonnegative numbers $\frac{a_1}{a_1 + b_1}, \ldots, \frac{a_n}{a_n + b_n}$:

$$\left(\frac{a_1}{a_1 + b_1} \cdots \frac{a_n}{a_n + b_n}\right)^{1/n} \le \frac{1}{n}\sum_{i=1}^{n} \frac{a_i}{a_i + b_i}.$$

Apply the same inequality to $\frac{b_1}{a_1+b_1}, \ldots, \frac{b_n}{a_n+b_n}$:

$$\left(\frac{b_1}{a_1+b_1} \cdots \frac{b_n}{a_n+b_n}\right)^{1/n} \leq \frac{1}{n}\sum_{i=1}^{n} \frac{b_i}{a_i+b_i}.$$

Add the two inequalities. On the right, for each i, $\frac{a_i}{a_i+b_i} + \frac{b_i}{a_i+b_i} = 1$ exactly, so the sum of the two right-hand sides is $\frac{1}{n}\sum_{i=1}^{n} 1 = 1$. So

$$\left(\prod_{i=1}^{n} \frac{a_i}{a_i+b_i}\right)^{1/n} + \left(\prod_{i=1}^{n} \frac{b_i}{a_i+b_i}\right)^{1/n} \leq 1.$$

Both product terms split as a ratio of products, $\left(\prod a_i\right)^{1/n} / \left(\prod (a_i+b_i)\right)^{1/n}$ and likewise for b, since each factor $a_i + b_i > 0$. Multiplying through by $\left(\prod_{i=1}^{n}(a_i+b_i)\right)^{1/n}$ gives exactly

$$(a_1 \cdots a_n)^{1/n} + (b_1 \cdots b_n)^{1/n} \leq \left[(a_1+b_1)\cdots(a_n+b_n)\right]^{1/n}.$$

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains reach for AM-GM immediately — the shape of the problem all but names the tool. They split on *what* AM-GM is applied to. The natural attempt applies it to $a_1, \ldots, a_n$ and to $b_1, \ldots, b_n$ as they stand, producing a bound in terms of $\sum(a_i + b_i)$ — the wrong mean of the wrong quantity. The model chain first divides every a_i, b_i by $a_i + b_i$, and applies AM-GM to those ratios instead.

> *If I divide a_i and b_i by their own sum $a_i + b_i$ before doing anything else, I've turned two numbers that live on an arbitrary scale into two numbers that are pieces of a partition of 1 — and now the average I get out of AM-GM is a number I can compute exactly, not one I still have to bound, because $\frac{a_i}{a_i+b_i} + \frac{b_i}{a_i+b_i} = 1$ for every single i, so summing over i and dividing by n gives exactly 1, not an unknown quantity I'd have to control separately.*

Layer 3 THE WHY

Why does normalising by $a_i + b_i$ turn a weak, wrong-direction bound into an exact one? AM-GM says (geometric mean) $\leq$ (arithmetic mean); the inequality is only as useful as the arithmetic mean it produces. Applied to raw a_i, the arithmetic mean is $\frac{1}{n}\sum a_i$ — a quantity with no fixed relationship to the target's normalising factor $\left[(a_1+b_1)\cdots(a_n+b_n)\right]^{1/n}$, which is why the direct approach stalls: it bounds the left side by something incommensurate with the right side, not by a piece of it. The target inequality, read as "GM of a plus GM of $b \leq$ GM of $a+b$", becomes, after dividing every term by $\left(\prod(a_i+b_i)\right)^{1/n}$, exactly "GM of $\frac{a_i}{a_i+b_i}$ plus GM of $\frac{b_i}{a_i+b_i} \leq 1$" — and 1 is not an arbitrary target, it is the one number that $\frac{a_i}{a_i+b_i} + \frac{b_i}{a_i+b_i}$ is guaranteed to sum to, termwise, for every i, by definition of the two ratios. So the normalisation is not a trick pulled from nowhere: it is forced by asking what quantity would make the two separate AM-GM applications add up to something known rather than something that still needs bounding, and the problem's own structure — a product of sums on the right, matching a sum of products on the left — supplies exactly that quantity, $a_i + b_i$, as the natural per-index denominator.

The edge case $a_i + b_i = 0$ for some i is not a technicality to wave past: if some $a_i = b_i = 0$, both products on the left vanish (since one factor of each is 0), so the inequality reads $0 \leq \left[(a_1+b_1)\cdots\right]^{1/n}$, which holds since the right side is a nonnegative real number. Handling it

separately, rather than trying to force it through a division by zero, is part of the model chain's discipline, not an afterthought.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the mismatch between the two sides' arithmetic: the left is a sum of two products, the right is a single product of two sums, with the sums matching index-by-index ($a_i + b_i$). On seeing it, write down: this per-index sum is a candidate normalising denominator, since dividing by it converts each raw pair into complementary weights.

2. That cue suggests applying AM-GM twice — once to $a_i/(a_i + b_i)$, once to $b_i/(a_i + b_i)$ — rather than once each to a_i and b_i directly.

3. Hesitation occurs right after the first (unnormalised) application of AM-GM produces a true but useless bound, $(a_1 \cdots a_n)^{1/n} + (b_1 \cdots b_n)^{1/n} \leq \frac{1}{n}\sum(a_i + b_i)$. This is reinforced away by checking the bound against the target: it compares to the *arithmetic* mean of $a_i + b_i$, which is always $\geq$ the geometric mean the problem wants, so the direction is wrong for the purpose and a different quantity must be fed into AM-GM.

4. The trap is treating the $n = 2$ case (which does yield to direct squaring and a Cauchy–Schwarz-flavoured identity) as a template for general n by pairing indices — this only closes for n a power of 2 and silently fails for e.g. $n = 3$.

5. Recognise this family faster by checking, whenever a target has a product of sums on one side matching a sum of products on the other with the sums aligned index-by-index, whether dividing through by that per-index sum turns the raw quantities into weights that sum to 1.

6. Reusable rule: to combine two applications of AM-GM (or any mean inequality) by addition, first normalise so that the two resulting arithmetic means are guaranteed to sum to a fixed, known constant (often 1, via complementary weights) — otherwise the addition only produces a bound in terms of a new, uncontrolled quantity.

▷ **Cue:** a target inequality has a product of index-wise sums on one side and a sum of products on the other, with the sums matching term-by-term $\longrightarrow$ **Move:** divide every quantity by its own matching sum, converting raw values into complementary weights that sum to 1 per index, then apply AM-GM to each normalised sequence and add

▷ **Cue:** combining two separate AM-GM bounds by simple addition $\longrightarrow$ **Move:** check first whether the two resulting arithmetic means are guaranteed to sum to a fixed known constant; if not, normalise until they do, rather than adding two uncontrolled bounds

2003 A3

2003 A3. Find the minimum value of

$$|\sin x + \cos x + \tan x + \cot x + \sec x + \csc x|$$

for real numbers x.

Fields: trigonometric algebra, AM-GM (calculus-free optimisation) **Position:** A3 **Answer:** The minimum value is $2\sqrt{2} - 1$, attained where $\sin x + \cos x = 1 - \sqrt{2}$.

Layer 1 THE WHAT — surface analysis

The natural attempt. Six trigonometric functions are stacked in one expression, and the reflex on seeing "find the minimum" is to differentiate: set $f(x) = \sin x + \cos x + \tan x + \cot x + \sec x + \csc x$ and compute $f'(x)$. Carried out honestly, this means differentiating $\tan x$, $\cot x$, $\sec x$, $\csc x$ individually — $\sec^2 x$, $-\csc^2 x$, $\sec x \tan x$, $-\csc x \cot x$ — and setting the sum of six unrelated-looking terms to zero:

$$\cos x - \sin x + \sec^2 x - \csc^2 x + \sec x \tan x - \csc x \cot x = 0.$$

This equation mixes powers and products of $\sin x, \cos x$ up to degree -2 with no shared substitution in sight; clearing denominators turns it into a high-degree polynomial in $\sin x, \cos x$ whose roots are not identifiable by inspection. Worse, before any of this is usable the domain has to be tracked separately — f is undefined wherever $\sin x = 0$ or $\cos x = 0$ — and the resulting critical-point equation offers no obvious reason to expect a clean closed form like $2\sqrt{2} - 1$ to fall out. This is the calculus attack pursued honestly, and it does not collapse; it is six terms' worth of derivative computation with no structural reason to expect simplification, which is the signature of having differentiated before simplifying rather than after.

The model chain. Do not differentiate the six-term expression in x at all. Instead notice that every term is a rational function of two symmetric quantities: $s = \sin x + \cos x$ and $p = \sin x \cos x$, related by the single identity

$$p = \frac{s^2 - 1}{2},$$

obtained by squaring $s^2 = 1 + 2\sin x \cos x$. Then

$$\tan x + \cot x = \frac{\sin^2 x + \cos^2 x}{\sin x \cos x} = \frac{1}{p}, \qquad \sec x + \csc x = \frac{\cos x + \sin x}{\sin x \cos x} = \frac{s}{p}.$$

So the whole expression collapses to a function of s alone:

$$f = s + \frac{1}{p} + \frac{s}{p} = s + \frac{1+s}{p} = s + \frac{2(1+s)}{s^2 - 1} = s + \frac{2}{s-1},$$

valid whenever $s \neq \pm 1$ (where $p = 0$ and the original expression is undefined), for $s = \sin x + \cos x \in [-\sqrt{2}, \sqrt{2}]$.

Now minimise $|f(s)| = \left| s + \frac{2}{s-1} \right|$ without calculus. Split on the sign of $s - 1$.

$s < 1$: write $u = 1 - s > 0$, so $u \in (0, 1 + \sqrt{2}] \setminus \{2\}$ (excluding $u = 2$, i.e. $s = -1$). Then

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

By AM-GM, $u + \frac{2}{u} \geq 2\sqrt{2}$ for all $u > 0$, with equality at $u = \sqrt{2}$ (i.e. $s = 1 - \sqrt{2}$, which lies in $[-\sqrt{2}, \sqrt{2}]$ and is not ± 1, so it is attained). Hence $f \leq 1 - 2\sqrt{2} < 0$ throughout this branch, so

$$|f| = \left(u + \frac{2}{u} \right) - 1 \geq 2\sqrt{2} - 1,$$

with equality exactly at $u = \sqrt{2}$.

$s > 1$: write $t = s - 1 \in (0, \sqrt{2} - 1]$. Then $f = 1 + t + \frac{2}{t}$. Since $t \leq \sqrt{2} - 1 < \sqrt{2}$, the function $t + 2/t$ is decreasing on this whole interval (its unconstrained minimum at $t = \sqrt{2}$ lies outside the range), so it is minimised at the right endpoint $t = \sqrt{2} - 1$, giving $t + 2/t = 3\sqrt{2} + 1$ there and $f = 3\sqrt{2} + 2 > 6$, far larger in absolute value than $2\sqrt{2} - 1$.

So the global minimum of $|f|$, and hence of the original expression, is $2\sqrt{2} - 1$, attained at $s = 1 - \sqrt{2}$.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains treat "minimise" as a call to compare values of a function of one real variable. They split at the very first move: the natural attempt keeps the function written in x, with all six trigonometric terms present, and differentiates that. The model chain first asks what smaller set of variables the expression actually depends on, finds it is a rational function of $s = \sin x + \cos x$ alone (via $p = (s^2 - 1)/2$), and only then optimises — in s, by AM-GM, with no derivative taken at all.

> *Six trig functions all built from* $\sin x$ *and* $\cos x$ *is a strong signal that the expression is really a function of a much smaller set of symmetric combinations —* $s = \sin x + \cos x$ *and* $p = \sin x \cos x$ *— linked by one identity, so before touching a derivative I should rewrite everything in terms of* s *and* p *and see how many variables are actually left.*

> *Once the problem is down to a single-variable expression of the shape (linear in* s*) plus (constant)/(linear in* s*), with a fixed sign on each side of the pole, I don't need calculus to bound it — I need to isolate the shifted variable (*$u = 1 - s$ *or* $t = s - 1$*) so the expression becomes const* $\pm (u + k/u)$*, and AM-GM bounds* $u + k/u$ *directly, cheaply, and without any domain-of-derivative bookkeeping.*

Layer 3 THE WHY

Why does the substitution $s = \sin x + \cos x$, $p = (s^2 - 1)/2$ succeed where direct differentiation does not? Because $\tan x, \cot x, \sec x, \csc x$ are not independent functions of x once you note they are all built from the same two building blocks $\sin x, \cos x$ — and $\sin x + \cos x$ and $\sin x \cos x$ are exactly the elementary symmetric functions of $\sin x, \cos x$, which satisfy the one algebraic identity $(\sin x + \cos x)^2 = 1 + 2 \sin x \cos x$ coming straight from $\sin^2 x + \cos^2 x = 1$. Any rational expression in $\sin x, \cos x$ that is itself symmetric — invariant under swapping $\sin x \leftrightarrow \cos x$ — must be expressible purely in s and p, and this expression manifestly is symmetric (swapping $\sin x$ and $\cos x$ permutes $\tan x \leftrightarrow \cot x$ and $\sec x \leftrightarrow \csc x$ and fixes the sum). Differentiating in x before checking this symmetry throws the reduction away and forces six independent derivative computations to somehow re-discover, term by painful term, a cancellation pattern that the symmetry already guarantees exists.

Once reduced to $f(s) = s + 2/(s - 1)$, the shape "linear plus a constant divided by a linear factor" is precisely the shape AM-GM is built to bound: writing $u = 1 - s$ turns f into $1 - (u + 2/u)$, and $u + 2/u \geq 2\sqrt{2}$ is the single most basic instance of AM-GM (applied to u and $2/u$, whose product is the constant 2). The pole at $s = 1$ is not a nuisance to be handled by limits or sign charts — it is exactly the location the substitution $u = 1 - s$ is centred on, and the domain exclusion $s \neq \pm 1$ (from $p \neq 0$) is precisely what keeps $u \neq 0$ and $u \neq 2$, so the AM-GM minimiser $u = \sqrt{2}$ survives every domain restriction untouched. Every piece of the calculus-free argument is therefore forced by the algebraic shape the substitution produces, not chosen for convenience after the fact.

It is also worth noting, as an efficiency point rather than a correction, that the official solution introduces a further substitution $y = \pi/4 - x$ and $c = \sqrt{2} \cos y$ before arriving at $f = c + 2/(c - 1)$ — algebraically the same function reached here directly in s, one substitution sooner. The extra layer is not wrong, but it is not load-bearing either: s itself already ranges over $[-\sqrt{2}, \sqrt{2}]$ and already reduces f to the needed one-variable rational function, so introducing y buys no further collapse. The reusable point is to stop substituting once the expression is already a function of a single real variable ranging over a known interval — a second change of variable that does not shrink the variable count further is not paying for itself.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is that all six terms — $\sin x, \cos x, \tan x, \cot x, \sec x, \csc x$ — are built from just $\sin x$ and $\cos x$, and the whole expression is symmetric under $\sin x \leftrightarrow \cos x$. On seeing it, write down: this is a function of $s = \sin x + \cos x$ and $p = \sin x \cos x$ alone, linked by $p = (s^2 - 1)/2$.
2. That cue suggests substituting $\tan x + \cot x = 1/p$ and $\sec x + \csc x = s/p$ to collapse the six-term expression into the single-variable rational function $f(s) = s + 2/(s - 1)$ over $s \in [-\sqrt{2}, \sqrt{2}] \setminus \{\pm 1\}$, then minimising $|f|$ by AM-GM rather than calculus.
3. Hesitation occurs at the very first line, where "find the minimum" reflexively suggests differentiating the expression as written. This is reinforced away by counting the cost:

differentiating $\tan x, \cot x, \sec x, \csc x$ individually produces four dissimilar terms with no shared substitution visible, a strong sign the representation is wrong before the calculus is even attempted.

4. The trap is forgetting to exclude $s = \pm 1$ (where $p = 0$ and the original expression is undefined) when minimising $f(s)$ over the interval $[-\sqrt{2}, \sqrt{2}]$ — an unconstrained AM-GM minimiser that happened to land on an excluded point would need separate handling, and here it must be checked, not assumed, that $u = \sqrt{2}$ (i.e. $s = 1 - \sqrt{2}$) survives the exclusion.

5. Recognise this family faster by checking, on any expression combining several trig functions of the same angle, whether it is symmetric under $\sin x \leftrightarrow \cos x$; if so it is a function of $s = \sin x + \cos x$ and $p = \sin x \cos x = (s^2 - 1)/2$ alone, and that reduction should happen before any derivative is taken.

6. Reusable rule: before optimising a multi-term trigonometric (or otherwise symmetric multi-variable) expression by calculus, check whether it reduces to a single-variable expression via the elementary symmetric functions of the underlying variables; once reduced to the shape (linear) + (constant)/(linear), bound it with AM-GM on the shifted variable rather than differentiating, and stop substituting as soon as the variable count has actually reached one.

▷ **Cue:** an expression combines several trigonometric functions of the same angle and is symmetric under $\sin x \leftrightarrow \cos x$ $\longrightarrow$ **Move:** rewrite in terms of $s = \sin x + \cos x$ and $p = \sin x \cos x = (s^2 - 1)/2$ before differentiating

▷ **Cue:** a single-variable target has the shape (linear in u) + (constant)/(linear in u) after centring on its pole $\longrightarrow$ **Move:** apply AM-GM to u and k/u directly rather than differentiating

2003 A4

2003 A4. Suppose that a, b, c, A, B, C are real numbers, $a \neq 0$ and $A \neq 0$, such that

$$|ax^2 + bx + c| \leq |Ax^2 + Bx + C|$$

for all real numbers x. Show that

$$|b^2 - 4ac| \leq |B^2 - 4AC|.$$

Fields: algebra (quadratics, inequalities); quadratic forms **Position:** A4 — mid-paper. No single hard step, but three genuinely different regimes, and the problem punishes anyone who tries to force one argument through all of them. **Answer:** A proof, by cases on the sign of $B^2 - 4AC$: real roots force P to be a multiple of Q; otherwise $Q \pm P \geq 0$ and the parallelogram law for the discriminant, or a comparison of minimum values, finishes.

Layer 1 THE WHAT — surface analysis

Notation for the whole analysis: $P(x) = ax^2 + bx + c$, $Q(x) = Ax^2 + Bx + C$, and for any quadratic-or-lower $\alpha x^2 + \beta x + \gamma$ write $D(\alpha, \beta, \gamma) = \beta^2 - 4\alpha\gamma$, so $D(P) = b^2 - 4ac$ and $D(Q) = B^2 - 4AC$.

The natural attempt. You square the hypothesis to kill the absolute values: $Q(x)^2 - P(x)^2 \geq 0$ for all x. That is a quartic, and you try to extract something about discriminants from "a quartic is non-negative". Expanding gives $(A^2 - a^2)x^4 + 2(AB - ab)x^3 + \cdots$, and the conditions for a quartic to be non-negative are not something you can write down in an exam. So you retreat to the factorisation $(Q - P)(Q + P) \geq 0$, which is better, but a product of two quadratics being non-negative does not force each factor to be non-negative — they could both change sign at the same points. You try to push on with "the discriminant measures the distance between the roots", which works when both quadratics have real roots, and then stall on the case where Q

has no real roots: there is no distance to compare, and $|b^2 - 4ac|$ could come from a P with real roots or without. The wall is that you are looking for one inequality chain, and there isn't one.

The model chain. First, $|a| \le |A|$: divide the hypothesis by x^2 and let $x \to \infty$. Replacing (A, B, C) by $(-A, -B, -C)$ changes neither the hypothesis nor $D(Q)$, and likewise for P; so assume $A > 0$ and $a > 0$.

Case 1: $D(Q) > 0$. Then Q has distinct real roots $r_1 \neq r_2$, and $|P(r_i)| \le |Q(r_i)| = 0$ forces $P(r_1) = P(r_2) = 0$. A quadratic with $a \neq 0$ and roots r_1, r_2 is $a(x - r_1)(x - r_2) = \frac{a}{A}Q(x)$. Since D is homogeneous of degree 2 in the coefficients, $D(P) = \frac{a^2}{A^2}D(Q)$, so $|D(P)| \le D(Q) = |D(Q)|$.

Case 2: $D(Q) \le 0$. Now $A > 0$ gives $Q(x) \ge 0$ for all x, so the hypothesis reads $-Q \le P \le Q$, i.e. both $Q - P$ and $Q + P$ are non-negative on all of $\mathbb{R}$. A polynomial $\alpha x^2 + \beta x + \gamma$ that is non-negative everywhere has $D \le 0$: if $\alpha > 0$ that is the usual criterion; $\alpha < 0$ is impossible; if $\alpha = 0$ then $\beta = 0$ (a non-constant linear function changes sign), so $D = 0$.

Case 2a: $D(P) > 0$. D is a quadratic form in the coefficient vector, so it obeys the parallelogram law

$$D(Q + P) + D(Q - P) = 2D(Q) + 2D(P).$$

The left side is ≤ 0, so $D(P) \le -D(Q)$, i.e. $|D(P)| \le |D(Q)|$.

Case 2b: $D(P) \le 0$. With $a > 0$, $P \ge 0$ and its minimum value is $-D(P)/(4a)$; with $A > 0$, Q has minimum value $-D(Q)/(4A)$, attained at some x_0. Then

$$\frac{-D(P)}{4a} \le P(x_0) \le Q(x_0) = \frac{-D(Q)}{4A},$$

so $-D(P) \le \frac{a}{A}(-D(Q)) \le -D(Q)$, using $0 < a \le A$ and $-D(Q) \ge 0$. That is $|D(P)| \le |D(Q)|$.

The three cases are exhaustive. □

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains reach $(Q - P)(Q + P) \ge 0$ (†). The natural attempt tries to deduce sign information about the factors from the product and gets nowhere. The model does not use (†) at all: it first splits on whether Q has real roots, because that decides whether $|Q|$ can be replaced by Q.

> *The hypothesis has $|Q|$ in it. I can't do anything with an absolute value until I know the sign of what's inside. So the first question is: does Q change sign? That's exactly $D(Q) > 0$ or not — so the case split is forced by the hypothesis, not by the conclusion.*

> *If Q has real roots, $|P| \le |Q|$ pins P to zero there. Two roots and $a \neq 0$: P is a scalar multiple of Q. Nothing more to do.*

> *If $Q \ge 0$, then $|P| \le Q$ means $Q - P \ge 0$ and $Q + P \ge 0$ separately. Two non-negative quadratics: two discriminants ≤ 0. The discriminant is a quadratic form, so the sum of $D(Q + P)$ and $D(Q - P)$ is $2D(Q) + 2D(P)$.*

Layer 3 THE WHY

Why is the case split not a cop-out? Because an absolute value is a two-case definition: $|y| = y$ if $y \ge 0$, $-y$ otherwise. The hypothesis $|P| \le |Q|$ has two absolute values in it and the conclusion has two more, one of which, $|D(Q)|$, is literally the question "does Q have real roots?". The problem is telling you that the sign of $D(Q)$ matters. Splitting on it is reading the problem, not giving up on elegance.

Why does real-roots-of-Q make everything collapse? Because $|P| \le |Q|$ is weakest where $|Q|$ is large and strongest where $|Q|$ is small; at a root of Q it becomes an equation, $P = 0$. A quadratic is determined up to scale by its two roots. That is the whole of Case 1: the hypothesis is at its strongest exactly where Q vanishes.

Why the parallelogram law in Case 2a? Look at what the discriminant is: $D(\alpha, \beta, \gamma) = \beta^2 - 4\alpha\gamma$ is a homogeneous quadratic in the three coefficients — a quadratic form, in the same sense that $x^2 + y^2$ is. Every quadratic form q satisfies $q(u+v) + q(u-v) = 2q(u) + 2q(v)$, because the cross terms appear with opposite signs and cancel. You know this identity from vectors: $|u+v|^2 + |u-v|^2 = 2|u|^2 + 2|v|^2$. And the hypothesis, once $Q \ge 0$, is precisely a statement about $Q + P$ and $Q - P$. So the one identity that relates $D(Q \pm P)$ to $D(Q)$ and $D(P)$ is the identity you already own. Kedlaya's Case 3 adds the two discriminant inequalities "and adding these together yields" — here is where that comes from: adding is the parallelogram law, and the parallelogram law is what quadratic forms do.

Why does the parallelogram law fail to finish Case 2b? It gives $D(P) \le -D(Q)$, which is an upper bound on $D(P)$. When $D(P) \le 0$ you need a lower bound, $-D(P) \le -D(Q)$, i.e. "P's trough is no deeper than Q's". Now read $-D/(4 \cdot \text{leading coefficient})$ as what it is: the minimum value of the quadratic. The inequality you want is then a comparison of minima, and $P \le Q$ pointwise compares minima for free — evaluate at Q's minimiser. The only leak is the leading coefficients, which is why $|a| \le |A|$ had to be extracted at the start. Kedlaya instead shifts to $B = 0$ and compares constant terms; that is the same comparison of values at Q's vertex, disguised by the change of variable.

So each case uses the one fact about quadratics that the regime makes available: roots (Case 1), discriminant as a quadratic form (2a), discriminant as a minimum value (2b).

Caveat. Verified: the three-case proof above was checked line by line, and the statement was sampled numerically (200,000 random coefficient sets, keeping those satisfying the hypothesis on a grid) with no counterexample. The degenerate sub-case $A = \pm a$, where $Q \mp P$ has degree < 2, is covered by the "$\alpha = 0$ forces $\beta = 0$" remark.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is $|Ax^2 + Bx + C|$ on the right of the hypothesis alongside $|B^2 - 4AC|$ in the conclusion. Write down: "Case on sign of $B^2 - 4AC$: does Q change sign?"

2. Case analysis driven by the absolute values; in each case, the characterisation of quadratics suited to that case — roots, non-negativity criterion, minimum value.

3. Hesitation comes in Case 2 when the parallelogram step gives an inequality in the wrong direction for $D(P) < 0$. It is reinforced away by asking what $-D$ means geometrically (the depth of the trough, times $4\times$leading coefficient) and comparing troughs directly.

4. Squaring to $Q^2 - P^2 \ge 0$ and trying to handle a quartic. It removes the absolute values at the cost of doubling the degree, and loses the information that Q has a fixed sign in Case 2.

5. Slogan: *an absolute value in the hypothesis is a request to split on the sign of what's inside.* And: *a discriminant is a quadratic form in the coefficients, so it satisfies the parallelogram law.*

6. Decision rule: **when the hypothesis is $|P| \le |Q|$, first decide whether Q vanishes somewhere; if it does, the hypothesis becomes equations at the zeros; if it does not, remove the absolute value to get $-Q \le P \le Q$, i.e. two one-sided inequalities $Q \pm P \ge 0$, and combine them with whatever identity the target quantity satisfies under $\pm$.**

▷ **Cue:** a hypothesis $|P(x)| \le |Q(x)|$ with Q allowed to vanish $\longrightarrow$ **Move:** evaluate at the zeros of Q: P must vanish there too

▷ **Cue:** $Q \ge 0$ and $|P| \le Q$ $\longrightarrow$ **Move:** rewrite as the pair $Q - P \ge 0$, $Q + P \ge 0$

▷ **Cue:** an expression like $b^2 - 4ac$ to be compared for Q, P, $Q \pm P$ $\longrightarrow$ **Move:** it is a quadratic form in the coefficients: $D(Q+P) + D(Q-P) = 2D(Q) + 2D(P)$

▷ **Cue:** an inequality between discriminants of two non-negative quadratics ⟶ **Move:** read $-D/(4a)$ as the minimum value and compare the functions at one minimiser
▷ **Cue:** two quadratics with leading coefficients a, A and $|P| \le |Q|$ for all x ⟶ **Move:** let $x \to \infty$: $|a| \le |A|$

2003 A5

2003 A5. A Dyck n-path is a lattice path of n upsteps $(1,1)$ and n downsteps $(1,-1)$ that starts at the origin O and never dips below the x-axis. A return is a maximal sequence of contiguous downsteps that terminates on the x-axis. For example, the Dyck 5-path illustrated has two returns, of length 3 and 1 respectively.

[Figure in the original: the path through $(0,0), (1,1), (2,2), (3,1), (4,2), (5,3), (6,2), (7,1), (8,0), (9,1), (10,$ i.e. the step sequence $UUDUUDDDUD$ with $U = (1,1)$, $D = (1,-1)$.]

Show that there is a one-to-one correspondence between the Dyck n-paths with no return of even length and the Dyck $(n-1)$-paths.

Fields: combinatorics (lattice paths, Catalan objects, bijections) **Position:** A5 — hard end of the paper, but the difficulty is entirely in finding the right decomposition; once found, every verification is one line. **Answer:** The correspondence exists. Every Dyck n-path with no even return is $U\,Q\,D\,R$ with UQD its first arch; send it to the $(n-1)$-path QR. The inverse cuts an $(n-1)$-path after its rightmost even return (or prepends UD if there is none).

Layer 1 THE WHAT — surface analysis

Notation: $U = (1,1)$, $D = (1,-1)$. An *arch* is a Dyck path that touches the axis only at its two ends; every non-empty Dyck path factors uniquely as a concatenation of arches (cut at each visit to the axis), and every arch is $U\,Q\,D$ with Q a Dyck path (possibly empty). The returns of a path are in one-to-one correspondence with its arches: each arch ends with exactly one return.

The natural attempt. You count first. Small cases: $n = 1$: UD, one path, return length 1; $C_0 = 1$. $n = 2$: $UDUD$ (returns 1, 1) and $UUDD$ (return 2, even) — one survivor, $C_1 = 1$. $n = 3$: five paths, and the survivors are $UDUDUD$ and $UUDDUD$? No: $UUDDUD$ has a return DD of length 2. The survivors are $UDUDUD$ and $UUDUDD$ — wait, $UUDUDD$ ends DD, length 2 again. You recheck: $UUUDDD$ (length 3, odd) survives, $UDUDUD$ survives. Two, $= C_2$. The numbers $1, 1, 2, 5, 14$ appear and you are confident the claim is a Catalan identity. Now you try to prove $X_n = C_{n-1}$ by a recursion on the first arch, find that the first arch's return is odd exactly when its inside Q ends in an even return, introduce a second count for "last return even", and end with a coupled system of convolution recurrences. Carried through (as in Bernstein's official second solution), it works — but it costs a page of bookkeeping, needs an auxiliary sequence, and the error rate in the hand enumeration above is a warning of what the bookkeeping costs under time pressure. And the problem said "one-to-one correspondence"; the count proves existence but you finish without having seen why.

The model chain. Structure of a path with no even return. Let T be a Dyck n-path, $n \ge 1$, and write $T = U\,Q\,D\,R$ where UQD is its first arch, Q a Dyck path and R a Dyck path. The return of the first arch consists of the final D together with the maximal run of downsteps ending Q *if* that run reaches the axis of Q, i.e. is Q's last return. So

$$\text{first return length} = \begin{cases} 1, & Q \text{ empty}, \\ 1 + (\text{length of the last return of } Q), & Q \text{ non-empty}. \end{cases}$$

The returns of T after the first are exactly the returns of R (the step after the first arch is U, so maximality is unaffected). Hence T has no even return iff

$$(Q \text{ is empty or } Q\text{'s last return is even}) \text{ and } (R \text{ has no even return}). \qquad (*)$$

The map g. Send $T = UQDR$ to $g(T) = QR$, a Dyck $(n-1)$-path (a concatenation of Dyck paths).

The inverse f. Given a Dyck $(n-1)$-path S: if S has no even return, put $f(S) = UD\,S$. Otherwise let the rightmost even return of S end at step j; it ends on the axis, so $S = Q\,R$ with Q the first j steps (a Dyck path whose last return is this even one) and R the rest (a Dyck path, starting with U if non-empty, having no even return by the choice of "rightmost"). Put $f(S) = U\,Q\,D\,R$.

In both cases $f(S)$ satisfies $(*)$, so $f(S)$ is an n-path with no even return. $g(f(S)) = QR = S$. Conversely, for $T = UQDR$ with no even return, $(*)$ says Q is empty or ends in an even return, and R has none; in QR the last return of Q is still a return (it is followed by U or by nothing), so it is the rightmost even return of QR, or $QR = R$ has none when Q is empty. Either way $f(g(T)) = T$. So f and g are inverse bijections. □

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains decompose at the first arch, $T = UQDR$, and both observe that the first return is odd iff Q is empty or ends in an even return $(\dagger)$. At $(\dagger)$ the natural attempt turns the observation into a recurrence and counts. The model asks what information T contains beyond the $(n-1)$-path QR — and finds that it is none.

> *Deleting the first U and the D that closes the first arch always leaves a Dyck $(n-1)$-path. So the map exists for free. The only question is whether I can recover where the cut was.*

> *Where was the cut? At the end of Q. And Q is exactly the part that ends in an even return, while everything after it, R, has none. So the cut is after the* last *even return of QR — the position is written in the path itself.*

> *If QR has no even return at all, Q must have been empty. That's the one extra case, and it's UD in front.*

Layer 3 THE WHY

Why the first arch? Because the definition of a return is local to arches: a return is the downward run that closes an arch, so "no return of even length" is a condition on each arch separately. Any argument must therefore look at arches, and the first one is the one you can strip off while leaving a Dyck path behind. This is the standard first-return decomposition of Catalan objects, $C_n = \sum C_k C_{n-1-k}$, and it is the reason C_{n-1} is plausible in the first place: removing one U and one D drops n by one.

Why does "delete U and D" lose no information? Deleting two steps from a $2n$-step path forgets exactly one thing: where the deleted D was. A map is injective iff the forgotten data can be read off the image. So the right question is not "how many paths survive?" but "given QR, where is the seam between Q and R?". The problem's hypothesis supplies the answer: evenness is forbidden in T, so its traces in QR must be confined to one place. Parity of the first return is $1 +$ (last return of Q), so Q's final return is forced to be even; R is forced to have no even return at all. An even return in QR is therefore a marker that sits exactly at the seam. The hypothesis that looked like an arbitrary restriction is precisely the information that makes the deletion reversible.

Kedlaya's first solution states the inverse ("let $a_i \cdots a_j$ denote the rightmost even-length return") and asserts "evidently $f \circ g$ and $g \circ f$ are identity maps". The evidence is $(*)$, the computation of the first return's length from Q's last return, which the official text leaves unstated. That computation is the entire content: once you know the first return has length $1 + \ell(Q)$, "rightmost even return" is not a guess — it is the only marker available.

Why does the counting route work but feel heavier? It encodes the same $(*)$ as a recurrence. Writing $O(x), E(x)$ for generating functions of non-empty paths whose last return is odd, even, and $C(x)$ for all Dyck paths, $(*)$ becomes $O = xC(1+E)$, $E = xC\,O$, and the no-even-return paths have generating function $1/(1-x(1+E))$; with $C = 1+xC^2$ this reduces to $1+xC$, i.e. $X_n = C_{n-1}$. Correct, but it proves equality of sizes and then leaves you to believe a bijection exists. The bijective route makes $(*)$ do the work once.

Caveat. Verified: the maps f and g above were implemented and checked to be mutually inverse bijections between Dyck $(n-1)$-paths and Dyck n-paths with no even return for $n = 1, \ldots, 11$ (counts $1, 1, 2, 5, 14, \ldots, 16796 = C_{10}$). The generating-function reduction in the WHY layer was checked by hand from $C = 1 + xC^2$.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the definition "a return is a maximal sequence of contiguous downsteps that terminates on the x-axis" — a return closes an arch. Write down $T = U\,Q\,D\,R$ and the length of the first return in terms of Q.

2. First-arch decomposition, then a deletion map $UQDR \mapsto QR$, then ask how to recover the seam.

3. Hesitation: after seeing that the first return is odd iff Q's last return is even, it is tempting to start counting. Reinforce it away by asking, before any count, "the target has size $n-1$; what natural operation removes one U and one D?"

4. Hand enumeration of small cases (it is easy to misclassify $UUDUDD$), and the recurrence route, which proves equal cardinality but hides the bijection the question names.

5. Slogan: *a deletion map is a bijection when the forbidden pattern in the source marks where the deletion happened in the target.*

6. Decision rule: **when asked for a bijection from a restricted family of size-n objects to all objects of size $n-1$, try the most natural "remove one unit" map first, then ask what information it forgets; the restriction in the hypothesis is almost always what lets you recover it — usually as a "first" or "last" occurrence of some pattern.**

▷ **Cue:** a condition on the "returns" (or blocks, or cycles) of a Catalan-type object $\longrightarrow$ **Move:** decompose at the first return: $T = UQDR$, and express the first block's property through Q

▷ **Cue:** a restricted family at size n equinumerous with an unrestricted family at size $n-1$ $\longrightarrow$ **Move:** find a map deleting one unit; the restriction must encode where the unit was

▷ **Cue:** a deletion map whose inverse must locate a seam $\longrightarrow$ **Move:** look for the leftmost or rightmost occurrence of a forbidden pattern — extremality makes it unique

▷ **Cue:** a small-case table that matches $1, 1, 2, 5, 14$ $\longrightarrow$ **Move:** the objects are Catalan; use the first-return decomposition before counting

2003 A6

2003 A6. For a set S of nonnegative integers, let $r_S(n)$ denote the number of ordered pairs (s_1, s_2) such that $s_1 \in S$, $s_2 \in S$, $s_1 \neq s_2$, and $s_1 + s_2 = n$. Is it possible to partition the nonnegative integers into two sets A and B in such a way that $r_A(n) = r_B(n)$ for all n ?

Fields: combinatorics / number theory (additive representation functions); generating functions; binary expansions **Position:** A6 — the last problem of the morning. Hard because the answer has to be found, not only proved, and the object that works (the Thue–Morse partition) is not guessable from the statement unless you let the condition generate it. **Answer:** Yes. Put n in A if the binary expansion of n has an even number of 1s, in B otherwise (the Thue–Morse partition). It is the unique such partition with $0 \in A$.

Layer 1 THE WHAT — surface analysis

Notation: $s_2(n)$ is the number of 1s in the binary expansion of n; $\varepsilon_n = +1$ if $n \in A$, -1 if $n \in B$.

The natural attempt. You try to build A and B greedily. Put $0 \in A$ (by symmetry). $n = 1$: the only pair is $\{0,1\}$; for $r_A(1) = r_B(1) = 0$ you need $1 \in B$. $n = 2$: the pair $\{0,2\}$; $0 \in A$, so $2 \in B$... then $r_B(2) = 0 = r_A(2)$ needs $2 \notin A$, and $2 \in B$ gives no B-pair since $0 \notin B$. Fine. $n = 3$: pairs $\{0,3\},\{1,2\}$; $\{1,2\} \subset B$ gives $r_B(3) = 2$, so you need $\{0,3\} \subset A$: $3 \in A$. You continue: $A = \{0,3,5,6,9,10,12,15,\dots\}$, $B = \{1,2,4,7,8,11,13,14,\dots\}$. Each step is forced — flipping the new element m changes exactly one of $r_A(m)$, $r_B(m)$ by 2 — so the construction never has a choice. But it also never tells you it will not get stuck: the worry is that at some n both placements fail. A strong student spots the pattern ("even number of binary 1s") and then faces a second problem: proving the pattern works for every n. Checking by cases on the last few binary digits leads into carries, and carries are where this attempt bogs down.

The model chain. Work with formal power series. Let $F(x) = \sum_{n\in A} x^n$ and $G(x) = \sum_{n\in B} x^n$. The coefficient of x^n in $F(x)^2$ counts ordered pairs $(s_1, s_2) \in A^2$ with $s_1 + s_2 = n$, including $s_1 = s_2 = n/2$; $F(x^2)$ counts exactly those diagonal pairs. So

$$\sum_n r_A(n)x^n = F(x)^2 - F(x^2), \qquad \sum_n r_B(n)x^n = G(x)^2 - G(x^2).$$

Hence $r_A \equiv r_B$ iff $F(x)^2 - G(x)^2 = F(x^2) - G(x^2)$. Put $H(x) = F(x) - G(x) = \sum_n \varepsilon_n x^n$. Since A, B partition $\mathbb{Z}_{\geq 0}$, $F + G = \frac{1}{1-x}$, and the condition becomes

$$\frac{H(x)}{1-x} = H(x^2) \iff H(x) = (1-x)\,H(x^2).$$

Comparing coefficients of x^{2m} and x^{2m+1} on both sides:

$$\varepsilon_{2m} = \varepsilon_m, \qquad \varepsilon_{2m+1} = -\varepsilon_m \qquad (m \geq 0). \tag{$*$}$$

(For $m = 0$ the first equation is empty and the second says $\varepsilon_1 = -\varepsilon_0$.) So a partition works *if and only if* $(*)$ holds. $(*)$ determines every ε_n from ε_0 by stripping the last binary digit: appending a 0 keeps the sign, appending a 1 flips it. Taking $\varepsilon_0 = 1$ gives $\varepsilon_n = (-1)^{s_2(n)}$, which does satisfy $(*)$ since $s_2(2m) = s_2(m)$ and $s_2(2m+1) = s_2(m) + 1$. So the partition $A = \{n : s_2(n) \text{ even}\}$, $B = \{n : s_2(n) \text{ odd}\}$ works, and it is the only one with $0 \in A$. Equivalently, $H(x) = \prod_{k\geq 0}(1 - x^{2^k})$. The answer is **yes**. □

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains reach the same data: the forced placements $0,3,5,6 \in A$, $1,2,4 \in B$. They split at the question (†) "how do I show this never breaks down?". The natural attempt guesses the pattern and tries to verify it pair by pair. The model encodes the condition for *all* n at once as one equation between generating functions and solves that equation.

> *$r_S(n)$ counts pairs from S adding to n — that is a coefficient of a square. The condition is "for all n", so I want one identity of power series, not infinitely many checks.*

> *I have two unknown series but one constraint ties them: $F + G = 1/(1-x)$. The condition involves $F^2 - G^2$, which factors as $(F+G)(F-G)$. So the only real unknown is $H = F - G$, and the equation $H(x) = (1-x)H(x^2)$ is a recursion in disguise.*

> *$H(x)$ in terms of $H(x^2)$: even-indexed coefficients from index m, odd-indexed from m with a sign. That is "delete the last binary digit". So the answer lives in binary.*

Layer 3 THE WHY

Why generating functions? Because the definition of $r_S(n)$ is a convolution: the number of ways to write $n = s_1 + s_2$ with $s_1, s_2 \in S$. A convolution is the coefficient rule for multiplying power series; that is what multiplication of polynomials *is*. Whenever a quantity is "the number of ways to write n as a sum of elements of prescribed sets", its generating function is a product. The small correction $s_1 \neq s_2$ removes the diagonal term, which contributes x^{2s} for each $s \in S$ — hence $F(x^2)$. Nothing here is a trick; it is the definition of r_S rewritten in the language designed for it.

Why difference of squares? The partition condition says $F + G$ is completely known. A problem with two unknowns and one known sum is a problem about their difference: set $H = F - G$, and $F = \frac{1}{2}(\frac{1}{1-x} + H)$. The condition $F^2 - F(x^2) = G^2 - G(x^2)$ rearranges to $F^2 - G^2 = F(x^2) - G(x^2)$, and $F^2 - G^2$ is a difference of two squares whose sum-factor is the known quantity. You would make exactly the same move with two real numbers $u + v = s$, $u^2 - v^2 = t$.

Why is $(*)$ binary? The functional equation relates x to x^2. Substituting $x \mapsto x^2$ sends x^m to x^{2m}: it doubles indices. Doubling an integer in binary appends a 0; the extra factor $(1 - x)$ supplies the odd indices $2m + 1$ (append a 1) with a minus sign. Any functional equation linking $H(x)$ and $H(x^2)$ is a statement about binary expansions — this is as structural as "a circle is the locus equidistant from the centre". Kedlaya's first solution says the partition is discovered "by simply attempting to place the first few numbers by hand and noticing the resulting pattern". Here is what that hides: the pattern is forced, $(*)$ is why it is forced, and $(*)$ is simultaneously the proof. The noticing is replaced by solving.

It is worth also knowing why the official bijective proof works, because it is a different mechanism. Given $a_1 + a_2 = n$ with $a_1 \neq a_2$, both in A, find the lowest binary place where they differ; one has a 1 there and the other a 0. Swap those two bits. The sum is unchanged (one gains 2^k, the other loses 2^k), each number's digit count changes by exactly one, so both land in B. The lower places agree, so the lowest differing place is the same after the swap, making the map an involution between A-pairs and B-pairs. This is the combinatorial shadow of $H(x) = (1 - x)H(x^2)$ applied repeatedly.

Caveat. Verified: $r_A(n) = r_B(n)$ for the Thue–Morse partition was checked by direct computation for all $n < 1500$. The equivalence of the partition condition with $(*)$ is an if-and-only-if derivation over formal power series and needs no convergence; uniqueness up to swapping A, B follows from it and agrees with the Note in the official solutions.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the phrase "the number of ordered pairs $(s_1, s_2) \ldots$ such that $s_1 + s_2 = n$". Write down $\sum r_S(n)x^n = F(x)^2 - F(x^2)$.

2. Generating functions, then reduce two unknown series to one via the known sum $F + G = \frac{1}{1-x}$.

3. Hesitation arrives after the greedy placements look consistent but unproven. It is reinforced away by recognising that "for all n" plus "pairs summing to n" is a single power-series identity: stop checking cases and write that identity.

4. The trap is to guess the pattern and then prove it by digit cases, which runs into carries. The other trap is to believe the answer is no because greedy construction "must eventually fail"; it cannot fail, because each step changes $r_A(m) - r_B(m)$ by ± 2 with exactly one choice making it zero.

5. Slogan: *counts of representations as sums are coefficients of products. And an equation linking $H(x)$ with $H(x^2)$ is a statement about binary digits.*

6. Decision rule: **when a condition must hold for every n and involves the number of ways to write n as a sum, convert it to one identity of generating functions; if two unknown series have a known sum or product, pass to their difference or ratio; if the resulting equation relates x and x^k, read off a recursion on base-k digits.**

▷ **Cue:** the number of pairs $(s_1, s_2) \in S^2$ with $s_1 + s_2 = n$ $\longrightarrow$ **Move:** it is the coefficient of x^n in $F_S(x)^2$; subtract $F_S(x^2)$ to remove $s_1 = s_2$

▷ **Cue:** two unknown sets partitioning $\mathbb{Z}_{\geq 0}$ $\longrightarrow$ **Move:** $F + G = \frac{1}{1-x}$; work with $H = F - G = \sum \pm x^n$

▷ **Cue:** a functional equation relating $H(x)$ and $H(x^2)$ $\longrightarrow$ **Move:** compare even and odd coefficients: a recursion on the last binary digit

▷ **Cue:** a greedy construction where each step is forced $\longrightarrow$ **Move:** the forced object is unique; find the equation that forces it and solve that equation instead of guessing

▷ **Cue:** need to change a pair with fixed sum while changing a digit-parity statistic $\longrightarrow$ **Move:** swap the lowest differing binary digit between the two numbers

Putnam 2003 — Paper B

2003 B1

2003 B1. Do there exist polynomials $a(x), b(x), c(y), d(y)$ such that

$$1 + xy + x^2y^2 = a(x)c(y) + b(x)d(y)$$

holds identically?

Fields: linear algebra, polynomial algebra **Position:** B1 — opener of Paper B **Answer:** No.

Layer 1 THE WHAT — surface analysis

The natural attempt. You see a two-variable identity and reach for an ansatz. The left side has bidegree $(2,2)$, so you try a, b of degree ≤ 2 in x and c, d of degree ≤ 2 in y. To hit the x^2y^2 term you set $a(x) = x^2$, $c(y) = y^2$; then $a(x)c(y) = x^2y^2$ exactly, and $b(x)d(y)$ has to supply the rest, $1 + xy$. But $1 + xy$ is not itself a product of a function of x alone and a function of y alone — setting $x = 0$ gives the value 1, setting $x = 1$ gives $1 + y$, and these are not proportional as functions of y, so no single b, d pair works. You try another split: $a(x) = 1$, $c(y)$ absorbing a piece of everything, $b(x) = x$, and so on. Each attempt either misses a term or reproduces the same obstruction in a different guise. Twelve unknown coefficients ($a_0, a_1, a_2, b_0, b_1, b_2$ and likewise for c, d) against nine target coefficients (the entries of the 3×3 array of x^iy^j coefficients) looks, on the face of it, generous — plenty of freedom, not enough equations. And yet nothing you try closes the loop. The trouble with this attempt is structural: guess-and-check can only ever report failures of specific choices, never rule out every choice of every degree at once.

The model chain. Read the identity as a statement about a matrix. Index the coefficient of x^iy^j in a bidegree-$(2,2)$ polynomial by its position in a 3×3 array (rows $i = 0, 1, 2$, columns $j = 0, 1, 2$). The left side $1 + xy + x^2y^2$ has coefficient 1 at positions $(0,0)$, $(1,1)$, $(2,2)$ and 0 everywhere else — its array is exactly the 3×3 identity matrix I_3, whose rank is 3. Now examine what a product $a(x)c(y)$ looks like in this array: if a has coefficient vector (a_0, a_1, a_2) and c has coefficient vector (c_0, c_1, c_2), the coefficient of x^iy^j in $a(x)c(y)$ is a_ic_j — the array is precisely the outer product of the two coefficient vectors, and an outer product is, by definition, a rank-1 matrix (or the zero matrix, rank 0, if either factor is zero). So $a(x)c(y) + b(x)d(y)$ corresponds to a sum of two matrices each of rank at most 1. Since rank is subadditive — the column space of a sum of matrices is contained in the sum of the individual column spaces, so $\operatorname{rank}(A + B) \leq \operatorname{rank}(A) + \operatorname{rank}(B)$ — the right side of the proposed identity has rank at most 2. But the left side has rank 3. No polynomials a, b, c, d, of any degree whatsoever, can produce a rank-3 array as a sum of two rank-1 arrays. The identity cannot hold.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths agree, almost immediately, that this is really a question about the coefficient array of $1 + xy + x^2y^2$, not about x and y as numbers. The split happens at what you do with that array. The natural attempt keeps trying to fill in the nine target entries with two guessed products, solving case by case and hoping a lucky choice of degree closes the gap.

The model chain instead asks: whatever a, b, c, d turn out to be, what invariant of the whole array does a sum of two products $a(x)c(y) + b(x)d(y)$ have to satisfy?

> *Forget solving for coefficients — $a(x)c(y)$ is a rank-one object no matter what a and c are, so I should ask what rank the left side has as a matrix, not whether I can match nine numbers.*

Layer 3 THE WHY

A bivariate polynomial of bidegree (n, n) and the space of $(n+1) \times (n+1)$ matrices are the same vector space under two different names: the correspondence $x^i y^j \leftrightarrow$ "entry (i, j)" is a linear isomorphism. Under that isomorphism, the product $a(x)c(y)$ is not merely *like* a rank-one matrix — it *is* one, by the definition of an outer product: entry (i, j) equals $a_i \cdot c_j$, which is exactly what "rank one" means (every row is a scalar multiple of every other row, namely of the vector $(c_0, \ldots, c_n)$). Rank subadditivity is not a special fact about this problem; it follows from the definition of matrix rank as the dimension of the column space, and the column space of $A + B$ is contained in $\text{colspace}(A) + \text{colspace}(B)$ purely because every column of $A + B$ is a sum of a column of A and a column of B. So a sum of two rank-≤ 1 matrices has rank ≤ 2, full stop, independent of the matrices' size. The right side of the identity, read as a matrix, is I_3 — and the rank of I_3 is 3 for the most basic possible reason: its three columns are the standard basis vectors e_1, e_2, e_3, which are linearly independent by definition. Three does not fit inside two. Padding a, b, c, d with extra high-degree terms only enlarges the array with rows and columns of zeros around the same 3×3 block; it changes nothing about the rank contributed by that block, so no amount of extra freedom in degree can ever rescue the construction. This is also why the official solution's trick of substituting $y = -1, 0, 1$ works: it produces the three polynomials $1 - x + x^2$, 1, $1 + x + x^2$, each forced to lie in the (at most) 2-dimensional span of $a(x)$ and $b(x)$ — and these three specific polynomials are linearly independent, so no 2-dimensional space can hold all three. That is the identical rank bound, viewed through evaluation rather than through the coefficient array.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the shape of the right side itself: a sum of $r = 2$ products, each a function of x times a function of y. Any time you see that shape, the word to write down immediately is *separable*, and the next word after it should be *rank*.

2. That cue suggests representing the target bivariate polynomial as its coefficient matrix and bounding the rank of a sum of r rank-1 pieces by r.

3. Hesitation occurs right after an ansatz fails: the instinct is "maybe I haven't tried the right degree or combination yet." It is reinforced away by noticing that rank is a property of the entire matrix and is invariant under padding with extra rows and columns of zeros, so raising the degree of a, b, c, d can never increase the rank contributed by a fixed block.

4. The trap is the coincidence that the left side has only three nonzero coefficients, which tempts the false belief that two "double" terms $a(x)c(y)$ and $b(x)d(y)$ ought to be plenty for three targets. Rank counting shows the true ceiling is 2 regardless of how sparse the target is, because all six of the other, zero, entries of the 3×3 array cost the argument nothing.

5. To recognise this faster next time: the instant an identity has the shape $f(x, y) = \sum_{k=1}^{r} g_k(x) h_k(y)$, write down f's coefficient (or value) matrix and compare its rank to r, before attempting any coefficient chase.

6. The reusable decision rule: a two-variable function is a sum of r separable (single-variable-product) terms only if its coefficient, or value, matrix has rank at most r — and this is a necessary condition regardless of the degree, or even the polynomial-ness, of the pieces involved.

▷ **Cue:** an identity of the form $f(x, y) = \sum_{k=1}^{r} g_k(x) h_k(y)$ $\longrightarrow$ **Move:** write f as its coefficient matrix and bound its rank by r

2003 B2

2003 B2. Let n be a positive integer. Starting with the sequence $1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n}$, form a new sequence of $n-1$ entries $\frac{3}{4}, \frac{5}{12}, \ldots, \frac{2n-1}{2n(n-1)}$ by taking the averages of two consecutive entries in the first sequence. Repeat the averaging of neighbors on the second sequence to obtain a third sequence of $n-2$ entries, and continue until the final sequence produced consists of a single number x_n. Show that $x_n < 2/n$.

Fields: combinatorics, algebra, analysis **Position:** B2 **Answer:** $x_n = \dfrac{2^n-1}{n\,2^{n-1}} = \dfrac{2}{n} - \dfrac{1}{n\,2^{n-1}} < \dfrac{2}{n}$.

Layer 1 THE WHAT — surface analysis

The natural attempt. You compute small cases to get a feel for it. For $n=2$: $1, \frac{1}{2} \to x_2 = \frac{3}{4}$, and $2/n = 1$, so $\frac{3}{4} < 1$. For $n=3$: $1, \frac{1}{2}, \frac{1}{3} \to \frac{3}{4}, \frac{5}{12} \to x_3 = \frac{7}{12}$, and $2/n = \frac{2}{3} \approx 0.667$, so $\frac{7}{12} \approx 0.583$ works. Encouraged, you reach for induction on n, the reflex for any statement indexed by a positive integer. But the induction stalls immediately: what would "x_{n-1}" even mean inside a proof about x_n? The process for n starts from $1, \frac{1}{2}, \ldots, \frac{1}{n}$; the process for $n-1$ starts from the different sequence $1, \frac{1}{2}, \ldots, \frac{1}{n-1}$. There is no single averaging step that turns one process into the other — you cannot "peel off the last stage" of the n-process and land on the $(n-1)$-process, because the seed sequence itself is a different length before any averaging has happened. Pushing harder, you try to track the whole triangular array of partial averages explicitly and bound its bottom entry directly by crude estimates (every stage of averaging keeps every entry $\le$ the maximum of its two parents, so $x_n \le$ max of the original sequence $= 1$ — true, but far too weak to give $2/n$). The recursive handle that induction wants is not there in this form, and the direct triangular bookkeeping grows faster than it illuminates.

The model chain. Treat one averaging step as a linear operator on sequences: for a sequence a, define $L(a)_j = \frac{1}{2}(a_j + a_{j+1})$. The whole process is exactly $n-1$ applications of L to the original sequence, and x_n is the single surviving entry, $(L^{n-1}a)_1$. Write $L = \frac{1}{2}(I+S)$, where I is the identity operator and S is the shift $(Sa)_j = a_{j+1}$. Since I and S are linear operators acting on the same sequence, they commute, so the binomial theorem applies to their sum exactly as it would to two numbers:

$$L^{n-1} = \frac{1}{2^{n-1}}(I+S)^{n-1} = \frac{1}{2^{n-1}} \sum_{i=0}^{n-1} \binom{n-1}{i} S^i.$$

Reading off the first entry, with $a_{1+i} = \dfrac{1}{i+1}$,

$$x_n = (L^{n-1}a)_1 = \frac{1}{2^{n-1}} \sum_{i=0}^{n-1} \binom{n-1}{i} \frac{1}{i+1}.$$

Now use the identity $\binom{n-1}{i}\frac{1}{i+1} = \binom{n}{i+1}\frac{1}{n}$ (both sides equal $\dfrac{(n-1)!}{(i+1)!\,(n-1-i)!}$, since $\binom{n}{i+1}/n = \frac{n!}{(i+1)!(n-1-i)!} \cdot \frac{1}{n}$), to get

$$x_n = \frac{1}{n\,2^{n-1}} \sum_{i=0}^{n-1} \binom{n}{i+1} = \frac{1}{n\,2^{n-1}} \sum_{m=1}^{n} \binom{n}{m} = \frac{1}{n\,2^{n-1}}(2^n - 1),$$

using $\sum_{m=0}^{n} \binom{n}{m} = 2^n$ and discarding the $m=0$ term, which is 1. So

$$x_n = \frac{2^n-1}{n\,2^{n-1}} = \frac{2}{n} - \frac{1}{n\,2^{n-1}} < \frac{2}{n},$$

the last step being immediate since $\frac{1}{n2^{n-1}} > 0$.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths agree that the process is linear and can, in principle, be tracked stage by stage through $n-1$ rounds of averaging. The split occurs at the question each side asks next. The natural attempt asks "how does x_n relate to x_{n-1}?" — the wrong question, because the seed sequence changes length with n, so there is no clean recursive descent from the n-problem to the $(n-1)$-problem. The model chain instead asks "what does applying this same linear operator k times look like in closed form?", which needs no reference to x_{n-1} at all.

> *This isn't really a sequence problem, it's the binomial theorem in disguise — averaging twice in a row is exactly what a coin flipped twice does to a starting position, so I should write the final answer as a sum weighted by binomial coefficients, not chase a recursion in n.*

Layer 3 THE WHY

The operator $L(a)_j = \frac{1}{2}(a_j + a_{j+1})$ is linear, and it equals $\frac{1}{2}(I+S)$ where S is the shift. That I and S commute is not a special feature of this problem; it is true of any two linear operators where one of them is the identity, or more generally whenever the order of application does not change the composite map — here true immediately because S acts on the whole sequence and I leaves it untouched, so $IS = SI = S$. Once operators commute, the binomial theorem for $(I+S)^k$ is not an analogy borrowed from arithmetic; it is a direct consequence of the same distributive expansion that proves the binomial theorem for numbers, applied formally to operators instead. This is exactly the mechanism that produces Pascal's triangle whenever a two-outcome step ("keep this entry" vs. "shift to the next one") is repeated independently k times: it is the same structure as k fair coin flips, where the number of heads after k flips is Binomial$(k, \frac{1}{2})$, and $(L^k a)_1$ is literally the expectation of a_{1+X} for $X \sim$ Binomial$(k, \frac{1}{2})$. That is why Kedlaya's official solution can assert, in one terse line, that "the j-th entry of the k-th sequence is $\sum_i \binom{k-1}{i-1}/(2^{k-1}(i+j-1))$" and prove it "by induction": the binomial coefficients are not discovered by pattern-matching small cases, they are forced the moment you recognise repeated pairwise averaging as the operator $\frac{1}{2}(I+S)$ raised to a power. The archive's job is to supply the derivation the official line skips: there is no way to expand a repeated 2-term linear combination rule without producing exactly these weights.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the phrase "repeat the averaging of neighbours" itself: any process built from a fixed 2-term linear combination rule, applied over and over, should immediately trigger writing that rule as a linear operator.
2. That cue suggests writing the single step as $\frac{1}{2}(I+S)$ and expanding the k-fold iterate with the binomial theorem, rather than attempting to track individual entries by hand.
3. Hesitation happens at the instinctive move to induct on n directly, since almost every Putnam sequence problem yields to induction; it is reinforced away by noticing that the seed sequence for n and for $n-1$ are genuinely different objects, so there is no "smaller version of the same problem" sitting inside this one — only the same operator applied to different starting data.
4. The trap is computing a few small values of x_n and trying to spot the closed form $\dfrac{2^n-1}{n2^{n-1}}$ directly from the raw fractions $\frac{3}{4}, \frac{7}{12}, \ldots$ — the pattern does not announce itself without the operator viewpoint, so time spent guessing at the arithmetic pattern is time not spent finding the structural reason for it.
5. To recognise this faster next time: whenever a problem describes "repeat this simple two-term rule k times," write the single step as $(I+S)/2$ (or whatever the relevant identity-plus-shift combination is) before doing any arithmetic.
6. The reusable decision rule: to find a closed form for the result of repeatedly applying a linear two-term averaging (or any linear recurrence built from a sequence and one shifted

copy of itself) k times, express the single step as $(I+S)/2$, expand the k-th power by the binomial theorem, and evaluate the resulting binomially-weighted sum against the original sequence.

▷ **Cue:** a rule applied repeatedly by combining each term with its neighbour (average, or any fixed 2-term linear rule) $\longrightarrow$ **Move:** write the single step as $(I+S)/2$ and expand the k-th iterate by the binomial theorem

2003 B3

2003 B3. Show that for each positive integer n,

$$n! = \prod_{i=1}^{n} \operatorname{lcm}\{1, 2, \ldots, \lfloor n/i \rfloor\}.$$

(Here lcm denotes the least common multiple, and $\lfloor x \rfloor$ denotes the greatest integer $\leq x$.)

Fields: number theory **Position:** B3 **Answer:** The identity holds; proved by matching, for every prime p, the exponent of p on both sides.

Layer 1 THE WHAT — surface analysis

The natural attempt. You fix a prime p — the right instinct, since $n!$ immediately calls Legendre's formula to mind: $v_p(n!) = \sum_{k\geq 1} \lfloor n/p^k \rfloor$. To match the right side you need the analogous formula for an lcm: $v_p(\operatorname{lcm}\{1,\ldots,m\})$ is the largest power of p that is $\leq m$, i.e. $\lfloor \log_p m \rfloor$. So the identity to prove, prime by prime, becomes

$$\sum_{i=1}^{n} \lfloor \log_p \lfloor n/i \rfloor \rfloor = \sum_{k\geq 1} \lfloor n/p^k \rfloor.$$

The natural next move is algebraic: attack the nested floor expression directly. You might split the sum over i into ranges where $\lfloor n/i \rfloor$ sits in $[p^k, p^{k+1})$ for each k, trying to invert the floor by locating exactly which i give $\lfloor n/i \rfloor \geq p^k$. That range turns out to be $i \leq n/p^k$ — but re-deriving that fact from inside the floor-algebra feels like proving the identity by restating it, and it's easy to lose the thread of what's been shown versus assumed. Pushed further, you retreat to checking small cases ($n = 4, 5, 6$, small p) and perhaps an induction on n: this can be made to work (it is essentially the official solution's second route), but it forces a case split on whether each divisor n/i of n is a prime power, a non-prime-power composite, or not a divisor at all — correct, but fragile, and it verifies the identity without ever showing why the two floor expressions had to coincide.

The model chain. Read every floor as a count, not a number. Two facts, both immediate from the definition of $\lfloor \cdot \rfloor$:

$$\lfloor n/p^k \rfloor = \#\{\, i \geq 1 : i \leq n/p^k \,\} = \#\{\, i \geq 1 : ip^k \leq n \,\},$$

$$\lfloor \log_p m \rfloor = \#\{\, k \geq 1 : p^k \leq m \,\},$$

the second because the set of k with $p^k \leq m$ is an initial segment $\{1, 2, \ldots, K\}$ (as p^k is increasing in k), so its size is its largest element K, which by definition is $\lfloor \log_p m \rfloor$. Set $m = \lfloor n/i \rfloor$. Since p^k is an integer, $p^k \leq \lfloor n/i \rfloor$ if and only if $p^k \leq n/i$, if and only if $ip^k \leq n$. So

$$\lfloor \log_p \lfloor n/i \rfloor \rfloor = \#\{\, k \geq 1 : ip^k \leq n \,\}.$$

Now sum. The left side of the target identity is $\sum_{i\geq 1} \#\{k \geq 1 : ip^k \leq n\}$, and the right side is $\sum_{k\geq 1} \#\{i \geq 1 : ip^k \leq n\}$. Both are the total number of pairs of positive integers (i, k) with

$ip^k \le n$ — the same finite set, counted once by rows and once by columns of the same 0/1 incidence array. They are equal by nothing more than reordering a finite sum. This holds for every prime p, hence $v_p(n!) = v_p\big(\prod_i \operatorname{lcm}\{1,\dots,\lfloor n/i\rfloor\}\big)$ for every p, so the two integers are equal.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths arrive at the same two floor expressions, $\sum_i \lfloor \log_p \lfloor n/i \rfloor \rfloor$ and $\sum_k \lfloor n/p^k \rfloor$. At that point the natural attempt keeps trying to prove an identity *between* two nested floor functions using floor-function algebra — bounding, casework on fractional parts, induction on n. The model chain instead asks what each floor is secretly counting, and the identity dissolves into "the same set, counted two ways."

> *A floor function bounding a count of positive integers is already a count — stop treating $\lfloor n/p^k \rfloor$ and $\lfloor \log_p m \rfloor$ as numbers to manipulate and start treating them as sizes of the same set of pairs (i,k) with $i \cdot p^k \le n$, just grouped by k or by i.*

Layer 3 THE WHY

Neither half of this identity is a coincidence waiting to be verified; both are counting statements the moment you look at their definitions correctly. Legendre's formula itself, $v_p(n!) = \sum_k \lfloor n/p^k \rfloor$, is normally *proved* by exactly this counting device — each multiple of p^k among $1,\dots,n$ contributes one to the count $\lfloor n/p^k \rfloor$, and a number divisible by p exactly j times gets counted once for each $k = 1,\dots,j$, so summing over k correctly totals up every factor of p in $n!$. Recognising that as a counting statement, rather than quoting it as a formula, is what makes the reinterpretation of $\lfloor n/p^k \rfloor$ as $\#\{i : ip^k \le n\}$ feel forced rather than guessed: it is just the definition of $\lfloor x \rfloor$ as "the number of positive integers not exceeding x," applied to $x = n/p^k$. The lcm side is the same fact from the other direction: $\operatorname{lcm}\{1,\dots,m\}$ contains p to the power of the *largest* power of p not exceeding m (any higher power isn't present among $1,\dots,m$; any lower power is dominated by that largest one), and "the largest k with $p^k \le m$" is, by the same floor-as-count logic, "the number of $k \ge 1$ with $p^k \le m$," because those k form an unbroken initial run $1,2,\dots,K$. Once both sides are seen as counts of the identical relation "i times p to the k is at most n," summing over i first or over k first is nothing but interchanging the order of a finite double sum — the discrete form of Fubini's theorem, which is itself just the fact that addition of finitely many numbers does not care what order you perform it in. There is no trick anywhere in this problem: both sides are the same count of the same relation, viewed from two directions, and the identity is forced by that alone.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is seeing a factorial equal a product of lcm's in the same line: factorial should trigger "fix a prime p, apply Legendre's formula," and lcm should trigger "the exponent of p is the largest power of p not exceeding the bound." Both should be written down before any algebra starts.
2. That cue suggests reducing the whole identity to one identity per prime p between the two resulting exponent expressions.
3. Hesitation sets in at the two nested floor expressions, which look like they demand to be manipulated as formulas. It is reinforced away by rewriting each floor as "the size of an explicit set of positive integers satisfying an inequality," which converts a floor-function identity into a counting identity.
4. The trap is reaching for induction on n as the default proof technique for a factorial identity: it works (as in the official second solution) but only after a delicate case split on whether each divisor n/i of n is a prime power, a composite non-prime-power, or not a divisor of n at all — a correct but fragile path, easy to mishandle at the boundary between cases.

5. To recognise this faster next time: whenever an identity has nested floor or floor-of-logarithm expressions summed over an index, ask whether each floor term is secretly the count of lattice points under a curve, and whether the two sides of the identity are that same region counted by rows versus by columns.

6. The reusable decision rule: to prove $\sum_i g(f(i)) = \sum_k h(k)$ for floor-type functions f, g, h, look for a single set of pairs (i, k) satisfying one inequality, such that $g(f(i))$ counts the pairs with that i and $h(k)$ counts the pairs with that k; the identity then follows automatically from interchanging the order of summation.

▷ **Cue:** an identity between a factorial-type product and an lcm-type product $\longrightarrow$ **Move:** fix a prime p, compute v_p on both sides via Legendre's formula and the lcm valuation, then look for a shared counting set to force the two sums equal

2003 B4

2003 B4. Let $f(z) = az^4 + bz^3 + cz^2 + dz + e = a(z - r_1)(z - r_2)(z - r_3)(z - r_4)$ where a, b, c, d, e are integers, $a \neq 0$. Show that if $r_1 + r_2$ is a rational number and $r_1 + r_2 \neq r_3 + r_4$, then $r_1 r_2$ is a rational number.

Fields: algebra (polynomials, Vieta's formulas); a whisper of Galois theory **Position:** B4 — nominally mid-paper, in practice one of the two easiest problems on the 2003 B paper once the roots are grouped in pairs **Answer:** Proof: with $g = r_1 + r_2$, $h = r_3 + r_4$, $u = r_1 r_2$, $v = r_3 r_4$, Vieta gives h, $u + v$ and $hu + gv$ rational, and $(g - h)u = g(u + v) - (hu + gv)$, so u is rational because $g \neq h$.

Layer 1 THE WHAT — surface analysis

The natural attempt. The strong student reads "rational", "integer coefficients", "quartic" and reaches for factorisation. The thought is: if $r_1 + r_2 = g$ is rational, perhaps $(z - r_1)(z - r_2) = z^2 - gz + u$ is a factor of f with rational coefficients, and then u is rational for free. So write

$$f(z) = a(z^2 - gz + u)(z^2 - hz + v)$$

and try to show the first factor lies in $\mathbb{Q}[z]$. That is precisely the thing to be proved, so the student looks for a divisibility argument: the gcd of f with something rational? There is no second rational polynomial in sight to take a gcd with. Next attempt: Galois-type reasoning ("conjugates of r_1 are among the r_i…"), which is correct in spirit but needs automorphisms of $\overline{\mathbb{Q}}$ and a case analysis the student may not be able to justify under time pressure. A third attempt expands all four Vieta relations and tries to solve for r_1, r_2 individually, meeting square roots of discriminants and getting lost. The wall: the student is trying to control the *roots*, when only symmetric combinations of pairs are ever rational.

The model chain. **Step 1: name the pair data.** Put

$$g = r_1 + r_2, \quad h = r_3 + r_4, \quad u = r_1 r_2, \quad v = r_3 r_4.$$

Then $(z - r_1)(z - r_2) = z^2 - gz + u$ and $(z - r_3)(z - r_4) = z^2 - hz + v$, so

$$f(z) = a(z^2 - gz + u)(z^2 - hz + v) = a\big(z^4 - (g + h)z^3 + (gh + u + v)z^2 - (gv + hu)z + uv\big).$$

Step 2: compare coefficients. Since $a \neq 0$ and $a, b, c, d, e \in \mathbb{Z}$,

$$g + h = -\tfrac{b}{a}, \qquad gh + u + v = \tfrac{c}{a}, \qquad gv + hu = -\tfrac{d}{a}, \qquad uv = \tfrac{e}{a}$$

are all rational.

Step 3: peel off the unknowns in order. g is rational by hypothesis, so $h = -b/a - g$ is rational. Then $u + v = c/a - gh$ is rational. Now u and v satisfy the linear system

$$\begin{pmatrix} 1 & 1 \\ h & g \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} c/a - gh \\ -d/a \end{pmatrix},$$

whose coefficients and right-hand side are rational. Eliminating v:

$$g(u + v) - (hu + gv) = (g - h)\, u.$$

The left side is rational. Since $g \neq h$ by hypothesis, $u = \dfrac{g(u+v) - (hu+gv)}{g-h}$ is rational. □

(The same computation gives v rational; the fourth relation $uv = e/a$ was never needed.)

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains write f as a product of two quadratics $a(z^2 - gz + u)(z^2 - hz + v)$. They split at the next line. (N) asks "how do I prove this factor has rational coefficients?" and goes looking for an external reason (gcds, conjugates, discriminants). (M) expands the product and reads the coefficients as *equations in the unknowns* h, u, v, noticing that once g is known they are solvable one at a time — and that the last one, for u and v, is *linear*.

> *I am told something about $r_1 + r_2$ and asked about $r_1 r_2$. Those are the coefficients of $(z - r_1)(z - r_2)$. So I split f into the pair $\{r_1, r_2\}$ and the pair $\{r_3, r_4\}$ and give their sums and products names: g, u and h, v.*

> *Vieta for the full quartic hands me four rational numbers built from g, h, u, v. I know g. Which unknowns can I peel off, in which order? h from the cubic coefficient, $u + v$ from the quadratic one, and then the linear coefficient is linear in u and v.*

> *I have two linear equations in u, v with determinant $g - h$. The hypothesis $r_1 + r_2 \neq r_3 + r_4$ is exactly "the determinant is nonzero". That is why it is there.*

- ▷ **Cue:** a hypothesis about a sum of some roots, a conclusion about their product ⟶ **Move:** group the roots into those pairs, name each pair's sum and product, and rewrite Vieta in terms of the pair data
- ▷ **Cue:** Vieta relations where some quantities are known and others unknown ⟶ **Move:** solve for the unknowns one at a time, looking for an equation that is linear in what remains
- ▷ **Cue:** a hypothesis of the form "$X \neq Y$" with no obvious role ⟶ **Move:** look for the division by $X - Y$ it licenses; it is usually a determinant or a denominator

Layer 3 THE WHY

Why is grouping the roots in pairs natural rather than inspired? Because of what the problem mentions. The hypothesis is about $r_1 + r_2$ and the conclusion about $r_1 r_2$: those are, by definition, the two non-leading coefficients of $(z - r_1)(z - r_2)$. The problem is literally asking about the quadratic factor $z^2 - (r_1 + r_2)z + r_1 r_2$. Once that quadratic is written down, its partner $z^2 - (r_3 + r_4)z + r_3 r_4$ is forced, and multiplying them out is the only thing to do with the given integer coefficients.

Why does the natural attempt stall? It treats rationality of $z^2 - gz + u$ as a property to be *proved from outside* (divisibility, conjugation) instead of *computed from inside*. The coefficients b, c, d are four-variable polynomials in g, h, u, v, and in each of them the unknowns u, v appear at most to the first power, apart from the product uv. Rationality questions about solutions

of linear systems with rational coefficients have one answer: if the determinant is nonzero, Cramer's rule gives rational solutions. So the problem is linear algebra over $\mathbb{Q}$ wearing a polynomial costume.

Why is the hypothesis $g \neq h$ there? Kedlaya's first solution writes "$g(u+v) - (gv+hu) = (g-h)u$ is rational. Since $g \neq h$, u is rational." Here is where that comes from: the pair of equations

$$u + v = \text{rational}, \qquad hu + gv = \text{rational}$$

has matrix $\begin{pmatrix} 1 & 1 \\ h & g \end{pmatrix}$ of determinant $g - h$. If $g = h$, the second equation is g times the first and says nothing new, so u and v are only known through $u + v$ and uv — that is, as the roots of a rational quadratic, which may be irrational. The official note gives the witness: $x^4 - 2 = (x^2 - \sqrt{2})(x^2 + \sqrt{2})$, where both pairs have sum 0 (rational and equal) while $u = -\sqrt{2}$ is irrational. So the hypothesis is not decoration; it is the exact condition for the linear system to be invertible, and the counterexample is what happens when it fails.

The Galois solution is the same fact in different clothes: an automorphism fixing $\mathbb{Q}$ either keeps the pair $\{r_1, r_2\}$ or swaps it with $\{r_3, r_4\}$ (it must preserve g, and the only other pair with sum g would need $h = g$). It never swaps, so u is fixed, so u is rational. That argument needs the theorem that a complex algebraic number fixed by every automorphism of $\overline{\mathbb{Q}}$ is rational — true, but far outside what the problem needs. Reach for the cheap tool first.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the pairing of hypothesis and conclusion: $r_1 + r_2$ given, $r_1 r_2$ wanted. On seeing it, write $f(z) = a(z^2 - gz + u)(z^2 - hz + v)$ with the four names defined.
2. Vieta's formulas in the pair variables, then elimination: solve for h, then $u + v$, then use the linear coefficient to isolate $(g - h)u$.
3. Hesitation occurs at the z-coefficient $gv + hu$: it mixes the unknowns with g and h and looks unhelpful. It is reinforced away by asking "which combination of my known quantities isolates u?" — multiply $u + v$ by g and subtract, which is Gaussian elimination on a 2×2 system.
4. The trap is trying to prove that $z^2 - gz + u$ is a rational factor by an external argument (gcds, irreducibility, conjugates), or solving for the individual roots. A second trap is ignoring the hypothesis $g \neq h$ and producing a "proof" that would also prove the false statement for $x^4 - 2$.
5. When the data are symmetric functions of a *subset* of the roots, recast the polynomial as a product over that subset and its complement, and treat the new coefficients as unknowns. Check which equations are linear in them.
6. Reusable decision rule: *to prove that a quantity lies in a field (here $\mathbb{Q}$), exhibit it as the solution of a system with coefficients in that field and nonzero determinant; every "$X \neq Y$" hypothesis is a candidate determinant, and the example with $X = Y$ tells you why it is needed.*

2003 B5

2003 B5. Let A, B, and C be equidistant points on the circumference of a circle of unit radius centered at O, and let P be any point in the circle's interior. Let a, b, c be the distance from P to A, B, C, respectively. Show that there is a triangle with side lengths a, b, c, and that the area of this triangle depends only on the distance from P to O.

Fields: Euclidean geometry; complex numbers; symmetric sums of equally spaced cosines **Position:** B5 — late in the paper, but two independent cheap routes exist; the difficulty is choosing a representation of "area from side lengths" that does not involve square roots **Answer:** With $d = OP < 1$, the triangle exists and has area $\frac{\sqrt{3}}{4}(1 - d^2)$.

Layer 1 THE WHAT — surface analysis

The natural attempt. Set O at the origin, $A = (1, 0)$, B, C at angles $\pm 2\pi/3$, and $P = (d\cos\varphi, d\sin\varphi)$. The cosine rule in triangle OPA gives

$$a^2 = 1 + d^2 - 2d\cos\varphi, \quad b^2 = 1 + d^2 - 2d\cos(\varphi - \tfrac{2\pi}{3}), \quad c^2 = 1 + d^2 - 2d\cos(\varphi + \tfrac{2\pi}{3}).$$

To show a triangle exists, the student tries to prove $a < b + c$ directly: square to get $a^2 < b^2 + c^2 + 2bc$, isolate bc, square again, and face a quartic trigonometric inequality in φ and d — and that is only one of three inequalities, with the ordering of a, b, c depending on φ. For the area, the student writes Heron as $\sqrt{s(s-a)(s-b)(s-c)}$ with $s = \frac{a+b+c}{2}$ and is immediately stuck: a, b, c are square roots of trigonometric expressions and $s - a$ does not simplify. That is the wall. Every quantity the student is manipulating is a *length*, and the lengths are square roots; the problem's data are *squared* lengths.

The model chain. **Step 1: squared distances.** With A, B, C at angles $0, \pm\frac{2\pi}{3}$ on the unit circle and P at distance $d < 1$ from O and angle φ, the cosine rule gives

$$a^2 = m - 2d\cos\alpha, \qquad b^2 = m - 2d\cos\beta, \qquad c^2 = m - 2d\cos\gamma,$$

where $m = 1 + d^2$ and $\alpha = \varphi$, $\beta = \varphi - \frac{2\pi}{3}$, $\gamma = \varphi + \frac{2\pi}{3}$.

Step 2: two facts about three equally spaced angles. $\cos\alpha + \cos\beta + \cos\gamma = \mathrm{Re}\left(e^{i\varphi}(1 + \omega + \omega^2)\right) = 0$ with $\omega = e^{2\pi i/3}$. And since $\cos^2 t = \frac{1}{2} + \frac{1}{2}\cos 2t$, and $2\alpha, 2\beta, 2\gamma$ are again three angles spaced by $\frac{4\pi}{3} \equiv -\frac{2\pi}{3}$,

$$\cos^2\alpha + \cos^2\beta + \cos^2\gamma = \tfrac{3}{2}.$$

Hence

$$a^2 + b^2 + c^2 = 3m, \qquad a^4 + b^4 + c^4 = 3m^2 - 4dm \cdot 0 + 4d^2 \cdot \tfrac{3}{2} = 3m^2 + 6d^2.$$

Step 3: Heron in squared form. For any nonnegative a, b, c put

$$Q = (a+b+c)(-a+b+c)(a-b+c)(a+b-c) = (a^2+b^2+c^2)^2 - 2(a^4+b^4+c^4).$$

(Expand: $(b+c)^2 - a^2$ times $a^2 - (b-c)^2$ gives $2a^2b^2 + 2b^2c^2 + 2c^2a^2 - a^4 - b^4 - c^4$.) Here

$$Q = 9m^2 - 6m^2 - 12d^2 = 3\big((1+d^2)^2 - 4d^2\big) = 3(1-d^2)^2.$$

Step 4: existence. Since P is interior, $d < 1$ and $Q > 0$. Also $a, b, c > 0$ (P is not on the circle), so $a + b + c > 0$ and the product xyz of $x = -a + b + c$, $y = a - b + c$, $z = a + b - c$ is positive. Then either all three are positive or exactly two are negative. But $x + y = 2c > 0$, $y + z = 2a > 0$, $z + x = 2b > 0$, so no two are negative. Hence $x, y, z > 0$: the three strict triangle inequalities hold, and a triangle with sides a, b, c exists.

Step 5: area. By Heron, $16K^2 = Q$, so

$$K = \frac{\sqrt{Q}}{4} = \frac{\sqrt{3}}{4}(1 - d^2),$$

which depends only on $d = OP$. □

Second model (the construction the official solution uses). Put A, B, C at $1, \omega, \omega^2$ and P at $x \in \mathbb{C}$, $|x| = d$. Then

$$(x-1) + \omega(x-\omega) + \omega^2(x-\omega^2) = x(1+\omega+\omega^2) - (1+\omega^2+\omega^4) = 0,$$

because $\omega^4 = \omega$. So the vectors $x-1$, $\omega(x-\omega)$, $\omega^2(x-\omega^2)$ — that is, $\overrightarrow{AP}$, and $\overrightarrow{BP}$, $\overrightarrow{CP}$ rotated by 120° and 240° — close up into a triangle with side lengths a, b, c. With $v_1 = x-1$, $v_2 = \omega(x-\omega)$, the area is $\frac{1}{4}|v_1\overline{v_2} - \overline{v_1}v_2|$, and

$$v_1\overline{v_2} - \overline{v_1}v_2 = (\omega^2-\omega)\left(|x|^2-1\right) = i\sqrt{3}\,(1-|x|^2),$$

using $\omega^2 - \omega = -i\sqrt{3}$. Area $\frac{\sqrt{3}}{4}(1-d^2)$ again, nonzero because $d<1$. (The official text prints the middle quantity as $i\sqrt{3}(|x|^2-1)$; the sign is off by one factor of -1, which the absolute value absorbs.)

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains start from the cosine rule for a^2, b^2, c^2. They split at the choice of what to compute with. (N) takes square roots and works with a, b, c and $s = \frac{a+b+c}{2}$. (M) never takes a square root: it writes both the existence question and the area as the *single* quantity $Q = 2\sum a^2b^2 - \sum a^4$, a symmetric polynomial in a^2, b^2, c^2, which the cosine rule feeds directly.

> *The distances come to me squared — the cosine rule gives a^2, not a. So I want the area in terms of a^2, b^2, c^2 only. Heron multiplied out does that:* $16K^2 = 2(a^2b^2 + b^2c^2 + c^2a^2) - (a^4 + b^4 + c^4)$.

> *Three points equally spaced round the circle means three cosines of angles $2\pi/3$ apart. Their sum is 0 and the sum of their squares is $3/2$, whatever φ is. Anything symmetric of degree two in a^2, b^2, c^2 will lose its φ. That is the "depends only on OP".*

> *The same Q decides existence: $Q > 0$ with positive sides forces all three triangle inequalities. I do not need to prove them one at a time.*

- ▷ **Cue:** lengths given as distances from a moving point, so their squares come from the cosine rule or $|z-w|^2$ ⟶ **Move:** work with the squares; use Heron in the form $16K^2 = 2\sum a^2b^2 - \sum a^4$
- ▷ **Cue:** "is there a triangle with sides a, b, c?" ⟶ **Move:** compute $(a+b+c)(-a+b+c)(a-b+c)(a+b-c)$; positive means yes, zero means degenerate
- ▷ **Cue:** a symmetric function of $\cos(\varphi + 2\pi k/n)$, $k = 0, \ldots, n-1$ ⟶ **Move:** power sums of equally spaced cosines of low degree are independent of φ; use $1 + \omega + \cdots + \omega^{n-1} = 0$
- ▷ **Cue:** three vectors of lengths a, b, c that can be rotated independently ⟶ **Move:** look for rotations (multiplications by roots of unity) making them sum to zero; then they form a triangle

Layer 3 THE WHY

Why squared lengths? Because of what a distance is. In coordinates, the distance from P to A is defined through $a^2 = (x_P - x_A)^2 + (y_P - y_A)^2$: the square is the polynomial object, the length is a square root laid on top. Every tool that relates the position of P to a — cosine rule, dot product, $|z-w|^2 = (z-w)(\overline{z}-\overline{w})$ — delivers a^2. So a solver who wants to avoid radicals should ask: *is there a version of the target that is a polynomial in the squares?* For area there is, because area squared is a polynomial in the squared sides (Heron's formula squared out is a polynomial in a^2, b^2, c^2: the four factors pair up into $\left((b+c)^2 - a^2\right)\left(a^2 - (b-c)^2\right)$, which is a difference of two squares, $(2bc)^2 - (b^2+c^2-a^2)^2$, and the odd term $2bc$ appears only squared). The natural attempt used Heron in its s-form, which is the right theorem in the wrong coordinates.

Why does the answer depend only on d? Rotating P about O by $2\pi/3$ permutes a, b, c cyclically, and area is symmetric in the sides; so the area is automatically invariant under a rotation of order 3. That alone would not kill all dependence on φ. What does is degree: a^2, b^2, c^2

are *linear* in the three cosines, and Q is quadratic in the squares, so Q involves only $\sum\cos$ and $\sum\cos^2$ (and $\sum\cos\beta\cos\gamma$, which is determined by these). For three equally spaced angles, the first and second power sums are constants — they are the real parts of $e^{ik\varphi}(1+\omega^k+\omega^{2k})$ for $k=1,2$, and $1+\omega^k+\omega^{2k}=0$ unless $3\mid k$. A degree-three quantity would have seen $\cos 3\varphi$. The problem is designed so that area sits exactly at the degree where the symmetry of the equilateral triangle wipes out φ.

Why does "interior" matter? $Q=3(1-d^2)^2$ vanishes exactly on the circle. There the triangle is degenerate: the largest of a,b,c equals the sum of the other two (this is van Schooten's theorem, a special case of Ptolemy for the cyclic quadrilateral $PABC$). So the hypothesis is not decorative; it is the boundary of the region where the triangle exists.

Kedlaya's first solution begins with the identity $(x-1)+\omega(x-\omega)+\omega^2(x-\omega^2)=0$. Here is where that comes from. You want three vectors of lengths a,b,c that sum to zero. The available vectors of those lengths are $x-1$, $x-\omega$, $x-\omega^2$, and multiplying by a unit complex number rotates without changing length. Their sum $3x-(1+\omega+\omega^2)=3x$ is not zero, but the x-terms can be killed by weighting with $1,\omega,\omega^2$ (whose sum is zero); the constant terms become $1+\omega^2+\omega^4=1+\omega^2+\omega$, also zero. Both cancellations are the single fact $1+\omega+\omega^2=0$, used twice. The area then comes from the cross product $\mathrm{Im}(v_1\overline{v_2})$, where the $|x|^2$ term and the constant term survive and the terms linear in x cancel — the complex-number version of "degree two, so φ drops out".

Layer $+\alpha$ BACKTESTING

1. The first cue is "area of a triangle with side lengths a,b,c" where a,b,c are distances. On seeing it, write $16K^2=2\sum a^2b^2-\sum a^4$ and the cosine-rule expressions for a^2,b^2,c^2.
2. Squared-length computation: Heron in squared form plus the power sums $\sum\cos=0$, $\sum\cos^2=\frac{3}{2}$ for equally spaced angles. Alternatively, complex numbers with $\omega=e^{2\pi i/3}$.
3. Hesitation occurs at existence: proving three triangle inequalities looks like a separate, harder job. It is reinforced away by recognising that the Heron product Q is the existence test, since two of its three non-trivial factors cannot be negative at once.
4. The trap is the s-form of Heron and direct attacks on $a<b+c$, both of which force square roots of trigonometric expressions. A quieter trap is forgetting to use $d<1$ and so missing that the triangle is degenerate on the circle.
5. Ask, before computing: *does my formula take the data in the form the problem supplies it?* Distances arrive squared; roots of unity arrive as sums that vanish. Choose the version of the target formula that consumes exactly those.
6. Reusable decision rule: *when lengths are defined as distances, compute with their squares and choose the squared form of every formula (Heron, triangle inequality test, cross product); when points are equally spaced on a circle, low-degree symmetric expressions are independent of the rotation angle.*

2003 B6

2003 B6. Let $f(x)$ be a continuous real-valued function defined on the interval $[0,1]$. Show that

$$\int_0^1\int_0^1 |f(x)+f(y)|\,dx\,dy \geq \int_0^1 |f(x)|\,dx.$$

Fields: real analysis; inequalities; discretisation **Position:** B6 — the last problem, and hard: the constant 1 is sharp, so no lossy estimate survives, and the proof must track signs **Answer:** Proof: split the domain by the sign of f; the inequality reduces to a four-number inequality in the measures p,q of the positive and negative sets and the masses X,Y of $|f|$ on them, namely $(p-q)(X-Y)+2|qX-pY|\geq 0$.

Layer 1 THE WHAT — surface analysis

The natural attempt. The strong student reaches for the triangle inequality, in each of its forms,

and integrates.

- $|f(x)+f(y)| \geq |f(x)|-|f(y)|$. Integrating over the square gives $\int|f|-\int|f| = 0$. Useless.
- $\iint|f(x)+f(y)| \geq \left|\iint(f(x)+f(y))\right| = 2\left|\int f\right|$. This bounds by $2|\int f|$, which can be 0 while $\int|f|>0$ (take $f(x)=x-\frac{1}{2}$). Useless in general.
- $|f(x)+f(y)| \geq \big||f(x)|-|f(y)|\big|$, then Cauchy–Schwarz or Jensen on the right. Every such bound depends only on $|f|$, and it cannot work: when $f(x)=-f(y)$ the left side is 0, so any pointwise lower bound in terms of $|f(x)|, |f(y)|$ alone must vanish on the diagonal $|f(x)|=|f(y)|$. For $f\equiv 1$ such a bound then yields a lower bound ≤ 0 for a left side that is really 2, against a target of 1.

A fourth attempt tries to test extreme cases. If $f\geq 0$ everywhere, the left side is $2\int f$, twice the right. If f is $+1$ on half of $[0,1]$ and -1 on the other half (smoothed near the jump to be continuous), the left side is $\frac{1}{2}\cdot 2+\frac{1}{2}\cdot 0 = 1$ and the right side is 1. So the inequality is *tight*, and the student now knows why the pointwise attempts failed but not what to do. That is the wall: a sharp inequality whose equality case mixes signs in a specific proportion, attacked with tools that forget the signs.

The model chain. **Step 1: a discrete version.** We first prove: for any real $x_1,\dots,x_n$,

$$\frac{1}{n^2}\sum_{i,j=1}^{n}|x_i+x_j| \;\geq\; \frac{1}{n}\sum_{i=1}^{n}|x_i|. \tag{D}$$

Let $P=\{i : x_i\geq 0\}$ and $N=\{i : x_i<0\}$, and put

$$p=\frac{|P|}{n},\quad q=\frac{|N|}{n},\quad X=\frac{1}{n}\sum_{i\in P}|x_i|,\quad Y=\frac{1}{n}\sum_{i\in N}|x_i|,$$

so $p+q=1$ and the right side of (D) is $X+Y$.

Step 2: split the double sum by signs. For $i,j\in P$, $|x_i+x_j|=|x_i|+|x_j|$, so

$$\frac{1}{n^2}\sum_{i,j\in P}|x_i+x_j| = \frac{1}{n^2}\big(|P|\textstyle\sum_P|x_i|+|P|\sum_P|x_j|\big) = 2pX.$$

Likewise the $N\times N$ block contributes $2qY$. For $i\in P$, $j\in N$, $|x_i+x_j| = \big||x_i|-|x_j|\big|$, so by the triangle inequality for sums,

$$\frac{1}{n^2}\sum_{i\in P,\,j\in N}|x_i+x_j| \;\geq\; \frac{1}{n^2}\Big|\sum_{i\in P,\,j\in N}(|x_i|-|x_j|)\Big| = \big|qX-pY\big|,$$

and the $N\times P$ block is the same sum with i,j exchanged. Hence

$$\text{LHS of (D)} \;\geq\; 2pX+2qY+2|qX-pY|.$$

Step 3: reduce to four numbers. Since $p+q=1$, the right side of (D) is $(p+q)(X+Y)$, and

$$2pX+2qY+2|qX-pY|-(p+q)(X+Y) = (p-q)(X-Y)+2|qX-pY|.$$

Lemma. For $p,q,X,Y\geq 0$, $(p-q)(X-Y)+2|qX-pY|\geq 0$.

Proof. If $(p-q)(X-Y)\geq 0$ there is nothing to do. Otherwise the two factors have opposite strict signs. The expression is unchanged by the swap $(p,X)\leftrightarrow(q,Y)$, so we may assume $p>q$ and $Y>X$. Then

$$2(pY-qX)-(p-q)(Y-X) = pY+pX+qY-3qX\geq 0,$$

because each of pY, pX, qY is at least qX (using $p > q \geq 0$, $Y > X \geq 0$). So $2|qX - pY| \geq 2(pY - qX) \geq (p-q)(Y-X) = -(p-q)(X-Y)$. ■

This proves (D).

Step 4: pass to the limit. Apply (D) with $x_i = f(i/n)$. The right side is a Riemann sum for $\int_0^1 |f|$, since $|f|$ is continuous. The left side is a Riemann sum, over the grid of n^2 squares of side $1/n$, for $\iint_{[0,1]^2} |f(x) + f(y)|\, dx\, dy$, since $(x, y) \mapsto |f(x) + f(y)|$ is continuous on the closed square, hence uniformly continuous, hence Riemann integrable with Riemann sums converging as the mesh tends to 0. Letting $n \to \infty$ in (D) preserves the non-strict inequality. □

(With Lebesgue measure one can run Steps 1–3 directly on $E_+ = \{f \geq 0\}$ and $E_- = \{f < 0\}$, as the official first solution does; the discrete route avoids needing to integrate over sets that need not be finite unions of intervals.)

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains begin with the triangle inequality. They split at *where* it is applied. (N) applies one inequality pointwise on the whole square, and every such application discards the sign of f. (M) first cuts the square into four blocks by the signs of $f(x)$ and $f(y)$, computes the two same-sign blocks *exactly* (no inequality at all), and spends the triangle inequality only on the mixed blocks — and there applies it to the *integral*, not the integrand: $\iint \big||f(x)| - |f(y)|\big| \geq \big|\iint (|f(x)| - |f(y)|)\big|$.

> *Equality holds for "+1 on half, −1 on half". So I cannot afford to lose anything where $f(x), f(y)$ have the same sign, and I must keep track of how much of $[0,1]$ is positive. I split by sign.*

> *On the same-sign blocks $|f(x) + f(y)| = |f(x)| + |f(y)|$ exactly. Only the mixed blocks need an estimate, and there the cheapest estimate that still sees the sizes is to pull the absolute value outside the integral.*

> *After the split everything is expressed through four numbers: the measure of the positive set, the measure of the negative set, and the mass of $|f|$ on each. The problem has become an inequality in four real variables. Check it by cases on which is bigger.*

- ▷ **Cue:** an inequality about $|\cdot|$ of a sum, where signs can cancel ⟶ **Move:** split the domain by sign; same-sign pieces are exact, only mixed pieces need an estimate
- ▷ **Cue:** a sharp integral inequality whose equality case is a step function ⟶ **Move:** discretise (or pass to indicator-type data), and reduce to finitely many summary parameters
- ▷ **Cue:** $\iint g(x)h(y)$ or $\iint(\alpha(x) - \beta(y))$ over a product set ⟶ **Move:** it factorises into (measure)×(integral) products; name those products
- ▷ **Cue:** a continuous integral inequality that is easier for finite sums ⟶ **Move:** prove the finite version for all n and take Riemann sums; continuity on a closed box guarantees convergence

Layer 3 THE WHY

Why split by sign, and why is that not a trick? Go to the definition of absolute value: $|t| = t$ for $t \geq 0$ and $|t| = -t$ for $t < 0$. It is linear on each half-line and the only difficulty lives at the switch. $|f(x) + f(y)|$ is therefore linear in $(|f(x)|, |f(y)|)$ on each block where the signs of $f(x)$ and $f(y)$ are fixed and the sign of the sum is determined; the only block where the sign of the sum is *not* determined is the mixed one. So the definition itself tells you the domain splits into four pieces, two of which (the same-sign ones) are exact.

Why were the pointwise attempts doomed? Because of sharpness. The step function $\mathbf{1}_{[0,1/2]} - \mathbf{1}_{(1/2,1]}$ (a limit of continuous functions) gives equality. At equality every estimate in

the proof must be tight. For that function the mixed blocks contribute 0, and a bound that also gives 0 there is forced; that happens exactly when the positive and negative parts balance ($qX = pY$). A pointwise bound which is blind to sign cannot know whether it is sitting in a mixed block or a same-sign block, so it must use the mixed-block value everywhere and loses the factor 2 that the same-sign blocks carry. That is why the model estimates nothing on the same-sign blocks.

Why does the problem reduce to four numbers? Because on a product set $E \times F$ the integrand $|f(x)|$ depends on x alone, so $\iint_{E\times F} |f(x)| = \mu(F) \int_E |f|$: measure of one set times mass on the other. Every term in the split is of this form. The shape of f *within* each sign class is irrelevant; only its total mass and the size of the class matter. That is the WHY behind the official solution's $\mu_\pm$ and $I_\pm$: they are not chosen, they are what the product structure leaves behind.

Kedlaya's first solution ends by saying that "for each of the possible comparisons between μ_+ and μ_-, and between I_+ and I_-, one of the three terms above is manifestly nonnegative." Here is what that compresses. After the split the target is $(p-q)(X-Y) + 2|qX - pY| \geq 0$. If more of the interval is positive *and* the positive part carries more mass (or both are smaller), the product term is already nonnegative. The dangerous case is the mismatch: a large set with small mass against a small set with large mass. But then the mixed term $|qX - pY|$ — which compares "positive mass weighted by negative measure" against the reverse — is large, and the case check shows it is large enough. The whole difficulty of B6 is that bookkeeping; the analysis is only the discretisation in Step 4.

Why discretise? Because (D) is a statement about finitely many real numbers, where splitting into sign classes is counting, and the limit step is standard for continuous integrands on a closed box. The official second solution discretises too, for the same reason. It then runs a different finite argument (moving the smallest absolute value to reduce the number of distinct values), which is sound but longer; the four-number reduction above is the cheaper reroute.

Caveat. The lemma's case check and the Riemann-sum passage have been verified by hand; the discrete inequality (D) was additionally spot-checked numerically on 20,000 random vectors with $n \leq 8$ (minimum observed slack positive). The optional sharper bound $(1 + (1-2\mu)^2) \int |f|$ from Savitt's solution is quoted, not re-proved, here.

The sharp form is worth knowing: with μ the smaller of the measures of $\{f > 0\}$ and $\{f < 0\}$, Savitt shows the left side is at least $(1 + (1-2\mu)^2) \int |f|$, which equals $\int |f|$ only at $\mu = \frac{1}{2}$ — the balanced step function again.

Layer $+\alpha$ BACKTESTING

1. The first cue is $|f(x) + f(y)|$ with f allowed to change sign, and a target in which the constant is 1. On seeing it, test $f \equiv 1$ (ratio 2) and the ± 1 step function (ratio 1), and write down: *sharp; equality needs balanced signs; split by sign.*

2. Sign decomposition into $E_+ \times E_+$, $E_- \times E_-$ and the mixed blocks; exact evaluation on the first two, one triangle inequality on the integral of the mixed block; reduction to four parameters; finite version plus Riemann sums.

3. Hesitation occurs at the four-number inequality, which looks as though it might fail when the larger set carries the smaller mass. It is reinforced away by checking that mismatched case directly: pY, pX, qY each dominate qX, and those three terms are exactly what the case needs.

4. The trap is applying the triangle inequality pointwise on the whole square. Every version either integrates to 0 or to $2|\int f|$, and a sign-blind pointwise bound cannot reach a sharp constant whose equality case mixes signs.

5. Before estimating, find the equality case. If it is a step function (or an extreme configuration), the proof must be exact on the pieces where that function is constant, and estimates are allowed only where it has slack.

6. Reusable decision rule: *for inequalities involving |sum|, partition by sign; compute same-sign pieces exactly; on product domains reduce everything to (measure)×(mass) parameters, then prove the resulting finite-dimensional inequality by cases. If the continuous setting is awkward, prove the discrete analogue for every n and pass to the limit with Riemann sums.*