DAWN PUTNAM SESSION

Putnam 2001 — Papers A and B

Putnam 120 Archive

Contents

Putnam 2001 — Paper A **5**
2001 A1 5
2001 A2 7
2001 A3 10
2001 A4 13
2001 A5 15
2001 A6 18

Putnam 2001 — Paper B **23**
2001 B1 23
2001 B2 25
2001 B3 28
2001 B4 31
2001 B5 33
2001 B6 36

Putnam 2001 — Paper A

2001 A1

2001 A1. Consider a set S and a binary operation $*$, i.e., for each $a, b \in S$, $a * b \in S$. Assume $(a * b) * a = b$ for all $a, b \in S$. Prove that $a * (b * a) = b$ for all $a, b \in S$.

Fields: abstract algebra (binary operations, substitution identities) **Position:** A1 — the opener **Answer:** The identity holds for all $a, b \in S$; it follows from the hypothesis by two further substitutions into itself, with no assumption of associativity, commutativity, or an identity element.

Layer 1 THE WHAT — surface analysis

The natural attempt. The hypothesis is $(p * q) * p = q$ for all $p, q \in S$, and the target looks like a mirror image of it: $a * (b * a) = b$ has the same three letters a, a, b but the inner pair reversed. The obvious first move is to substitute directly toward that shape. Fix a, and in the hypothesis set $q := b * a$ (keeping $p := a$):

$$(a * (b * a)) * a = b * a.$$

Write $X = a * (b * a)$, the exact expression the problem wants shown to equal b. The line above reads $X * a = b * a$. This feels like real progress – the target expression X is now pinned down by an equation, and if $*$ behaved like ordinary multiplication or a group operation, the next line would be to cancel a from both sides and read off $X = b$.

But nothing in the hypothesis grants cancellation. S is only a set with one operation satisfying one identity; there is no stated inverse, no identity element, no associativity. Trying to justify $X * a = Y * a \implies X = Y$ from $(p * q) * p = q$ alone leads nowhere directly – substituting $p := a$ into the hypothesis with different second arguments produces $(a * X) * a = X$ and $(a * Y) * a = Y$, which says something about $a * X$ and $a * Y$, not about $X * a$ and $Y * a$. The natural attempt has correctly reduced the problem to a cancellation law, and then discovers it has no way to prove that cancellation law without already knowing the answer. It stalls holding $X * a = b * a$, one substitution short of the goal, with no legal next move.

The model chain. Work entirely from the hypothesis $(p * q) * p = q$, valid for every $p, q \in S$, and use it three times with carefully chosen substitutions – never introducing cancellation, inverses, or any property not given.

First, substitute $p := b$, $q := a$:

$$(b * a) * b = a. \tag{$*$}$$

This is just the hypothesis read with the two letters swapped; it says nothing new yet, but it produces a fact about the compound expression $b * a$ that will be needed.

Second, substitute $p := b * a$, $q := b$ – that is, let the compound expression $b * a$ itself play the role of the hypothesis's first argument:

$$\big((b * a) * b\big) * (b * a) = b. \tag{$**$}$$

This is legitimate: the hypothesis holds for *every* pair of elements of S, and $b * a$ is an element of S like any other, so it is entitled to sit in the first slot.

Now substitute $(*)$ into $(**)$: the left-hand factor $(b * a) * b$ equals a, so

$$a * (b * a) = b.$$

This is exactly the claim, established for arbitrary $a, b \in S$.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains begin from the same first move: substituting $q := b * a$ with $p := a$ into $(p * q) * p = q$, since the target's outer structure $a * (\cdots)$ makes $p := a$ the obvious anchor. That substitution is common ground and gives $\big(a * (b * a)\big) * a = b * a$ in both cases. They split immediately afterward, on what to substitute *next*. The natural attempt tries to finish by cancelling a from $X * a = b * a$ – a move the hypothesis never licenses. The model chain abandons that line entirely and instead substitutes into the hypothesis a second time with $p := b * a$, promoting the compound expression $b * a$ to the first argument of a fresh instance of the identity.

> *The hypothesis is a statement about* every *pair of elements of S – it does not care whether the first argument is a bare letter like a or b, or a compound expression built out of them, such as $b * a$. So instead of trying to cancel a out of $X * a = b * a$, which the hypothesis gives me no tool to do, I should feed the hypothesis a new pair of inputs: let $b * a$ itself be the "p" and b be the "q", and see what identity that produces.*

> *Once I have $((b * a) * b) * (b * a) = b$, the left-hand side contains the block $(b * a) * b$ – and that block is itself an instance of the* original *hypothesis, with $p := b$, $q := a$, which reads $(b * a) * b = a$. I don't need to solve for anything; I can substitute this known value directly into the equation I just derived.*

Layer 3 THE WHY

Why does promoting $b * a$ to the first argument work, when cancelling does not? Because the hypothesis $(p * q) * p = q$ is a universally quantified statement over *all* of $S \times S$ – it makes no reference to how p was built. An element of S that happens to be written as $b * a$ is, as far as the hypothesis is concerned, exactly as eligible for the first slot as an element called a. Cancellation, by contrast, is not a consequence of this quantifier; it is an extra algebraic property (right-cancellativity) that the problem never assumes and that cannot be manufactured from a single identity of this shape without already knowing what is being proved. The lesson is structural: when a hypothesis is "for all p, q", the entire content available to the solver is the freedom to choose p and q – including choosing compound expressions already produced by earlier steps – not any inherited arithmetic intuition (cancellation, associativity, commutativity) that ordinary operations happen to have.

The second load-bearing fact is that $(*)$, namely $(b * a) * b = a$, is not a new discovery but literally the hypothesis itself with the letters relabelled: $(p * q) * p = q$ with $p := b, q := a$. This is why the proof needs no case analysis and no auxiliary lemma outside the given axiom – everything used is one identity, invoked three times at three different pairs of inputs $(a, b * a)$, (b, a), $(b * a, b)$, chosen so that the output of one invocation is exactly the sub-expression needed to simplify another. Seeing the target as "a plug for $(*)$" rather than as "an equation to be solved for X" is what makes the substitutions line up.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the shape of the hypothesis itself: $(p * q) * p = q$, a "for all p, q" identity with no other structure assumed. On seeing that, the thing to write down is: my only tool is choosing what to substitute for p and q – and those substitutions can be arbitrary elements of S, including expressions already built from a and b.

2. That cue suggests treating the proof as a substitution-chasing exercise: write the hypothesis with the target's outer shape substituted in, see what sub-expression appears, and check whether that sub-expression is itself an instance of the hypothesis under some other choice of p, q.

3. Hesitation sets in exactly at $X * a = b * a$, where cancelling a looks like the only way to finish and feels almost automatic from years of manipulating equations. What reinforces the switch away from it is asking explicitly: is right-cancellation actually one of my given facts? It is not stated, and the only tool I have is the identity itself – so the next legal move has to be another substitution into that identity, not an algebraic law borrowed from elsewhere.

4. The trap is treating $*$ as though it inherits properties of familiar operations (associativity, commutativity, invertibility) just because it is written with an infix symbol. None of these are given, and using any of them silently produces a proof that looks fine but rests on an unstated assumption.

5. Recognise this shape faster by noticing when a hypothesis and a goal are built from the same three symbols in a different bracketing order (here: (a, b, a) under $*$, twice). That is the signal to substitute compound expressions – not just bare variables – into the hypothesis, using an earlier consequence of the hypothesis as the "q" or "p" of a fresh instance.

6. Reusable rule: for a single "for all x, y" identity with no other assumed structure, prove a derived identity by chaining substitutions – plug earlier-derived equalities into new instances of the same hypothesis at cleverly chosen arguments – rather than reaching for cancellation, inverses, or any property the operation was never given.

▷ **Cue:** a hypothesis and a goal are built from the same symbols in a different order, and the operation has no assumed structure beyond the one given identity $\longrightarrow$ **Move:** substitute compound expressions (not just bare variables) into the hypothesis, using the output of one instance of the identity as an input to another

▷ **Cue:** an equation of the form $X * a = Y * a$ (or similarly $a * X = a * Y$) appears mid-proof and cancelling a would finish it $\longrightarrow$ **Move:** check whether cancellation is actually a given or provable property before using it; if not, look for a different substitution into the same hypothesis instead

2001 A2

2001 A2. You have coins $C_1, C_2, \ldots, C_n$. For each k, C_k is biased so that, when tossed, it has probability $1/(2k+1)$ of falling heads. If the n coins are tossed, what is the probability that the number of heads is odd? Express the answer as a rational function of n.

Fields: probability, generating functions **Position:** A2 **Answer:** $P_n = \dfrac{n}{2n+1}$. Built from the probability generating function $f(x) = \prod_{k=1}^{n} (p_k + q_k x)$, evaluated at $x = 1$ and $x = -1$, where the second evaluation telescopes.

Layer 1 THE WHAT — surface analysis

The natural attempt. The coins are independent, so the natural first move is to build the answer up one coin at a time. Let P_n be the probability that $C_1, \ldots, C_n$ together give an odd number of heads. Condition on the last coin, C_n, which is heads with probability $q_n = 1/(2n+1)$ and tails with probability $p_n = 2n/(2n+1)$. The total is odd exactly when (first $n-1$ coins give an even count and C_n is heads) or (first $n-1$ coins give an odd count and C_n is tails):

$$P_n = (1 - P_{n-1})q_n + P_{n-1}p_n = \left(\frac{2n-1}{2n+1}\right) P_{n-1} + \frac{1}{2n+1}.$$

With $P_1 = 1/3$ this recurrence is completely determined. Computing a few terms, $P_2 = \frac{2}{5}$, $P_3 = \frac{3}{7}$, and the pattern $P_n = n/(2n+1)$ suggests itself immediately. Proving it then requires an induction: assume $P_{n-1} = (n-1)/(2n-1)$ and substitute,

$$P_n = \frac{2n-1}{2n+1} \cdot \frac{n-1}{2n-1} + \frac{1}{2n+1} = \frac{n-1}{2n+1} + \frac{1}{2n+1} = \frac{n}{2n+1},$$

which closes the induction. This works completely – the recurrence is genuinely correct and the induction genuinely closes – but notice the shape of the argument: the closed form was *guessed* from three computed values and only afterward verified. Nothing in the recurrence itself explains, before the pattern-spotting, why the answer should be a clean ratio $n/(2n+1)$ rather than some messier rational function; the proof is a check, not a derivation.

The model chain. Rather than building P_n coin by coin as a running probability, encode the whole distribution of the number of heads in a single polynomial. For coin C_k, write $p_k = 2k/(2k+1)$ (tails) and $q_k = 1/(2k+1)$ (heads), and set

$$f(x) = \prod_{k=1}^{n} (p_k + q_k x).$$

Because the coins are independent, expanding this product and collecting the coefficient of x^i gives exactly the probability of getting i heads in total: each term in the expansion picks, for every coin, either its "tails" contribution (p_k, contributing x^0) or its "heads" contribution ($q_k x$, contributing one power of x), and by independence the probability of any specific pattern of heads/tails is the product of the individual probabilities. So $f(x) = \sum_{i=0}^{n} c_i x^i$ with $c_i = \Pr(\text{exactly } i \text{ heads})$.

The probability of an odd number of heads is $\sum_{i \text{ odd}} c_i$. This is extracted from f by the standard device $\sum_{i \text{ odd}} c_i = \dfrac{f(1) - f(-1)}{2}$, since $f(1) = \sum c_i$ sums every coefficient with sign $+1$ while $f(-1) = \sum c_i (-1)^i$ sums every coefficient with sign $(-1)^i$; subtracting cancels all even-index terms and doubles all odd-index terms.

Now evaluate. $f(1) = \prod_{k=1}^{n} (p_k + q_k) = \prod_{k=1}^{n} 1 = 1$, since each coin's probabilities sum to 1. And

$$f(-1) = \prod_{k=1}^{n} (p_k - q_k) = \prod_{k=1}^{n} \frac{2k-1}{2k+1} = \frac{1}{3} \cdot \frac{3}{5} \cdot \frac{5}{7} \cdots \frac{2n-1}{2n+1} = \frac{1}{2n+1},$$

a telescoping product in which every factor's numerator cancels the previous factor's denominator, leaving only the first denominator's "1" (from the empty numerator before $k = 1$) and the last denominator $2n+1$. Hence

$$P_n = \frac{f(1) - f(-1)}{2} = \frac{1 - \frac{1}{2n+1}}{2} = \frac{\frac{2n}{2n+1}}{2} = \frac{n}{2n+1}.$$

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains correctly identify that the coins are independent and that the answer must come from combining n individual probabilities into one number. They split on *how* that combination is organised. The natural attempt organises it sequentially, one coin at a time, which produces a first-order recurrence in n – true, but a recurrence must then either be solved in closed form (hard, in general) or have its closed form guessed and checked by induction. The model chain organises the same independence structure all at once, as a single product $\prod(p_k + q_k x)$, so that "probability of odd count" becomes a request about a polynomial's coefficients rather than a request about a sequence.

> *I don't actually want the probability of exactly i heads for each i – I only want the sum over* odd *i. Building the full distribution coin by coin (a recurrence in n) answers a harder question than the one being asked. What I want is a way to sum "every other coefficient" of a generating function directly, without ever finding the individual coefficients.*

> *Summing coefficients at even positions with $+$ and odd positions with $-$ is exactly what evaluating a polynomial at $x = -1$ does; evaluating at $x = 1$ sums every coefficient with $+$. So $(f(1) - f(-1))/2$ isolates the odd-coefficient sum without ever expanding the product – I should therefore never expand $f(x)$ at all, only evaluate it at two convenient points.*

Layer 3 THE WHY

Why does the product $\prod(p_k + q_k x)$ correctly encode the distribution of the head-count, and why is that the right object to reach for? Because for independent trials, the probability generating function of a sum of independent random variables is the *product* of their individual probability generating functions – this is a direct consequence of independence meaning $\Pr(\text{joint pattern}) = \prod \Pr(\text{individual outcome})$, and the coefficient of x^i in a product of linear factors $(p_k + q_k x)$ is precisely the sum, over all ways of choosing "heads" from i of the n factors and "tails" from the rest, of the corresponding probability products – which is exactly $\Pr(\text{exactly } i \text{ heads})$ by definition. Once that identification is made, the question "probability the count is odd" is no longer a probability question at all; it is the purely algebraic question "sum of the odd-degree coefficients of this specific polynomial", and that question has a standard, general answer, $(f(1) - f(-1))/2$, that has nothing to do with coins.

Why does $f(-1)$ telescope so cleanly, rather than leaving an ugly product? Because $p_k - q_k = \dfrac{2k}{2k+1} - \dfrac{1}{2k+1} = \dfrac{2k-1}{2k+1}$, and the specific bias $1/(2k+1)$ chosen by the problem was engineered so that the denominator of one factor, $2k+1$, is exactly the numerator of the next factor's denominator's neighbour, $2(k+1) - 1 = 2k+1$ – in other words the problem's numbers were chosen to make this particular telescope collapse. Recognising the generating-function route is what exposes that design; the coin-by-coin recurrence never brings $f(-1)$'s product into view at all, so it never gets to exploit the telescope and instead has to rediscover the same cancellation implicitly, one step of induction at a time, without ever seeing why it is happening.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is "probability that the count is odd" for a sum of independent $0/1$ trials. On seeing that phrase, the thing to write down immediately is: I want an odd-indexed coefficient sum, not a full distribution, so I should reach for a generating function rather than a running probability.

2. That cue suggests building the probability generating function $f(x) = \prod_k (p_k + q_k x)$ for the individual coins, then using the fact that summing coefficients of odd degree is $(f(1) - f(-1))/2$.

3. Hesitation sets in at the very first coin, where "condition on the last toss" feels like the default, textbook way to handle a sequence of independent trials, and it does produce a correct recurrence with almost no thought. What reinforces the switch away from it is asking: am I being asked for the whole distribution, or only a single aggregate statistic (odd vs. even) of it? An aggregate parity statistic is the specific signature of the $x = \pm 1$ trick.

4. The trap is treating the recurrence-and-guess route as though it were the intended solution because it works. It does work, but it conceals the fact that the numbers $1/(2k+1)$ were chosen so that $f(-1)$ telescopes; a solver who only ever manipulates P_n numerically never sees that design, and would be stuck if the bias sequence had been slightly different and non-telescoping.

5. Recognise this shape faster by treating "probability of an odd count among independent binary trials" as a fixed trigger for $(f(1) - f(-1))/2$ on the product generating function, in the same family as "probability of an even count" ($(f(1) + f(-1))/2$) and other parity-extraction problems.

6. Reusable rule: when a question asks for the sum of probabilities over indices of one parity (odd or even count of successes among independent trials), build the product generating function of the individual trials and evaluate it at $x = 1$ and $x = -1$; do not compute the full distribution or set up a recurrence in the number of trials unless the parity trick has already been ruled out.

▷ **Cue:** the question asks for the probability that a count of independent binary outcomes is odd (or even), not the full distribution $\longrightarrow$ **Move:** form the product probability generating function $f(x) = \prod(p_k + q_k x)$ and extract the answer as $(f(1) - f(-1))/2$ (odd) or $(f(1) + f(-1))/2$ (even), evaluating at the two points rather than expanding

▷ **Cue:** a product of fractions with numerators and denominators built from consecutive odd (or consecutive integer) values appears, e.g. $\prod(2k-1)/(2k+1)$ $\longrightarrow$ **Move:** write out the first few factors explicitly and check for telescoping cancellation before attempting to bound or estimate the product

2001 A3

2001 A3. For each integer m, consider the polynomial

$$P_m(x) = x^4 - (2m+4)x^2 + (m-2)^2.$$

For what values of m is $P_m(x)$ the product of two non-constant polynomials with integer coefficients?

Fields: polynomial algebra, algebraic numbers **Position:** A3 **Answer:** P_m factors over $\mathbb{Z}$ exactly when m is a perfect square or twice a perfect square. Found by solving the biquadratic for its four roots $\pm\sqrt{m} \pm \sqrt{2}$ in closed form, then checking which groupings of those roots give rational (hence, by monicity, integer) elementary symmetric functions.

Layer 1 THE WHAT — surface analysis

The natural attempt. P_m has no odd-degree terms, so any quadratic×quadratic factorization over $\mathbb{Z}$ must respect that symmetry. Try $P_m(x) = (x^2 + ax + b)(x^2 - ax + c)$ – opposite-sign linear terms, since otherwise the x^3 coefficient would not automatically vanish. Expanding,

$$x^4 + (b + c - a^2)x^2 + a(c - b)x + bc.$$

Matching coefficients against $P_m(x) = x^4 - (2m+4)x^2 + (m-2)^2$ (no x term) forces $a(c-b) = 0$: either $a = 0$, or $c = b$.

If $a = 0$: the factorization is $(x^2 + b)(x^2 + c)$ with $b + c = -(2m+4)$ and $bc = (m-2)^2$. This is a genuine two-variable integer system in b, c for each m – solvable in principle by treating b, c as roots of $t^2 + (2m+4)t + (m-2)^2 = 0$ and demanding its discriminant be a perfect square, which is itself a further Diophantine condition on m.

If $c = b$: the factorization is $(x^2 + ax + b)(x^2 - ax + b) = x^4 + (2b - a^2)x^2 + b^2$, giving $b^2 = (m-2)^2$ (so $b = \pm(m-2)$) and $a^2 = 2b + 2m + 4$ – again a perfect-square condition on m, different from the first.

Both subcases are honest and each does eventually produce a condition equivalent to a part of the true answer. But notice the cost: there are two structurally different subcases to track ($a = 0$ versus $c = b$), each ending in its own discriminant-is-a-perfect-square condition that must be solved separately, and nothing yet confirms that quadratic×quadratic is even the only

shape of factorization to check – a linear×cubic factorization has not been ruled out and would need its own, different coefficient system. The approach is not wrong, but it multiplies casework rather than exposing why the two final conditions are "m is a perfect square" and "m is twice a perfect square" and nothing else.

The model chain. Solve $P_m(x) = 0$ directly. Treating P_m as a quadratic in x^2 and applying the quadratic formula,

$$x^2 = (m+2) \pm \sqrt{(m+2)^2 - (m-2)^2} = (m+2) \pm \sqrt{8m} = m + 2 \pm 2\sqrt{2m}.$$

Both expressions under this outer square root are perfect squares of a sum: $(\sqrt{m} + \sqrt{2})^2 = m + 2 + 2\sqrt{2m}$ and $(\sqrt{m} - \sqrt{2})^2 = m + 2 - 2\sqrt{2m}$. So $x^2 = (\sqrt{m} \pm \sqrt{2})^2$, and the four roots of P_m are exactly

$$S = \{\sqrt{m} + \sqrt{2},\ \sqrt{m} - \sqrt{2},\ -\sqrt{m} + \sqrt{2},\ -\sqrt{m} - \sqrt{2}\} = \{\pm\sqrt{m} \pm \sqrt{2}\}.$$

P_m is monic with integer coefficients, so any factorization into non-constant integer polynomials corresponds to splitting S into two non-empty groups, each group being the root-set of one (monic, WLOG) factor. A monic polynomial with rational coefficients whose roots are algebraic integers must itself have integer coefficients (Gauss's Lemma), so it suffices to find, for each possible grouping, whether the elementary symmetric functions of that group are rational.

Groups of size 1. A single root $r \in S$ must be rational. Since $r^2 = m + 2 \pm 2\sqrt{2m}$, rationality of r forces $\sqrt{2m}$ rational, so $2m = k^2$ for some integer k; since $2m$ is even, $k = 2n$ and $m = 2n^2$. Then $\sqrt{m} = n\sqrt{2}$ and $r = (n \pm 1)\sqrt{2}$, rational only if $n = \mp 1$, giving $m = 2$.

Groups of size 2 (three shapes, up to complement).

- $\{\sqrt{m} + \sqrt{2}, \sqrt{m} - \sqrt{2}\}$: sum $2\sqrt{m}$, product $m - 2$. Product is always an integer; sum is rational iff m is a perfect square.
- $\{\sqrt{m} + \sqrt{2}, -\sqrt{m} + \sqrt{2}\}$: sum $2\sqrt{2}$, never rational – impossible for any m.
- $\{\sqrt{m} + \sqrt{2}, -\sqrt{m} - \sqrt{2}\}$: sum 0 (rational), product $-(m + 2 + 2\sqrt{2m})$, rational iff $2m$ is a perfect square, i.e. $m = 2n^2$ for some integer n – exactly the "twice a perfect square" family, which already contains the size-1 case $m = 2$.

Groups of size 3 are complements of groups of size 1, and give an integer cubic factor automatically once the linear factor is integral (polynomial division of an integer polynomial by an integer monic factor stays integral), so they add no new values of m.

Combining every case: P_m factors over $\mathbb{Z}$ into non-constant integer polynomials exactly when m is a perfect square or twice a perfect square.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains are forced to begin the same way, by treating P_m as a quadratic in x^2 and applying the quadratic formula – that is the only way to get any explicit handle on a biquadratic, and it is common ground. The split happens immediately after: at $x^2 = (m+2) \pm \sqrt{8m}$, the natural attempt sets this expression aside and returns to guessing a factorization *shape* $(x^2 + ax + b)(x^2 - ax + c)$, then grinds coefficient-matching. The model chain instead keeps working on $x^2 = (m+2) \pm \sqrt{8m}$ for one more line, recognising it as a disguised perfect square, and gets the four roots in fully explicit closed form before ever thinking about factorization shapes.

> *The expression $m + 2 \pm 2\sqrt{2m}$ under the outer square root is exactly the expansion of $(\sqrt{m} \pm \sqrt{2})^2$. I should not leave x^2 as a nested radical and go back to blind coefficient matching – I should finish simplifying it, because once the four roots are written explicitly as $\pm\sqrt{m} \pm \sqrt{2}$, the factorization question becomes a question about* which finite set of these four labelled numbers *can be grouped*

> *together, which is something I can just enumerate.*

> *A monic integer polynomial factoring over $\mathbb{Z}$ is the same statement as its root-multiset splitting into groups whose elementary symmetric functions are rational – and since there are only four roots, there are only a handful of ways to split them (one root alone, or two together) up to taking complements. I don't need to guess which factorization shape (linear×cubic vs. quadratic×quadratic) is the right one to try; checking every grouping of the four explicit roots covers every possible shape automatically.*

Layer 3 THE WHY

Why does going via explicit roots beat guessing a factorization shape? Because factoring a polynomial over $\mathbb{Z}$ and grouping its roots are, by the fundamental theorem of algebra together with Gauss's Lemma, literally the same operation viewed two ways: a factor corresponds to a subset of the roots (with multiplicity), the factor's coefficients are (up to sign) the elementary symmetric functions of that subset, and those coefficients are integers exactly when they are rational, because a root of a monic integer polynomial is by definition an algebraic integer, and a rational algebraic integer is an ordinary integer. Once the roots are known explicitly, checking rationality of symmetric functions of a 4-element set is finite case-checking – there are only $\binom{4}{1} + \binom{4}{2}/2 = 6$ genuinely different non-trivial subsets up to complementation. Guessing a coefficient shape instead requires guessing correctly that the shape is exhaustive (is quadratic×quadratic really the only option? what about linear×cubic?) and then separately solving a fresh Diophantine system for each guessed shape, with no guarantee in advance that the two resulting conditions are the clean "square" and "twice a square" that the roots reveal immediately.

The second thing worth grounding is why $m + 2 \pm 2\sqrt{2m}$ was recognisable as $(\sqrt{m} \pm \sqrt{2})^2$ at all: it is because $(\sqrt{m} + \sqrt{2})^2 = m + 2\sqrt{2m} + 2$ is the single most natural thing to compute once two square roots $\sqrt{m}$ and $\sqrt{2}$ are in view together – the cross term $2\sqrt{m}\sqrt{2} = 2\sqrt{2m}$ is exactly what the quadratic formula produced under its own square root. The problem's coefficients $2m + 4$ and $(m - 2)^2$ were not arbitrary; they were built so that P_m's roots are $\pm\sqrt{m} \pm \sqrt{2}$ for a clean reason, and noticing the completed-square shape is what exposes that design instead of fighting it with undetermined coefficients.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is a quartic with no odd-degree terms. On seeing that, the thing to write down immediately is that it is a quadratic in x^2, so x^2 can be solved for explicitly via the quadratic formula – this is forced, not optional.
2. That cue suggests going all the way to explicit closed-form roots (simplifying any nested radical that results, checking for a perfect-square pattern), rather than stopping at $x^2 = (m + 2) \pm \sqrt{8m}$ and switching to coefficient-matching on a guessed factorization shape.
3. Hesitation sets in right after the quadratic formula produces a nested radical $(m + 2) \pm \sqrt{8m}$ – it looks messy, and returning to "guess a factorization shape and match coefficients" feels more concrete. What reinforces pushing through the radical instead is asking: does $\sqrt{8m} = 2\sqrt{2m}$ look like a cross term $2\sqrt{A}\sqrt{B}$ from expanding $(\sqrt{A} \pm \sqrt{B})^2$? If the constant term under the radical splits naturally into two square-root pieces, it almost always does.
4. The trap is assuming, without checking, that quadratic×quadratic is the only factorization shape worth testing. A linear×cubic factorization is a genuinely different case (a single rational root), and the coefficient-matching route has no built-in reason to notice it needs separate treatment, whereas working from the explicit roots handles size-1 and size-3 groupings in the very same framework as size-2.
5. Recognise this shape faster by treating "quartic in x^2, no odd-degree terms, asked to characterise integer factorizability by a parameter" as a standing cue for: solve explicitly for the

four roots first, then classify subsets of the roots by rationality of their symmetric functions, rather than by guessing coefficient shapes.

6. Reusable rule: for a monic integer polynomial whose roots can be found in closed radical form, integer-factorizability is equivalent to some non-trivial subset of the roots having rational elementary symmetric functions (rational forces integer, by Gauss's Lemma, since roots of a monic integer polynomial are algebraic integers) – find the roots first, then enumerate subsets, instead of guessing the shape of the factors and matching coefficients.

▷ **Cue:** a polynomial in one variable is a quartic (or higher even degree) with no odd-degree terms and asks for integer factorizability depending on a parameter $\longrightarrow$ **Move:** treat it as a polynomial in x^2, solve for x^2 via the quadratic formula (or further substitution), and push the resulting radical all the way to fully explicit closed-form roots

▷ **Cue:** a monic integer polynomial's factorizability over $\mathbb{Z}$ is in question and its roots are known explicitly in closed radical form $\longrightarrow$ **Move:** enumerate the non-trivial subsets of the roots (up to complementation) and check whether each subset's elementary symmetric functions are rational; rational forces integer for a monic factor by Gauss's Lemma

2001 A4

2001 A4. Triangle ABC has an area 1. Points E, F, G lie, respectively, on sides BC, CA, AB such that AE bisects BF at point R, BF bisects CG at point S, and CG bisects AE at point T. Find the area of the triangle RST.

Fields: Euclidean geometry (affine); barycentric coordinates; iteration of a decreasing map **Position:** A4 — mid-paper. The geometry is routine once it is written in the right coordinates; the one real idea is showing the three unknown ratios are equal rather than assuming it. **Answer:** $[RST] = \dfrac{7-3\sqrt{5}}{4} \approx 0.0729$.

Layer 1 THE WHAT — surface analysis

The natural attempt. You draw the picture, and it looks symmetric, so you place coordinates: $A = (0,0)$, $B = (1,0)$, $C = (0,2)$ (area 1). You parametrise $E = (1-\lambda, 2\lambda)$ on BC, $F = (0, 2\mu)$ on CA, $G = (\nu, 0)$ on AB. The midpoint of BF is $(\frac{1}{2}, \mu)$; it must lie on line AE, whose equation is $2\lambda x - (1-\lambda)y = 0$, giving $\lambda = (1-\lambda)\mu$. Then the midpoint of CG must lie on BF, and the midpoint of AE on CG: two more equations, each with a different shape because your coordinates treat A, B, C unequally. At this point most people do one of two things. Either they grind through the three equations (it can be done, but the asymmetry of the coordinates hides that the three equations are the same equation cycled), or — far more common — they write "by symmetry E, F, G divide their sides in the same ratio" and solve one equation. The second is the trap: the configuration is *not* given as symmetric. The hypothesis is cyclic, and cyclic hypotheses can have non-symmetric solutions. An answer obtained that way is correct but unproved.

The model chain. Write every point as a weighted average of A, B, C (barycentric coordinates: $P = \alpha A + \beta B + \gamma C$ with $\alpha + \beta + \gamma = 1$). Put

$$E = rB + (1-r)C, \qquad F = sC + (1-s)A, \qquad G = tA + (1-t)B, \qquad r, s, t \in (0,1).$$

(So $EC = r\,BC$, $FA = s\,CA$, $GB = t\,AB$, the official parametrisation.)

Step 1: the midpoint condition. "AE bisects BF at R" means the midpoint $R = \frac{1}{2}(B+F) = \frac{1}{2}\big((1-s)A + B + sC\big)$ lies on line AE. A point $\alpha A + \beta B + \gamma C$ lies on line AE exactly when its B- and C-weights are in the ratio in which E mixes B and C, namely $\beta : \gamma = r : (1-r)$. So $1 : s = r : (1-r)$, i.e.

$$r(1+s) = 1.$$

The hypothesis is invariant under the cycle $A \to B \to C \to A, E \to F \to G \to E, r \to s \to t \to r$, so the same computation gives $s(1+t) = 1$ and $t(1+r) = 1$.

Step 2: the ratios are equal. Let $f(x) = 1/(1+x)$ on $[0, \infty)$. The three equations say $r = f(s)$, $s = f(t)$, $t = f(r)$, so $r = f(f(f(r)))$. Compose directly:

$$f(f(x)) = \frac{1+x}{2+x}, \qquad f(f(f(x))) = \frac{2+x}{3+2x}.$$

So $r(3+2r) = 2+r$, i.e. $r^2 + r - 1 = 0$, whose only positive root is $z = \frac{\sqrt{5}-1}{2} \in (0,1)$. Since $z^2 + z = 1$, $f(z) = z$; so $t = f(r) = z$ and then $s = f(t) = z$. Hence $r = s = t = z$. This is now proved, not assumed.

Step 3: the area. With $r = s = t = z$,

$$R = \tfrac{1}{2}\big((1-z)A + B + zC\big), \quad S = \tfrac{1}{2}\big(zA + (1-z)B + C\big), \quad T = \tfrac{1}{2}\big(A + zB + (1-z)C\big).$$

The area of a triangle with barycentric rows P_1, P_2, P_3 is $|\det[P_1; P_2; P_3]| \cdot [ABC]$. The matrix $\frac{1}{2}\begin{pmatrix} 1-z & 1 & z \\ z & 1-z & 1 \\ 1 & z & 1-z \end{pmatrix}$ is circulant; for entries c_0, c_1, c_2 the determinant is $c_0^3 + c_1^3 + c_2^3 - 3c_0c_1c_2 = (c_0+c_1+c_2)\big(c_0^2 + c_1^2 + c_2^2 - c_0c_1 - c_1c_2 - c_2c_0\big)$. Here $c_0 + c_1 + c_2 = 2$ and the second factor is $(2 - 2z + 2z^2) - (1 + z - z^2) = 1 - 3z + 3z^2 = 4 - 6z$ using $z^2 = 1 - z$. So

$$[RST] = \tfrac{1}{8} \cdot 2(4 - 6z) = 1 - \tfrac{3}{2}z = 1 - \tfrac{3(\sqrt{5}-1)}{4} = \frac{7 - 3\sqrt{5}}{4}.$$

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains translate the first condition correctly into one equation. They split at (†): "the other two conditions are the same by symmetry." The natural attempt reads that as *therefore $r = s = t$*. The model reads it as *therefore the equations are $r = f(s)$, $s = f(t)$, $t = f(r)$ for one function f* — and then proves equality from the properties of f.

> *Cyclic symmetry of the hypothesis gives me cyclic equations, not equal unknowns. Before I set $r = s = t$ I owe a proof. The equations say $r = f(f(f(r)))$; I compose f three times and see what the fixed points are.*

> *"Bisects" means "midpoint", and a midpoint is an average of two points. I should be writing points as weighted averages of A, B, C, not as (x, y) pairs.*

Layer 3 THE WHY

Why barycentric coordinates? Look at what the hypotheses are made of: points on sides (a weighted average of two vertices), midpoints (an equal-weight average of two points), collinearity, and a ratio of areas. Every one of those is preserved by any affine map — a shear or stretch of the plane sends midpoints to midpoints, lines to lines, and multiplies all areas by the same factor. So the answer cannot depend on the shape of ABC, only on its area. The coordinates that see exactly this much structure and no more are the weights (α, β, γ). Cartesian coordinates carry extra structure (lengths, angles) that the problem never uses, and you pay for it by writing three differently-shaped equations for one condition. Once you write $E = rB + (1-r)C$, the collinearity test "B-weight : C-weight equals $r : (1-r)$" is a direct consequence of the definition of line AE as $\{\lambda A + (1-\lambda)E\}$: mixing in A changes only the A-weight and scales B and C together.

Why is the symmetry step a genuine gap? The hypothesis is invariant under a cyclic relabelling. That tells you the *set* of solutions (r, s, t) is invariant under cycling — not that

each solution is fixed by it. A cyclic system $x = g(y)$, $y = g(z)$, $z = g(x)$ has non-constant solutions whenever g has a genuine 3-cycle. So the question you must ask is whether $f \circ f \circ f$ has fixed points other than those of f. Kedlaya's solution answers it with monotonicity: f is decreasing, so $f \circ f$ is increasing and $f \circ f \circ f$ is decreasing; a decreasing function crosses the line $y = x$ at most once. That is the axiom-level reason — a decreasing map composed an odd number of times is decreasing, and a decreasing function has at most one fixed point. The direct composition above is the cheaper check here because f is a Möbius map ($f^{\circ 3}$ is again a ratio of linear functions, so its fixed-point equation is a quadratic you can solve). Both routes end in the same place; the monotonicity route is the one that transfers to any decreasing f.

Finally, the area. Kedlaya's route subtracts $[ABR] + [BCS] + [CAT] = \frac{3}{2}z$ from 1. That works because R, S, T happen to be arranged so the three corner triangles and RST tile ABC — which you have to check from a picture. The determinant route has no such case analysis; it is the definition of area in coordinates. I would take the determinant in the exam: it cannot silently give the wrong sign or miss an overlap, and the absolute value is visible.

Layer $+\alpha$ BACKTESTING

1. The word "bisects". Write down immediately: $R = \frac{1}{2}(B + F)$. Then write E, F, G as weighted averages of the two endpoints of their sides.

2. Barycentric (or vector) coordinates relative to A, B, C; affine invariance means the shape of the triangle is irrelevant.

3. Hesitation comes after the first equation $r(1 + s) = 1$, at the temptation to declare $r = s = t$. It is reinforced away by writing the system as $r = f(s)$, $s = f(t)$, $t = f(r)$ and composing.

4. The trap is "by symmetry". The hypothesis is cyclic, the picture looks symmetric, and the correct answer does come out of assuming $r = s = t$ — so nothing warns you that the proof is missing a step.

5. Slogan: *a symmetric hypothesis gives a symmetric solution set, not symmetric solutions.* Whenever you are about to set cyclically-related unknowns equal, look for the map that links them.

6. Decision rule: **for a cyclic system $x_1 = f(x_2)$, $x_2 = f(x_3)$, …, $x_k = f(x_1)$, reduce to $x_1 = f^{\circ k}(x_1)$; if f is monotone increasing, or decreasing with k odd, the fixed points of $f^{\circ k}$ are those of f, so all x_i are equal.** For configuration problems built from midpoints, ratios and areas, use barycentric coordinates.

▷ **Cue:** a point "bisects" or is the midpoint of a segment, or a point lies on a side in a given ratio ⟶ **Move:** write it as a weighted average of the vertices; collinearity with a vertex becomes a ratio of the other two weights

▷ **Cue:** the hypothesis involves only midpoints, ratios along lines, collinearity and area ⟶ **Move:** it is affine invariant: use barycentric coordinates or place the triangle anywhere convenient

▷ **Cue:** a cyclic system of equations and the urge to write "by symmetry, all equal" ⟶ **Move:** write it as $x = f(y)$, $y = f(z)$, $z = f(x)$ and study the fixed points of $f \circ f \circ f$

▷ **Cue:** you need the fixed points of an iterate of a decreasing function ⟶ **Move:** an odd iterate of a decreasing function is decreasing, hence has at most one fixed point, which is the fixed point of f

▷ **Cue:** area of a triangle whose vertices are given in barycentric coordinates ⟶ **Move:** area ratio $= |\det|$ of the 3×3 coordinate matrix

2001 A5

2001 A5. Prove that there are unique positive integers a, n such that $a^{n+1} - (a+1)^n = 2001$.

Fields: number theory (exponential Diophantine equations; congruences) **Position:** A5 — late in the paper, but short once the right moduli are chosen. The difficulty is that there are two unknowns and

one of them is an exponent, so inequalities alone cannot finish. **Answer:** $(a, n) = (13, 2)$: $13^3 - 14^2 = 2197 - 196 = 2001$, and no other pair works.

Layer 1 THE WHAT — surface analysis

The natural attempt. You try small n. For $n = 1$: $a^2 - (a+1) = 2001$, i.e. $a^2 - a - 2002 = 0$, discriminant 8009, which lies strictly between $89^2 = 7921$ and $90^2 = 8100$: no integer root. For $n = 2$: $a^3 - (a+1)^2 = 2001$; $a = 12$ gives $1728 - 169 = 1559$, $a = 13$ gives $2197 - 196 = 2001$. Found it. Existence is done.

Uniqueness is where the wall is. You try to argue by size: for fixed n, $a^{n+1} - (a+1)^n$ is eventually increasing in a, so each n has at most one a — but that is one *a per n*, infinitely many cases. For fixed a, the sign of $a^{n+1} - (a+1)^n$ depends on whether $a\,(a/(a+1))^n$ exceeds 1; the expression grows and then collapses, so for each a there can be several n in play. You try $n = 3$: $a^4 - (a+1)^3 = 2001$ gives $a = 7$: $2401 - 512 = 1889$, $a = 8$: $4096 - 729 = 3367$; skip. $n = 4$: $a = 5$: $3125 - 1296 = 1829$, $a = 6$: $7776 - 2401 = 5375$; skip. Each n is a fresh search, and nothing in this approach ever ends. Growth estimates compare the two terms; they cannot certify that a difference is *never exactly* 2001 across an infinite two-parameter family.

The model chain. Suppose $a^{n+1} - (a+1)^n = 2001$ with $a, n \geq 1$.

Modulo a. Since $a + 1 \equiv 1$, the equation reads $0 - 1 \equiv 2001 \pmod{a}$, so $a \mid 2002 = 2 \cdot 7 \cdot 11 \cdot 13$.

Modulo $a+1$. Since $a \equiv -1$ and $n \geq 1$, it reads $(-1)^{n+1} - 0 \equiv 2001 \pmod{a+1}$.

- If n is even: $-1 \equiv 2001$, so $(a+1) \mid 2002$.
- If n is odd: $1 \equiv 2001$, so $(a+1) \mid 2000 = 2^4 \cdot 5^3$.

The candidates. The divisors of 2002 are $1, 2, 7, 11, 13, 14, 22, 26, 77, 91, 143, 154, 182, 286, 1001, 2002$. Consecutive pairs $a, a+1$ both dividing 2002: only $(1, 2)$ and $(13, 14)$. Divisors a of 2002 with $a + 1 \mid 2000$: $a+1 \in \{2, 8\}$, i.e. $a \in \{1, 7\}$ (checking the list: $3, 12, 14, 15, 23, 27, 78, 92, 144, 155, 183, 287, 1002, 2003$ do not divide 2000). So

$$(a,\ n \bmod 2) \in \{(1, \text{any}),\ (13, \text{even}),\ (7, \text{odd})\}.$$

Eliminate $a = 1$. $1 - 2^n < 0 < 2001$.

Eliminate $a = 7$, n odd. Modulo 3: $7 \equiv 1$, $8 \equiv -1$, so $7^{n+1} - 8^n \equiv 1 - (-1)^n = 2 \pmod{3}$, while $2001 = 3 \cdot 667 \equiv 0$. Contradiction.

Pin down n for $a = 13$. n is even. $n = 2$ works. If $n \geq 4$, then $2^n \mid 14^n$, so $14^n \equiv 0 \pmod{8}$; and $n + 1$ is odd, so $13^{n+1} \equiv 5^{n+1} \equiv 5 \pmod{8}$ (as $5^2 = 25 \equiv 1$). Then $13^{n+1} - 14^n \equiv 5 \not\equiv 1 \equiv 2001 \pmod{8}$. Contradiction.

Hence $(a, n) = (13, 2)$ is the unique solution. □

Caveat. Independently checked: a computer search over $1 \leq a < 3000$, $1 \leq n < 80$ finds only $(13, 2)$; the residues $7^{n+1} - 8^n \equiv 2 \pmod{3}$ (n odd) and $13^{n+1} - 14^n \equiv 5 \pmod{8}$ ($n \geq 4$ even) were confirmed for many n. The proof above does not depend on the search.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains reach $13^3 - 14^2 = 2001$ by trial. They split at (†): "now show nothing else works." The natural attempt reaches for size, n by n. The model reaches for divisibility, and picks the modulus by asking which reduction makes an unknown-exponent power stop depending on its exponent.

> *I have two unknowns and one is an exponent. Inequalities won't give me a finite list; divisibility will. The bases are a and $a+1$, so modulo a the first term is 0 and the second is 1 — whatever n is. That tells me a divides 2002.*

> *The same trick works the other way round: modulo $a+1$, a is -1, and $(-1)^{n+1}$ only depends on the parity of n. So $a+1$ divides 2002 or 2000. Two consecutive divisors — that's a very short list.*

> *Each survivor has an exponent of known parity. I kill it with a small modulus in which one base is 0 or ± 1 and the other has a short period.*

Layer 3 THE WHY

Why reduce modulo a and $a+1$? Go to the definition of a power. x^k is x multiplied by itself k times, and modulo m its value depends on k *unless* $x \equiv 0$ (then $x^k \equiv 0$ for $k \geq 1$), $x \equiv 1$ (then $x^k \equiv 1$) or $x \equiv -1$ (then $x^k \equiv \pm 1$ by parity). Those three are the residues whose powers are exponent-free in *every* modulus (other exponent-free cases, such as $3^k \equiv 3 \pmod 6$, depend on the particular modulus and cannot be arranged in advance when the base is unknown). With n unknown and unbounded, you need a modulus in which each base is one of $0, 1, -1$. The bases here are a and $a+1$; they differ by 1, so modulo a they are 0 and 1, and modulo $a+1$ they are -1 and 0. The problem's phrasing — two consecutive integers as bases — is the cue that both reductions will be exponent-free. Nothing about this is a trick: it is the kind of modulus that can say something about an equation with an unknown base and an unbounded exponent at the same time.

Why divisibility and not size for uniqueness? A uniqueness claim over an infinite set is a claim that every other case fails. Size arguments compare magnitudes and are good at excluding *ranges*, but $a^{n+1} - (a+1)^n$ oscillates in sign and magnitude across the (a, n) plane, so no single inequality excludes everything. Divisibility gives a *finite* list in one move: $a \mid 2002$ caps a at 2002. After that, every survivor is a single case.

Kedlaya's official solution uses $a \mid 2002$ and then works modulo 3 and 4 to force n even, a odd and $a \equiv 1 \pmod 4$, filtering the divisors of 1001 down to 13. That is correct, but the choice of 3 and 4 comes from the specific number $2001 = 3 \cdot 23 \cdot 29 \equiv 1 \pmod 4$, and in the arena you have no way to know in advance which small moduli will bite. The Note at the end of his solution — use that $a+1$ divides 2002 — is the symmetric partner of $a \mid 2002$, and it is the version I prefer, because it is chosen by the structure of the left-hand side, not by the digits of 2001. The price is that it splits on the parity of n, which is exactly the information $(-1)^{n+1}$ carries.

The finishing moduli are chosen by the same principle. For $a = 7$: modulo 3, $7 \equiv 1$ and $8 \equiv -1$ — both bases become exponent-free, and $2001 \equiv 0 \pmod 3$ is the fact you check against. For $a = 13$ with $n \geq 4$: modulo 8, 14^n vanishes once $n \geq 3$ (a high power of 2 swallows it), and $13 \equiv 5$ has order 2, so $13^{\text{odd}} \equiv 5$. The general rule: to separate one exponent from larger ones, choose a modulus that is a power of a prime dividing one base, so that base vanishes for all large exponents.

Layer $+\alpha$ BACKTESTING

1. The bases a and $a+1$ are consecutive. Write down at once: modulo a, $a+1 \equiv 1$; modulo $a+1$, $a \equiv -1$.
2. Congruences modulo the bases themselves, giving $a \mid 2002$ and $a+1 \mid 2002$ or 2000; then small moduli to kill the finitely many survivors.
3. Hesitation comes after finding $(13, 2)$, when you do not know how to prove uniqueness. It is reinforced away by the question "which modulus makes n disappear?"

4. The trap is size-based case analysis: fixing n and bounding a, one n at a time. It feels like progress and never terminates. A second trap is stopping at "$a \mid 2002$" and testing all sixteen divisors against all n — a finite list in a but still infinite in n.

5. Slogan: *with an unknown exponent, work modulo something that makes the base* 0 *or* ± 1.

6. Decision rule: **for an exponential Diophantine equation, reduce modulo each base (or each base ± 1) to get a finite list of bases; then, for each survivor, choose a modulus that is a power of a prime dividing one base (to separate small exponents from large) or in which both bases have short order.**

▷ **Cue:** an equation with an unknown exponent, and bases differing by 1 $\longrightarrow$ **Move:** reduce modulo each base; the other base becomes ± 1 and the exponent disappears (up to parity)

▷ **Cue:** a uniqueness claim for a Diophantine equation in two unknowns $\longrightarrow$ **Move:** look for a divisibility that caps one unknown and produces a finite list; size arguments rarely finish

▷ **Cue:** a residue of the form $(-1)^{n+1}$ $\longrightarrow$ **Move:** split on the parity of n and carry the parity forward as information

▷ **Cue:** need to show a candidate base works for only one exponent $\longrightarrow$ **Move:** use a modulus p^k with p dividing the other base, so that base vanishes for all large exponents

2001 A6

2001 A6. Can an arc of a parabola inside a circle of radius 1 have a length greater than 4?

Fields: real analysis (arc length, divergent integrals, asymptotics) **Position:** A6 — last problem of the morning. The computation is A-level integration; the difficulty is believing the answer is "yes" by a margin of under 0.3%, and knowing what to subtract so that the margin becomes visible. **Answer:** Yes. The arc of $y = Ax^2$ inside $x^2 + (y-1)^2 = 1$ has length > 4 for all $A > 34.74$ (numerically); the maximum is about 4.0027, at $A \approx 94.1$. A proof-grade choice is $A = 400$.

Layer 1 THE WHAT — surface analysis

The natural attempt. You think geometrically first. The circle has diameter 2. A parabola is convex and U-shaped; the longest arc should have its vertex at the bottom of the circle and its two arms climbing towards the top. Each arm rises at most 2 vertically, so the arc is "about 4". Then the doubt: the arms cannot be vertical, the parabola is round at the bottom and wastes height there, the arms end below the top of the circle. So you guess *no*, and try to prove it.

You try the convex-hull argument: a convex arc is no longer than any path enclosing it. The half-arc from the vertex O to the endpoint Q lies inside the triangle formed by the chord OQ and the tangents at O and Q, so its length is at most the sum of the two tangent segments. For $y = Ax^2$ the tangent at $Q = (X, AX^2)$ meets the x-axis at $(X/2, 0)$, so the bound is $X/2 + \sqrt{X^2/4 + Y^2}$, where $Y = AX^2$. For large A this exceeds 2: the term $X/2 \approx 1/\sqrt{2A}$ outweighs the shortfall $2 - Y = 1/A$. The bound does not prove "no". You try a general parabola with vertex, axis direction and focal parameter all free, write the length as an integral, and are faced with a three-parameter optimisation of an integral whose endpoints are roots of a quartic. That is the wall: nothing forces the answer either way, and the honest computation is too heavy to finish.

The model chain. Take the circle $x^2 + (y-1)^2 = 1$ (centre $(0,1)$, radius 1) and the parabola $y = Ax^2$ with $A > \frac{1}{2}$. Substituting, $x^2 + (Ax^2 - 1)^2 - 1 = x^2\big(A^2x^2 - (2A-1)\big)$, which is negative for $0 < x^2 < (2A-1)/A^2$. So the arc with $|x| \le X := \sqrt{2A-1}/A$ lies in the closed disc, touching the circle only at $O = (0,0)$ and at $(\pm X, Y)$ with $Y = AX^2 = 2 - \frac{1}{A}$.

By symmetry the arc length is $2L$ with

$$L = \int_0^X \sqrt{1 + 4A^2x^2}\,dx = \frac{1}{2A}\int_0^U \sqrt{1+u^2}\,du, \qquad u = 2Ax,\ U = 2AX = 2\sqrt{2A-1}.$$

Split off the vertical rise. Along the arc, $dy = 2Ax\,dx = \frac{u\,du}{2A}$, so $Y = \frac{1}{2A}\int_0^U u\,du = \frac{U^2}{4A} = 2 - \frac{1}{A}$. Hence

$$L = Y + \frac{1}{2A}\int_0^U \left(\sqrt{1+u^2} - u\right) du = 2 + \frac{1}{2A}\big(I(U) - 2\big), \qquad I(U) := \int_0^U \left(\sqrt{1+u^2} - u\right) du.$$

So $L > 2$ if and only if $I(U) > 2$.

I diverges. For $u \geq 0$,

$$\sqrt{1+u^2} - u = \frac{1}{\sqrt{1+u^2}+u} > \frac{1}{2\sqrt{1+u^2}} \geq \frac{1}{2(1+u)},$$

the first inequality because $u < \sqrt{1+u^2}$, the second because $\sqrt{1+u^2} \leq 1+u$. Therefore $I(U) > \frac{1}{2}\ln(1+U)$, which is unbounded.

A concrete A. We need $\frac{1}{2}\ln(1+U) \geq 2$, i.e. $U \geq e^4 - 1 \approx 53.60$. Take $A = 400$: $U = 2\sqrt{799} \approx 56.53$. Then $I(U) > 2$, so $L > 2$ and the arc length $2L > 4$.

Strictly inside. If "inside" is read as the open disc, shrink the whole picture by a factor $\lambda < 1$ about the centre $(0,1)$. The image of a parabola under a similarity is a parabola, the arc now lies in the open disc, and its length is $2\lambda L$, which still exceeds 4 for λ close enough to 1. So the answer is yes under either reading. □

Caveat. Independently checked by numerical quadrature: $2L = 3.9680$ at $A = 10$, 4.00005 at $A = 35$, 4.00154 at $A = 400$; $2L = 4$ at $A \approx 34.7386$; maximum $2L \approx 4.002670$ at $A \approx 94.09$. These agree with the Note in the official solution. The proof above uses only the certified inequality and does not depend on the numerics. Whether some parabola *not* of this symmetric, vertex-on-circle form does better is not needed for the question and was not investigated.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains reach the same geometric picture: vertex at the bottom of the circle, arms climbing towards the top, length "about 4". They split at (†): "the arms end below the top and the vertex wastes height, so the length is less than 4." The natural attempt counts only the vertical climb and the losses. The model also counts what the arc gains by travelling sideways, and measures the gain against the loss.

> *Why 4? Because 4 is two diameters: each arm can climb almost 2. So the length is (vertical climb) + (the extra from moving sideways). I should subtract the vertical climb from the arc length and see whether what is left beats the shortfall $2 - Y$.*

> *The shortfall is $1/A$, the sideways surplus is $\frac{1}{2A}\int_0^U(\sqrt{1+u^2} - u)\,du$. Both have the factor $1/A$, so it's a race between 2 and an integral whose integrand behaves like $1/(2u)$. That integral grows like a logarithm — slowly, but without bound. So the answer is yes.*

Layer 3 THE WHY

Kedlaya's solution says it has "artificially introduced $-x$ into the integrand". Here is where that comes from, and why it is not artificial at all. Arc length is $\int \sqrt{dx^2 + dy^2}$. The vertical rise of the same arc is $\int dy$. Their difference is

$$\int \left(\sqrt{dx^2 + dy^2} - dy\right),$$

the length the curve gains over a vertical segment of the same height. In the variable $u = 2Ax$ (the slope), $dy \propto u\,du$ and $ds \propto \sqrt{1+u^2}\,du$ with the same constant $\frac{1}{2A}$, so the difference is $\frac{1}{2A}\int(\sqrt{1+u^2}-u)\,du$. The "$-x$" is the vertical rise. The reason you subtract it is written in the problem's own number: 4 is the total vertical travel available, so the question "is the length more than 4?" is the question "does the sideways surplus exceed the vertical shortfall?". Once you frame it that way, you know what to subtract before you write a single integral.

Why does the surplus win? At slope u, the surplus per unit of slope is $\sqrt{1+u^2}-u = 1/(\sqrt{1+u^2}+u) \approx \frac{1}{2u}$: a steep arc is almost vertical, and the extra length it earns per step shrinks like $1/u$. The harmonic integral $\int \frac{du}{u}$ diverges — this is the same fact as the divergence of $\sum \frac{1}{k}$. Making A large makes the arms steeper (larger U), so the integral runs further up the harmonic tail; its value grows like $\frac{1}{2}\ln U \approx \frac{1}{4}\ln A$. The shortfall $2 - Y = 1/A$ and the prefactor $\frac{1}{2A}$ scale identically, so the comparison is between a constant (2) and something that grows logarithmically. A logarithm is slow, but it beats any fixed constant eventually, and here the opponent is a fixed constant.

Why is the margin so thin, and why does that matter? Because $I(U)$ grows like a logarithm and is multiplied by $1/(2A)$, the excess $2L - 4 \approx \frac{1}{A}\left(\frac{1}{2}\ln(2U) + \frac{1}{4} - 2\right)$ is tiny for every A: positive only for $A \gtrsim 35$, with $2L$ never above about 4.0027. No drawing, and no rough estimate, will show this. That is exactly why intuition says no. The lesson is that a borderline number in a yes/no question is an instruction to compute the exact difference, not to estimate each side.

Why the vertex on the circle and the axis along a diameter? Because 4 is two diameters, and to have any chance you need each arm to climb as nearly the full diameter as possible. Starting the vertex at the lowest point of the circle and pointing the axis at the centre is how you give both arms the whole height. Symmetry then halves the work.

Layer $+\alpha$ BACKTESTING

1. The number 4 in a circle of radius 1. Write down: $4 = 2\times$diameter, so the natural candidate climbs a full diameter twice.

2. Place the vertex at the bottom of the circle, axis vertical; compute the half-arc length as an integral and subtract the vertical rise to isolate the surplus.

3. Hesitation comes when the rough picture suggests "slightly less than 4" and the estimate is inconclusive. It is reinforced away by refusing to estimate both sides separately: compute *length minus height* exactly.

4. The trap is the visual intuition that the round bottom and the short arms must lose length, leading to a confident "no". A second trap: bounding the arc by a polygonal path and hoping the bound comes out below 2 per arm — it does not.

5. Slogan: *when a length or sum is compared to a natural benchmark, subtract the benchmark inside the integral and study what is left.*

6. Decision rule: **when a quantity is compared with a threshold that it nearly attains, write quantity − threshold as a single integral (or sum) and look for a divergent tail, typically harmonic ($\int du/u$); a slowly divergent gain beats any constant loss once the parameter is pushed far enough.**

▷ **Cue:** a question "can the length (or sum) exceed N?" where N is a natural benchmark such as two diameters $\longrightarrow$ **Move:** subtract the benchmark integrand (here the vertical rise dy) and study the difference integral

▷ **Cue:** an integrand $\sqrt{1+u^2}-u$ or any difference of two nearly equal quantities $\longrightarrow$ **Move:** rationalise: $\sqrt{1+u^2}-u = 1/(\sqrt{1+u^2}+u)$, and compare with $1/(2u)$ or $1/(2(1+u))$

▷ **Cue:** an integrand behaving like c/u at infinity $\longrightarrow$ **Move:** the integral diverges logarithmically; it beats any fixed constant for large enough upper limit

▷ **Cue:** a yes/no extremal question whose answer is within a fraction of a percent $\longrightarrow$ **Move:** do not trust the picture; compute the exact excess and find a certified parameter value

▷ **Cue:** "inside a circle" with the construction touching the boundary $\longrightarrow$ **Move:** shrink by a similarity; strict inequalities survive small perturbations

Putnam 2001 — Paper B

2001 B1

2001 B1. Let n be an even positive integer. Write the numbers $1, 2, \ldots, n^2$ in the squares of an $n \times n$ grid so that the k-th row, from left to right, is

$$(k-1)n+1, (k-1)n+2, \ldots, (k-1)n+n.$$

Color the squares of the grid so that half of the squares in each row and in each column are red and the other half are black (a checkerboard coloring is one possibility). Prove that for each coloring, the sum of the numbers on the red squares is equal to the sum of the numbers on the black squares.

Fields: Combinatorics; invariance / linear-decomposition arguments **Position:** B1 – opening problem of the second session **Answer:** True for every admissible coloring: the value written in cell (i, j) splits additively into a row-part $(i-1)n$ and a column-part j, and because every row and every column contains exactly $n/2$ red and $n/2$ black cells, each part sums to the same total on red as on black – independently of the shape of the coloring.

Layer 1 THE WHAT — surface analysis

The natural attempt. You check $n = 2$ by hand. The grid is $\begin{smallmatrix} 1 & 2 \\ 3 & 4 \end{smallmatrix}$; "half a row red" forces exactly one red and one black cell per row and per column, which means the coloring is one of the two diagonals. Both give $5 = 5$. Reassured, you reach for the coloring the problem names explicitly: the checkerboard. On a checkerboard, every red cell sits directly beside a black cell, so you tile the grid into $n^2/2$ dominoes, each domino one red and one black square whose entries differ by exactly 1 (horizontally adjacent) or exactly n (vertically adjacent, depending on how you tile). If you could show the domino differences cancel in total, you would be done – and for the checkerboard this can be made to work, with some case-splitting on whether $n/2$ is even or odd to get the horizontal and vertical dominoes to balance.

But the problem does not say "for the checkerboard coloring." It says *for each coloring* satisfying the half-red-half-black-per-line constraint. So you try to extend the domino argument: take an arbitrary admissible coloring and attempt to build a red–black matching – pair each red cell with some black cell so that the matched pairs' differences cancel, or at least sum in a controllable way. This is where the wall appears. Nothing forces red and black cells in a general coloring to be adjacent, or even close together. A coloring can put all of a row's red cells on the left half in one row and on the right half in the next, so there is no local domino structure to exploit at all. You are trying to construct an explicit pairing between two sets of $n^2/2$ numbers that need not have any geometric relationship to each other, using only the row/column count constraint – and a matching argument built for adjacency does not survive the loss of adjacency. The checkerboard proof was real progress on a special case, but it does not generalise, because it was never using the actual hypothesis (half per row and column); it was using a stronger, unstated one (red and black cells are locally adjacent).

The model chain. Write the value in cell (i, j) (row i, column j, both from 1 to n) as

$$v(i, j) = (i - 1)n + j.$$

This is exactly the labeling rule given: row i starts at $(i - 1)n + 1$ and counts up by 1 across the columns. Fix any admissible coloring and let R, Bk denote the sets of red and black cells. Then

$$\sum_{(i,j)\in R} v(i, j) = n \sum_{(i,j)\in R} (i - 1) \; + \; \sum_{(i,j)\in R} j,$$

and likewise for Bk. Treat the two pieces separately.

Row part. $\sum_{(i,j)\in R}(i - 1)$ groups by row: row i contributes $(i - 1)$ once for every red cell it contains. By hypothesis every row contains exactly $n/2$ red cells, so row i contributes $(i - 1) \cdot (n/2)$ to the red sum – regardless of which columns those red cells occupy. Hence

$$\sum_{(i,j)\in R} (i - 1) = \frac{n}{2} \sum_{i=1}^{n} (i - 1) = \sum_{(i,j)\in Bk} (i - 1),$$

since the same computation with the same $n/2$ gives the identical total for black. The row part of the red sum equals the row part of the black sum, and in fact both equal $\frac{n}{2} \sum_i (i - 1)$ exactly – this quantity does not depend on the coloring at all.

Column part. By the symmetric argument, $\sum_{(i,j)\in R} j$ groups by column: column j contributes j once for every red cell it contains, and every column contains exactly $n/2$ red cells, so

$$\sum_{(i,j)\in R} j = \frac{n}{2} \sum_{j=1}^{n} j = \sum_{(i,j)\in Bk} j.$$

Adding the two equalities, $\sum_R v = \sum_{Bk} v$, for every admissible coloring.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths start from the same object: the value $v(i, j)$ written in each cell. They split at the very first question asked about that object. The natural attempt asks "what pairs of actual cells can I match up so their values cancel or combine nicely?" – a question about the *geometry* of the coloring. The model chain instead asks "how does $v(i, j)$ split as a function of i and j separately?" – a question about the *formula*, before any coloring is even chosen.

> *I don't need to pair individual red cells with individual black cells at all – I need the* sum *over red to equal the* sum *over black, and if the thing I'm summing splits additively into a row-only piece and a column-only piece, I can prove each piece separately equal without ever matching a single cell to another.*

> *The row-part sum over red cells doesn't depend on which columns the red cells sit in – only on how many red cells each row has – so the moment I know "$n/2$ red cells per row," the row-part sum is already fixed, for any coloring satisfying that constraint, not just this one.*

Layer 3 THE WHY

Why does the decomposition $v(i, j) = (i - 1)n + j$ do the entire job here, when a geometric pairing argument does not? Because v is not an arbitrary function of position – it is an *affine function of the two coordinates separately*, $v(i, j) = n \cdot (\text{row index}) + (\text{column index})$, up to the constant shift. Any sum of an affine-in-two-variables quantity over a subset S of cells decomposes exactly the same way: it depends on S only through two numbers, the multiset of row indices touched (with multiplicity) and the multiset of column indices touched (with

multiplicity) – never through the finer question of *which* cell in a row is paired with *which* cell in a column. This is the structural fact the problem's phrasing is built on: "half of the squares in *each row* and in *each column*" is a constraint stated entirely in terms of row-counts and column-counts. It is, verbatim, a constraint on exactly the two summary statistics that the row/column decomposition of v needs and nothing more. The problem was never asking for a geometric matching; the hypothesis was handed to you already phrased in the coordinates that make the affine decomposition trivial to close. Kedlaya–Bhargava–Ng's published solution goes directly to this decomposition (writing $v = f(s)n + g(s) + 1$ with f row-dependent, g column-dependent) without commentary on why – this is where that choice comes from: the hypothesis is row/column-local, so the proof should be too, and the only way to make v row/column-local is to split it additively along exactly those two axes.

Caveat. This part-file's model chain was derived independently before comparison and matches the structure of Kedlaya–Bhargava–Ng's published solution (B–1) exactly, including the row/column additive split; the write-up above (the affine-decomposition framing in Layer 3) is this author's own gloss on why that split is forced, which the official solution states without justification.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the phrase "half of the squares in each row and in each column" – a constraint stated purely in terms of row-counts and column-counts. On seeing it, write down: whatever proof strategy I use had better only need to know how many red cells are in each row and column, not where they are.
2. That cue suggests decomposing whatever quantity is being summed into a part that depends only on the row index and a part that depends only on the column index – because the hypothesis only constrains sums along rows and columns, an additive row/column split is the natural way to make the hypothesis directly usable.
3. Hesitation occurs right after the checkerboard case succeeds via a domino-pairing argument that feels like real progress. The reinforcement away from it: ask explicitly "does this pairing use anything beyond the stated hypothesis?" – here it silently uses adjacency, which is not given, and that mismatch is the signal to abandon it.
4. The trap is solving an easier, unstated version of the problem (the checkerboard, or any coloring with local structure) so convincingly that the method feels general, when it in fact depends on a property the problem deliberately does not grant.
5. To recognise this faster: whenever a hypothesis is stated as "the same count in every row and every column," check immediately whether the quantity being summed can be written as (row-dependent term) + (column-dependent term). If it can, the proof is usually two one-line symmetric arguments, not a construction.
6. The reusable rule: a sum of an affine-in-two-coordinates quantity over a set depends on that set only through its row-count and column-count profile. Whenever a combinatorics hypothesis is phrased row-by-row-and-column-by-column, look for this decomposition before reaching for an explicit matching or induction.

2001 B2

2001 B2. Find all pairs of real numbers (x, y) satisfying the system of equations

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

Fields: Algebra (symmetric-function manipulation, binomial recognition) **Position:** B2 – second problem of the second session **Answer:** The unique solution is $x = \frac{3^{1/5}+1}{2}$, $y = \frac{3^{1/5}-1}{2}$, found by turning the system into $(x+y)^5 = 3$ and $(x-y)^5 = 1$.

Layer 1 THE WHAT — surface analysis

The natural attempt. You have two equations, each with $1/x$ and $1/(2y)$ on the left and a quartic in x, y on the right. The reflex is to clear denominators immediately: multiply the first equation by $2xy$ to get

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

and multiply the second by $2xy$ to get

$$2y - x = 4xy(y^4 - x^4).$$

This does remove the fractions, but look at what you now have: a degree-1 left side against a degree-5 right side on each equation, with the right sides fully multiplied out (the first right side alone expands to $2xy(3x^4 + 10x^2y^2 + 3y^4)$, six terms once distributed). You could add and subtract these two new equations, hoping for cancellation, but the terms don't match up cleanly across the two – one right side is built from $(x^2 + 3y^2)(3x^2 + y^2)$, an even-symmetric product, and the other from $y^4 - x^4$, an odd-under-$x \leftrightarrow -x$ difference, and clearing by $2xy$ on both sides has now buried whatever relationship they have to each other under fully expanded polynomials in x and y separately. You can grind through the algebra – expand everything, add, subtract, try to factor – but nothing you've done so far has told you *why* these two particular quartics were chosen, so you have no compass for which grouping of the resulting mess will simplify. This is workable in principle but it is zombie maths: manipulation without a target, on a system that was visibly constructed to collapse into something clean.

The model chain. Do not clear denominators yet. Instead add the two original equations directly:

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

Expand only the right side: $(x^2 + 3y^2)(3x^2 + y^2) = 3x^4 + 10x^2y^2 + 3y^4$, so

$$\frac{2}{x} = 3x^4 + 10x^2y^2 + 3y^4 + 2y^4 - 2x^4 = x^4 + 10x^2y^2 + 5y^4.$$

Subtracting the two original equations instead of adding gives, by the same expansion,

$$\frac{1}{y} = 5x^4 + 10x^2y^2 + y^4.$$

Now look at the coefficients: $1, 10, 5$ and $5, 10, 1$. Those are binomial coefficients $\binom{5}{0}, \binom{5}{2}, \binom{5}{4}$ and their mirror – exactly the odd-degree terms of $(x+y)^5$ and $(x-y)^5$ split by which variable carries the odd power. Multiply the first identity by x and the second by y:

$$2 = x^5 + 10x^3y^2 + 5xy^4, \qquad 1 = 5x^4y + 10x^2y^3 + y^5.$$

Add these:

$$3 = x^5 + 5x^4y + 10x^3y^2 + 10x^2y^3 + 5xy^4 + y^5 = (x+y)^5.$$

Subtract them:

$$1 = x^5 - 5x^4y + 10x^3y^2 - 10x^2y^3 + 5xy^4 - y^5 = (x-y)^5.$$

Fifth powers of reals are strictly increasing, hence injective, so $x+y = 3^{1/5}$ and $x-y = 1$ are forced, uniquely. Solving,

$$x = \frac{3^{1/5}+1}{2}, \qquad y = \frac{3^{1/5}-1}{2}.$$

One checks $x \neq 0$, $y \neq 0$ (both terms are positive/nonzero since $3^{1/5} \approx 1.246 > 1$), so this pair lies in the domain of the original system, and it is the unique solution.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both paths begin with the same two equations and both correctly compute $2/x = x^4 + 10x^2y^2 + 5y^4$ and $1/y = 5x^4 + 10x^2y^2 + y^4$ if pursued carefully – the difference is what happens *before* that point. The natural attempt clears the fractions on the original system immediately, at the moment the equations are still in their most awkward form, and only then adds/subtracts. The model chain adds/subtracts the fraction-form equations *first*, while $1/x$ and $1/y$ are still isolated and simple, and only clears denominators afterward, by multiplying by x and y respectively rather than by $2xy$ jointly.

> *Clearing denominators is not free – it should be the last move, not the first, because once I've multiplied out a degree-5 polynomial I've destroyed whatever structure let me see it as a sum of simpler pieces; I should add and subtract the two equations* in fraction form *first, while the left sides are still just* $1/x$ *and* $1/y$*, and only then decide how to clear.*

> *Coefficients* $1, 10, 5$ *appearing next to powers* x^4, x^2y^2, y^4 *are not incidental –* $10 = \binom{5}{2}$ *and* $5 = \binom{5}{1}$ *are binomial coefficients of a fifth power, so if I multiply this equation by* x *(or the mirrored one by* y*) and add to its partner, I should get exactly the missing odd-degree terms of* $(x+y)^5$*, collapsing six monomials into one fifth power.*

Layer 3 THE WHY

Caveat. This chain – add/subtract in fraction form, multiply by x and y, add/subtract again to reach $(x+y)^5 = 3$, $(x-y)^5 = 1$ – was derived and checked independently by full expansion (verifying $(x+y)^5$ term by term against $x^5 + 5x^4y + 10x^3y^2 + 10x^2y^3 + 5xy^4 + y^5$) before comparison, and it matches Bhargava–Kedlaya–Ng's published B2 solution exactly. No step is unverified.

Why is adding-before-clearing the natural move here, rather than a trick to be memorised? Because the problem hands you two equations whose left sides are already additively separated – $1/x + 1/(2y)$ and $1/x - 1/(2y)$ – and addition/subtraction of a "sum" equation and a "difference" equation is precisely the operation that isolates one variable's reciprocal at a time ($2/x$, then $1/y$). The problem's phrasing is doing the work of telling you the intended first move: two equations built from $A+B$ and $A-B$ are a standing invitation to add and subtract, regardless of what A and B are. Clearing denominators before taking that invitation buries it. The second thought – recognising $1, 10, 5$ as binomial coefficients – is likewise not a spot of luck but a consequence of what a fifth-power expansion *is*: $(x+y)^5$ has exactly six monomials with coefficients $1, 5, 10, 10, 5, 1$, alternating in a real proof between even-power and

odd-power terms in x. Once the equation $2/x = x^4 + 10x^2y^2 + 5y^4$ is sitting in front of you with those coefficients on the even-degree terms, multiplying by x is the only operation that turns "even-degree terms with binomial coefficients" into "odd-degree terms with binomial coefficients" – which is exactly what is missing to complete the expansion of $(x+y)^5$. The structural fact underneath is that $(x+y)^5 + (x-y)^5$ and $(x+y)^5 - (x-y)^5$ each isolate half the binomial terms (even-power-in-y terms in the first, odd-power-in-y terms in the second) – this problem is built by running that decomposition backwards, and the two given equations are x times and y times the two halves.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the shape of the left-hand sides: $1/x + 1/(2y)$ in one equation, $1/x - 1/(2y)$ in the other – an $A+B$, $A-B$ pair. On seeing it, write down: add and subtract these two equations before touching the right-hand sides at all.
2. That cue suggests the standard "sum and difference" elimination method, here applied not to isolate a linear unknown but to isolate $1/x$ and $1/y$ separately – the same idea, one level removed from where it's usually taught.
3. Hesitation occurs at the instant the equations still look intractable (degree-1 versus degree-4, fractions present) and clearing denominators feels like the only way to "do something." The reinforcement away from it: resist any denominator-clearing move until the equations have been combined as far as possible in their given form – clearing should be the last simplification, not the first.
4. The trap is treating "remove the fractions" as a free, harmless, always-correct first step. It is algebraically valid but strategically wrong here: it destroys the additive structure ($A+B/A-B$) that the problem was built to exploit, before you've had a chance to use it.
5. To recognise this faster: whenever a system presents two equations as an explicit sum-form and difference-form of the same two quantities, add and subtract immediately, before any other algebraic manipulation, no matter how the right-hand sides look.
6. The reusable rule: coefficients that are recognisably binomial ($1, 5, 10, 10, 5, 1$ or any row of Pascal's triangle) sitting next to matching powers of two variables are a standing invitation to multiply through by one variable and add to a mirrored equation, reconstructing a perfect n-th power that collapses many monomials into one clean equation.

2001 B3

2001 B3. For any positive integer n, let $\langle n \rangle$ denote the closest integer to $\sqrt{n}$. Evaluate

$$\sum_{n=1}^{\infty} \frac{2^{\langle n \rangle} + 2^{-\langle n \rangle}}{2^n}.$$

Fields: Analysis (series manipulation), number theory (nearest-integer function) **Position:** B3 – third problem of the second session **Answer:** The sum equals 3. Grouping the terms by the constant value $k = \langle n \rangle$ turns the series into a telescoping sum of the form $2^{1-(k-1)^2} - 2^{1-(k+1)^2}$, which collapses to its first two terms.

Layer 1 THE WHAT — surface analysis

The natural attempt. The summand $\left(2^{\langle n \rangle} + 2^{-\langle n \rangle}\right)/2^n$ has 2^{-n} in it, so the reflex is that this should behave like a geometric series and sum to something clean – and $2^{\langle n \rangle}$ looks close to $2^{\sqrt{n}}$, so maybe there's a slick closed form via a known generating function. You compute the first several terms directly. $\langle 1 \rangle = 1$ (since $\sqrt{1} = 1$), $\langle 2 \rangle = 1$ (since $\sqrt{2} \approx 1.41$, closer to 1), $\langle 3 \rangle = 2$ (since $\sqrt{3} \approx 1.73$, closer to 2), $\langle 4 \rangle = 2$, $\langle 5 \rangle = 2$, $\langle 6 \rangle = 2$ ($\sqrt{6} \approx 2.45$), $\langle 7 \rangle = 3, \ldots$ You get partial sums $n = 1$: $(2 + 1/2)/2 = 1.25$; adding a few more terms, the running total creeps toward

something that looks like it might be 3. That is a real observation, but it is a guess, not a proof, and you have no way yet to justify it beyond "the calculator says so." To make it rigorous you would need to sum the series exactly, and summing $\sum 2^{\langle n\rangle - n}$ term by term, treating $\langle n\rangle$ as an opaque function of n, gives you no algebraic handle: $\langle n\rangle$ is neither linear in n nor expressible as a simple closed form of n (it grows like $\sqrt{n}$), so none of the standard geometric-series or generating-function machinery applies directly to the sum written this way. You are stuck computing more and more terms, more and more confident of the numerical answer, with no route from "the numbers suggest 3" to "the numbers are 3."

The model chain. The obstruction is that $\langle n\rangle$ is not a nice function of n termwise, but it *is* constant on long, exactly-describable blocks of consecutive n. Pin down those blocks first. Since $\langle n\rangle = k$ means k is the integer nearest $\sqrt{n}$, i.e. $|\sqrt{n} - k| \le 1/2$, square the (nonnegative) bound: $\langle n\rangle = k$ iff

$$\left(k - \tfrac{1}{2}\right)^2 \le n \le \left(k + \tfrac{1}{2}\right)^2 \iff k^2 - k + 1 \le n \le k^2 + k,$$

using that n is an integer and $(k - \frac{1}{2})^2 = k^2 - k + \frac{1}{4}$, $(k + \frac{1}{2})^2 = k^2 + k + \frac{1}{4}$. This is a block of exactly $2k$ consecutive integers, and as k runs over $1, 2, 3, \ldots$ these blocks partition all of $\mathbb{N}$ with no gaps or overlaps (block k ends at $k^2 + k$; block $k+1$ starts at $(k+1)^2 - (k+1) + 1 = k^2 + k + 1$). So swap the order of summation, grouping by k instead of by n:

$$\sum_{n=1}^{\infty} \frac{2^{\langle n\rangle} + 2^{-\langle n\rangle}}{2^n} = \sum_{k=1}^{\infty} \left(2^k + 2^{-k}\right) \sum_{n=k^2-k+1}^{k^2+k} 2^{-n}.$$

The inner sum is a finite geometric series with $2k$ terms, first term $2^{-(k^2-k+1)}$, ratio $\frac{1}{2}$:

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

Multiply by $(2^k + 2^{-k})$ and expand: the cross terms cancel,

$$(2^k + 2^{-k})\left(2^{-(k^2-k)} - 2^{-(k^2+k)}\right) = 2^{-(k^2-2k)} - 2^{-k^2} + 2^{-k^2} - 2^{-(k^2+2k)} = 2^{-(k^2-2k)} - 2^{-(k^2+2k)}.$$

Complete the square in each exponent: $k^2 - 2k = (k-1)^2 - 1$ and $k^2 + 2k = (k+1)^2 - 1$, so the k-th term is

$$T_k = 2^{1-(k-1)^2} - 2^{1-(k+1)^2}.$$

Setting $a_m := 2^{1-(m-1)^2}$ for $m \ge 1$, this is $T_k = a_k - a_{k+2}$: a telescoping sum with lag 2. Summing over $k = 1, \ldots, N$ leaves only the first two and last two terms:

$$\sum_{k=1}^{N} T_k = a_1 + a_2 - a_{N+1} - a_{N+2}.$$

Since $a_m = 2^{1-(m-1)^2} \to 0$ as $m \to \infty$, letting $N \to \infty$ gives

$$\sum_{k=1}^{\infty} T_k = a_1 + a_2 = 2^{1-0} + 2^{1-1} = 2 + 1 = 3.$$

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both approaches begin by computing $\langle n\rangle$ for small n and both correctly sense the answer is close to 3. They split at the question of what to do with the observation that $\langle n\rangle$ is locally constant. The natural attempt notices this only informally, while still summing term-by-term over n, and never converts it into an algebraic tool. The model chain converts it into an exact statement – $\langle n\rangle = k \iff k^2 - k + 1 \le n \le k^2 + k$ – and uses that statement to change the entire order of summation, from indexing by n to indexing by k.

> *$\langle n \rangle$ is a nightmare to sum against directly because it's not a formula in n, it's a step function – so instead of summing over n with $\langle n \rangle$ appearing inside each term, I should sum over the* values *$k = \langle n \rangle$ takes, and for each value k, sum over the block of n's that produce it; that turns an unmanageable function of n into a constant on each piece.*

> *Once I've isolated the block for a fixed k, the inner sum is just a finite geometric series in 2^{-n} that I know how to sum exactly – so the entire difficulty of the problem is now concentrated in finding the block's endpoints correctly, via $|\sqrt{n} - k| \leq 1/2$ squared to $k^2 - k + 1 \leq n \leq k^2 + k$, not in the summation itself.*

Layer 3 THE WHY

Caveat. This derivation was carried out independently, including the completing-the-square step ($k^2 \mp 2k = (k \mp 1)^2 - 1$) and the telescoping-with-lag-2 argument, and checked numerically against partial sums of the original series for small n; it agrees with Bhargava–Kedlaya–Ng's published B3 solution (which reaches the same T_k via a slightly different but equivalent exponent bookkeeping) and with their listed alternate solution. No step is unverified.

Why does grouping by $k = \langle n \rangle$ rather than manipulating the series in n directly turn out to be forced, not merely convenient? Because $\langle n \rangle$ is, by its very definition, a function whose entire behaviour is "constant on intervals determined by squares" – the nearest-integer-to-$\sqrt{\cdot}$ function is precisely the inverse of $n \mapsto n^2$ read coarsely. Any sum with $\langle n \rangle$ inside it is, structurally, a sum over a sequence that has already been organised into blocks by the problem's own definition; the work of finding those blocks (the inequality $k^2 - k + 1 \leq n \leq k^2 + k$) is not extra machinery imported from outside, it is simply making explicit the partition the definition of $\langle \cdot \rangle$ already imposes. Once the sum is reorganised along that partition, the inner sum over each block is an ordinary finite geometric series – the one piece of the problem that was never hard – and the outer sum over k collapses because $(2^k + 2^{-k})$ times a difference of two powers of 2 centred at k^2 is, after completing the square, *exactly* a difference of two terms of a single sequence $a_m = 2^{1-(m-1)^2}$ two steps apart. That telescoping is not a coincidence either: $2^k \cdot 2^{-(k^2-k)} = 2^{-(k^2-2k)} = 2^{-((k-1)^2-1)}$ is the general pattern that any expression of the form (power of 2 in k) times (power of 2 in k^2) will complete a square and land on a shifted perfect square in the exponent – which is exactly the shape a telescoping series needs. The problem was built by running this construction backwards: start from a telescoping sequence in $2^{-(m-1)^2}$, and the $\langle n \rangle$ definition is what disguises it as a sum over n.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is that $\langle n \rangle$ is a step function of n defined via $\sqrt{n}$ – it is constant over stretches of n, not smoothly varying. On seeing it, write down: find the exact range of n for which $\langle n \rangle = k$, for general k, before trying to sum anything.
2. That cue suggests swapping the order of summation: instead of summing over n with $\langle n \rangle$ buried inside each term, sum over k, with an inner finite geometric sum over the block of n's where $\langle n \rangle = k$.
3. Hesitation occurs at the point of translating "k is the nearest integer to $\sqrt{n}$" into a clean inequality in n; it is tempting to guess the bounds are k^2 to $(k+1)^2 - 1$ (i.e. floor-based) rather than working out $(k - \frac{1}{2})^2 \leq n \leq (k + \frac{1}{2})^2$ from the actual nearest-integer condition. The reinforcement: derive the bound explicitly from $|\sqrt{n} - k| \leq 1/2$ rather than pattern-matching to a floor function.

4. The trap is confusing "nearest integer to $\sqrt{n}$" with "floor of $\sqrt{n}$" – they give different, off-by-one block boundaries, and using the wrong one silently corrupts every subsequent computation without any obvious symptom until the final telescoping fails to close cleanly.

5. To recognise this faster: any series indexed by n whose summand contains a step function of n defined via a nearest-integer or floor condition on some $g(n)$ should immediately be re-indexed by the value of that step function, with the block boundaries found by solving the defining inequality exactly.

6. The reusable rule: a sum $\sum_n a_{f(n)} \cdot r^n$ where f is constant on known blocks of n reorganises into $\sum_k a_k \sum_{n \in \text{block}_k} r^n$, an outer sum against a closed-form finite geometric inner sum; look for the resulting outer sum to telescope by completing the square in the block-boundary formula whenever the blocks are centred at k^2.

2001 B4

2001 B4. Let S denote the set of rational numbers different from $\{-1, 0, 1\}$. Define $f : S \to S$ by $f(x) = x - 1/x$. Prove or disprove that

$$\bigcap_{n=1}^{\infty} f^{(n)}(S) = \emptyset,$$

where $f^{(n)}$ denotes f composed with itself n times.

Fields: number theory (rationals in lowest terms); descent / height functions; iteration of maps **Position:** B4 — mid-paper; short once the right measuring stick is chosen, and almost impossible to finish without one **Answer:** True: the intersection is empty. With $H(p/q) = |p| + |q|$ in lowest terms, every application of f raises H by at least 2, so $f^{(n)}(S)$ contains only rationals of height at least $2n + 3$, and no rational survives every n.

Layer 1 THE WHAT — surface analysis

The natural attempt. The strong student first checks that f really maps S into S, then starts computing. $f(2) = 3/2$, $f(3/2) = 5/6$, $f(5/6) = -11/30$, $f(-11/30) = 779/330$. The numbers get uglier fast, which suggests the intersection is empty: surely nothing can be hit by arbitrarily long chains of preimages.

So the student tries to prove it by fixing $y \in S$ and asking whether $y = f^{(n)}(x)$ is solvable for every n. One step back: $x - 1/x = y$ means $x^2 - yx - 1 = 0$, so $x = \left(y \pm \sqrt{y^2 + 4}\right)/2$, and x is rational exactly when $y^2 + 4$ is the square of a rational. Two steps back: the preimage must itself have $x^2 + 4$ a rational square, and so on. The student now has a tower of conditions "$y^2 + 4 = \square$, and one of the two roots x has $x^2 + 4 = \square$, and …", which is a problem about rational points on a sequence of curves. That is the wall. The backward direction is a branching tree of Diophantine conditions, and nothing in the student's toolkit says the tree dies out. The instinct ("the numbers get uglier") was right; the direction of travel was wrong.

The model chain. **Step 0: f maps S to S, and preserves lowest terms.** Write $x = p/q$ in lowest terms with $q > 0$; $x \in S$ means $p \neq 0$ and $|p| \neq |q|$. Then

$$f(x) = \frac{p}{q} - \frac{q}{p} = \frac{p^2 - q^2}{pq}.$$

This fraction is in lowest terms: a prime dividing pq divides p or q, say p; if it also divided $p^2 - q^2$ it would divide q^2, hence q, contradicting $\gcd(p, q) = 1$ (the case of q is symmetric under swapping the roles of p, q, since $p^2 - q^2 = -(q^2 - p^2)$). The numerator is nonzero since

$|p| \neq |q|$. And $f(x) \neq \pm 1$: that would need $p^2 \mp pq - q^2 = 0$, whose discriminant in p is $5q^2$, not a perfect square for $q \neq 0$, so no integer p exists. Hence $f(S) \subseteq S$.

Step 1: define the height. For $x = p/q \in S$ in lowest terms put $H(x) = |p| + |q|$. For every $x \in S$, $|p|, |q| \geq 1$ and $|p| \neq |q|$, so $H(x) \geq 1 + 2 = 3$.

Step 2: f raises height by at least 2. By Step 0, $H(f(x)) = |p^2 - q^2| + |pq|$. Now

$$|p^2 - q^2| = \big||p| - |q|\big| \cdot \big(|p| + |q|\big) \geq 1 \cdot 3 = 3,$$

because $|p| \neq |q|$ are integers and $|p| + |q| \geq 3$. Therefore

$$H(f(x)) - H(x) \;\geq\; 3 + |p||q| - |p| - |q| \;=\; (|p| - 1)(|q| - 1) + 2 \;\geq\; 2.$$

Step 3: conclude. If $y \in f^{(n)}(S)$ then $y = f^{(n)}(x)$ for some $x \in S$, and n applications of Step 2 give $H(y) \geq H(x) + 2n \geq 2n + 3$. Given any $y \in S$, choose n with $2n + 3 > H(y)$; then $y \notin f^{(n)}(S)$, so y is not in the intersection. The intersection is empty. □

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains agree that the iterates "get uglier". They split at the choice of direction. The natural attempt fixes the output y and runs *backwards*, solving $x^2 - yx - 1 = 0$ and meeting square-root conditions. The model runs *forwards*: it turns "uglier" into a number, H, and proves that number strictly climbs. Forwards, f is a single rational function with an explicit formula in p, q; backwards, it is a two-valued map with an irrationality test at every node.

> *I want to show no point is in the image of $f^{(n)}$ for all n. That is the same as: every point has only boundedly many ancestors in a chain. So I need a quantity that goes up every time I apply f and is finite for every point — something like a size of a rational.*

> *"Size" of p/q in lowest terms means how big p and q are. Let me compute $f(p/q) = (p^2 - q^2)/(pq)$ and check it is still in lowest terms; if it is, the new numerator and denominator are roughly the squares of the old ones, so $|p| + |q|$ has to go up.*

> *I need a strict increase by a fixed amount, not just "bigger". Where could it fail? Only when $|p|$ or $|q|$ is 1. Factor: $|pq| - |p| - |q| + 1 = (|p| - 1)(|q| - 1) \geq 0$, and $|p^2 - q^2|$ gives the rest.*

- ▷ **Cue:** "prove the intersection of the images $f^{(n)}(S)$ is empty" (or: no point has arbitrarily long preimage chains) ⟶ **Move:** find a function $H \geq 0$ with only finitely many points below each bound, and show $H(f(x)) \geq H(x) + c$ with $c > 0$
- ▷ **Cue:** a map on rationals given by a rational function with integer coefficients ⟶ **Move:** write $x = p/q$ in lowest terms, compute the image as a fraction in p, q, check coprimality, and read off a height
- ▷ **Cue:** iterates of a map "look like they get more complicated" ⟶ **Move:** quantify "complicated" as a height and prove monotonicity forwards rather than solving backwards
- ▷ **Cue:** need $|pq| - |p| - |q|$ bounded below ⟶ **Move:** add 1 and factor as $(|p| - 1)(|q| - 1)$

Layer 3 THE WHY

Why a height, and why run forwards? Go to what the set $\bigcap_n f^{(n)}(S)$ actually is. A point y lies in $f^{(n)}(S)$ when there is a chain $x_0 \mapsto x_1 \mapsto \cdots \mapsto x_n = y$ inside S. So y is in the intersection when it has ancestor chains of every length. To rule that out you need a reason chains cannot be long, and the only general reason a chain in a discrete world cannot be long is that something is

being *used up* or *built up* along it. That is the principle of descent (Fermat's infinite descent, or the well-ordering of $\mathbb{N}$): a strictly monotone quantity with values in the non-negative integers cannot move by at least 2 indefinitely without passing any fixed finite value. So the problem is really asking you to invent a monovariant, and the only freedom is which one.

Why $|p| + |q|$? Because the objects are rationals and the only intrinsic integer data a rational carries is its lowest-terms numerator and denominator. The decimal value $|x|$ will not do: $f(3/2) = 5/6$ is smaller than $3/2$, and f moves points both towards and away from zero. What *does* grow is arithmetic complexity: the formula $(p^2 - q^2)/(pq)$ is quadratic in (p, q), so if no cancellation happens the numerator and denominator roughly square. The whole proof therefore rests on one fact, the coprimality check of Step 0 — that is where the phrase "lowest terms" earns its place. Without it, H of the image could collapse.

Why does the problem exclude $\{-1, 0, 1\}$? Look at the formula. $x = 0$ has no image; $f(\pm 1) = 0$, which would then be stuck. These are exactly the points where $|p| = |q|$ or $p = 0$, which are exactly the cases where the estimate $|p^2 - q^2| \geq 3$ fails. The excluded set is the problem-setter telling you where the height argument would break, and so it is a cue, not decoration. The inequality needs two ingredients from the hypotheses: $|p| \neq |q|$ (so $||p| - |q|| \geq 1$) and $|p|, |q| \geq 1$ (so $|p| + |q| \geq 3$ and $(|p| - 1)(|q| - 1) \geq 0$).

The official solution (Bhargava, Kedlaya, Ng) begins "define its height $H(p/q)$ to be $|p| + |q|$". Here is where that comes from: the question "is anything in every image?" is a question about the length of backward chains; lengths are bounded by a monovariant; and on $\mathbb{Q}$ the natural integer-valued measure is the size of the lowest-terms pair. Their closing note says the key properties are that only finitely many rationals have height below any bound, and that H is algebraic enough to compare $H(f(x))$ with $H(x)$. Those are exactly the two things a descent argument needs: finiteness below every level (so every point has finite height) and a computable increment.

The natural attempt was not stupid. It asked the right question ("what can map to y?"), but it answered it by computing preimages, and preimages under a quadratic map come with square roots. The shortcut was to stop asking about any particular y and to prove a uniform statement about all chains at once.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the shape $\bigcap_n f^{(n)}(S)$ together with "rational numbers". On seeing it, write $x = p/q$ in lowest terms and $f(p/q) = (p^2 - q^2)/(pq)$.

2. A monovariant (height) argument: find an integer-valued H on S that increases by a fixed positive amount under f; then membership in $f^{(n)}(S)$ forces $H \geq 2n + 3$.

3. Hesitation comes at whether the new fraction is in lowest terms, and whether the increase is *strict and uniform* when $|p|$ or $|q|$ is 1. Reinforce it away by doing the prime-divisor check explicitly and by the factorisation $(|p| - 1)(|q| - 1) + 2$.

4. The trap is running backwards: solving $x^2 - yx - 1 = 0$ and chasing "$y^2 + 4$ is a rational square" through a tree of preimages. A secondary trap is using the absolute value $|x|$ as the measure, which is not monotone under f.

5. Recognise faster by translating the set-theoretic statement into one about chains: "in every image" = "has arbitrarily long ancestor chains". Long chains in a discrete set are killed by monovariants.

6. Reusable decision rule: *when asked whether some point survives infinitely many applications of a map (or has infinitely long preimage chains), do not solve backwards; build an integer-valued height that the map strictly increases (or decreases), and let well-ordering finish. On rationals, try* $|p| + |q|$ *or* $\max(|p|, |q|)$ *in lowest terms.*

2001 B5

2001 B5. Let a and b be real numbers in the interval $(0, 1/2)$, and let g be a continuous real-valued function such that $g(g(x)) = ag(x) + bx$ for all real x. Prove that $g(x) = cx$ for some constant c.

Fields: real analysis (continuity, monotonicity); functional equations; linear recurrences **Position:** B5 — hard; the obstacle is not a computation but knowing which object to study **Answer:** $g(x) = r_1 x$ or $g(x) = r_2 x$, where $r_1 > 0 > r_2$ are the roots of $t^2 - at - b = 0$: r_1 if g is increasing, r_2 if decreasing.

Layer 1 THE WHAT — surface analysis

The natural attempt. The strong student guesses first: try $g(x) = cx$, get $c^2 x = acx + bx$, so $c^2 - ac - b = 0$, two real roots of opposite sign. Good — the answer is known. Now to prove nothing else works.

The next instincts are the standard functional-equation moves. Substitute $x = 0$: $g(g(0)) = ag(0)$. Differentiate? g is only continuous. Try to show $g(x)/x$ is constant: put $h(x) = g(x)/x$, then $h(g(x))\,h(x) = a\,h(x) + b$, a relation between h at two different points which does not close up. Try to show g is additive, or odd, or that $g(2x) = 2g(x)$ — none of these follow from the equation, because the equation only ever relates x to $g(x)$, never x to $2x$ or $x + y$.

A student who has seen recurrences may notice $g(g(x)) - ag(x) - bx = 0$ looks like $t^2 - at - b$, and write $x_{n+2} = ax_{n+1} + bx_n$ for $x_n = g^{(n)}(x)$. Then $x_n = c_1 r_1^n + c_2 r_2^n$, and since $|r_1|, |r_2| < 1$ (that is what $a, b < \frac{1}{2}$ is for, surely), $x_n \to 0$. And then… nothing. Forward orbits of *every* g with this property decay regardless of whether g is linear, so forward decay cannot distinguish linear from non-linear. That is the wall: the recurrence is the right object, but looked at in the direction where both modes die, it carries no information.

The model chain. Let $r_1 = \frac{1}{2}\left(a + \sqrt{a^2 + 4b}\right)$ and $r_2 = \frac{1}{2}\left(a - \sqrt{a^2 + 4b}\right)$ be the roots of $t^2 - at - b$. Since $a, b > 0$: $r_1 r_2 = -b < 0$, so $r_1 > 0 > r_2$; and $r_1 + r_2 = a > 0$, so $r_1 > |r_2| > 0$.

Step 1: g is a strictly monotone bijection of $\mathbb{R}$. If $g(x) = g(y)$ then $g(g(x)) = g(g(y))$, so $bx = by$ and $x = y$ (as $b \neq 0$). A continuous injective function on an interval is strictly monotone (otherwise three points $u < v < w$ with $g(v)$ not between $g(u)$ and $g(w)$ would let the intermediate value theorem produce two points with equal value). If $g(x)$ tended to a finite limit L as $x \to +\infty$, then $g(g(x)) \to g(L)$ by continuity, and $ag(x) \to aL$, so $bx = g(g(x)) - ag(x)$ would stay bounded — false. Likewise as $x \to -\infty$. A monotone function with no finite limit at either end is unbounded above and below, and by the intermediate value theorem it is onto $\mathbb{R}$.

Step 2: $g(0) = 0$. Put $y = g(0)$; the equation at $x = 0$ gives $g(y) = ay$. If g is increasing and $y > 0$, then $g(y) > g(0) = y$, but $g(y) = ay < y$. If g is increasing and $y < 0$, then $g(y) < y$ but $ay > y$. If g is decreasing, then $\phi(x) = g(x) - x$ is strictly decreasing, continuous, and tends to $\mp\infty$ as $x \to \pm\infty$, so g has a unique fixed point p; then $p = g(g(p)) = ap + bp$, so $p(1 - a - b) = 0$, and since $a + b < 1$, $p = 0$. In every case $g(0) = 0$.

Consequently, since g is strictly monotone with $g(0) = 0$: if g is increasing, $g(x)$ has the same sign as x; if decreasing, the opposite sign.

Step 3: the two-sided orbit. Fix $x_0 \neq 0$. Define $x_n = g^{(n)}(x_0)$ for $n \geq 0$ and $x_n = (g^{-1})^{(-n)}(x_0)$ for $n < 0$ (possible by Step 1). The functional equation applied at x_n says $x_{n+2} = ax_{n+1} + bx_n$ for every $n \in \mathbb{Z}$. Choose c_1, c_2 with $c_1 + c_2 = x_0$ and $c_1 r_1 + c_2 r_2 = x_1$ (possible as $r_1 \neq r_2$). The sequence $c_1 r_1^n + c_2 r_2^n$ satisfies the same recurrence, which, because $b \neq 0$, can be run backwards as $x_n = (x_{n+2} - ax_{n+1})/b$; two sequences obeying it and agreeing at $n = 0, 1$ agree for all $n \in \mathbb{Z}$. So $x_n = c_1 r_1^n + c_2 r_2^n$ for all integers n.

Step 4: sign pattern kills one mode.

g increasing. All x_n have the sign of x_0 (Step 2, applied to g and g^{-1}, which is also increasing and fixes 0). Suppose $c_2 \neq 0$. As $n \to -\infty$, $|r_2|^n / r_1^n = (r_1/|r_2|)^{-n} \to \infty$, so $x_n = r_2^n\left(c_2 + c_1 (r_1/r_2)^n\right)$ has the sign of $c_2 r_2^n$ for n sufficiently negative, which alternates because $r_2 < 0$. Contradiction; so $c_2 = 0$ and $g(x_0) = x_1 = c_1 r_1 = r_1 x_0$.

g decreasing. Consecutive x_n have opposite signs. Suppose $c_1 \neq 0$. As $n \to +\infty$, $(|r_2|/r_1)^n \to 0$, so x_n has the sign of $c_1 r_1^n$, which is constant. Contradiction; so $c_1 = 0$ and $g(x_0) = r_2 x_0$.

As $x_0 \neq 0$ was arbitrary and $g(0) = 0$, $g(x) = cx$ with $c = r_1$ or $c = r_2$. □

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains write down $x_{n+2} = ax_{n+1} + bx_n$ and solve it. They split at the direction of the orbit. The natural attempt runs it forwards, where both r_1^n and r_2^n decay and no contradiction is visible. The model first proves g is invertible, so the orbit extends to $n \to -\infty$, where the *smaller* root r_2 takes over; then it compares the resulting sign pattern with the sign pattern monotonicity forces.

> *Guessing $g(x) = cx$ gives $c^2 = ac + b$ — a characteristic equation. So the functional equation is a second-order linear recurrence in disguise, and the object to study is the orbit x, $g(x)$, $g(g(x))$, ... Linearity of g is the statement that one of the two modes is absent from every orbit.*

> *Forwards both modes shrink; I cannot see the small one. Is g invertible? $g(x) = g(y)$ forces $bx = by$ — yes, injective, and continuous so monotone. Then I can run the orbit backwards, and going backwards the root of* smaller *modulus blows up fastest and dominates.*

> *What does monotonicity say about an orbit? If g is increasing and fixes 0, it never changes sign; if decreasing, it always changes sign. A mode with a negative root alternates in sign; a mode with a positive root does not. That is my contradiction.*

- ▷ **Cue:** a functional equation of the form $g(g(x)) = ag(x) + bx$ (an iterate of g expressed linearly in lower iterates) ⟶ **Move:** set $x_n = g^{(n)}(x)$; the equation becomes a linear recurrence with characteristic polynomial $t^2 - at - b$
- ▷ **Cue:** $g(x) = g(y)$ can be fed back into the equation to extract $x = y$ ⟶ **Move:** g is injective; with continuity, strictly monotone; check limits at $\pm\infty$ for surjectivity
- ▷ **Cue:** a linear recurrence whose forward behaviour is uninformative ⟶ **Move:** if the step is invertible, run it backwards: the smallest-modulus root dominates as $n \to -\infty$
- ▷ **Cue:** a monotone map fixing 0 and an orbit written as $c_1 r_1^n + c_2 r_2^n$ with roots of opposite sign ⟶ **Move:** compare sign patterns: increasing preserves sign, decreasing alternates it; a negative root alternates, a positive root does not

Layer 3 THE WHY

Why is the orbit the right object, and not some property of g at a point? Go back to what the equation constrains. It relates the three numbers x, $g(x)$, $g(g(x))$ and nothing else. There is no $g(x + y)$, no $g(2x)$, so no substitution will ever connect two points that are not on a common orbit. The only information the equation contains is information *along orbits*, and along an orbit it says exactly $x_{n+2} = ax_{n+1} + bx_n$. So the orbit is not a clever choice; it is the only thing the hypothesis talks about. And the claim "$g(x) = cx$" is, orbit by orbit, the claim $x_1 = cx_0$, i.e. that the orbit is a pure geometric sequence, i.e. that one of c_1, c_2 vanishes. Once you say it that way, the proof must be a mechanism for killing one mode.

Why run backwards? In $c_1 r_1^n + c_2 r_2^n$, the root of larger modulus dominates as $n \to +\infty$ and the root of smaller modulus dominates as $n \to -\infty$. A contradiction that kills c_2 (the smaller mode) can only come from looking where c_2 is visible, which is $n \to -\infty$. Forwards, r_2^n is swamped by r_1^n, and the forward orbit of a hypothetical bad g would look exactly like the orbit of $r_1 x$. That is why the natural attempt stalled: it was looking at the region where the evidence is invisible. Going backwards is legitimate only because g is a bijection, which is why Step 1 is not bookkeeping but the load-bearing wall.

Why signs? Because continuity gives monotonicity, and monotonicity is exactly a statement about order; the cheapest order information about an orbit is its sign pattern. A negative root is the only thing in the closed form that can produce alternation, and alternation is exactly what an increasing sign-preserving map forbids.

Kedlaya's solution (with Bhargava and Ng) compares x_n and x_{n+2} for n very negative and derives an order contradiction. Here is where that comes from: it is the same sign/order clash, done without first locating the fixed point. I have routed through $g(0) = 0$ because it makes the clash visible in one line: an increasing map fixing 0 cannot send a positive number to a negative one. (One small point on reading the official text: for n very negative the magnitudes $|x_n|$ grow as n decreases, so on a same-sign subsequence $x_n > x_{n+2} > 0$, not $x_n < x_{n+2}$; the contradiction survives either way, because x_{n+1} and x_{n+3} then have the opposite order.)

What is the hypothesis $a, b \in (0, \frac{1}{2})$ for? Audit the proof. It uses $b \neq 0$ (injectivity, backwards recurrence), $b > 0$ (roots of opposite sign), $a > 0$ (so $r_1 > |r_2|$, making the modes separate by size), and, in Step 2's decreasing case, $a + b \neq 1$. It never uses $r_1 < 1$. The upper bound $\frac{1}{2}$ guarantees $a + b < 1$ and $|r_1|, |r_2| < 1$, and the second of these is a red herring for this proof: it makes the forward decay tempting and irrelevant. (Step 2 can even avoid $a + b < 1$: for decreasing g, the orbit of a fixed point $p \neq 0$ would be constant, forcing $c_2 = 0$, $c_1 = p$, and then $x_1 = r_1 p = p$ requires $r_1 = 1$; this is still a special case to handle, so I keep the hypothesis-based argument.)

Layer $+\alpha$ BACKTESTING

1. The first cue is $g(g(x))$ written as a linear combination of $g(x)$ and x. On seeing it, write $x_n = g^{(n)}(x)$, $x_{n+2} = ax_{n+1} + bx_n$, and the characteristic polynomial $t^2 - at - b$ with roots $r_1 > 0 > r_2$, $r_1 > |r_2|$.

2. Linear recurrence on two-sided orbits, combined with monotonicity from continuity plus injectivity.

3. Hesitation comes after solving the recurrence forwards and seeing everything decay. It is reinforced away by asking whether g is invertible (it is, from the equation) and then asking which mode dominates as $n \to -\infty$.

4. The trap is the interval $(0, \frac{1}{2})$: it invites the thought "roots inside the unit disc, so orbits decay", which is true and useless. A second trap is the standard functional-equation toolkit (substitute special values, look for additivity), which cannot work because the equation only links points on the same orbit.

5. Ask: *which points does the equation actually connect?* If it only connects x to its iterates, the problem lives on orbits, and on orbits a linear equation is a linear recurrence.

6. Reusable decision rule: *for a functional equation linking iterates of g linearly, pass to orbits and solve the recurrence; to show a mode is absent, go to the end of the orbit where that mode dominates (backwards for the small root, which needs g invertible), and contradict a qualitative property of g — monotonicity, sign, boundedness — that continuity has already given you.*

2001 B6

2001 B6. Assume that $(a_n)_{n \geq 1}$ is an increasing sequence of positive real numbers such that $\lim a_n/n = 0$. Must there exist infinitely many positive integers n such that $a_{n-i} + a_{n+i} < 2a_n$ for $i = 1, 2, \ldots, n-1$?

Fields: real analysis (sequences, sublinear growth); convexity / supporting lines; extremal principle
Position: B6 — the last problem of the paper; hard to start, short to finish once the right object is on the page **Answer:** Yes. For each slope $m > 0$ let $n(m)$ be the largest index maximising $a_k - mk$; every such $n(m)$ has the property, and $n(m) \to \infty$ as $m \to 0^+$, so there are infinitely many. ("Increasing" must be read as strictly increasing: the constant sequence satisfies the other hypotheses and has no such n.)

Layer 1 THE WHAT — surface analysis

The natural attempt. The strong student reads $a_{n-i} + a_{n+i} < 2a_n$ as "a_n is above the average of its two neighbours at distance i", i.e. a concavity condition centred at n, and checks examples. For $a_n = \sqrt{n}$ or $a_n = \log(n+1)$ (strictly concave) every n works. So the question is whether some cunning non-concave sequence can dodge the condition from some point on. The natural move is to try to build a counterexample: a staircase that climbs steeply for a stretch, then almost flat for a long stretch, then steeply again. Inside a stretch that bends upwards (convex), $a_{n-1} + a_{n+1} > 2a_n$ and the condition fails already at $i = 1$; along a flat stretch, pick i reaching back to the previous steep part and a_{n-i} is small, which helps the inequality instead of hurting it. The student keeps finding that *some* n at the top of each steep stretch works, but cannot turn "the top of a stretch" into a definition.

The second natural attempt is contradiction by chaining. Suppose that for every $n \geq N$ there is an $i = i(n) \in \{1, \ldots, n-1\}$ with

$$a_{n-i} + a_{n+i} \;\geq\; 2a_n, \qquad \text{i.e.} \qquad a_{n+i} - a_n \;\geq\; a_n - a_{n-i}. \tag{$*$}$$

The student tries to iterate $(*)$: jump from n to $n + i(n)$, apply $(*)$ again, and hope to show the increments never shrink, so the sequence grows at least linearly, contradicting $a_n/n \to 0$. It does not close. The jumps $i(n)$ vary from step to step, so "the increment over a window of length $i(n)$" at one step cannot be compared with "the increment over a window of length $i(n')$" at the next. Each instance of $(*)$ is a two-point comparison with its own scale, and there is no way to glue them. That is the wall: the condition is stated one centre at a time, and the student is trying to fight it one centre at a time.

The model chain. Throughout, "increasing" means $a_1 < a_2 < a_3 < \cdots$. Positivity of the a_n will not be used.

Step 1: for each slope $m > 0$ the tilted sequence has a last maximum. Fix $m > 0$ and put $b_k = a_k - mk$. Since $a_k/k \to 0 < m$, there is K with $a_k/k < m/2$ for $k \geq K$, so $b_k < -mk/2 \to -\infty$. Hence only finitely many k have $b_k \geq b_1$; the maximum of b_k over all $k \geq 1$ is attained, and the set of maximisers is finite. Let $n = n(m)$ be the *largest* maximiser. Then

$$b_k \leq b_n \;\; (k < n), \qquad b_k < b_n \;\; (k > n).$$

Step 2: every $n(m)$ has the required property. Let $n = n(m)$ and $1 \leq i \leq n-1$ (so $n - i \geq 1$ and the terms exist). From Step 1,

$$a_{n-i} - m(n-i) \;\leq\; a_n - mn, \qquad a_{n+i} - m(n+i) \;<\; a_n - mn.$$

Adding, the m-terms cancel ($-m(n-i) - m(n+i) = -2mn$), and

$$a_{n-i} + a_{n+i} \;<\; 2a_n.$$

(If $n = 1$ the condition is vacuous; it does no harm.)

Step 3: the indices $n(m)$ are unbounded. Fix any $N \geq 1$. Since the sequence is strictly increasing, $d = a_{N+1} - a_N > 0$; choose m with $0 < m < d/N$. For every $k \leq N$,

$$b_k = a_k - mk \;\leq\; a_N - m \cdot 1,$$

because $a_k \leq a_N$ and $k \geq 1$. On the other hand

$$b_{N+1} = a_{N+1} - m(N+1) = a_N + d - mN - m \;>\; a_N - m,$$

since $d > mN$. So b_{N+1} beats every b_k with $k \leq N$, and the maximiser $n(m)$ is at least $N + 1$.

Conclusion. For each N there is an index $n > N$ with the property (Steps 2 and 3). Hence infinitely many such n exist. The answer is yes. □

The hypotheses are needed. If "increasing" allowed equality, $a_n \equiv 1$ would be a counterexample: $a_{n-i} + a_{n+i} = 2a_n$ always. If $a_n/n \to 0$ is weakened to a_n/n bounded, $a_n = n$ fails for the same reason; and a convex sequence such as $a_n = n^2$ (with $a_n/n \to \infty$) fails at every n with $i = 1$.

Layer 2 THE HOW — the bifurcation

Bifurcating point. Both chains start from the same reading: the inequality is concavity at n, tested against every pair $(n-i, n+i)$ that fits inside $[1, 2n-1]$. They split at the first line written after that reading. The natural attempt writes the *negation* $(*)$ at a single centre and tries to chain centres together. The model writes down a *line* $y = mx + c$ and asks which point (k, a_k) it touches last as it is lowered onto the sequence from above — that is, it maximises $a_k - mk$. From that line on, the condition is no longer something to be checked at n; it is inherited by n from a single global inequality.

> *"$a_{n-i} + a_{n+i} < 2a_n$ for every admissible i" is a statement about all points to the left of n at once. A local chain of inequalities will never produce that. What produces an inequality against every other point simultaneously? Being a maximum.*

> *Maximum of what? Not of a_k — that is increasing and has no maximum. Subtract a line: $a_k - mk$. Since $a_k/k \to 0$, this goes to $-\infty$ for every $m > 0$, so it has a maximum, and at the maximiser the point (n, a_n) sits on a line with every other point on or below it. Two points symmetric about n below a line through (n, a_n) average to less than a_n — the slope cancels.*

> *I need strict inequality, and ties could break it. Take the* last *maximiser: then everything to the right is strictly below, and one strict inequality in the sum is enough.*

> *Infinitely many: make the line flatter. As $m \to 0$ the tilt stops penalising late indices, and a strictly increasing sequence then wins by going further right.*

- ▷ **Cue:** a condition "$x_{n-i} + x_{n+i} < 2x_n$ for all i" (or any inequality comparing x_n with all other terms at once) $\longrightarrow$ **Move:** look for n as a maximiser of a modified quantity; a global maximum delivers inequalities against every competitor simultaneously
- ▷ **Cue:** a sequence with $a_k/k \to 0$ (sublinear growth) $\longrightarrow$ **Move:** for every slope $m > 0$, $a_k - mk \to -\infty$, so the tilted sequence attains its maximum: lower a line onto the graph and see where it touches
- ▷ **Cue:** a symmetric pair of points $(n \pm i, \cdot)$ and a line through (n, a_n) $\longrightarrow$ **Move:** add the two "below the line" inequalities; the slope terms cancel and leave the midpoint (concavity) inequality
- ▷ **Cue:** strict inequality required but the extremum may be tied $\longrightarrow$ **Move:** choose the largest (or smallest) extremiser, so that one side is strict; one strict term in a sum suffices
- ▷ **Cue:** a family of extremisers depending on a parameter, and "infinitely many" is required $\longrightarrow$ **Move:** show the extremiser escapes every bound as the parameter tends to its limit, by comparing one explicit competitor with everything below the bound

Layer 3 THE WHY

Kedlaya's solution (with Bhargava and Ng) begins by taking the convex hull of the points (n, a_n) and looking for vertices on its upper boundary. Here is where that comes from, and why it is the natural object rather than a piece of geometry pulled from a hat.

Start from the definition of concavity. A function is concave on an interval when the chord between any two points lies below the graph — equivalently, when at every point there is a line through the graph with the whole graph beneath it (a supporting line). The midpoint inequality $f(x-h)+f(x+h) \le 2f(x)$ is the chord condition in its most symmetric form. So when the problem asks for $a_{n-i}+a_{n+i} < 2a_n$ for *every* i up to $n-1$, it is asking for the chord through $(n-i, a_{n-i})$ and $(n+i, a_{n+i})$ to pass below (n, a_n) for every symmetric pair. The cheapest sufficient condition for "every chord through n passes below (n, a_n)" is the other half of the definition: a supporting line at n. If some line through (n, a_n) has every other point below it, then every symmetric chord is below that line at its midpoint, hence below a_n. That is Step 2, and it is not a trick; it is the second of the two equivalent definitions of concavity, used on a sequence that is not concave.

The problem's phrase "for $i = 1, 2, \ldots, n-1$" is the cue that points there. The window reaches all the way back to index 1. The test at n involves every earlier term. A property of n that involves every earlier term is a global property, and the most economical global properties are extremal ones. Once you ask "n is the maximiser of what?", the answer is forced: a_k itself is increasing and has no maximum, so you must subtract something, and the only subtraction that is invisible to the symmetric sum $a_{n-i}+a_{n+i}-2a_n$ is a linear function $mk+c$ (because $m(n-i)+m(n+i)-2mn = 0$). Linear functions are exactly the functions whose second difference is zero. That is why the tilt is by a line and not by anything else.

Why does $a_n/n \to 0$ matter? It says every line of positive slope eventually overtakes the sequence — geometrically, the graph has "slope 0 at infinity". That is precisely what makes $a_k - mk$ attain a maximum for every $m > 0$, i.e. what guarantees that a line of any positive slope, lowered from above, touches the graph somewhere instead of sliding off to infinity. With $a_n = n$ the slope at infinity is 1 and lines of slope below 1 never touch; with $a_n = n^2$ no line touches at all.

Why does strict increase matter? It is the engine of Step 3. A flatter line penalises late indices less; strict increase means the sequence always has something to gain by moving right, so the touching point escapes to infinity. With a constant sequence every flat line touches everywhere and the strictness in Step 2 disappears. (Positivity of the a_n is never used: the problem is invariant under adding a constant to every term.)

Why fix the slope first rather than the point? The official solution fixes a vertex n, finds the next vertex r by maximising the slope $(a_k - a_n)/(k-n)$ over $k > n$, and then needs a line through (r, a_r) strictly above every other point. It asserts that slope $m - \varepsilon$ works "for suitably small ε". That is true, but it owes an argument: for $k > r$ one needs $\sup_{k>r}(a_k - a_r)/(k-r) < m$, which holds because those slopes are each $< m$ and tend to $0 < m$, so their supremum is attained and is $< m$; for $k < n$ one uses that the supporting slope at n exceeds m. My route fixes m first, so the maximiser exists by one limit argument and strictness comes free from choosing the last maximiser. I do not want to spend the exam controlling an ε over infinitely many indices when a tie-break does the same job. That reroute is the cheapness the problem rewards.

Layer $+\alpha$ BACKTESTING

1. The first useful cue is the range $i = 1, \ldots, n-1$: the inequality at n tests a_n against every earlier term. On seeing it, write "this is a global property of n; look for n as an extremiser", and translate $a_{n-i}+a_{n+i} < 2a_n$ as "(n, a_n) lies above the chord of every symmetric pair".

2. Supporting lines: maximise $a_k - mk$ over k for a fixed slope $m > 0$, then vary m. Equivalently, the upper convex hull of the points (k, a_k).

3. Hesitation comes after testing concave examples and failing to build a counterexample, when the student starts chaining the negation $(*)$ across centres with varying $i(n)$. It is reinforced away by asking "what single inequality would give me n versus every other index at once?" — whose answer is always a maximum.

4. Two traps. First, the local reading: treating the condition as a second-difference condition at n and trying to control it with neighbouring centres, which cannot work because each

centre has its own window. Second, the strictness trap: the maximiser of $a_k - mk$ may be tied, and a careless proof ends with $\leq$; the fix is the last maximiser. A third, quieter trap is the word "increasing": the answer is yes only for strictly increasing sequences, and the proof must use strictness somewhere (it is used in Step 3).

5. Recognise faster by listening for a condition that compares one term with *all* others (here, all symmetric pairs back to index 1). Such conditions are almost never proved by local manipulation; they are inherited from an extremal choice. And whenever an expression is invariant under adding a linear function (as any second difference is), tilt by a line before maximising.

6. Reusable decision rule: *when you must find infinitely many n satisfying an inequality against every other index, choose n as the (last) maximiser of the sequence minus a function to which the inequality is blind (a line, for second differences); use the growth hypothesis to make the maximum exist, a tie-break to make it strict, and a limit in the parameter to push the maximiser to infinity.*